Linear Models and Regression

Lutz Dümbgen
University of Bern

June 6, 2023

Bibliography

[1] A.D. BARBOUR and LOUIS H.Y. CHEN (2005). *An Introduction to Stein's Method.* Institute for Mathematical Sciences, University of Singapore, Lecture Notes Series, Volume 4.

[2] G. BASSET, JR. and R. KOENKER (1982). An empirical quantile function for linear models with iid errors. *J. Amer. Statist. Assoc.* **77**, 407-415.

[3] R. BERAN and G.R. DUCHARME (1991). *Asymptotic Theory for Bootstrap Methods in Statistics.* Les Publications CRM, Montreal.

[4] P.J. BICKEL and D.A. FREEDMAN (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196-1217.

[5] P.J. BICKEL and D.A. FREEDMAN (1983). Bootstrapping regression models with many parameters. In: *A Festschrift for Erich Lehmann (P. Bickel, K. Doksum and J.L. Hodges, eds.)*, Wadsworth, Belmont, CA, pp. 28-48.

[6] C. DE BOOR (2002). *A Practical Guide to Splines (revised edition).* Springer-Verlag.

[7] L. DÜMBGEN (1993). On nondifferentiable functions and the bootstrap. *Prob. Theory Rel. Fields* ***95****(1)*, 125-140.

[8] L. DÜMBGEN (2003). *Stochastik für Informatiker.* Springer-Verlag.

[9] L. DÜMBGEN (2015). *Einführung in die Statistik.* Birkhäuser, Basel. (English translation available upon request.)

[10] L. DÜMBGEN (2006/2007). *Wahrscheinlichkeitstheorie.* Lecture notes, Univ. of Bern.

[11] L. DÜMBGEN (2021). *Optimization Methods.* Lecture notes, Univ. of Bern.

[12] B. EFRON (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1-26.

[13] J. FAN and I. GIJBELS (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall.

[14] P. GROENEBOOM and G. JONGBLOED (2014). *Nonparametric Estimation under Shape Constraints.* Cambridge University Press.

[15] A. HENZI, A. MÖSCHING and L. DÜMBGEN (2022). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. *Methodol. Comput. Appl. Probab.* ***24****(4)*, 2633-2645.

[16] D.W. HOSMER and S. LEMESHOW (1989). *Applied Logistic Regression.* John Wiley & Sons.

[17] K.-H. JÖCKEL (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Ann. Statist.* ***14****(1)*, 336-347.

[18] J.K. LINDSEY (1997). *Applying Generalized Linear Models.* Springer-Verlag.

[19] R.Y. LIU (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* ***16***, 1696-1708.

[20] E. MAMMEN (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* ***21***, 255-285.

[21] R.G. MILLER (1981). *Simultaneous Statistical Inference (2nd edition).* Springer-Verlag.

[22] A. MÖSCHING and *L. Dümbgen* (2020). Monotone least squares and isotonic quantiles. *Electron. J. Stat.* ***14****(1)*, 24-49.

[23] G. OPFER (1994). *Numerische Mathematik für Anfänger.* Fr. Vieweg & Sohn.

[24] A.B. OWEN (1990). Empirical likelihood confidence regions. *Ann. Statist.* ***18***, 90-120.

[25] A.B. OWEN (1991). Empirical likelihood for linear models. *Ann. Statist.* ***19***, 1725-1747.

[26] A.B. OWEN (2001). *Empirical Likelihood.* CRC Press.

[27] R.L. PRENTICE and R. PYKE (1979). Logistic disease incidence models and case-control studies. *Biometrika* ***66***, 403-411.

[28] J. RICE (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* ***12***, 1215-1230.

[29] T. ROBERTSON, F.T. WRIGHT and R.L. DYKSTRA (1988). *Order Restricted Statistical Inference.* John Wiley & Sons.

[30] T.P. RYAN (1997). *Modern Regression Methods.* John Wiley & Sons.

[31] S. SCHACH UND T. SCHÄFER (1978). *Regressions- und Varianzanalyse - Eine Einführung.* Springer-Verlag

[32] H. SCHEFFÉ (1959). *Analysis of Variance.* John Wiley & Sons.

[33] L. SCHUMAKER (1981). *Spline Functions: Basic Theory.* John Wiley & Sons.

[34] D. URBAN (1993). *Logit-Analyse - Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen.* G. Fischer Verlag.

[35] C.F.J. WU (1986). Jackknife, bootstrap, and other resampling methods in regression analysis (with discussion). *Ann. Statist.* ***14***, 1261-1295 (1295-1350).

Acknowledgements. Over the last two decades, José Araujo, Sebastian Arnold, Yves Bartels, Claire Descombes, Gabriel Fischer, Andrea Fraefel, Livio Käslin, Dirk Klingbiel, Thomas Loubejac, Werner Luginbühl, Géraldine Oppliger, Paul Ruppen, Günter Sawitzki, Ben Spycher and Niki Zumbrunnen provided numerous hints to different versions of these lecture notes. Dominic Schuhmacher, Christof Strähl, Anja Mühlemann, Federico Pianoforte, Johanna Ziegel and Michael Vock helped a lot with the more recent editions. Thanks a lot to all of them!

I am indebted to Dietrich W. Müller and Günter Sawitzki who introduced me to linear models and many other branches of statistics at the University of Heidelberg.

Contents

1 Introduction **11**

1.1 Definition of a Linear Model 11

1.2 Examples of Linear Models 12

2 Estimation of Parameters **15**

2.1 Vector and Matrix Representation 15

2.2 Estimation of $\boldsymbol{\theta}$ 18

2.2.1 Least Squares Estimation 18

2.2.2 Examples for $\widehat{\boldsymbol{\theta}}$ 22

2.2.3 The Coefficient of Determination 27

2.2.4 The Precision of $\widehat{\boldsymbol{\theta}}$ 29

2.3 Estimation of $\boldsymbol{\sigma}$ 32

2.4 The Gauss–Markov Theorem and Standard Errors 34

2.5 Properties of the Estimators in Misspecified Models 39

2.6 Special Parametrizations and Interactions 43

2.6.1 One-Way Analysis of Variance 43

2.6.2 Two-Way Analysis of Variance 45

2.6.3 Interactions 47

2.6.4 Paper and Blackboard Notation 51

3 Tests and Confidence Regions **53**

3.1 Multivariate Gaussian Distributions 53

3.2 Special Univariate Distributions 55

3.3 The Joint Distribution of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$ 57

3.4 Student Confidence Intervals and Tests 59

3.4.1 Student Confidence Regions 59

3.4.2 Student Tests . . . 62
3.5 F Confidence Regions and Tests . . . 64
3.5.1 F Tests and Confidence Ellipsoids . . . 65
3.5.2 Simultaneous Confidence Intervals via F Tests . . . 66
3.5.3 Generalizations . . . 73
3.5.4 A Geometrical Approach to F Tests . . . 76
3.6 Alternative Simultaneous Confidence Intervals . . . 88
3.6.1 The Bonferroni Method . . . 88
3.6.2 Tukey's Method . . . 89
3.6.3 Examples of Simultaneous Confidence Regions . . . 90
3.6.4 Comparison of the Methods . . . 95
3.7 Non-Central F Distributions and Approximation Errors . . . 98
3.8 Calibration . . . 103
3.9 Random Effects . . . 106

4 Regression Diagnostics **111**
4.1 Leverage . . . 111
4.2 An Application of the Central Limit Theorem . . . 112
4.3 Residual Analysis . . . 115
4.3.1 Q-Q-Plots for Normality . . . 115
4.3.2 Plots of Residuals versus Functions of the Covariates or the Fit . . . 118
4.4 Transformations . . . 123

5 Nonparametric Regression **125**
5.1 Spline Regression . . . 125
5.1.1 Definition of Splines . . . 125
5.1.2 Polynomial Representation and a First Basis . . . 126
5.1.3 B Splines . . . 127
5.1.4 Precision in Case of Linear Splines . . . 129
5.2 Local Polynomials . . . 131
5.2.1 Examples for the Weights $w_i(x)$ . . . 132
5.2.2 Explicit Computation . . . 132
5.2.3 Precision of Locally Linear Estimators . . . 134

5.3 Regularization . . . 137
5.3.1 Smoothing Splines . . . 137
5.3.2 A Related Approach . . . 141
5.3.3 Choosing the Regularization Parameter . . . 142

6 General Considerations about Estimation 145

6.1 Means and Quantiles as Optimal Predictors . . . 145
6.2 Loss Functions and Risks . . . 149
6.3 Maximum Likelihood Estimation . . . 152
6.4 Application to Regression Problems . . . 155

7 Logistic Regression and Related Models 159

7.1 Logistic Regression . . . 159
7.1.1 Maximum Likelihood Estimation . . . 160
7.1.2 The Asymptotic Behavior of the Log-Likelihood Function . . . 166
7.1.3 Likelihood-Based Statistical Procedures . . . 170
7.1.4 Returning from Asymptopia . . . 175
7.1.5 A Data Example . . . 175
7.1.6 Case–Control Studies . . . 181
7.2 General Asymptotic Considerations . . . 182
7.3 Methods for a Multicategorical Response . . . 187
7.3.1 Multinomial Logit Models . . . 187
7.3.2 The Ordinal Logit Model . . . 188
7.4 Poisson Regression . . . 191
7.5 Complements . . . 195
7.5.1 Generalized Linear Models . . . 195
7.5.2 Exact Confidence Regions for f_* . . . 196
7.5.3 Permutation Tests of Association . . . 197

8 Bootstrap Methods 199

8.1 Bootstrap Methods for I.I.D. Observations . . . 199
8.2 Bootstrap Methods for Regression Models . . . 206
8.2.1 Logistic and Poisson Regression . . . 206
8.2.2 Bootstrap Methods for Linear Models . . . 208

8.2.3 The Residual Bootstrap . . . 209
8.2.4 Wild Bootstrap . . . 213
8.3 Exact Tests and Confidence Regions in the Linear Model . . . 217
8.4 Bootstrap Failures and Subsampling . . . 218

9 Empirical Likelihood **225**
9.1 Empirical Likelihood for I.I.D. Observations . . . 225
9.1.1 Analytical Properties of an Empirical Likelihood Function . . . 226
9.1.2 Inference about the Mean . . . 232
9.2 Empirical Likelihood for Linear Regression . . . 232

10 Isotonic Regression **239**
10.1 The Pool-Adjacant-Violators Algorithm (PAVA) . . . 240
10.2 Further Considerations for Isotonic Least Squares . . . 244
10.3 Isotonic Distributional Regression . . . 251

A Miscellaneous Auxiliary Results **253**
A.1 The QR Decomposition . . . 253
A.2 Expected Values and Covariances . . . 255
A.3 Monte Carlo Estimators for Tukey's Method . . . 258
A.4 B Splines . . . 259
A.5 Weak Convergence of Distributions . . . 265
A.6 Lindeberg's Central Limit Theorem . . . 269
A.6.1 The Univariate Case . . . 270
A.6.2 The Multivariate Case . . . 270
A.7 Iteratively Reweighted Least Squares . . . 277
A.8 Couplings and Mallows Distances . . . 278
A.8.1 Optimal Transport . . . 278
A.8.2 Optimal Transport on the Real Line . . . 280
A.8.3 Mallows Distances . . . 282
A.9 An Inequality for Sums of Independent Random Vectors . . . 287
A.10 Stochastic Landau Symbols . . . 290

Chapter 1

Introduction

Linear Models play an important role in applications of statistics. By means of such models one can model and analyse data from various disciplines. Typically the main question is as follows: Let (X, Y) be a generic observation consisting of a *covariate* X or a *tuple* X *of several covariates* with values in some set $\mathcal{X}$ and a *response (variable)* Y with values in some set $\mathcal{Y}$. Is there an association between X and Y? More precisely, what is the *conditional distribution of* Y, *given* X? This type of question is called *regression (analysis)*.

The first part of this course treats the important special case of a numerical response Y, that means, $\mathcal{Y} = \mathbb{R}$. In the second part we shall also consider integer-valued or categorical responses Y. Regression models for a vector-valued response Y are sometimes treated in Multivariate Statistics.

In case of a tuple $X = (X(j))_{j=1}^{d}$ of covariates, the single components are sometimes called *independent variables* whereas Y is called the *dependent variable*. Here the adjective "independent" has nothing to do with stochastic independence, it refers rather to the asymmetrical viewpoint of considering the conditional distribution of Y depending on X. Categorical covariates are sometimes called *factors*. The variety of nomenclature is due to the many different fields and communities utilizing regression models.

1.1 Definition of a Linear Model

Depending on the specific application, X may be viewed as a random variable or as a fixed quantity or tuple, possibly to be chosen haphazardly by an experimenter. In the present course, we shall treat X mostly as a fixed quantity or tuple. In settings with genuinely random X, we always condition on the actual value of X.

- We assume that Y may be written as

$$Y \;=\; f(X) + \varepsilon$$

with an unknown *regression function* $f : \mathcal{X} \to \mathbb{R}$ and a *random error* ε such that

$$\mathbb{E}(\varepsilon) \;=\; 0.$$

The distribution of ε may depend on X. For instance, one could think about $\varepsilon = \sigma(X)Z$ with a certain function $\sigma : \mathcal{X} \to [0, \infty)$ and a standard Gaussian random variable Z.

• We often assume that the standard deviation of ε is finite and does *not* depend on X. In that case, we write

$$\sigma \ := \ \mathrm{Std}(\varepsilon) = \sqrt{\mathrm{Var}(\varepsilon)}$$

and talk about *homoscedastic errors*. Otherwise we talk about *heteroscedastic errors*.

• Concerning the regression function f, we assume that it belongs to a given family $\mathcal{F}$ of functions. Moreover, $\mathcal{F}$ is a *finite-dimensional linear space* (i.e. a finite-dimensional real vector space) of functions. The latter property is the reason for the term "linear model".

1.2 Examples of Linear Models

The subsequent examples have somewhat mysterious names which arose historically. The reader should not be frustrated if he or she does not see any coherent scheme (yet).

Example 1.1 (One-way analysis of variance (One-way ANOVA)). Let X be a categorical covariate with values in, say, $\mathcal{X} = \{1, 2, \ldots, L\}$. The set $\mathcal{F}$ of all real-valued functions on $\mathcal{X}$ is a linear space of dimension L. It corresponds to $\mathbb{R}^L$ if we identify a function $f \in \mathcal{F}$ with the vector $(f(x))_{x=1}^L$.

An specific example is the yield of a certain agricultural crop on a field of given area. The covariate X could stand for different types of soil, irrigation schemes, treatments of the seeds etc. Thus, we assume that the yield Y is equal to a number $f(X)$ plus some random fluctuation ε which is for instance due to variations in the weather conditions or other environmental influences.

Example 1.2 (Simple linear regression). Let X be a numerical covariate, i.e. $X \in \mathbb{R}$. Often one assumes that there is a linear relationship between X and Y, that means,

$$Y \ = \ a + bX + \varepsilon \tag{1.1}$$

with unknown parameters a and b. Thus, we consider the family $\mathcal{F}$ of all affine functions on $\mathbb{R}$, which is a two-dimensional linear space.

A specific example are indirect measurements and calibration lines. Suppose that X stands for a physical or chemical parameter, e.g. the concentration of a certain substance in a fluid. Suppose that the value X could be determined precisely with an expensive or time-consuming method. Suppose further that an indirect and less precise measurement Y is comparatively easy to obtain, e.g. the measurement of light absorption of a fluid. If the relationship between X and Y is given by (1.1) with $b > 0$, one could estimate the parameters a and b by means of a calibration experiment in which complete observation pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with different values $X_1, \ldots, X_n$ are determined, leading to $\widehat{a}$ and $\widehat{b}$. Then for any future observation (X, Y) of which only Y is observed, one could estimate X by $\widehat{X} = (Y - \widehat{a})/\widehat{b}$. We shall come back to this specific example later.

Example 1.3 (Polynomial regression). As in Example 1.2 let $X \in \mathbb{R}$. Instead of a linear relationship between X and Y one could assume that f is a polynomial of given order $d \geq 1$. That means, we consider the $(d+1)$-dimensional vector space of all functions f of type

$$f(x) \;=\; a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d \tag{1.2}$$

with real parameters $a_0, a_1, \ldots, a_d$. In case of $d = 1$ we have simple linear regression, the cases $d = 2$ and $d = 3$ are quadratic and cubid regression, respectively.

This model is often used with $d \geq 2$ to check the plausibility of simple linear regression (Example 1.2). That means, one tests whether the coefficients a_j with $j > 1$ are really needed.

The special model of quadratic regression ($d = 2$) can be used, for instance, to model the effect of the dose X of some ingredient (e.g. a fertilizer) on a certain response Y (e.g. the yield of an agricultural crop). In case of $a_1 > 0 > a_2$, the function f in (1.2) describes a parabola with unique maximum at $x = a_1/(-2a_2) > 0$.

Example 1.4 (One-way analysis of covariance (One-way ANCOVA)). Soppose that X consists of a categorical covariate $C \in \{1, \ldots, L\}$ and a numerical covariate $W \in \mathbb{R}$. A possible model $\mathcal{F}$ consists of all functions f of type

$$f(c, w) \;=\; a(c) + bw$$

with real parameters $a(1), \ldots, a(L)$ and b. Thus one combines one-way ANOVA (Example 1.1) with simple linear regression (Example 1.2). The term "covariance" has nothing to do with covariance in the sense of stochastics. It is rather a wordplay combining "variance" as in ANOVA with "(numerical) covariate".

A specific example is the cholesterol level of adults (Y). This is known to be dependent on their age (W), but it may also depent on other factors such as gender (C).

Another specific example is the log-income (Y) of persons in relation to their age (W) and profession or type of education (C).

Example 1.5 (Multiple linear regression). Suppose that $X = (X(j))_{j=1}^d$ is a vector of d numerical (or $\{0,1\}$-valued) covariates. A simple linear model for the relationship between X and Y assumes that

$$Y \;=\; a + \sum_{j=1}^{d} b(j) X(j) + \varepsilon \tag{1.3}$$

with real parameters a and $b(1), \ldots, b(d)$.

Again, a specific example is the cholesterol level of adults (Y) and numerical covariates such as age, body height and weight as well as gender, coded by a number in $\{0,1\}$.

Exercise 1.6. For a wild cherry tree (prunus avium) we consider the three variables

Y : its timber yield,

$X(1)$: the height of its trunk,

$X(2)$: its maximal diameter (at waist height).

(a) Which connection do you expect between Y and the pair $(X(1), X(2))$?

(b) What does your model look like if you replace Y with $\log Y$?

Exercise 1.7 (Periodic signals). For $t \in \mathbb{Z}$ let

$$y_t \;=\; \mu + A\cos(\omega t - \phi)$$

with certain parameters $\mu \in \mathbb{R}$, $A > 0$, $\omega \in \mathbb{R} \setminus (2\pi\mathbb{Z})$ and $\phi \in \mathbb{R}$.

(a) Show that for suitable coefficients $a, b_1, b_2 \in \mathbb{R}$,

$$y_t \;=\; a + b_1 y_{t-1} + b_2 y_{t-2} \quad \text{for all } t \in \mathbb{Z}.$$

(Are these coefficients a, b_1, b_2 unique?)

(b) Show that for arbitrary integers s and $T > 0$,

$$\frac{1}{T}\sum_{t=s+1}^{s+T} y_t \;\to\; \mu \quad \text{as } T \to \infty, \text{ uniformly in } s.$$

Chapter 2

Estimation of Parameters

This chapter is about estimation of the regression function $f \in \mathcal{F}$ and, in case of homoscedastic errors, the standard deviation of the error $\varepsilon = Y - f(X)$. The available data are n observation pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$. The values $X_1, \ldots, X_n \in \mathcal{X}$ are considered as fixed while

$$Y_i \;=\; f(X_i) + \varepsilon_i$$

with random errors $\varepsilon_1, \ldots, \varepsilon_n$ such that $\mathbb{E}(\varepsilon_i) = 0$.

2.1 Vector and Matrix Representation

It is useful to represent data and model by means of vectors and matrices. We define the *response vector*

$$\boldsymbol{Y} \;:=\; \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \;\in\; \mathbb{R}^n$$

and the (unobserved) *error vector*

$$\boldsymbol{\varepsilon} \;:=\; \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \;\in\; \mathbb{R}^n.$$

For the model $\mathcal{F}$, we choose basis functions $f_1, \ldots, f_p$. Then, any function $f \in \mathcal{F}$ may be written as

$$f(x) \;=\; \sum_{j=1}^{p} \theta_j f_j(x)$$

with a *parameter vector*

$$\boldsymbol{\theta} \;=\; \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \;\in\; \mathbb{R}^p.$$

The basis functions f_j and the observations X_i yield the so-called *design matrix*

$$\boldsymbol{D} \;:=\; \begin{bmatrix} f_1(X_1) & \cdots & f_p(X_1) \\ f_1(X_2) & \cdots & f_p(X_2) \\ \vdots & & \vdots \\ f_1(X_n) & \cdots & f_p(X_n) \end{bmatrix} \;\in\; \mathbb{R}^{n\times p}.$$

With these ingredients we may write

$$\boldsymbol{Y} \;=\; \boldsymbol{D\theta} + \boldsymbol{\varepsilon}.$$

With the tuple $\boldsymbol{X} := (X_i)_{i=1}^n \in \mathcal{X}^n$ and the convention $g(\boldsymbol{X}) := (g(X_i))_{i=1}^n \in \mathbb{R}^n$ for functions $g : \mathcal{X} \to \mathbb{R}$, one may also write

$$\boldsymbol{D} \;=\; \big[f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})\big] \quad \text{and} \quad \boldsymbol{D\theta} \;=\; f(\boldsymbol{X}).$$

Note that many authors and most textbooks denote the design matrix with $\boldsymbol{X}$ rather than $\boldsymbol{D}$. We prefer the symbol $\boldsymbol{D}$ to emphasize the dependence of the design matrix on both, the observations X_i and the choice of the basis functions f_j.

Examples for this parametrization. We illustrate the vector and matrix representation with the examples from Section 1.2.

One-way ANOVA (Example 1.1). For the space $\mathcal{F}$ of all functions on $\{1, \ldots, L\}$ we choose the basis functions $f_j(x) := 1_{[x=j]}$, $1 \le j \le L$. (If one identifies $\mathcal{F}$ with $\mathbb{R}^L$, the functions $f_1, \ldots, f_L$ correspond to the standard basis of $\mathbb{R}^L$.) In this case the design matrix contains the entries

$$D_{ij} \;=\; 1_{[X_i=j]} \;\in\; \{0, 1\}.$$

Hence, the i-th row contains $L-1$ zeros and one entry 1 in column number X_i. The corresponding parameter vector $\boldsymbol{\theta}$ for $f \in \mathcal{F}$ is just $\boldsymbol{\theta} = (f(j))_{j=1}^L$.

Suppose we have arranged the observation pairs (X_i, Y_i) such that

$$\boldsymbol{X} \;=\; \big(\underbrace{1, \ldots, 1}_{n(1) \text{ times}}, \underbrace{2, \ldots, 2}_{n(2) \text{ times}}, \ldots, \underbrace{L, \ldots, L}_{n(L) \text{ times}}\big)^\top.$$

Then, the design matrix $\boldsymbol{D}$ equals

$$\boldsymbol{D} \;=\; \begin{bmatrix} 1 & 0 & & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & \cdots & 0 & 1 \\ \vdots & & & \vdots & \vdots \\ 0 & & \cdots & 0 & 1 \end{bmatrix} \;\in\; \mathbb{R}^{n\times L}.$$

Simple linear regression (Example 1.2). Here one can take the basis function $f_1(x) := 1$ and $f_2(x) := x$, leading to the design matrix

$$\boldsymbol{D} \;=\; \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \;=\; [\mathbf{1}\, \boldsymbol{X}] \;\in\; \mathbb{R}^{n\times 2}$$

with $\mathbf{1} := (1, 1, \ldots, 1)^\top \in \mathbb{R}^n$. If $f(x) = a + bx$, then the corresponding parameter vector is $\boldsymbol{\theta} = (a, b)^\top$.

Polynomial regression (Example 1.3). If we choose the basis functions $f_j(x) := x^{j-1}$, $1 \le j \le d+1$, then

$$\boldsymbol{D} \;=\; \begin{bmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^d \\ 1 & X_2 & X_2^2 & \cdots & X_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^d \end{bmatrix} \;\in\; \mathbb{R}^{n\times(d+1)}.$$

If $f(x) = \sum_{j=0}^{d} a_j x^j$, then the corresponding parameter vector is $\boldsymbol{\theta} = (a_0, a_1, \ldots, a_d)^\top$.

One-way ANCOVA (Example 1.4). Similarly as in Example 1.1, suppose that the observations (C_i, W_i, Y_i) have been rearranged such that

$$(C_1, C_2, \ldots, C_n) \;=\; \big(\underbrace{1, \ldots, 1}_{n(1)\text{ times}}, \underbrace{2, \ldots, 2}_{n(2)\text{ times}}, \ldots, \underbrace{L, \ldots, L}_{n(L)\text{ times}}\big).$$

With the basis functions $f_j(c, w) := 1_{[c=j]}$ for $1 \le j \le L$ and $f_{L+1}(c, w) := w$ we obtain the design matrix

$$\boldsymbol{D} \;=\; \begin{bmatrix} 1 & 0 & & \cdots & 0 & W_1 \\ \vdots & \vdots & & & \vdots & \vdots \\ 1 & 0 & & \cdots & 0 & W_{n(1)} \\ 0 & 1 & 0 & \cdots & 0 & W_{n(1)+1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 & W_{n(1)+n(2)} \\ & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & & 0 & 1 & W_{n-n(L)+1} \\ \vdots & & & \vdots & \vdots & \vdots \\ 0 & \cdots & & 0 & 1 & W_n \end{bmatrix} \;\in\; \mathbb{R}^{n\times(L+1)},$$

and $f(c, w) = a(c) + bw$ corresponds to the parameter vector $\boldsymbol{\theta} = \big(a(1), \ldots, a(L), b\big)^\top$.

Multiple linear regression (Example 1.5). Here we observe $X_i = (X_i(j))_{j=1}^d \in \mathbb{R}^d$. The basis functions $f_1(x) := 1$ and $f_{1+j}(x) := x(j)$ for $1 \le j \le d$ yield the design matrix

$$\boldsymbol{D} \;=\; \begin{bmatrix} 1 & X_1(1) & \cdots & X_1(d) \\ 1 & X_2(1) & \cdots & X_2(d) \\ \vdots & \vdots & & \vdots \\ 1 & X_n(1) & \cdots & X_n(d) \end{bmatrix} \;=\; \begin{bmatrix} 1 & X_1^\top \\ 1 & X_2^\top \\ \vdots & \vdots \\ 1 & X_n^\top \end{bmatrix} \;\in\; \mathbb{R}^{n\times(d+1)},$$

and $f(x) = a + \sum_{j=1}^d b_j x(j)$ corresponds to the parameter vector $\boldsymbol{\theta} = (a, b_1, \ldots, b_d)^\top$.

2.2 Estimation of $\boldsymbol{\theta}$

From now on, we assume that the design matrix has linearly independent colums, that means,

$$\operatorname{rank}(\boldsymbol{D}) \;=\; p \;\le\; n. \tag{2.1}$$

In particular, $n \ge p$. That means, for arbitrary vectors $\boldsymbol{\eta} \in \mathbb{R}^p \setminus \{0\}$, the vector $\boldsymbol{D\eta} \neq \mathbf{0}$, so

$$0 \;<\; \|\boldsymbol{D\eta}\|^2 \;=\; \boldsymbol{\eta}^\top \boldsymbol{D}^\top \boldsymbol{D}\,\boldsymbol{\eta}.$$

Here and throughout this course, $\|\cdot\|$ denotes standard Euclidean norm. Consequently, Condition (2.1) is equivalent to

$$\boldsymbol{D}^\top \boldsymbol{D} \text{ is positive definite.} \tag{2.2}$$

This fact will be used frequently.

Exercise 2.1. For real numbers $X_1, X_2, \ldots, X_n$ and an integer $d \ge 1$, let

$$\boldsymbol{D} \;:=\; \begin{bmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^d \\ 1 & X_2 & X_2^2 & \cdots & X_2^d \\ \vdots & \vdots & & \vdots & \\ 1 & X_n & X_n^2 & \cdots & X_n^d \end{bmatrix}$$

be the design matrix for polynomial regression of order d. What is a necessary and sufficient condition on the numbers X_i for (2.1) to hold?

Hint: One can use determinants. Alternatively, one can think about the meaning of $\boldsymbol{D\eta} = \mathbf{0}$ in terms of the function $\mathbb{R} \ni x \mapsto g(x) := \sum_{j=0}^d \eta_{j+1} x^j$.

2.2.1 Least Squares Estimation

A vector $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^p$ is called *least squares estimator (LSE)* of $\boldsymbol{\theta}$ if

$$\|\boldsymbol{Y} - \boldsymbol{D}\widehat{\boldsymbol{\theta}}\|^2 \;=\; \min_{\boldsymbol{\eta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{D\eta}\|^2.$$

In other words, one chooses $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^p$ such that the sum of squares

$$\sum_{i=1}^n \Big(Y_i - \sum_{j=1}^p \widehat{\theta}_j f_j(X_i)\Big)^2$$

becomes minimal. The function $\widehat{f} := \sum_{j=1}^p \widehat{\theta}_j f_j$ is a LSE of the true regression function f.

Lemma 2.2. *Under Condition* (2.1) *there exists a unique LSE of* $\boldsymbol{\theta}$*, namely,*

$$\widehat{\boldsymbol{\theta}} \;=\; (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{Y}.$$

First proof of Lemma 2.2 (analytical). Let $Q(\boldsymbol{\eta}) := \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\eta}\|^2$. For arbitrary vectors $\boldsymbol{\eta}, \boldsymbol{v} \in \mathbb{R}^p$,

$$\begin{aligned} Q(\boldsymbol{\eta} + \boldsymbol{v}) \;&=\; \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\eta} - \boldsymbol{D}\boldsymbol{v}\|^2 \\ &=\; \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\eta}\|^2 - 2(\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\eta})^\top \boldsymbol{D}\boldsymbol{v} + \|\boldsymbol{D}\boldsymbol{v}\|^2 \\ &=\; Q(\boldsymbol{\eta}) - 2(\boldsymbol{D}^\top \boldsymbol{Y} - \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\eta})^\top \boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{v}. \end{aligned}$$

This shows that the gradient of Q at $\boldsymbol{\eta}$ is given by $-2(\boldsymbol{D}^\top \boldsymbol{Y} - \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\eta})$, and the Hessian matrix (2nd derivative) equals $2\boldsymbol{D}^\top \boldsymbol{D}$ everywhere. The latter is positive definite by assumption (2.2), whence Q is a strictly convex function. The gradient is zero if and only if $\boldsymbol{D}^\top \boldsymbol{Y} = \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\eta}$, so $\boldsymbol{\eta} = (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{Y}$. Hence, Q has a unique local minimum at $(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{Y}$, and by convexity of Q, this point is a global minimum. □

Second proof of Lemma 2.2 (quadratic completion). One may write

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\eta}\|^2 \;&=\; \|\boldsymbol{Y}\|^2 - 2\boldsymbol{\eta}^\top \boldsymbol{D}^\top \boldsymbol{Y} + \boldsymbol{\eta}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\eta} \\ &=\; \|\boldsymbol{Y}\|^2 - 2\boldsymbol{\eta}^\top (\boldsymbol{D}^\top \boldsymbol{D})(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{Y} + \boldsymbol{\eta}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\eta} \\ &=\; \|\boldsymbol{Y}\|^2 - 2\boldsymbol{\eta}^\top (\boldsymbol{D}^\top \boldsymbol{D})\boldsymbol{\eta}_o + \boldsymbol{\eta}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\eta} \\ &=\; \|\boldsymbol{Y}\|^2 - \boldsymbol{\eta}_o^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\eta}_o + (\boldsymbol{\eta} - \boldsymbol{\eta}_o)^\top (\boldsymbol{D}^\top \boldsymbol{D})(\boldsymbol{\eta} - \boldsymbol{\eta}_o), \end{aligned}$$

where $\boldsymbol{\eta}_o := (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{Y}$. Together with (2.2) this implies that $\boldsymbol{\eta}_o$ is the unique LSE of $\boldsymbol{\theta}$. □

Remark 2.3 (Numerical computation). The formula in Lemma 2.2 is useful for theoretical considerations. For the explicit calculation of $\widehat{\boldsymbol{\theta}}$, it is possibly problematic because the quadratic matrix $\boldsymbol{D}^\top \boldsymbol{D}$ can be rather ill-conditioned in the sense that the ratio of its smallest and largest eigenvalues is very small. Numerically more stable procedures are based on good choices of basis functions, as illustrated later in three particular settings, or, on the QR decomposition of $\boldsymbol{D}$; see Section A.1 in the appendix.

Remark 2.4 (Geometric interpretation). For a better understanding of the properties of $\widehat{\boldsymbol{\theta}}$ and other prodedures introduced later, the following consideration is useful: We assume that the vector $\boldsymbol{Y}$ is equal to $f(\boldsymbol{X}) = \boldsymbol{D}\boldsymbol{\theta}$ plus some random error $\boldsymbol{\varepsilon}$. The vector $f(\boldsymbol{X})$ is a point in the linear subspace

$$\boldsymbol{M} \;:=\; \{g(\boldsymbol{X})\colon g \in \mathcal{F}\} \;=\; \operatorname{span}\big(f_1(\boldsymbol{X}), f_2(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})\big) \;=\; \{\boldsymbol{D}\boldsymbol{\eta}\colon \boldsymbol{\eta} \in \mathbb{R}^p\}$$

of $\mathbb{R}^n$, the so-called *model space*. By definition of the LSE,

$$\widehat{\boldsymbol{Y}} := \widehat{f}(\boldsymbol{X}) = \boldsymbol{D}\widehat{\boldsymbol{\theta}} \;=\; \operatorname*{arg\,min}_{\boldsymbol{w} \in \boldsymbol{M}} \|\boldsymbol{Y} - \boldsymbol{w}\|.$$

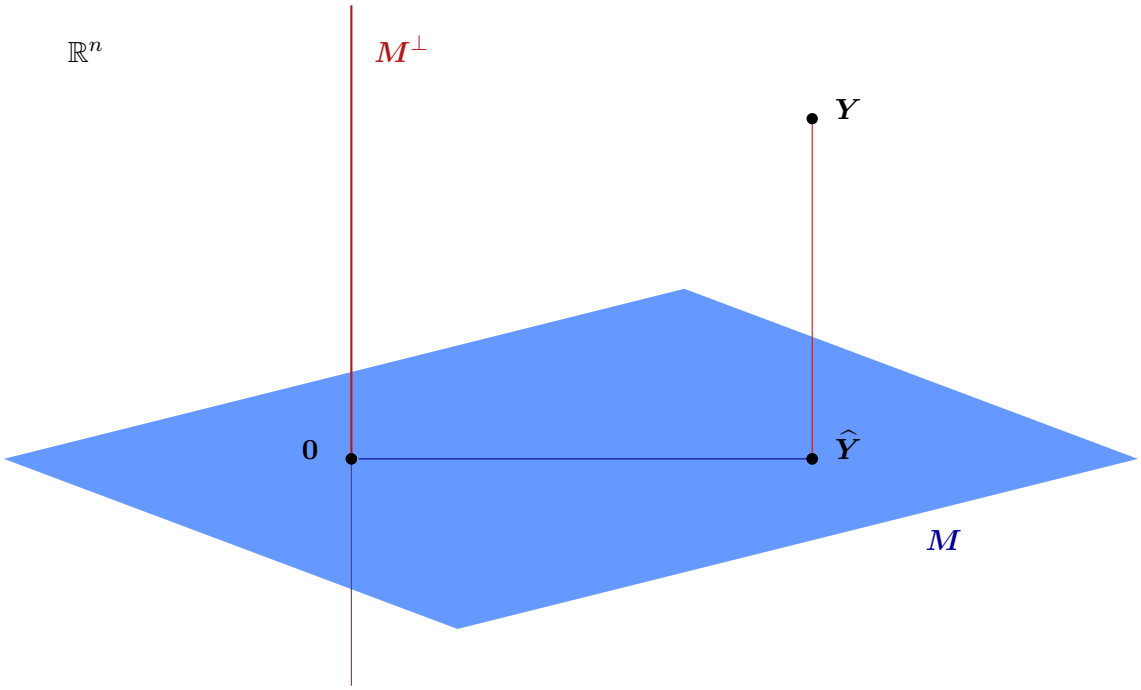

Figure 2.1: Data vector $\boldsymbol{Y}$ and its projection $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ onto the model space $\boldsymbol{M}$.

Hence, the "fitted vector" (vector of fitted values) $\widehat{\boldsymbol{Y}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the model space $\boldsymbol{M}$; see figure 2.1. That $\boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ is perpendicular to $\boldsymbol{M}$ follows from the following standard argument: Since $\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|$ minimizes $\|\boldsymbol{Y} - \boldsymbol{w}\|$ over all $\boldsymbol{w} \in \boldsymbol{M}$, for any direction $\boldsymbol{v} \in \boldsymbol{M}$,

$$0 \;=\; \frac{d}{dt}\Big|_{t=0} \|\boldsymbol{Y} - (\widehat{\boldsymbol{Y}} + t\boldsymbol{v})\|^2 \;=\; -2(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^\top \boldsymbol{v}.$$

If one uses a different parametrization of the linear model $\mathcal{F}$, that means, different basis functions f_j, both the design matrix $\boldsymbol{D}$ and the parameter vectors $\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}$ change, but the model space $\boldsymbol{M}$, the vectors $f(\boldsymbol{X}), \widehat{\boldsymbol{Y}} = \widehat{f}(\boldsymbol{X})$ and the function $\widehat{f}$ remain the same!

Remark 2.5 (Hat matrix). By means of Lemma 2.2, one may write the fitted vector $\widehat{\boldsymbol{Y}} = \boldsymbol{D}\widehat{\boldsymbol{\theta}}$ as

$$\widehat{\boldsymbol{Y}} \;=\; \boldsymbol{H}\boldsymbol{Y}$$

with the so-called *hat matrix*

$$\boldsymbol{H} \;:=\; \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top \;\in\; \mathbb{R}^{n\times n}.$$

This matrix describes the orthogonal projection of $\mathbb{R}^n$ onto the model space $\boldsymbol{M}$. The name indicates, that multiplying $\boldsymbol{Y}$ with $\boldsymbol{H}$ results in "putting a hat on $\boldsymbol{Y}$".

Exercise 2.6 (Projections and orthogonal projections). With this exercise we recall some facts from linear algebra.

(a) Let $\mathbb{X}$ be a real vector space, and suppose that $H : \mathbb{X} \to \mathbb{X}$ is a linear mapping. With $I(x) := x$ we define $\bar{H} := I - H$. Then, any point $x \in \mathbb{X}$ may be decomposed as $x = x_1 + x_2$ with $x_1 := H(x)$ and $x_2 := \bar{H}(x)$. Moreover, $\mathbb{X} = \mathbb{X}_1 + \mathbb{X}_2$ with $\mathbb{X}_1 := H(\mathbb{X})$ and $\mathbb{X}_2 := \bar{H}(\mathbb{X})$. Show that the following three statements are equivalent:

(a.1) $H^2 = H$;
(a.2) $\bar{H}^2 = \bar{H}$;
(a.3) $\mathbb{X}_1 \cap \mathbb{X}_2 = \{0\}$.

In case of (a.1-3), H is called a *(linear) projection*. Verify that in this case,

$$H(x) = x \text{ and } \bar{H}(x) = 0 \quad \text{if } x \in \mathbb{X}_1,$$
$$H(x) = 0 \text{ and } \bar{H}(x) = x \quad \text{if } x \in \mathbb{X}_2.$$

(b) Now let $\mathbb{X} = \mathbb{R}^2$ and $H(\boldsymbol{x}) := (x_1 - x_2, 0)^\top$. Show that $H^2 = H$, and determine the subspaces $\mathbb{X}_1, \mathbb{X}_2$.

(c) Now let $\mathbb{X} = \mathbb{R}^n$. A linear mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ corresponds to a matrix $\boldsymbol{H} \in \mathbb{R}^{n \times n}$. Show that the following statements are equivalent:

(c.1) $\boldsymbol{H} = \boldsymbol{H}^\top$ and $\boldsymbol{H}^2 = \boldsymbol{H}$;
(c.2) $\boldsymbol{H} = \boldsymbol{H}^\top \boldsymbol{H}$;
(c.3) $\mathbb{X}_1 \perp \mathbb{X}_2$.

In case of (c.1-3), H is called an *orthogonal projection*.

Exercise 2.7 (An abuse of regression?). In various contexts, people try to use linear regression in a somewhat unusual way to obtain "adjusted data". Suppose there are n units (for instance social services in n different municipalities) which should be ranked in terms of their costs. Let $Y_1, Y_2, \ldots, Y_n$ be the units' raw cost measures (for instance, expenditure for social welfare per year and inhabitant). To avoid comparing apples and oranges, one takes into account vectors $X_1, X_2, \ldots, X_n \in \mathcal{X}$ of covariates describing the units' circumstances (for instance, percentage of single parents, percentage of foreigners). The idea is that

$$Y_i = f_o(X_i) + \pi_i$$

with an unknown function $f_o : \mathcal{X} \to \mathbb{R}$ describing the costs to be expected under given circumstances and individual performances $\pi_1, \pi_2, \ldots, \pi_n$ which are the units' true contributions to the costs, low or high values of π_i meaning strong or poor performance, respectively. In order to reconstruct

$$\pi_i \;=\; Y_i - f_o(X_i),$$

one estimates f_o, for a given linear model $\mathcal{F}$, by

$$\widehat{f} \;\in\; \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

and estimates π_i by the residual

$$\widehat{\pi}_i \;:=\; Y_i - \widehat{f}(X_i).$$

What is your gut feeling about this approach?

(a) Suppose that our choice of $\mathcal{F}$ is appropriate, that means, $f_o \in \mathcal{F}$. Under what condition on $\boldsymbol{\pi}$ is

$$\widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}?$$

(b) Could you imagine potential reasons for this assumption to be violated?

2.2.2 Examples for $\widehat{\theta}$

In this subsection we derive the LSE in some specific models.

One-dimensional models. The simplest linear model is certainly

$$Y \;=\; \theta + \varepsilon$$

with an unknown parameter $\theta = \mathbb{E}(Y) \in \mathbb{R}$. In case of n observations, $\boldsymbol{D} = \mathbf{1} = (1, 1, \ldots, 1)^\top$, so $\boldsymbol{D}^\top \boldsymbol{Y} = \sum_{i=1}^n Y_i$, $\|\boldsymbol{D}\|^2 = n$ and $\widehat{\boldsymbol{\theta}}$ is just the sample mean of $\boldsymbol{Y}$,

$$\widehat{\boldsymbol{\theta}} \;=\; \bar{Y}.$$

Here and throughout these notes, we write $\bar{v} := n^{-1} \sum_{i=1}^n v_i$ for a vector $\boldsymbol{v} = (v_i)_{i=1}^n$.

Now, more generally, suppose that

$$Y \;=\; \theta f_1(X) + \varepsilon$$

with a given function $f_1 : \mathcal{X} \to \mathbb{R}$ and an unknown parameter $\theta \in \mathbb{R}$. In this case, $\boldsymbol{D} = f_1(\boldsymbol{X}) \in \mathbb{R}^n$, and

$$\widehat{\boldsymbol{\theta}} \;=\; \frac{\boldsymbol{D}^\top \boldsymbol{Y}}{\|\boldsymbol{D}\|^2}.$$

We encountered a specific example for such a one-dimensional model in Exercise 1.6: There, we considered the timber yield Y of a wild cherry tree in relation to the height $X(1)$ and maximal diameter $X(2)$ of its trunk. A simple geometrical consideration led to the linear model

$$Y \;=\; \theta X(1) X(2)^2 + \varepsilon,$$

so $f_1(X) = X(1)X(2)^2$.

One-way ANOVA (Example 1.1) Here $X \in \{1, 2, \ldots, L\}$ and $\boldsymbol{\theta} = (f(j))_{j=1}^L$. A natural estimator seems to be the vector $(\bar{Y}(j))_{j=1}^L$ of group-wise sample means

$$\bar{Y}(j) \;:=\; n(j)^{-1} \sum_{i\colon X_i = j} Y_i$$

with the group sizes

$$n(j) \;:=\; \#\{i\colon X_i = j\}.$$

Indeed, this is the least squares estimator: One can easily verify that

$$\boldsymbol{D}^\top \boldsymbol{D} \;=\; \operatorname{diag}\big(n(1), n(2), \ldots, n(L)\big) \;=\; \begin{bmatrix} n(1) & 0 & \cdots & 0 \\ 0 & n(2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n(L) \end{bmatrix}$$

and

$$\boldsymbol{D}^\top \boldsymbol{Y} \;=\; \Big(\sum_{i:\, X_i = j} Y_i \Big)_{j=1}^{L}.$$

The fitted vector $\widehat{\boldsymbol{Y}}$ is given by

$$\widehat{\boldsymbol{Y}} \;=\; \big(\bar{Y}(X_1), \bar{Y}(X_2), \ldots, \bar{Y}(X_n)\big)^\top.$$

The corresponding model space consists of all vectors $g(\boldsymbol{X}) = (g(X_i))_{i=1}^n$ with an arbitrary function $g : \{1, \ldots, L\} \to \mathbb{R}$.

Simple linear regression (Example 1.2). Here $\boldsymbol{D} = [\mathbf{1}, \boldsymbol{X}]$, so

$$\begin{aligned}
\boldsymbol{D}^\top \boldsymbol{D} \;&=\; \begin{bmatrix} \mathbf{1}^\top \mathbf{1} & \mathbf{1}^\top \boldsymbol{X} \\ \mathbf{1}^\top \boldsymbol{X} & \|\boldsymbol{X}\|^2 \end{bmatrix} \;=\; n \begin{bmatrix} 1 & \bar{X} \\ \bar{X} & n^{-1}\|\boldsymbol{X}\|^2 \end{bmatrix}, \\
(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \;&=\; \big(\|\boldsymbol{X}\|^2 - n\bar{X}^2\big)^{-1} \begin{bmatrix} n^{-1}\|\boldsymbol{X}\|^2 & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix}, \\
\boldsymbol{D}^\top \boldsymbol{Y} \;&=\; \begin{bmatrix} \mathbf{1}^\top \boldsymbol{Y} \\ \boldsymbol{X}^\top \boldsymbol{Y} \end{bmatrix} \;=\; \begin{bmatrix} n\bar{Y} \\ \boldsymbol{X}^\top \boldsymbol{Y} \end{bmatrix}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} \;&=\; \big(\|\boldsymbol{X}\|^2 - n\bar{X}^2\big)^{-1} \begin{bmatrix} n^{-1}\|\boldsymbol{X}\|^2 & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix} \begin{bmatrix} n\bar{Y} \\ \boldsymbol{X}^\top \boldsymbol{Y} \end{bmatrix} \\
&=\; \big(\|\boldsymbol{X}\|^2 - n\bar{X}^2\big)^{-1} \begin{bmatrix} \|\boldsymbol{X}\|^2 \bar{Y} - \boldsymbol{X}^\top \boldsymbol{Y} \bar{X} \\ \boldsymbol{X}^\top \boldsymbol{Y} - n\bar{X}\bar{Y} \end{bmatrix} \\
&=\; \big(\|\boldsymbol{X}\|^2 - n\bar{X}^2\big)^{-1} \begin{bmatrix} \big(\|\boldsymbol{X}\|^2 - n\bar{X}^2\big)\bar{Y} - \big(\boldsymbol{X}^\top \boldsymbol{Y} - n\bar{X}\bar{Y}\big)\bar{X} \\ \boldsymbol{X}^\top \boldsymbol{Y} - n\bar{X}\bar{Y} \end{bmatrix},
\end{aligned}$$

leading to

$$\widehat{a} \;=\; \bar{Y} - \widehat{b}\bar{X} \quad \text{and} \quad \widehat{b} \;=\; \frac{\boldsymbol{X}^\top \boldsymbol{Y} - n\bar{X}\bar{Y}}{\|\boldsymbol{X}\|^2 - n\bar{X}^2}. \tag{2.3}$$

This was a brute-force calculation using Lemma 2.2. The resulting fomula for $\widehat{b}$ should be used with care, because small rounding errors in $\bar{X}$ or $\bar{Y}$ may lead to strong aberrations in $\widehat{b}$.

With a bit more geometry, one can derive equivalent formulae more elegantly: The computation of the LSE is rather easy if the columns of the design matrix are orthogonal, because then $\boldsymbol{D}^\top \boldsymbol{D}$ is a diagonal matrix. This can always be achieved by orthogonalizing the columns of $\boldsymbol{D}$, e.g. via the Gram–Schmidt procedure. In the present example this works as follows:

We replace the vector $\boldsymbol{X}$ with

$$\tilde{\boldsymbol{X}} \;:=\; \boldsymbol{X} - \frac{\mathbf{1}^\top \boldsymbol{X}}{\mathbf{1}^\top \mathbf{1}} \mathbf{1} \;=\; \boldsymbol{X} - \bar{X}\mathbf{1},$$

because this vector is perpendicular to $\mathbf{1}$. In other words, we rewrite the model equation for a generic observation (X, Y) as

$$Y \;=\; a + bX + \varepsilon \;= \widetilde{a} + b(X - \bar{X}) + \varepsilon,$$

where $\widetilde{a} := a + b\bar{X}$. Hence, we consider the new basis functions $f_1(x) = 1$ and $f_2(x) = x - \bar{X}$. The corresponding design matrix equals

$$\widetilde{\boldsymbol{D}} \;=\; \big[\mathbf{1},\, \tilde{\boldsymbol{X}}\big] \;=\; \begin{bmatrix} 1 & X_1 - \bar{X} \\ 1 & X_2 - \bar{X} \\ \vdots & \vdots \\ 1 & X_n - \bar{X} \end{bmatrix},$$

and

$$\widetilde{\boldsymbol{D}}^\top \widetilde{\boldsymbol{D}} \;=\; \begin{bmatrix} n & 0 \\ 0 & \|\tilde{\boldsymbol{X}}\|^2 \end{bmatrix} \;=\; \begin{bmatrix} n & 0 \\ 0 & \|\boldsymbol{X}\|^2 - n\bar{X}^2 \end{bmatrix},$$
$$\widetilde{\boldsymbol{D}}^\top \boldsymbol{Y} \;=\; \begin{bmatrix} n\bar{Y} \\ \tilde{\boldsymbol{X}}^\top \boldsymbol{Y} \end{bmatrix} \;=\; \begin{bmatrix} n\bar{Y} \\ \boldsymbol{X}^\top \boldsymbol{Y} - n\bar{X}\bar{Y} \end{bmatrix}.$$

Thus

$$\begin{bmatrix} \widehat{\widetilde{a}} \\ \widehat{b} \end{bmatrix} \;=\; \begin{bmatrix} \bar{Y} \\ \tilde{\boldsymbol{X}}^\top \boldsymbol{Y} / \|\tilde{\boldsymbol{X}}\|^2 \end{bmatrix} \;=\; \begin{bmatrix} \bar{Y} \\ (\boldsymbol{X}^\top \boldsymbol{Y} - n\bar{X}\bar{Y})/(\|\boldsymbol{X}\|^2 - n\bar{X}^2) \end{bmatrix}.$$

Since $\widehat{\widetilde{a}} = \widehat{a} + \widehat{b}\bar{X}$, these formulae imply the expressions (2.3).

Remark: The estimated regression function $\widehat{f}$ may be written as

$$\widehat{f}(x) \;=\; \bar{Y} + \widehat{b}(x - \bar{X}).$$

In particular, $\widehat{f}(\bar{X}) = \bar{Y}$, so the regression line contains the barycenter $(\bar{X}, \bar{Y})$ of all data pairs (X_i, Y_i).

With the sample standard deviation

$$S(\boldsymbol{V}) \;:=\; \sqrt{(n-1)^{-1} \sum_{i=1}^{n} (V_i - \bar{V})^2}$$

of an arbitrary vector $\boldsymbol{V} \in \mathbb{R}^n$ and the sample correlation coefficient

$$r(\boldsymbol{X}, \boldsymbol{Y}) \;:=\; \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \;\in\; [-1, 1]$$

one can also write

$$\widehat{b} \;=\; r(\boldsymbol{X}, \boldsymbol{Y})\, \frac{S(\boldsymbol{Y})}{S(\boldsymbol{X})}.$$

Hence, the regression line always lies between the graphs of the functions

$$x \;\mapsto\; \bar{Y} \pm \frac{S(\boldsymbol{Y})}{S(\boldsymbol{X})}(x - \bar{X}); \tag{2.4}$$

see Figure 2.2.

The inequality $r(\boldsymbol{X}, \boldsymbol{Y}) \in [-1, 1]$ is a consequence of the Cauchy–Schwarz inequality. For $r(\boldsymbol{X}, \boldsymbol{Y})$ is the standard inner product of the unit vectors $\boldsymbol{u} := \|\boldsymbol{X} - \bar{X}\mathbf{1}\|^{-1}(\boldsymbol{X} - \bar{X}\mathbf{1})$ and $\boldsymbol{v} := \|\boldsymbol{Y} - \bar{Y}\mathbf{1}\|^{-1}(\boldsymbol{Y} - \bar{Y}\mathbf{1})$. The extremal cases $r(\boldsymbol{X}, \boldsymbol{Y}) = \pm 1$ correspond to $\boldsymbol{v}$ being a positive or negative multiple of $\boldsymbol{u}$, and this is equivalent to all observations (X_i, Y_i) lying on a straight line with positive or negative slope, respectively.

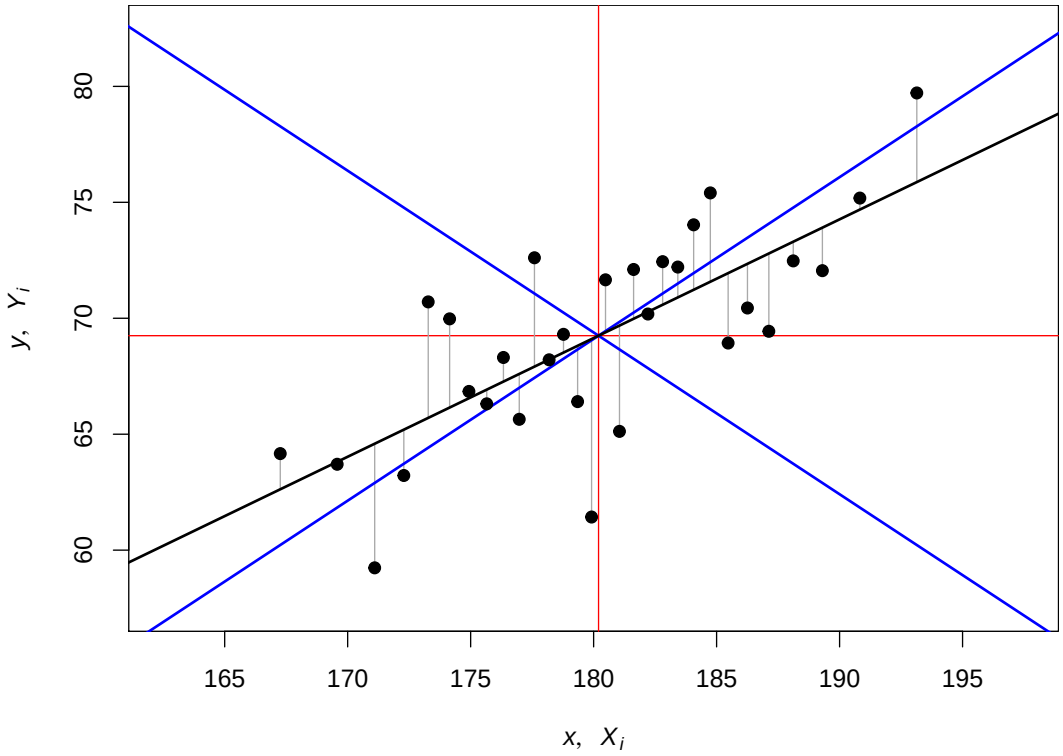

Figure 2.2: Position of the regression line (black). The red lines indicate $\bar{X}$ and $\bar{Y}$, the blue lines are given by (2.4).

Exercise 2.8 (Orthogonal polynomials, I). The Gram–Schmidt procedure is just one of several possibilities to orthogonalize the columns of the design matrix. In particular, for polynomial regression there is an elegant alternative based on so-called three-term recursions.

Preliminary consideration: Let $p_0(x) := 1$ and $p_1(x) := x - b_0$ for some $b_0 \in \mathbb{R}$. Now, we define inductively

$$p_{k+1}(x) \;:=\; xp_k(x) - b_k p_k(x) - c_k p_{k-1}(x)$$

for $k = 1, 2, 3, \ldots$ with certain real numbers b_k, c_k. One can verify that for each $k \in \mathbb{N}_0$, the function $p_k(x)$ is a polynomial of degree k with leading coefficient 1. In particular, each polynomial $p(x)$ of order k is a linear combination of $p_0(x), \ldots, p_k(x)$.

Now, let $\boldsymbol{X} \in \mathbb{R}^n$ with $\#\{X_1, X_2, \ldots, X_n\} \geq d + 1$, where $d \in \mathbb{N}$. Show that the constants b_0 and b_k, c_k $(1 \leq k < d)$ can be chosen such that the vectors $p_0(\boldsymbol{X})$, $p_1(\boldsymbol{X})$, …, $p_d(\boldsymbol{X})$ are orthogonal.

Hint: The considerations for simple linear regression show already that $b_0 = \bar{X}$. Now suppose that for some $1 \leq k < d$ the vectors $p_0(\boldsymbol{X}), \ldots, p_k(\boldsymbol{X})$ are orthogonal.

(i) Show that $p_{k+1}(\boldsymbol{X})^\top p_j(\boldsymbol{X}) = 0$ for $0 \leq j \leq k - 2$, no matter how b_k and c_k are chosen.

(ii) Determine b_k and c_k such that $p_{k+1}(\boldsymbol{X})^\top p_k(\boldsymbol{X}) = p_{k+1}(\boldsymbol{X})^\top p_{k-1}(\boldsymbol{X}) = 0$.

Exercise 2.9 (Orthogonal polynomials, II). Write a computer program $\mathsf{OrthPoly}(\boldsymbol{X}, d)$ with input arguments $\boldsymbol{X} \in \mathbb{R}^n$ and $d \in \mathbb{N}$, performing the following tasks: First it checks whether

$\#\{X_1, X_2, \ldots, X_n\} \geq d+1$. If no, it returns a warning. If yes, it returns the design matrix

$$\boldsymbol{D} \;=\; [p_0(\boldsymbol{X}), p_1(\boldsymbol{X}), \ldots, p_d(\boldsymbol{X})]$$

with the polynomials $p_0, p_1, \ldots, p_d(\boldsymbol{X})$ as in Exercise 2.8.

In addition, it should return upper triangular matrices $\boldsymbol{B}, \boldsymbol{\Theta} \in \mathbb{R}^{(d+1)\times(d+1)}$ such that

$$p_k(x) \;\equiv\; \sum_{j=0}^{k} B_{j+1,k+1} x^j$$

and

$$x^k \;\equiv\; \sum_{j=0}^{k} \Theta_{j+1,k+1} p_j(x)$$

for $k = 0, 1, \ldots, d$. These requirements on $\boldsymbol{\Theta}, \boldsymbol{B}$ are equivalent to saying that for arbitrary vectors $\boldsymbol{\theta}, \boldsymbol{\beta} \in \mathbb{R}^{d+1}$,

$$\sum_{k=0}^{d} \theta_{k+1} p_k(x) \;\equiv\; \sum_{j=0}^{d} (\boldsymbol{B}\boldsymbol{\theta})_{j+1} x^j,$$

and

$$\sum_{k=0}^{d} \beta_{k+1} x^k \;\equiv\; \sum_{j=0}^{d} (\boldsymbol{\Theta}\boldsymbol{\beta})_{j+1} p_j(x).$$

One-way ANCOVA (Example 1.4). Here, $X = (C, W)$ with the categorial covariate $C \in \{1, 2, \ldots, L\}$ and the numerical covariate $W \in \mathbb{R}$. The basis functions $f_j(c, w) := 1_{[c=j]}$ for $1 \leq j \leq L$ and $f_{L+1}(c, w) := w$ yield the design matrix $\boldsymbol{D}$ with columns

$$\boldsymbol{D}_j \;=\; \big(1_{[C_i=j]}\big)_{i=1}^{n}, \quad 1 \leq j \leq L,$$

and

$$\boldsymbol{D}_{L+1} \;=\; \boldsymbol{W} \;=\; (W_i)_{i=1}^{n}.$$

As mentioned for one-way ANOVA, the columns $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_L$ are orthogonal. So the Gram–Schmidt procedure would replace $\boldsymbol{W}$ with

$$\widetilde{\boldsymbol{W}} \;:=\; \boldsymbol{W} - \sum_{j=1}^{L} \frac{\boldsymbol{D}_j^\top \boldsymbol{W}}{\|\boldsymbol{D}_j\|^2} \boldsymbol{D}_j \;=\; \boldsymbol{W} - \sum_{j=1}^{L} \bar{W}(j) \boldsymbol{D}_j \;=\; \begin{bmatrix} W_1 - \bar{W}(C_1) \\ W_2 - \bar{W}(C_2) \\ \vdots \\ W_n - \bar{W}(C_n) \end{bmatrix},$$

where

$$\bar{W}(j) \;:=\; n(j)^{-1} \sum_{i:\, C_i = j} W_i \quad \text{and} \quad n(j) := \#\{i \colon C_i = j\}. \tag{2.5}$$

Replacing $\boldsymbol{W}$ with $\widetilde{\boldsymbol{W}}$ means to rewrite the model equation for (Y, C, W) as follows:

$$Y \;=\; a(C) + bW + \varepsilon \;=\; \widetilde{a}(C) + b(W - \bar{W}(C)) + \varepsilon$$

with

$$\widetilde{a}(c) \;:=\; a(c) + b\bar{W}(c).$$

The corresponding design matrix $\widetilde{\boldsymbol{D}}$ has the orthogonal columns $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_L, \widetilde{\boldsymbol{W}}$, and the LSE for $\widetilde{\boldsymbol{\theta}} = \big(\widetilde{a}(1), \ldots, \widetilde{a}(L), b\big)^\top$ is given by

$$\widehat{\widetilde{\boldsymbol{\theta}}} \;=\; \begin{bmatrix} \bar{Y}(1) \\ \vdots \\ \bar{Y}(L) \\ \widetilde{\boldsymbol{W}}^\top \boldsymbol{Y} / \|\widetilde{\boldsymbol{W}}\|^2 \end{bmatrix},$$

where $\bar{Y}(j)$ is defined as $\bar{W}(j)$ in (2.5) with $\boldsymbol{Y}$ in place of $\boldsymbol{W}$. Consequently,

$$\widehat{a}(j) \;=\; \bar{Y}(j) - \widehat{b}\bar{W}(j), \quad 1 \le j \le L,$$

and

$$\widehat{b} \;=\; \widetilde{\boldsymbol{W}}^\top \boldsymbol{Y} / \|\widetilde{\boldsymbol{W}}\|^2 \;=\; \sum_{i=1}^n (W_i - \bar{W}(C_i))Y_i \Big/ \sum_{i=1}^n (W_i - \bar{W}(C_i))^2.$$

Hence, the groupwise means of Y are corrected for the potential influence of W on Y and take into account potential differences of the groupwise means of W.

2.2.3 The Coefficient of Determination

A descriptive measure of the fit $\widehat{\boldsymbol{Y}}$ is the *coefficient of determination*

$$R^2 \;:=\; 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \;=\; 1 - \frac{\big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2}{\big\|\boldsymbol{Y} - \bar{Y}\mathbf{1}\big\|^2},$$

also called *'R-squared'*. The differences

$$Y_i - \widehat{Y}_i$$

are the so-called *residuals*, so R^2 compares the *residual sum of squares* $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$ with the *total sum of squares* $\sum_{i=1}^n (Y_i - \bar{Y})^2$ of the Y-values. One could say, R^2 is the percentage of variability in the Y-values which can be "explained" by the covariate X.

Obviously, $R^2 \le 1$, where $R^2 = 1$ if and only if $\widehat{\boldsymbol{Y}} = \boldsymbol{Y}$. Typically, $R^2 \ge 0$. The latter inequality is guaranteed if the linear space $\mathcal{F}$ contains the constant functions. A bit more generally, $R^2 \ge 0$ whenever the model space $\boldsymbol{M}$ contains the constant vector $\mathbf{1}$. Because then, the vector $\widehat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$ lies in the model space $\boldsymbol{M}$, and this implies that it is perpendicular to $\boldsymbol{Y} - \widehat{\boldsymbol{Y}} \in \boldsymbol{M}^\perp$. Hence,

$$\big\|\boldsymbol{Y} - \bar{Y}\mathbf{1}\big\|^2 \;=\; \big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2 + \big\|\widehat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\big\|^2.$$

The complete geometrical picture is that $\boldsymbol{Y}$ may be represented as a sum of three pairwise orthogonal vectors,

$$\boldsymbol{Y} \;=\; \underbrace{\bar{Y}\mathbf{1}}_{\in\, \mathrm{span}(\mathbf{1})} + \underbrace{\widehat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}}_{\in\, \boldsymbol{M} \cap \mathbf{1}^\perp} + \underbrace{\boldsymbol{Y} - \widehat{\boldsymbol{Y}}}_{\in\, \boldsymbol{M}^\perp}.$$

In particular,

$$R^2 = 1 - \frac{\big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2}{\big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2 + \big\|\widehat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\big\|^2} \;=\; \frac{\big\|\widehat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\big\|^2}{\big\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\big\|^2 + \big\|\widehat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\big\|^2} \;\in\; [0, 1].$$

Exercise 2.10. Show that in the model of simple linear regression, the coefficient of determination equals the square of the sample correlation coefficient,

$$R^2 \;=\; r(\boldsymbol{X},\boldsymbol{Y})^2.$$

When comparing different models for the same data set, it is advisable to work with the *adjusted coefficient of determination* (*adjusted R-squared*). It takes into account the complexity of a linear model $\mathcal{F}$, i.e. its dimension. For if we increase the set of basis functions, we usually decrease the residual sum of squares as well. The residual sum of squares $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$ involves p parameters, while the total sum of squares $\sum_{i=1}^n (Y_i - \bar{Y})^2$ corresponds to a one-dimensional model. Thus, we define the adjusted R-squared to be

$$R^2_{\text{adj}} \;:=\; 1 - \frac{(n-p)^{-1}\sum_i (Y_i - \widehat{Y}_i)^2}{(n-1)^{-1}\sum_i (Y_i - \bar{Y})^2} \;=\; 1 - \frac{(n-p)^{-1}\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2}{(n-1)^{-1}\|\boldsymbol{Y} - \bar{Y}\mathbf{1}\|^2}.$$

This formula has also a geometric interpretation: $\boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ is the orthogonal projection of the data vector $\boldsymbol{Y}$ onto the $(n-p)$-dimensional linear subspace $\boldsymbol{M}^\perp$ of $\mathbb{R}^n$, while $\boldsymbol{Y} - \bar{Y}\mathbf{1}$ is its orthogonal projection onto the $(n-1)$-dimensional linear subspace $\mathbf{1}^\perp$.

Exercise 2.11. The data set 'Trees.txt' contains for $n = 31$ wild cherry trees the values of the variables

$$\begin{aligned} Y: &\quad \text{its timber yield,}\\ X(1): &\quad \text{the height of its trunk,}\\ X(2): &\quad \text{its maximal diameter (at waist height).} \end{aligned}$$

(a) Determine by means of a suitable computer program the LSE and the values of R^2 and R^2_{adj} for each of the following models:

$$\begin{aligned} Y &= \theta\, X(1)X(2)^2 + \varepsilon,\\ Y &= \theta\, X(1)^3 + \varepsilon,\\ Y &= \theta\, X(2)^3 + \varepsilon,\\ Y &= a + b(1)X(1) + b(2)X(2) + \varepsilon,\\ \log Y &= a + b(1)\log X(1) + b(2)\log X(2) + \varepsilon,\\ \log Y &= a + \log X(1) + 2\log X(2) + \varepsilon. \end{aligned}$$

Compare and discuss the results.

(b) Which of the two covariates $X(1)$ and $X(2)$ is more reliable to predict the response Y? Can you imagine a biological reason for that? Or a technical one?

Exercise 2.12. Apply the model of one-way ANCOVA to the data set 'Goats.txt'. The latter is about the weight gain of goats in relation to their initial weight and the variant of an anti-worm treatment. Compute the LSE by means of the formulae provided in the lecture, and compare your results with the output of some statistics software.

Exercise 2.13. The data set 'Exam.txt' contains the exam results of $n = 88$ students in five different subjects. Analyze to what extent the results in one subject may be predicted by an affine function of the results in the other four subjects.

Exercise 2.14. The data set 'BrainSize.txt' contains for $n = 40$ students three different IQ scores plus the values of other covariates such as gender, body height, size and density of brain (based on magnetic resonance images). Analyze the (apparent) connection between one of the IQ scores (Y) ane the covariates gender, body height and brain size. What happens if you leave out some of the covariates? Specify your model equation for each analysis and interpret the results.

2.2.4 The Precision of $\widehat{\theta}$

Throughout these lecture notes, we use standard definitions and properties of expectations and covariances of matrix- and vector-valued random variables. These are collected in Section A.2 in the appendix.

The assumption that $\mathbb{E}(\varepsilon_i) = 0$ for all i may be rewritten as

$$\mathbb{E}(\boldsymbol{\varepsilon}) \;=\; \mathbf{0}.$$

If we assume that the errors ε_i are uncorrelated with the same finite standard deviation $\sigma \geq 0$, then

$$\mathrm{Var}(\boldsymbol{\varepsilon}) \;=\; \sigma^2 \boldsymbol{I}.$$

This has the following consequences for the LSE:

Lemma 2.15. *If $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, then $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$, that means,*

$$\mathbb{E}(\widehat{\boldsymbol{\theta}}) \;=\; \boldsymbol{\theta}.$$

In case of $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$, it holds that

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) \;=\; \sigma^2 (\boldsymbol{D}^\top \boldsymbol{D})^{-1}.$$

Proof of Lemma 2.15. Recall that $\widehat{\boldsymbol{\theta}} = \boldsymbol{A}\boldsymbol{Y}$ with

$$\boldsymbol{A} \;:=\; (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \;\in\; \mathbb{R}^{p \times n}.$$

If $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, it follows from the general rules of expected values and $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ that

$$\begin{aligned}
\mathbb{E}(\widehat{\boldsymbol{\theta}}) \;&=\; \mathbb{E}(\boldsymbol{A}\boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{A}\boldsymbol{\varepsilon}) \\
&=\; \boldsymbol{A}\boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{A}\,\mathbb{E}(\boldsymbol{\varepsilon}) \\
&=\; (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\theta} \\
&=\; \boldsymbol{\theta}.
\end{aligned}$$

Moreover, in case of $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$,

$$\begin{aligned}
\mathrm{Var}(\widehat{\boldsymbol{\theta}}) &= \mathrm{Var}(\boldsymbol{A}\boldsymbol{Y}) \\
&= \boldsymbol{A}\,\mathrm{Var}(\boldsymbol{Y})\boldsymbol{A}^\top \\
&= \sigma^2 \boldsymbol{A}\boldsymbol{A}^\top \\
&= \sigma^2 (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{D} (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \\
&= \sigma^2 (\boldsymbol{D}^\top \boldsymbol{D})^{-1}.
\end{aligned}$$

□

Simple linear regression (Example 1.2). To avoid tedious calculations, we apply the orthogonalization trick and write

$$Y \;=\; a + bX + \varepsilon \;=\; \widetilde{a} + b(X - \bar{X}) + \varepsilon.$$

With $\widetilde{\boldsymbol{X}} := (X_i - \bar{X})_{i=1}^n$ and $\widetilde{\boldsymbol{D}} := [\boldsymbol{1}, \widetilde{\boldsymbol{X}}]$, it follows from $\widetilde{\boldsymbol{D}}^\top \widetilde{\boldsymbol{D}} = \mathrm{diag}(n, \|\widetilde{\boldsymbol{X}}\|^2)$ that the LSE of $(\widetilde{a}, b)^\top$ equals

$$\begin{bmatrix} \widehat{\widetilde{a}} \\ \widehat{b} \end{bmatrix} \;=\; \begin{bmatrix} \bar{Y} \\ \widetilde{\boldsymbol{X}}^\top \boldsymbol{Y} / \|\widetilde{\boldsymbol{X}}\|^2 \end{bmatrix}.$$

Moreover, in case of homoscedastic errrors, Lemma 2.15 implies that the covariance matrix of this LSE $(\bar{Y}, \widehat{b})^\top$ is given by

$$\sigma^2 \begin{bmatrix} 1/n & 0 \\ 0 & 1/\|\widetilde{\boldsymbol{X}}\|^2 \end{bmatrix}.$$

This implies that the LSE for $\boldsymbol{\theta} = (a, b)^\top = (\widetilde{a} - b\bar{X}, b)^\top$ is given by

$$\widehat{\boldsymbol{\theta}} \;=\; \begin{bmatrix} \bar{Y} - \widehat{b}\bar{X} \\ \widehat{b} \end{bmatrix}$$

with covariance matrix

$$\begin{aligned}
\begin{bmatrix} \mathrm{Var}(\bar{Y} - \bar{X}\widehat{b}) & \mathrm{Cov}(\bar{Y} - \bar{X}\widehat{b},\, \widehat{b}) \\ \mathrm{Cov}(\bar{Y} - \bar{X}\widehat{b},\, \widehat{b}) & \mathrm{Var}(\widehat{b}) \end{bmatrix} &= \begin{bmatrix} \mathrm{Var}(\bar{Y}) + \bar{X}^2\,\mathrm{Var}(\widehat{b}) & -\bar{X}\,\mathrm{Var}(\widehat{b}) \\ -\bar{X}\,\mathrm{Var}(\widehat{b}) & \mathrm{Var}(\widehat{b}) \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} 1/n + \bar{X}^2/\|\widetilde{\boldsymbol{X}}\|^2 & -\bar{X}/\|\widetilde{\boldsymbol{X}}\|^2 \\ -\bar{X}/\|\widetilde{\boldsymbol{X}}\|^2 & 1/\|\widetilde{\boldsymbol{X}}\|^2 \end{bmatrix},
\end{aligned}$$

because $\mathrm{Cov}(\bar{Y}, \widehat{b}) = 0$. For a fixed number $x \in \mathbb{R}$, a natural estimator for $f(x)$ is given by

$$\widehat{f}(x) \;=\; \widehat{a} + \widehat{b}x \;=\; \bar{Y} + \widehat{b}(x - \bar{X}),$$

so $\mathbb{E}\,\widehat{f}(x) = f(x)$ and

$$\mathrm{Var}(\widehat{f}(x)) \;=\; \sigma^2\Big(\frac{1}{n} + \frac{(x - \bar{X})^2}{\|\widetilde{\boldsymbol{X}}\|^2}\Big) \;=\; \sigma^2\Big(\frac{1}{n} + \frac{(x - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\Big).$$

Hence, this variance is minimal at $x = \bar{X}$, and it is a strictly increasing, quadratic function of $|x - \bar{X}|$.

One-way ANCOVA (Example 1.4). Again we work with the modified model equation

$$Y \;=\; a(C) + bW + \varepsilon \;=\; \widetilde{a}(C) + b(W - \bar{W}(C)) + \varepsilon.$$

With the groupwise centered vector $\widetilde{\boldsymbol{W}} := (W_i - \bar{W}(C_i))_{i=1}^n$, the LSE of $\big(\widetilde{a}(1), \ldots, \widetilde{a}(L), b\big)^\top$ equals

$$\begin{bmatrix} \bar{Y}(1) \\ \vdots \\ \bar{Y}(L) \\ \widetilde{\boldsymbol{W}}^\top \boldsymbol{Y} / \|\widetilde{\boldsymbol{W}}\|^2 \end{bmatrix},$$

and its covariance matrix is

$$\sigma^2 \begin{bmatrix} \mathrm{diag}\,\big(n(1)^{-1}, \ldots, n(L)^{-1}\big) & 0 \\ 0 & \|\widetilde{\boldsymbol{W}}\|^{-2} \end{bmatrix}$$

(in case of homoscedastic errors). Hence, the LSE of $\boldsymbol{\theta} = \big(a(1), \ldots, a(L), b\big)^\top$ equals

$$\widehat{\boldsymbol{\theta}} \;=\; \begin{bmatrix} \bar{Y}(1) - \widehat{b}\bar{W}(1) \\ \vdots \\ \bar{Y}(L) - \widehat{b}\bar{W}(L) \\ \widehat{b} \end{bmatrix},$$

and its covariances are as follows:

$$\begin{aligned} \mathrm{Var}\big(\widehat{b}\big) \;&=\; \frac{\sigma^2}{\|\widetilde{\boldsymbol{W}}\|^2}, \\ \mathrm{Cov}\big(\widehat{a}(c), \widehat{a}(d)\big) \;&=\; \sigma^2\Big(\frac{1_{[c=d]}}{n(c)} + \frac{\bar{W}(c)\bar{W}(d)}{\|\widetilde{\boldsymbol{W}}\|^2}\Big), \\ \mathrm{Cov}\big(\widehat{a}(c), \widehat{b}\big) \;&=\; -\sigma^2 \frac{\bar{W}(c)}{\|\widetilde{\boldsymbol{W}}\|^2}. \end{aligned}$$

In particular, for two different categories $c, d \in \{1, \ldots, L\}$,

$$\begin{aligned} \mathrm{Var}\big(\widehat{a}(c) - \widehat{a}(d)\big) \;&=\; \mathrm{Var}\Big(\bar{Y}(c) - \bar{Y}(d) - \big(\bar{W}(c) - \bar{W}(d)\big)\widehat{b}\Big) \\ &=\; \sigma^2\Big(\frac{1}{n(c)} + \frac{1}{n(d)} + \frac{(\bar{W}(c) - \bar{W}(d))^2}{\|\widetilde{\boldsymbol{W}}\|^2}\Big). \end{aligned}$$

Exercise 2.16. Consider the model of one-way ANCOVA (Example 1.4). For two different categories $c, d \in \{1, \ldots, L\}$, we consider estimators $\widehat{\gamma} = \widehat{\gamma}(\text{data})$ of the difference $\gamma := a(c) - a(d)$ and quantify their imprecision with the mean squared error

$$\mathrm{MSE}(\widehat{\gamma}) \;:=\; \mathbb{E}\big((\widehat{\gamma} - \gamma)^2\big).$$

Compare the estimator $\widehat{a}(k) - \widehat{a}(j)$ with the naive estimator $\bar{Y}(c) - \bar{Y}(d)$ resulting from the model of one-way ANOVA, ignoring W. When is the LSE strictly better than the naive one in terms of MSE?

2.3 Estimation of σ

The observation vector $\boldsymbol{Y}$ can be written as

$$\boldsymbol{Y} \;=\; \widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\varepsilon}}.$$

Here $\widehat{\boldsymbol{Y}} = \widehat{f}(\boldsymbol{X}) = \boldsymbol{H}\boldsymbol{Y}$ is the fitted vector, with $\boldsymbol{H}$ denoting the hat matrix, and $\widehat{\boldsymbol{\varepsilon}}$ is the so-called *residual vector*

$$\widehat{\boldsymbol{\varepsilon}} \;:=\; \boldsymbol{Y} - \widehat{\boldsymbol{Y}} \;=\; (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

The i-th component of $\widehat{\boldsymbol{\varepsilon}}$ is $Y_i - \widehat{Y}_i$, called the i-th residual. Geometrically speaking, $\widehat{\boldsymbol{Y}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the model space $\boldsymbol{M}$, and $\widehat{\boldsymbol{\varepsilon}}$ is the orthogonal projection of $\boldsymbol{Y}$ onto the orthogonal complement $\boldsymbol{M}^{\perp}$ of $\boldsymbol{M}$.

The least squares estimator $\widehat{\boldsymbol{\theta}}$ depends only on $\widehat{\boldsymbol{Y}}$, because $\widehat{\boldsymbol{\varepsilon}}$ is perpendicular to $\boldsymbol{M}$, which is equivalent to $\boldsymbol{D}^{\top}\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$, whence

$$\widehat{\boldsymbol{\theta}} \;=\; (\boldsymbol{D}^{\top}\boldsymbol{D})^{-1}\boldsymbol{D}^{\top}(\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\varepsilon}}) \;=\; (\boldsymbol{D}^{\top}\boldsymbol{D})^{-1}\boldsymbol{D}^{\top}\widehat{\boldsymbol{Y}}.$$

On the other hand, $\widehat{\boldsymbol{\varepsilon}}$ depends only on the error vector $\boldsymbol{\varepsilon}$, because $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{D} = \boldsymbol{0}$, whence

$$\widehat{\boldsymbol{\varepsilon}} \;=\; (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \;=\; (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}.$$

Now let's assume that all errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have mean zero and variance $\sigma^2 < \infty$. If an oracle would give us the error vector $\boldsymbol{\varepsilon}$ or at least its Euclidean norm $\|\boldsymbol{\varepsilon}\|$, a natural unbiased estimator of σ^2 would be given by $\|\boldsymbol{\varepsilon}\|^2/n$. One could read this estimator as the "squared norm of the error vector divided by its dimension". Since we know at least the projection $\widehat{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon}$ onto $\boldsymbol{M}^{\perp}$, a feasible estimator for σ^2 is given by

$$\widehat{\sigma}^2 \;:=\; \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{\dim(\boldsymbol{M}^{\perp})} \;=\; \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{n-p} \;=\; \frac{\|\boldsymbol{Y}\|^2 - \|\widehat{\boldsymbol{Y}}\|^2}{n-p}.$$

The following Theorem provides some statistical properties of $\widehat{\sigma}^2$.

Theorem 2.17. *If $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$ with $0 \le \sigma < \infty$, then*

$$\mathbb{E}(\widehat{\sigma}^2) \;=\; \sigma^2.$$

If the errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent with $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ and $\mathbb{E}(\varepsilon_i^4) \le K\sigma^4$ for all i and some real constant K, then

$$\mathrm{Var}(\widehat{\sigma}^2) \;\le\; \frac{(K-3)^+ + 2}{n-p}\,\sigma^4.$$

Remark 2.18. In case of independent errors with a Gaussian distribution $\mathrm{N}(0, \sigma^2)$, we have the equation $\mathbb{E}(\varepsilon_i^4) = 3\sigma^4$. Then, the proof of Theorem 2.17 reveals that $\mathrm{Var}(\widehat{\sigma}^2) = 2\sigma^4/(n-p)$; see also the next chapter.

Proof of Theorem 2.17. With the hat matrix $\boldsymbol{H}$, the matrix $\bar{\boldsymbol{H}} := \boldsymbol{I} - \boldsymbol{H}$ describes the orthogonal projection onto $\boldsymbol{M}^\perp$; see also Exercise 2.6. In particular,

$$\bar{\boldsymbol{H}}^\top \;=\; \bar{\boldsymbol{H}} \;=\; \bar{\boldsymbol{H}}^2.$$

(This could also be verified by direct calculations.) Hence,

$$\|\widehat{\boldsymbol{\varepsilon}}\|^2 \;=\; \|\bar{\boldsymbol{H}}\boldsymbol{\varepsilon}\|^2 \;=\; \boldsymbol{\varepsilon}^\top \bar{\boldsymbol{H}}^\top \bar{\boldsymbol{H}}\boldsymbol{\varepsilon} \;=\; \boldsymbol{\varepsilon}^\top \bar{\boldsymbol{H}}\boldsymbol{\varepsilon} \;=\; \sum_{i,j=1}^n \bar{H}_{ij}\varepsilon_i\varepsilon_j.$$

Now, it follows from $\mathbb{E}(\varepsilon_i\varepsilon_j) = 0$ whenever $i \neq j$ and $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ that

$$\mathbb{E}\left(\|\widehat{\boldsymbol{\varepsilon}}\|^2\right) \;=\; \sum_{i,j=1}^n \bar{H}_{ij}\,\mathbb{E}(\varepsilon_i\varepsilon_j) \;=\; \sigma^2 \sum_{i=1}^n \bar{H}_{ii} \;=\; \sigma^2\,\mathrm{trace}(\bar{\boldsymbol{H}}) \;=\; \sigma^2(n-p),$$

and this yields the desired equation $\mathbb{E}(\widehat{\sigma}^2) = \sigma^2$. Here we used the equation

$$\mathrm{trace}(\bar{\boldsymbol{H}}) \;=\; \dim(\boldsymbol{M}^\perp) \;=\; n - p.$$

The latter may be verified by elementary calculations, but it is also a general fact for matrices describing orthogonal projections; see Exercise 2.19.

Concerning the variance of $\widehat{\sigma}^2$, the variance-covariance formula for weighted sums of random variables implies that

$$\mathrm{Var}\big(\|\widehat{\boldsymbol{\varepsilon}}\|^2\big) \;=\; \mathrm{Var}\Big(\sum_{i,j=1}^n \bar{H}_{ij}\varepsilon_i\varepsilon_j\Big) \;=\; \sum_{i,j,k,\ell=1}^n \bar{H}_{ij}\bar{H}_{k\ell}\,\mathrm{Cov}(\varepsilon_i\varepsilon_j, \varepsilon_k\varepsilon_\ell).$$

Elementary calculations (Exercise 2.20) show that

$$\mathrm{Cov}(\varepsilon_i\varepsilon_j, \varepsilon_k\varepsilon_\ell) \;=\; \begin{cases} \mathbb{E}(\varepsilon_i^4) - \sigma^4 & \text{if } i = j = k = \ell, \\ \sigma^4 & \text{if } i \neq j \text{ and } \{i,j\} = \{k,\ell\}, \\ 0 & \text{else.} \end{cases}$$

These formulae, together with symmetry of $\bar{\boldsymbol{H}}$, lead to

$$\begin{aligned} \mathrm{Var}\left(\|\widehat{\boldsymbol{\varepsilon}}\|^2\right) \;&=\; \sum_{i=1}^n \bar{H}_{ii}^2\left(\mathbb{E}(\varepsilon_i^4) - \sigma^4\right) + 2\sigma^4 \sum_{i,j=1}^n 1_{[i\neq j]}\bar{H}_{ij}^2 \\ &=\; \sum_{i=1}^n \bar{H}_{ii}^2\left(\mathbb{E}(\varepsilon_i^4) - 3\sigma^4\right) + 2\sigma^4 \sum_{i,j=1}^n \bar{H}_{ij}^2 \\ &=\; \sum_{i=1}^n \bar{H}_{ii}^2\left(\mathbb{E}(\varepsilon_i^4) - 3\sigma^4\right) + 2\sigma^4\,\mathrm{trace}(\bar{\boldsymbol{H}}^\top\bar{\boldsymbol{H}}) \\ &=\; \sum_{i=1}^n \bar{H}_{ii}^2\left(\mathbb{E}(\varepsilon_i^4) - 3\sigma^4\right) + 2\sigma^4\,\mathrm{trace}(\bar{\boldsymbol{H}}) \\ &=\; \sum_{i=1}^n \bar{H}_{ii}^2\left(\mathbb{E}(\varepsilon_i^4) - 3\sigma^4\right) + 2\sigma^4(n-p). \end{aligned}$$

But $0 \leq \bar{H}_{ii} \leq 1$, so $\bar{H}_{ii}^2 \leq \widehat{H}_{ii}$ (Exercise 2.19). Consequently,

$$\mathrm{Var}\left(\|\widehat{\boldsymbol{\varepsilon}}\|^2\right) \;\leq\; \big((K-3)^+ + 2\big)(n-p)\sigma^4,$$

whence $\mathrm{Var}(\widehat{\sigma}^2) = (n-p)^{-2}\,\mathrm{Var}\big(\|\widehat{\boldsymbol{\varepsilon}}\|^2\big)$ satisfies the asserted inequality. □

Exercise 2.19 (Orthogonal projections). Let $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ describe an orthogonal projection, that means, $\boldsymbol{H} = \boldsymbol{H}^2 = \boldsymbol{H}^\top$.

(a) Show that

$$\|\boldsymbol{H}\boldsymbol{x}\|^2 \begin{cases} \leq \|\boldsymbol{x}\|^2, & \\ = \|\boldsymbol{x}\|^2 & \text{if and only if } \boldsymbol{x} \in \boldsymbol{H}\mathbb{R}^n, \\ = 0 & \text{if and only if } \boldsymbol{x} \in (\boldsymbol{I}_n - \boldsymbol{H})\mathbb{R}^n. \end{cases}$$

Deduce from that the inequality

$$0 \;\leq\; H_{ii} \;\leq\; 1 \quad \text{for } 1 \leq i \leq n.$$

(b) Show that

$$\operatorname{trace}(\boldsymbol{H}) \;=\; \dim(\boldsymbol{H}\mathbb{R}^n).$$

Hint: Recall (or verify) that $\operatorname{trace}(\boldsymbol{a}\boldsymbol{b}^\top) = \boldsymbol{b}^\top\boldsymbol{a}$ for $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$ or, more generally, $\operatorname{trace}(\boldsymbol{A}\boldsymbol{B}^\top) = \operatorname{trace}(\boldsymbol{B}^\top\boldsymbol{A})$ for matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{k\times\ell}$.

Exercise 2.20. Show that in case of independent and homoscedastic errors $\varepsilon_1, \ldots, \varepsilon_n$ with finite fourth moments,

$$\operatorname{Cov}(\varepsilon_i\varepsilon_j, \varepsilon_k\varepsilon_\ell) \;=\; \begin{cases} \mathbb{E}(\varepsilon_i^4) - \sigma^4 & \text{if } i = j = k = \ell, \\ \sigma^4 & \text{if } i \neq j \text{ and } \{i,j\} = \{k,\ell\}, \\ 0 & \text{else.} \end{cases}$$

Remark 2.21 (Adjusted R-squared). With the variance estimator $\widehat{\sigma}^2$ in mind, one may re-interpret the adjusted coefficient of determination as follows:

$$R^2_{\text{adj}} = 1 - \frac{\sum_i (Y_i - \widehat{Y}_i)^2/(n-p)}{\sum_i (Y_i - \bar{Y})^2/(n-1)} \;=\; 1 - \frac{\widehat{\sigma}^2_{\text{full model}}}{\widehat{\sigma}^2_{\text{minimal model}}}.$$

Here $\widehat{\sigma}^2_{\text{full model}}$ is the estimator $\widehat{\sigma}^2$ described before while $\widehat{\sigma}^2_{\text{minimal modell}}$ is the sample variance of $\boldsymbol{Y}$, corresponding to the simplistic model equation $Y = \theta + \varepsilon$ with unknown $\theta \in \mathbb{R}$.

2.4 The Gauss–Markov Theorem and Standard Errors

Often one is not interested in the full vector $\boldsymbol{\theta}$ but rather in specific linear combinations of its components. Such a quantity may be written as $\boldsymbol{\psi}^\top\boldsymbol{\theta}$ with a non-zero vector $\boldsymbol{\psi} \in \mathbb{R}^p$.

Example 2.22 (Polynomial regression). Let $X_i \in \mathbb{R}$ and $Y_i = f(X_i) + \varepsilon_i$, where $f(x) = \sum_{j=0}^d \theta_{j+1}x^j$ with an unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$. Suppose we are interested in the value $f(x)$ for a particular point $x \in \mathbb{R}$. Then, we consider

$$f(x) \;=\; \boldsymbol{\psi}^\top\boldsymbol{\theta} \quad \text{with} \quad \boldsymbol{\psi} \;=\; (1, x, x^2, \ldots, x^d)^\top.$$

Suppose we are interested in the derivative $f'(x)$. Then, we consider

$$f'(x) \;=\; \boldsymbol{\psi}^\top\boldsymbol{\theta} \quad \text{with} \quad \boldsymbol{\psi} \;=\; (0, 1, 2x, \ldots, dx^{d-1})^\top.$$

Exercise 2.23. Consider $X_i \in \mathbb{R}$ and $Y_i = f(X_i) + \varepsilon_i$ with $f(x) = \sum_{j=0}^{d} \theta_{j+1} x^j$ with unknown $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$, where $d \geq 2$. Represent the three quantities

$$\frac{1}{n}\sum_{i=1}^{n} f'(X_i), \quad \int_a^b f(x)\,dx \quad \text{and} \quad f(b) - f(a) - (b-a)f'(a)$$

as $\boldsymbol{\psi}^\top \boldsymbol{\theta}$ with suitable vectors $\boldsymbol{\psi} \in \mathbb{R}^{d+1}$.

Example 2.24 (Simple linear regression). Let $X_i \in \mathbb{R}$ and $Y_i = a + bX_i + \varepsilon_i$ with unknown parameter vector $\boldsymbol{\theta} = (a, b)^\top$. Suppose we are interested only in the slope b, that means, in $\boldsymbol{\psi}^\top \boldsymbol{\theta}$ with $\boldsymbol{\psi} = (0, 1)^\top$. One could think of various estimators for b, for instance

$$\begin{aligned}
\widehat{b} &:= \boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} = \sum_{i=1}^{n} (X_i - \bar{X}) Y_i \Big/ \sum_{i=1}^{n} (X_i - \bar{X})^2, \\
\widehat{b} &:= \frac{Y_n - Y_1}{X_n - X_1} \quad (\text{if } X_1 \neq X_n), \\
\widehat{b} &:= \sum_{i,j=1}^{n} 1_{[X_i < X_j]} (Y_j - Y_i) \Big/ \sum_{i,j=1}^{n} 1_{[X_i < X_j]} (X_j - X_i).
\end{aligned}$$

All previous estimators $\widehat{b}$ may be written as $\boldsymbol{a}^\top \boldsymbol{Y}$ with a certain weight vector $\boldsymbol{a} \in \mathbb{R}^n$. Moreover, in all three cases one can show that $\mathbb{E}(\widehat{b}) = b$. Now an obvious question is whether there exists an optimal estimator of this type.

Exercise 2.25. Show that in the model of simple linear regression, the Gauss-Markov estimator $\widehat{b}$ for the slope may be written as

$$\widehat{b} = \frac{\sum_{1 \leq i < j \leq n} (X_i - X_j)(Y_i - Y_j)}{\sum_{1 \leq i < j \leq n} (X_i - X_j)^2}.$$

In general, $\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}}$ is a natural estimator for $\boldsymbol{\psi}^\top \boldsymbol{\theta}$, and it may be written as

$$\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} = \boldsymbol{a}_{\boldsymbol{\psi}}^\top \boldsymbol{Y} \quad \text{with} \quad \boldsymbol{a}_{\boldsymbol{\psi}} := \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}.$$

Moreover, it follows from Lemma 2.15 or from a direct calculation that $\mathbb{E}(\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}}) = \boldsymbol{\psi}^\top \boldsymbol{\theta}$. Thus, it is an unbiased linear estimator in the following sense:

Definition 2.26 (Linear and unbiased estimators). A *linear estimator of* $\boldsymbol{\psi}^\top \boldsymbol{\theta}$ is a linear form $\boldsymbol{a}^\top \boldsymbol{Y}$ with a fixed vector $\boldsymbol{a} \in \mathbb{R}^n$. Such an estimator is called *unbiased* if

$$\mathbb{E}(\boldsymbol{a}^\top \boldsymbol{Y}) = \boldsymbol{\psi}^\top \boldsymbol{\theta}$$

regardless of the actual value of $\boldsymbol{\theta}$. This is equivalent to the requirement

$$\boldsymbol{a}^\top \boldsymbol{D} \boldsymbol{\eta} = \boldsymbol{\psi}^\top \boldsymbol{\eta} \quad \text{for all } \boldsymbol{\eta} \in \mathbb{R}^p,$$

which may be expressed as

$$\boldsymbol{D}^\top \boldsymbol{a} = \boldsymbol{\psi}. \tag{2.6}$$

The following theorem shows that the so-called *Gauss-Markov estimator* $\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}}$ is the unique linear and unbiased estimator with minimal variance.

Theorem 2.27 (Gauss–Markov). *Suppose that* $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ *and* $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$. *A linear unbiased estimator* $\boldsymbol{Y} \mapsto \boldsymbol{a}^\top \boldsymbol{Y}$ *of* $\boldsymbol{\psi}^\top \boldsymbol{\theta}$ *has minimal variance if and only if* $\boldsymbol{a}$ *is contained in the model space* $\boldsymbol{M} = \boldsymbol{D}\mathbb{R}^p$, *that means* $\boldsymbol{H}\boldsymbol{a} = \boldsymbol{a}$. *There exists precisely one vector* $\boldsymbol{a}$ *with these properties, namely,*

$$\boldsymbol{a} \;=\; \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}.$$

Proof of Theorem 2.27. The variance of an arbitrary linear estimator $\boldsymbol{a}^\top \boldsymbol{Y}$ is equal to

$$\operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{Y}) \;=\; \boldsymbol{a}^\top \operatorname{Var}(\boldsymbol{Y}) \boldsymbol{a} \;=\; \sigma^2 \|\boldsymbol{a}\|^2.$$

Thus, we would like to minimize $\|\boldsymbol{a}\|^2$ under the constraint (2.6). Since

$$\boldsymbol{D}^\top \boldsymbol{H} \;=\; \boldsymbol{D}^\top \boldsymbol{D} (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \;=\; \boldsymbol{D}^\top,$$

property (2.6) for $\boldsymbol{a}$ carries over to $\boldsymbol{H}\boldsymbol{a}$. Hence

$$\begin{aligned} \operatorname{Var}(\boldsymbol{a}^\top \boldsymbol{Y}) \;&=\; \sigma^2 \|\boldsymbol{a}\|^2 \\ &=\; \sigma^2 \|\boldsymbol{H}\boldsymbol{a}\|^2 + \sigma^2 \|\boldsymbol{a} - \boldsymbol{H}\boldsymbol{a}\|^2 \\ &=\; \operatorname{Var}\big((\boldsymbol{H}\boldsymbol{a})^\top \boldsymbol{Y}\big) + \sigma^2 \|\boldsymbol{a} - \boldsymbol{H}\boldsymbol{a}\|^2 \\ &\geq\; \operatorname{Var}\big((\boldsymbol{H}\boldsymbol{a})^\top \boldsymbol{Y}\big) \end{aligned}$$

with equality if and only if $\boldsymbol{a} = \boldsymbol{H}\boldsymbol{a}$. Together with (2.6) the latter identity implies that

$$\boldsymbol{a} \;=\; \boldsymbol{H}\boldsymbol{a} \;=\; \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top \boldsymbol{a} \;=\; \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}.$$

□

Example 2.28 (Absorption spectra). Consider an aqueous solution of p different substances with unknown concentrations $\theta_1, \ldots, \theta_p$. To determine $\boldsymbol{\theta}$, one measures for given wavelengths $X_1 < X_2 < \cdots < X_n$ the absorptions $Y_1, Y_2, \ldots, Y_n$ of light. One assumes that

$$Y_i \;=\; \sum_{j=1}^{p} \theta_j f_j(X_i) + \varepsilon_i$$

with independent random errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ with mean 0 and standard deviation $\sigma > 0$, and $f_1, \ldots, f_p$ are the absorption spectra of the p substances, which have been determined in extensive experiments before. That means, $f_j(x) \geq 0$ is the mean absorption of light of wavelength x in a solution with unit concentration of substance j only.

Often, each spectrum f_k has a characteristic peak at a given frequency x_k, where $x_1, \ldots, x_p$ are different. Then, chemists or physicists often choose a moderate number ℓ and compute for $1 \leq k \leq p$ the local average

$$\bar{Y}_k \;:=\; \frac{1}{\ell} \sum_{s=1}^{\ell} Y_{i(k)+s}$$

over wavelengths $X_{i(k)+1}, \ldots, X_{i(k)+\ell}$ close to x_k. Then, they solve the linear equation system

$$\sum_{j=1}^{p} \bar{f}_{kj} \widehat{\theta}_j^o \stackrel{!}{=} \bar{Y}_k \quad \text{for } 1 \le k \le p$$

with the average

$$\bar{f}_{kj} := \frac{1}{\ell} \sum_{s=1}^{\ell} f_j(X_{i(k)+s})$$

of spectrum f_j over the ℓ frequencies for substance k. If the matrix

$$\boldsymbol{F} := (\bar{f}_{kj})_{k,j=1}^{p}$$

is nonsingular, the vector $\widehat{\boldsymbol{\theta}}^o = (\widehat{\theta}_j^o)_{j=1}^p$ is given by

$$\widehat{\boldsymbol{\theta}}^o = \boldsymbol{F}^{-1} (\bar{Y}_k)_{k=1}^p.$$

Note that this defines an unbiased linear estimator of $\boldsymbol{\theta}$. Moreover, the covariance matrix of this estimator $\widehat{\boldsymbol{\theta}}$ is given by

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}^o) = \sigma^2 \ell^{-1} (\boldsymbol{F}^\top \boldsymbol{F})^{-1},$$

provided that all $p\ell$ frequencies $X_{i(k)+s}$, $1 \le k \le p$ and $1 \le s \le \ell$, are different. In that case, the local averages $\bar{Y}_1, \ldots, \bar{Y}_p$ are independent, each with variance σ^2/ℓ.

The Gauss–Markov theorem, however, recommends to compute the LSE $\widehat{\boldsymbol{\theta}}$ with components

$$\widehat{\theta}_j = \boldsymbol{a}_j^\top \boldsymbol{Y},$$

where the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p \in \mathrm{span}(f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X}))$ are given by

$$[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p] = \big[f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})\big] \, (\boldsymbol{D}^\top \boldsymbol{D})^{-1}$$

with $\boldsymbol{D}^\top \boldsymbol{D} = \big(f_j(\boldsymbol{X})^\top f_k(\boldsymbol{X})\big)_{j,k=1}^p$. Here we know that

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = \sigma^2 (\boldsymbol{D}^\top \boldsymbol{D})^{-1}.$$

The following (artificial) example with $p = 3$ substances illustrates the benefit of using the LSE $\widehat{\boldsymbol{\theta}}$ rather than the ad hoc estimator $\widehat{\boldsymbol{\theta}}^o$. The upper panel of Figure 2.3 shows the spectral vectors $f_1(\boldsymbol{X}), f_2(\boldsymbol{X}), f_3(\boldsymbol{X})$ with peaks at $x_1 = X_{80}, x_2 = X_{100}, x_3 = X_{160}$. The lower panel shows the optimal "filter vectors" $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$. Explicit formulae are

$$\begin{aligned}
\boldsymbol{a}_1 &= 0.1830 \cdot f_1(\boldsymbol{X}) - 0.0620 \cdot f_2(\boldsymbol{X}) + 0.0002 \cdot f_3(\boldsymbol{X}), \\
\boldsymbol{a}_2 &= -0.0620 \cdot f_1(\boldsymbol{X}) + 0.0917 \cdot f_2(\boldsymbol{X}) - 0.0024 \cdot f_3(\boldsymbol{X}), \\
\boldsymbol{a}_2 &= 0.0002 \cdot f_1(\boldsymbol{X}) - 0.0024 \cdot f_2(\boldsymbol{X}) + 0.0371 \cdot f_3(\boldsymbol{X}).
\end{aligned}$$

Here and in what follows, numbers are rounded to four decimal places.

Now we compare $\widehat{\boldsymbol{\theta}}$ with the ad hoc estimator $\widehat{\boldsymbol{\theta}}^o$ based on the local averages

$$\bar{Y}_1 := \frac{1}{11} \sum_{i=75}^{85} Y_i, \quad \bar{Y}_2 := \frac{1}{11} \sum_{i=95}^{105} Y_i, \quad \bar{Y}_3 := \frac{1}{11} \sum_{i=155}^{165} Y_i$$

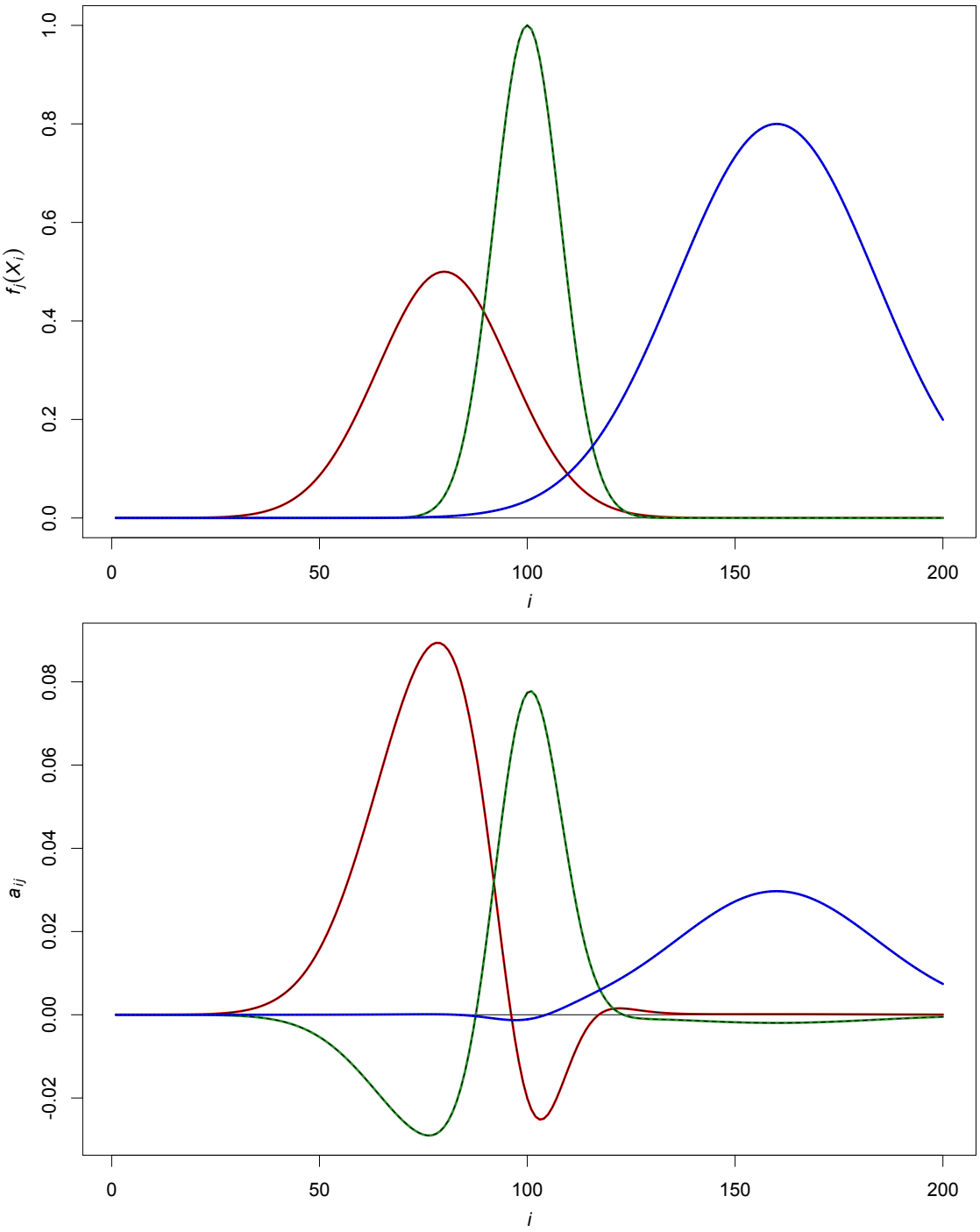

Figure 2.3: Three absorption spectra $f_j(\boldsymbol{X})$ (upper panel), and the corresponding filter vectors $\boldsymbol{a}_j = (a_{ij})_{i=1}^n$ (lower panel).

over $\ell = 11$ frequencies. The resulting covariance matrices are equal to

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) \;=\; \sigma^2 \begin{bmatrix} 0.1830 & -0.0620 & 0.0002 \\ -0.0620 & 0.0917 & -0.0024 \\ 0.0002 & -0.0024 & 0.0371 \end{bmatrix},$$

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}^o) \;=\; \sigma^2 \begin{bmatrix} 0.4050 & -0.1149 & -0.0003 \\ -0.1149 & 0.1381 & -0.0057 \\ -0.0003 & -0.0057 & 0.1445 \end{bmatrix}.$$

In particular, the standard deviations of the components of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^o$ are equal to

$$\begin{bmatrix} \mathrm{Std}(\widehat{\theta}_1) \\ \mathrm{Std}(\widehat{\theta}_2) \\ \mathrm{Std}(\widehat{\theta}_3) \end{bmatrix} \;=\; \sigma \begin{bmatrix} 0.4278 \\ 0.3027 \\ 0.1927 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathrm{Std}(\widehat{\theta}^o_1) \\ \mathrm{Std}(\widehat{\theta}^o_2) \\ \mathrm{Std}(\widehat{\theta}^o_3) \end{bmatrix} \;=\; \sigma \begin{bmatrix} 0.6364 \\ 0.3717 \\ 0.3802 \end{bmatrix}.$$

As predicted by the Gauss–Markov theorem, the estimator $\widehat{\boldsymbol{\theta}}$ is substantially more accurate than the ad hoc estimator $\widehat{\boldsymbol{\theta}}^o$.

To illustrate the LSE, we simulated noisy data $Y_i = f(X_i) + \varepsilon_i$ with independent random errors $\varepsilon_i \sim \mathrm{N}(0, 0.3^2)$, where $f(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \theta_3 f_3(x)$ with concentration vector $\boldsymbol{\theta} = (3, 2, 1)^\top$. Figure 2.4 shows the true spectrum $f(\boldsymbol{X})$ as well as the noisy data $\boldsymbol{Y}$ plus the fitted spectrum $\widehat{\boldsymbol{Y}} = \widehat{f}(\boldsymbol{X})$ and its constituents $\widehat{\theta}_j f_j(\boldsymbol{X})$. The LSE turned out to be $\widehat{\boldsymbol{\theta}} = (2.9524, 2.0442, 0.9649)^\top$.

Remark 2.29 (Standard errors). In case of $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$, the standard deviation $\mathrm{Std}(\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}})$ of the Gauss–Markov estimator of $\boldsymbol{\psi}^\top \boldsymbol{\theta}$ is equal to

$$\sigma_{\boldsymbol{\psi}} \;:=\; \sigma \sqrt{\boldsymbol{\psi}^\top (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}} \;=\; \sigma \|\boldsymbol{a}_{\boldsymbol{\psi}}\|$$

with $\boldsymbol{a}_{\boldsymbol{\psi}} := \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}$. This standard deviation $\sigma_{\boldsymbol{\psi}}$ involves the unknown standard deviation σ of the errors ε_i. If we replace σ with $\widehat{\sigma}$, we obtain the so-called *standard error*

$$\widehat{\sigma}_{\boldsymbol{\psi}} \;:=\; \widehat{\sigma} \sqrt{\boldsymbol{\psi}^\top (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}} \;=\; \widehat{\sigma} \|\boldsymbol{a}_{\boldsymbol{\psi}}\|.$$

In general, an estimated standard deviation is called a standard error.

2.5 Properties of the Estimators in Misspecified Models

Let us talk briefly about the behaviour of the estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$ in case of the true regression function f being *not necessarily* an element of the linear model $\mathcal{F}$. Geometrically this means, that the vector $f(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y})$ could be outside of the model space $\boldsymbol{M} = \boldsymbol{D}\mathbb{R}^p$.

As to $\widehat{\boldsymbol{\theta}}$ we could simply define

$$\boldsymbol{\theta} \;:=\; \mathbb{E}(\widehat{\boldsymbol{\theta}}) \;=\; (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{D}^\top f(\boldsymbol{X}),$$

with $\boldsymbol{D} = [f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})]$. Then, $\widehat{\boldsymbol{\theta}}$ is still an unbiased estimator of $\boldsymbol{\theta}$, and

$$\boldsymbol{D}\boldsymbol{\theta} \;=\; \boldsymbol{H} f(\boldsymbol{X}).$$

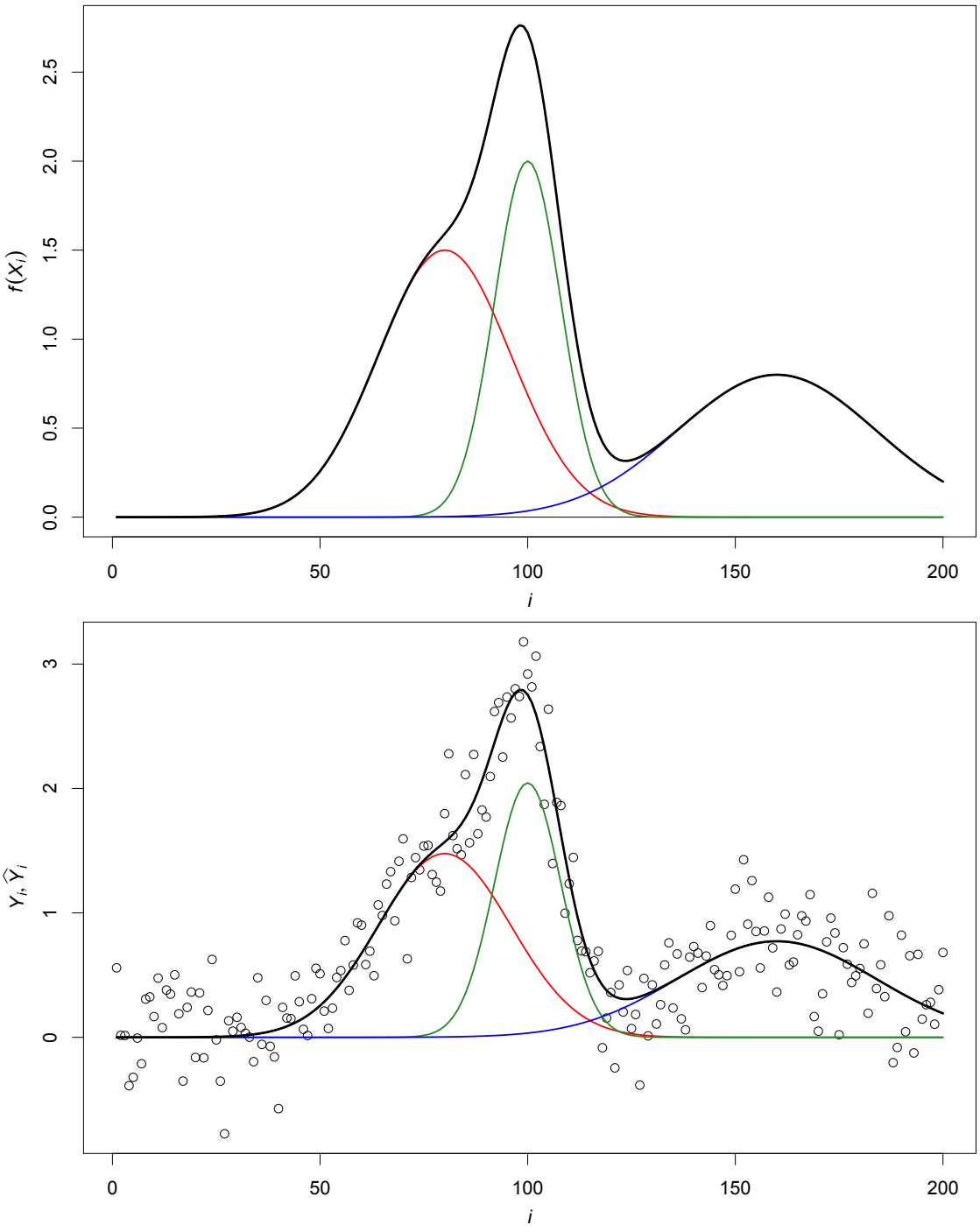

Figure 2.4: The spectrum $f(\boldsymbol{X})$ of a mixture (upper panel), and its noisy measurement $\boldsymbol{Y}$ (lower panel) together with the fitted spectrum $\widehat{\boldsymbol{Y}} = \sum_{j=1}^{3} \widehat{\theta}_j f_j(\boldsymbol{X})$ and the constituents $\widehat{\theta}_j f_j(\boldsymbol{X})$.

That means, $\boldsymbol{\theta}$ represents an approximation $\check{f} := \sum_{j=1}^{p} \theta_j f_j$ of the regression function f such that

$$\big\|f(\boldsymbol{X}) - \check{f}(\boldsymbol{X})\big\|^2 \;=\; \sum_{i=1}^{n} \big(f(X_i) - \check{f}(X_i)\big)^2$$

is minimal.

In case of homoscedastic errors ε_i with variance σ^2, misspecification of the model implies a positive bias of $\widehat{\sigma}^2$. Precisely,

$$\mathbb{E}(\widehat{\sigma}^2) \;=\; \sigma^2 + \frac{\big\|f(\boldsymbol{X}) - \check{f}(\boldsymbol{X})\big\|^2}{n-p}. \tag{2.7}$$

The proof of this equation is left to the reader as an exercise.

In case of polynomial regression, one can provide a rather accurate bound for the approximation error $\big\|f(\boldsymbol{X}) - \check{f}(\boldsymbol{X})\big\|$. It involves orthogonal polynomials as constructed in Exercise 2.8.

Theorem 2.30 (Approximation error in polynomial regression). *Let $\mathcal{X}$ be a real interval, and let $\boldsymbol{X} \in \mathcal{X}^n$ be a data vector with at least $d+2$ different components, $d \in \mathbb{N}_0$. Further let $f : \mathcal{X} \to \mathbb{R}$ be $d+1$ times differentiable. For $0 \le k \le d+1$, let $p_k(x)$ be a polynomial of order k with leading coefficient 1, such that the vectors $p_0(\boldsymbol{X}), p_1(\boldsymbol{X}), \ldots, p_{d+1}(\boldsymbol{X})$ are orthogonal. If $\mathcal{F}$ is the linear space of all polynomials of order d, then*

$$\big\|f(\boldsymbol{X}) - \check{f}(\boldsymbol{X})\big\| \;\le\; \frac{\sup_{x\in\mathcal{X}} |f^{(d+1)}(x)|}{(d+1)!}\, \big\|p_{d+1}(\boldsymbol{X})\big\|.$$

Equality holds if f is a polynomial of order $d+1$.

Proof of Theorem 2.30. We use a few well-known results about orthogonal polynomials and interpolation polynomials as presented, for instance, in the monograph of G. Opfer (1994).

Fact 1: The polynomial $p_{d+1}(x)$ has $d+1$ different zeros in $\mathcal{X}$!

Proof of Fact 1: Suppose there exist only $m \le d$ points $x_1 < \cdots < x_m$ in $\mathcal{X}$ at which p_{d+1} equals 0 and changes its sign. Then, for a suitable $\xi \in \{-1, 1\}$, $q(x) := \xi \prod_{i=1}^{m}(x - x_i)$ would be a polynomial of degree m with the particular property that $p_{d+1}(x)q(x) > 0$ for all $x \in \mathcal{X}$ such that $p_{d+1}(x) \neq 0$. (In case of $m = 0$ we set $q(x) := \xi$.) Since $q(\boldsymbol{X})$ is a linear combination of $p_0(\boldsymbol{X}), \ldots, p_d(\boldsymbol{X})$, all of which are orthogonal to $p_{d+1}(\boldsymbol{X})$, we arrive at the equation $0 = q(\boldsymbol{X})^\top p_{d+1}(\boldsymbol{X})$, so $p_{d+1}(\boldsymbol{X}) = \boldsymbol{0}$. Since $\boldsymbol{X}$ has at least $d+2$ different components, this would contradict our assumption that p_{d+1} has degree $d+1$.

Fact 2: Let $x_0 < x_1 < \cdots < x_d$ be the zeros of $p_{d+1}(x)$ in $\mathcal{X}$. Further let $p(x)$ be the unique polynomial of order d such that $p(x_i) = f(x_i)$ for $0 \le i \le d$. Then, for any $z \in \mathcal{X}$, there exists a point $\xi(z) \in \mathcal{X}$ such that

$$f(z) - p(z) \;=\; \frac{f^{(d+1)}(\xi(z))}{(d+1)!}\, p_{d+1}(z).$$

Proof of Fact 2: For $z \in \{x_0, x_1, \ldots, x_d\}$ there is nothing to be shown, because $f(z) - p(z) = 0 = p_{d+1}(z)$. Hence, let $z \notin \{x_0, x_1, \ldots, x_d\}$. Now we define

$$h(x) \;:=\; f(x) - p(x) - \gamma p_{d+1}(x)$$

with $\gamma := \big(f(z) - p(z)\big)/p_{d+1}(z)$. This defines a function $h : \mathcal{X} \to \mathbb{R}$ which is $d+1$ times differentiable and has at least $d+2$ different zeros on $\mathcal{X}$, namely, $z, x_0, x_1, \ldots, x_d$. Consequently, h' has at least $d+1$ different zeros on $\mathcal{X}$, and inductively one may conclude that $h^{(d+1)}$ has at least one zero $\xi(z) \in \mathcal{X}$. Since $p_{d+1}(x)$ has degree $d+1$ with leading coefficient 1, $p_{d+1}^{(d+1)} \equiv (d+1)!$, whereas $p^{(d+1)} \equiv 0$. Consequently,

$$0 \;=\; h^{(d+1)}(\xi(z)) \;=\; f^{(d+1)}(\xi(z)) - \gamma(d+1)!,$$

and this implies that $f(z) - p(z) = p_{d+1}(z) f^{(d+1)}(\xi(z))/(d+1)!$.

Proof of the theorem: With the interpolation polynomial p from Fact 2, the definition of $\check{f}$ implies that

$$\begin{aligned}
\big\|f(\boldsymbol{X}) - \check{f}(\boldsymbol{X})\big\|^2 \;&\le\; \big\|f(\boldsymbol{X}) - p(\boldsymbol{X})\big\|^2 \\
&=\; \sum_{i=1}^{n} \Big(\frac{f^{(d+1)}(\xi(X_i))}{(d+1)!}\Big)^2 p_{d+1}(X_i)^2 \\
&\le\; \Big(\frac{\sup_{x\in\mathcal{X}} |f^{(d+1)}(x)|}{(d+1)!}\Big)^2 \big\|p_{d+1}(\boldsymbol{X})\big\|^2.
\end{aligned}$$

Equality holds, for instance, if f is a polynomial of order $d+1$. For then $f \equiv p + c p_{d+1}$ for some $c \in \mathbb{R}$, and $\check{f} = p$, $f^{(d+1)} \equiv (d+1)!\, c$. □

Example 2.31. Suppose that $\boldsymbol{X} = (-5, -4, \ldots, 0, 1, \ldots, 5)^\top$, i.e. $n = 11$, and let $\mathcal{F}_d$ be the space of all polynomials of order d for a given $d \le 10$. By means of Exercises 2.8 and 2.9 one can determine the polynomials $p_0, p_1, \ldots, p_{10}$ mentioned in Theorem 2.30. The first seven are given by $p_0(x) = 1$, $p_1(x) = x$, and

$$\begin{aligned}
p_2(x) \;&=\; -10 + x^2, \\
p_3(x) \;&=\; -17.8\, x + x^3, \\
p_4(x) \;&=\; 72 - 25\, x^2 + x^4, \\
p_5(x) \;&=\; 190.667\, x - 31.667\, x^3 + x^5, \\
p_6(x) \;&=\; -436.364 + 342.182\, x^2 - 37.727\, x^4 + x^6.
\end{aligned}$$

with coefficients rounded to three decimals. Now suppose that

$$f(x) \;=\; \sin(x - 1).$$

Then, the best approximation of f by a function in $\mathcal{F}_d$ is given by

$$\check{f}(x) \;=\; \sum_{j=0}^{d} \theta_{j+1} p_j(x) \quad \text{with} \quad \theta_{j+1} \;=\; \sum_{i=1}^{n} f(X_i) p_j(X_i) \Big/ \sum_{i=1}^{n} p_j(X_i)^2.$$

The following table shows for $d = 0, 1, \ldots, 9$ the coefficients θ_{d+1}, the approximation errors $\|f(\boldsymbol{X}) - \check{f}_d(\boldsymbol{X})\|$ and the normalized squared norm $\|f(\boldsymbol{X}) - \check{f}_d(\boldsymbol{X})\|^2/(n-d-1)$ which

determines the bias of $\widehat{\sigma}^2$, all numbers rounded to five decimals:

d	θ_{d+1}	$\|f(\boldsymbol{X}) - \check{f}_d(\boldsymbol{X})\|$	$\frac{\|f(\boldsymbol{X}) - \check{f}_d(\boldsymbol{X})\|^2}{n-d-1}$
0	0.11258	2.26127	0.51133
1	−0.04655	2.20793	0.54166
2	0.01185	2.18047	0.59430
3	−0.01260	1.94260	0.53910
4	−0.00841	0.92950	0.14400
5	0.00148	0.56555	0.06397
6	0.00047	0.13640	0.00465
7	−0.00005	0.05296	0.00094
8	−0.00001	0.00671	0.00002
9	0.00000	0.00129	0.00000

Figure 2.5 depicts the true function f and the approximation $\check{f}$ for orders $d = 4$ and $d = 6$. Note that the polynomial approximations are useful within the range of $\boldsymbol{X}$ but problematic when extrapolated to values $x < \min(\boldsymbol{X})$ or $x > \max(\boldsymbol{X})$.

2.6 Special Parametrizations and Interactions

In connection with categorial covariates, so-called factors, one uses often special parametrizations. We shall explain these for the cases of one and two factors. Moreover, in the setting of multiple linear regression, an important concept are so-called interactions which will be introduced as well.

2.6.1 One-Way Analysis of Variance

As in Example 1.1, we consider a covariate $X \in \{1, 2, \ldots, L\}$. Instead of the model equation $Y = f(X) + \varepsilon$ one often writes

$$Y \;=\; \mu + a(X) + \varepsilon$$

with unknown parameters $\mu, a(1), \ldots, a(L)$ satisfying certain constraints:

Convention 1. For a given reference category j_o we require that $a(j_o) = 0$. Then, μ is the mean of Y in case of $X = j_o$, and $a(j)$ quantifies the difference between the category j and the category j_o with respect to the mean of Y. The connection to the function $f : \{1, \ldots, L\} \to \mathbb{R}$ is:

$$\mu \;=\; f(j_o) \quad \text{and} \quad a(j) \;=\; f(j) - f(j_o).$$

This convention is often appropriate in medical applications when X stands for potential medical treatments, and j_o refers to a standard treatment or placebo. Here $a(j)$ is the benefit of treatment j compared to treatment j_o.

This convention is used by most statistical software packages, and the user may specify the reference category j_o.

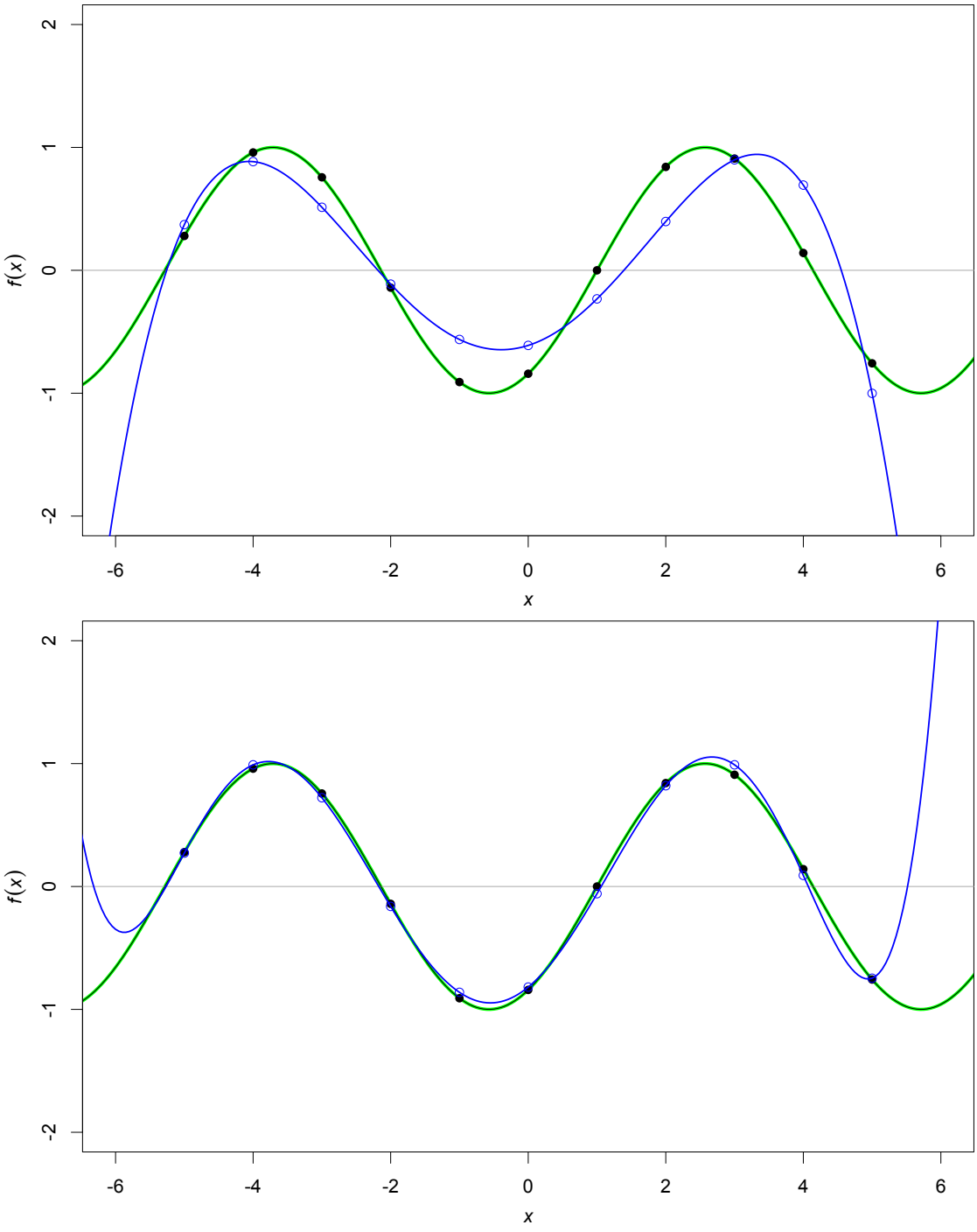

Figure 2.5: Approximating $f(x) = \sin(x - 1)$ (green/black) by a polynomial $\check{f}$ (blue) of order $d = 4$ (upper panel) and $d = 6$ (lower panel.)

Convention 2. We require that $\sum_{j=1}^{L} a(j) = 0$. In this case one may interpret μ as basic effect, and $a(j)$ is the effect of category j in relation to all others. Now, the connection to the function $f : \{1, \dots, L\} \to \mathbb{R}$ is:

$$\mu \;=\; \frac{1}{L}\sum_{j=1}^{L} f(j) \quad \text{and} \quad a(j) \;=\; f(j) - \mu.$$

Connection to multiple linear regression. Suppose we have arranged the values of X such that $j_o = L$. Then, one may interpret the model equation $Y = \mu + a(X) + \varepsilon$ as a special case of multiple linear regression. To this end we introduce so-called dummy variables

$$X(j) \;:=\; 1_{[X=j]}, \quad 1 \le j < L,$$

and then

$$Y \;=\; \mu + \sum_{j=1}^{L-1} a(j)X(j) + \varepsilon,$$

which is essentially the standard model for multiple linear regression with the covariate vector $(X(j))_{j=1}^{L-1} \in \{0,1\}^{L-1}$. Note that the value of X is uniquely determined by the latter vector and vice versa.

2.6.2 Two-Way Analysis of Variance

Suppose that X consists of two covariates $C \in \{1, \dots, L\}$ and $D \in \{1, \dots, M\}$. One could view X itself as a categorical covariate with $L \cdot M$ potential values. This leads to the model equation

$$Y \;=\; f(C, D) + \varepsilon$$

with an unknown regression function $f : \{1, \dots, L\} \times \{1, \dots, M\} \to \mathbb{R}$. From the perspective of users, however, it is often desirable to distinguish and highlight the influence of the two covariates separately. For that purpose, there are two different possibilities.

Cross classification

Instead of $Y = f(C, D) + \varepsilon$, we write

$$Y \;=\; \mu + a(C) + b(D) + h(C, D) + \varepsilon$$

with a "basic effect" μ, the "main effects" $a : \{1, \dots, L\} \to \mathbb{R}$ and $b : \{1, \dots, M\} \to \mathbb{R}$ of the two covariates and their "interactions"

$$h : \{1, \dots, L\} \times \{1, \dots, M\} \to \mathbb{R}.$$

That means, the impact of the main effects is purely additive, and the interactions describe the deviation of the regression function from a purely additive regression function with summands depending only on C or only on D. Again one needs certain conventions such that these parameters μ, $a(j)$, $b(k)$ and $h(j, k)$ are well-defined.

Convention 1. For given reference categories $j_o \in \{1, \ldots, L\}$ and $k_o \in \{1, \ldots, M\}$, it is required that

$$\begin{aligned} a(j_o) &= 0, \\ b(k_o) &= 0, \\ h(j_o, k) &= 0 \quad \text{for } 1 \le k \le M, \\ h(j, k_o) &= 0 \quad \text{for } 1 \le j \le L. \end{aligned}$$

Hence,

$$\begin{aligned} \mu &= f(j_o, k_o), \\ a(j) &= f(j, k_o) - f(j_o, k_o), \\ b(k) &= f(j_o, k) - f(j_o, k_o), \\ h(j, k) &= f(j, k) - f(j, k_o) - f(j_o, k) + f(j_o, k_o). \end{aligned}$$

Convention 2. We require that

$$\begin{aligned} \sum_{j=1}^{L} a(j) &= 0, \\ \sum_{k=1}^{M} b(k) &= 0, \\ \sum_{k=1}^{M} h(j, k) &= 0 \quad \text{for } 1 \le j \le L, \\ \sum_{j=1}^{L} h(j, k) &= 0 \quad \text{for } 1 \le k \le M. \end{aligned}$$

Here,

$$\begin{aligned} \mu &= \frac{1}{LM} \sum_{j=1}^{L} \sum_{k=1}^{M} f(j, k), \\ a(j) &= \frac{1}{M} \sum_{k=1}^{M} f(j, k) - \mu, \\ b(k) &= \frac{1}{L} \sum_{j=1}^{L} f(j, k) - \mu, \\ h(j, k) &= f(j, k) - a(j) - b(k) - \mu \\ &= f(j, k) - \frac{1}{M} \sum_{k'=1}^{M} f(j, k') - \frac{1}{L} \sum_{j'=1}^{L} f(j', k) + \mu. \end{aligned}$$

Hierarchical Modelling

Instead of $Y = f(C, D) + \varepsilon$, we write

$$Y = \mu + a(C) + b(C, D) + \varepsilon$$

with a "basic effect" μ, the "main effect" $a : \{1, \ldots, L\} \to \mathbb{R}$ of factor C and the "side effects" $b(j, \cdot) : \{1, \ldots, M\} \to \mathbb{R}$ of the factor D for $1 \leq j \leq L$. Again, we may achieve identifiability in two ways:

Convention 1: For given reference categories $j_o \in \{1, \ldots, L\}$ and $k_o \in \{1, \ldots, M\}$, we require that

$$\begin{aligned} a(j_o) &= 0, \\ b(j, k_o) &= 0 \quad \text{for } 1 \leq j \leq L. \end{aligned}$$

Then,

$$\begin{aligned} \mu &= f(j_o, k_o), \\ a(j) &= f(j, k_o) - f(j_o, k_o), \\ b(j, k) &= f(j, k) - f(j, k_o). \end{aligned}$$

Convention 2: We require that

$$\begin{aligned} \sum_{j=1}^{L} a(j) &= 0, \\ \sum_{k=1}^{M} b(j, k) &= 0 \quad \text{for } 1 \leq j \leq L. \end{aligned}$$

Here,

$$\begin{aligned} \mu &= \frac{1}{LM} \sum_{j=1}^{L} \sum_{k=1}^{M} f(j, k), \\ a(j) &= \frac{1}{M} \sum_{k=1}^{M} f(j, k) - \mu, \\ b(j, k) &= f(j, k) - \frac{1}{M} \sum_{\ell=1}^{M} f(j, \ell). \end{aligned}$$

2.6.3 Interactions

Let $X = (X(j))_{j=1}^{d}$ with $d \geq 2$ numerical or $\{0, 1\}$-valued covariates $X(1), \ldots, X(d)$. Some of the latter covariates could be the dummy variables of categorical covariates as described earlier. The standard model of multiple linear regression assumes a purely additive dependence of Y on X, i.e.

$$Y = \mu + \sum_{j=1}^{d} \beta_j X(j) + \varepsilon$$

with unknown real parameters $\mu, \beta_1, \ldots, \beta_d$ and a centered random error ε. The interpretation of the parameters is as follows:

- μ equals $\mathbb{E}(Y)$ in case of $X(j) = 0$ for $1 \leq j \leq d$;

- if $X(j)$ is increased by $c \in \mathbb{R}$ units, while the other covariates $X(k)$, $k \neq j$, remain fixed, the mean $\mathbb{E}(Y)$ changes by $\beta_j c$.

A more flexible model which allows for the influence of one covariate to be moderated by the others is as follows:

$$Y \;=\; \mu + \sum_{j=1}^{d} \beta_j X(j) + \sum_{(j,k)\in\mathcal{I}} \gamma_{jk} X(j)X(k) + \varepsilon$$

with unknown real parameters μ, $\beta_1, \ldots, \beta_d$ and γ_{jk}, $(j,k) \in \mathcal{I}$, and a centered random error ε. Here, $\mathcal{I}$ is a given set of index pairs (j,k) such that $1 \leq j < k \leq d$ and $X(j)X(k) \not\equiv 0$. Now, the parameters may be interpreted as follows:

- μ equals $\mathbb{E}(Y)$ in case of $X(j) = 0$ for $1 \leq j \leq d$;
- if $X(j)$ is increased by $c \in \mathbb{R}$ units, while $X(k) = 0$ for $k \neq j$, the mean $\mathbb{E}(Y)$ changes by $\beta_j c$;
- if $X(j)$ is increased by $c \in \mathbb{R}$ units, while the other covariates $X(k)$, $k \neq j$, remain fixed, the mean $\mathbb{E}(Y)$ changes by

$$\Big(\beta_j + \sum_{\ell<j:(\ell,j)\in\mathcal{I}} \gamma_{\ell j} X(\ell) + \sum_{k>j:(j,k)\in\mathcal{I}} \gamma_{jk} X(k)\Big)\, c.$$

That means, for any fixed $j \in \{1, \ldots, d\}$, the impact of the single covariate $X(j)$ on the response Y is affine, but both, the corresponding intercept and the slope parameters are affine functions of $X(-j) := (X(k))_{k\neq j}$:

$$Y \;=\; \mu_j(X(-j)) + \beta_j(X(-j)) \cdot X(j) + \varepsilon$$

with

$$\begin{aligned}
\mu_j(X(-j)) \;&:=\; \mu + \sum_{k\neq j} \beta_k X(k), \\
\beta_j(X(-j)) \;&:=\; \beta_j + \sum_{\ell<j:(\ell,j)\in\mathcal{I}} \gamma_{\ell j} X(\ell) + \sum_{k>j:(j,k)\in\mathcal{I}} \gamma_{jk} X(k).
\end{aligned}$$

Whenever one uses one of the previous models, with or without interactions, it is advisable to replace any raw covariate $X(j)$ with $X(j) - x_o(j)$, where $x_o(j)$ stands for a reasonable standard value of $X(j)$. Otherwise the parameters β_j and γ_{jk} may be difficult to interpret.

Example 2.32 (Boston housing data). The data set 'Boston' (available in the R package MASS) contains for 506 quarters in the Boston metropolitan area the median house price Y (in 1000 USD) and the values of various covariates, some of which on the level of municipality rather than quarter. The data set is from the 1970s. When checking the raw data, one realizes that for 16 observations the value of Y has been truncated from the right at 50, that means, for these very expensive quarters the precise value of Y is only known to be at least 50. Hence, we reduced the raw data to the $n = 490$ observations with $Y < 50$. From all covariates, we picked the following ones:

	NAME	meaning	reference value
$X(1)$	CRIM	per capita crime rate by town	0
$X(2)$	ZN	proportion of residential land zoned for lots over $25'000$ square feet	0
$X(3)$	INDUS	proportion of non-retail business acres per town	10
$X(4)$	NOX	nitrogen oxides concentration (parts per 10 million)	0.5
$X(5)$	RM	average number of rooms per dwelling	6
$X(6)$	AGE	proportion of owner-occupied units built prior to 1940	70
$X(7)$	DIS	weighted mean of distances to five Boston employment centres	3
$X(8)$	RAD	index of accessibility to radial highways	5
$X(9)$	TAX	full-value property-tax rate per $10'000$ USD	330
$X(10)$	PTRATIO	pupil-teacher ratio by town	19
$X(11)$	LSTAT	lower status of the population (percent)	12

The reference values have been chosen ad hoc by inspecting the respective range, median and mean. In the subsequent analyses, $X(j)$ stands for the raw covariate minus its reference value.

A standard multiple regression analysis, assuming that

$$Y \;=\; \mu + \sum_{j=1}^{11} \beta_j X(j) + \varepsilon,$$

yields the Gauss–Markov estimators $\widehat{\mu}, \widehat{\beta}_1, \ldots, \widehat{\beta}_{11}$ and corresponding standard errors (all rounded to three decimals) in Table 2.1. The value $\widehat{\mu}$ (intercept) estimates $\mathbb{E}(Y)$ for a quarter with all raw covariates being equal to their reference values, i.e. a quarter in a town with crime rate 0, with 6 rooms per house on average, with a nitrogen oxide concentration of 0.5, et cetera.

The main interest lies in the estimated parameters $\widehat{\beta}_j$, in particular, in their signs. As to be expected, the house prices decrease with the crime rate, increase with the number of rooms, decrease with the amount of air pollution, et cetera. The 't ratio' is the ratio of the estimator and its standard error. The computation and meaning of the 'p-values' will be explained in the next chapter. Each p-value is a strictly decreasing function of the modulus of the t ratio. Small p-values indicate that the corresponding parameter μ or β_j of the true regression function is significantly different from zero.

The left panel of Figure 2.6 shows a scatter plot of the values Y_i versus the fitted values $\widehat{Y}_i = \widehat{f}(X_i)$, together with the straight line '$y = x$'. In a later chapter, we shall investigate the interpretation of such and related plots in more detail. But note that the Y_i seem to be systematically larger than the predictions $\widehat{Y}_i$ whenever the latter are very large. Note also the opposite effect in case of $\widehat{Y}_i$ lying between 20 and 30. Another curiosity is that one fitted value $\widehat{Y}_i$ is negative. The adjusted coefficient of determination is equal to $R^2_{\text{adj}} = 0.7658$, and the estimated standard deviation equals $\widehat{\sigma} = 3.807$.

Next, we tried the standard linear model with all $\binom{11}{2} = 55$ interactions included,

$$Y \;=\; \mu + \sum_{j=1}^{11} \beta_j X(j) + \sum_{1 \le j < k \le 11} \gamma_{jk} X(j) X(k) + \varepsilon.$$

	estimate	st. error	t ratio	p-value
intercept	22.509	0.310	72.694	< 0.001
CRIM	-0.119	0.026	-4.545	< 0.001
ZN	0.036	0.011	3.130	0.002
INDUS	-0.048	0.050	-0.946	0.344
NOX	-13.066	3.082	-4.239	< 0.001
RM	3.630	0.360	10.090	< 0.001
AGE	-0.021	0.011	-1.970	0.050
DIS	-1.222	0.163	-7.515	< 0.001
RAD	0.241	0.053	4.517	< 0.001
TAX	-0.014	0.003	-4.791	< 0.001
PTRATIO	-0.831	0.106	-7.821	< 0.001
LSTAT	-0.372	0.043	-8.732	< 0.001

Table 2.1: Standard linear model output for the Boston housing data.

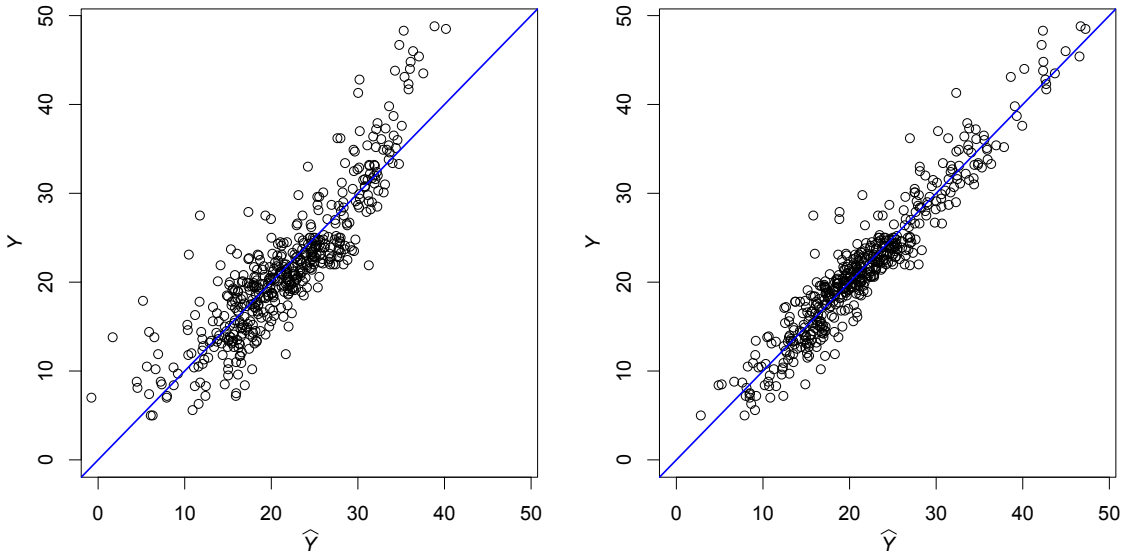

Figure 2.6: Observations versus fitted values for the Boston housing data, without interactions (left panel) and with interactions (right panel).

	estimate	st. error	t ratio	p-value
Intercept	21.819	0.454	48.059	< 0.001
CRIM	-0.544	0.726	-0.750	0.454
ZN	0.078	0.037	-2.112	0.036
INDUS	-0.039	0.121	-0.325	0.745
NOX	-12.234	8.091	-1.512	0.131
RM	5.757	0.656	8.774	< 0.001
AGE	-0.060	0.014	-4.339	< 0.001
DIS	-1.111	0.284	-3.914	< 0.001
RAD	0.437	0.177	2.471	0.014
TAX	-0.018	0.007	-2.449	0.015
PTRATIO	-0.589	0.209	-2.813	0.006
LSTAT	-0.126	0.073	-1.731	0.085
CRIM*NOX	-2.882	0.721	-3.998	< 0.001
CRIM*RM	0.183	0.039	4.655	< 0.001
CRIM*LSTAT	0.020	0.005	4.040	< 0.001
ZN*RAD	-0.013	0.006	-2.130	0.034
ZN*LSTAT	-0.008	0.005	-1.988	0.047
NOX*AGE	-0.392	0.198	-1.974	0.050
RM*AGE	-0.044	0.020	-2.209	0.028
RM*PTRATIO	-0.483	0.214	-2.259	0.025
RM*LSTAT	-0.215	0.036	-5.964	< 0.001
AGE*RAD	0.010	0.004	2.773	0.006
TAX*PTRATIO	0.007	0.002	3.337	< 0.001

Table 2.2: Standard linear model output for the Boston housing data, including interactions.

The right panel of Figure 2.6 shows the corresponding plot of observations Y_i versus the fitted values $\widehat{Y}_i$. Now $R^2_{\text{adj}} = 0.8946$ and $\widehat{\sigma} = 2.554$. The problematic features of the model fit without interactions disappeared. Table 2.2 shows the estimated parameters $\widehat{\mu}$, $\widehat{\beta}_1, \ldots, \widehat{\beta}_{11}$, and all $\widehat{\gamma}_{jk}$ such that the corresponding p-value is no larger than 0.05. Note the clear evidence that association between Y and covariates is modulated by the other covariates.

2.6.4 Paper and Blackboard Notation

In numerous papers and monographs, factors (i.e. categorical covariates) are often hidden by means of multiple subscripts. We illustrate this type of notation in three settings.

One-way ANOVA. Starting from $X_i \in \{1, \ldots, L\}$, let $Y_{j1}, Y_{j2}, \ldots, Y_{jn(j)}$ be those observations Y_i such that $X_i = j$. Then, we may write

$$Y_{js} \;=\; f_j + \varepsilon_{js}, \quad 1 \le j \le L,\, 1 \le s \le n(j),$$

or

$$Y_{js} \;=\; \mu + a_j + \varepsilon_{js}, \quad 1 \le j \le L,\, 1 \le s \le n(j).$$

Here, $f_1, \ldots, f_L$, or $\mu, a_1, \ldots, a_L$ are unknown parameters, and the ε_{js} are independent random errors with mean 0. The least squares estimators for $f_1, \ldots, f_j$ are given by the subsample means

$$\widehat{f_j} \;:=\; \bar{Y}_{j\cdot} := \frac{1}{n(j)} \sum_{s=1}^{n(j)} Y_{js}.$$

Two-way ANOVA. Starting from categorical covariates $C \in \{1, \ldots, L\}$ and $D \in \{1, \ldots, M\}$, let $Y_{jk1}, Y_{jk2}, \ldots, Y_{jkn(j,k)}$ be those observations Y_i such that $(C_i, D_i) = (j, k)$. Then, we may write

$$Y_{jks} \;=\; f_{jk} + \varepsilon_{jks}, \quad 1 \le j \le L,\ 1 \le k \le M,\ 1 \le s \le n(j,k),$$

or

$$Y_{jks} \;=\; \mu + a_j + b_k + h_{jk} + \varepsilon_{jks}, \quad 1 \le j \le L,\ 1 \le k \le M,\ 1 \le s \le n(j,k),$$

or

$$Y_{jks} \;=\; \mu + a_j + b_{jk} + \varepsilon_{jks}, \quad 1 \le j \le L,\ 1 \le k \le M,\ 1 \le s \le n(j,k),$$

Here, the parameters f_{jk}, or the parameters μ, a_j, b_k, h_{jk}, or the parameters μ, a_j, b_{jk} are unknown, and the ε_{jks} are independent random errors with mean 0. The least squares estimators for the f_{jk} are given by the subsample means

$$\widehat{f_{jk}} \;=\; \bar{Y}_{jk\cdot} := \frac{1}{n(j,k)} \sum_{s=1}^{n(j,k)} Y_{jks}.$$

One-way ANCOVA. Starting from $C_i \in \{1, \ldots, L\}$ and $W_i \in \mathbb{R}$, let (Y_{j1}, W_{j1}), (Y_{j2}, W_{j2}), $\ldots$, $(Y_{jn(j)}, W_{jn(j)})$ be those observation pairs (Y_i, W_i) such that $C_i = j$. Then, we may write

$$Y_{js} \;=\; a_j + bW_{js} + \varepsilon_{js}, \quad 1 \le j \le L,\ 1 \le s \le n(j),$$

with unknown parameters $a_1, \ldots, a_L, b$ and independent random errors ε_{js} with mean 0. The least squares estimators for the unknown parameters may be written as

$$\widehat{a}_j \;=\; \bar{Y}_{j\cdot} - \widehat{b}\bar{W}_{j\cdot} \quad \text{and} \quad \widehat{b} \;=\; \frac{\sum_{j=1}^{L}\sum_{s=1}^{n(j)}(W_{js} - \bar{W}_{j\cdot})Y_{js}}{\sum_{j=1}^{L}\sum_{s=1}^{n(j)}(W_{js} - \bar{W}_{j\cdot})^2}.$$

Chapter 3

Tests and Confidence Regions

Throughout this chapter, we study linear models under the assumption that the errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent and normally distributed random variables with mean zero and standard deviation $\sigma > 0$. In later chapters we shall discuss situations in which this assumption is violated. We also hold on to our assumption that the design matrix $\boldsymbol{D} \in \mathbb{R}^{n \times p}$ has rank $p < n$.

3.1 Multivariate Gaussian Distributions

Recall first the definition of univariate Gaussian distributions: A random variable Z follows a *standard normal* or *standard Gaussian* distribution if its distribution is given by the following Lebesgue density:

$$\phi(x) \;:=\; (2\pi)^{-1/2} \exp\Big(-\frac{x^2}{2}\Big).$$

In particular, $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = \mathbb{E}(Z^2) = 1$. This distribution is denoted by $\mathrm{N}(0,1)$.

For $\mu \in \mathbb{R}$ and $\sigma \geq 0$, the *Gaussian (or normal) distribution with mean μ and variance σ^2 (or standard deviation σ)* is defined as the distribution of $X := \mu + \sigma Z$, where $Z \sim \mathrm{N}(0,1)$. This distribution ist denoted by $\mathrm{N}(\mu, \sigma^2)$. In case of $\sigma > 0$ it has density function

$$\phi_{\mu,\sigma^2}(x) \;:=\; \sigma^{-1}\phi(\sigma^{-1}(x-\mu)) \;=\; (2\pi\sigma^2)^{-1/2} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big).$$

An essential property of Gaussian distributions is that the sum of independent random variables with Gaussian distributions follows again a Gaussian distribution. This can be verified, for instance, by means of characteristic functions (Fourier transform).

Now we consider a random vector $\boldsymbol{X} \in \mathbb{R}^k$. Its distribution is completely determined by the distribution of $\boldsymbol{b}^\top \boldsymbol{X}$, if $\boldsymbol{b}$ is running through all unit vectors in $\mathbb{R}^k$. This fact can also be verified with characteristic functions.

Definition 3.1 (Multivariate Gaussian distributions). Let $\boldsymbol{\mu}$ be a vector in $\mathbb{R}^k$, and let Σ be a symmetric, positive semidefinite matrix in $\mathbb{R}^{k \times k}$. A random vector $\boldsymbol{X}$ follows a *Gaussian (or normal) distribution with mean (vector) $\boldsymbol{\mu}$ and covariance (matrix) $\boldsymbol{\Sigma}$* if

$$\boldsymbol{b}^\top \boldsymbol{X} \;\sim\; \mathrm{N}\big(\boldsymbol{b}^\top \boldsymbol{\mu}, \boldsymbol{b}^\top \boldsymbol{\Sigma} \boldsymbol{b}\big)$$

for arbitrary vectors $\boldsymbol{b} \in \mathbb{R}^k$. The symbol for this distribution is $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The special distribution $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{I}_k)$ is called the *k-variate standard Gaussian (or normal) distribution.*

Remark 3.2 (Existence and simulation). For any vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and any symmetric, positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ there exists a random vector $\boldsymbol{X}$ with distribution $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and

$$\mathbb{E}(\boldsymbol{X}) \;=\; \boldsymbol{\mu}, \quad \mathrm{Var}(\boldsymbol{X}) \;=\; \boldsymbol{\Sigma}.$$

To see this, let $\boldsymbol{Z} = (Z_i)_{i=1}^k$ with stochastically independent components $Z_i \sim \mathrm{N}(0,1)$. This vector follows a standard Gaussian distribution. For if we fix $\boldsymbol{b} \in \mathbb{R}^k$, then $\boldsymbol{b}^\top \boldsymbol{Z} = b_1 Z_1 + \cdots + b_k Z_k$ is a sum of independent random variables with Gaussian distributions. Hence, $\boldsymbol{b}^\top \boldsymbol{Z}$ has Gaussian distribution with mean 0 and variance $b_1^2 + \cdots + b_k^2 = \boldsymbol{b}^\top \boldsymbol{b}$.

Now we define

$$\boldsymbol{X} \;:=\; \boldsymbol{\mu} + \boldsymbol{F}\boldsymbol{Z}$$

with a matrix $\boldsymbol{F} \in \mathbb{R}^{k\times k}$ such that $\boldsymbol{F}\boldsymbol{F}^\top = \boldsymbol{\Sigma}$, for instance, $\boldsymbol{F} = \boldsymbol{\Sigma}^{1/2}$. This random vector $\boldsymbol{X}$ has the desired distribution, because for any vector $\boldsymbol{b} \in \mathbb{R}^k$,

$$\boldsymbol{b}^\top \boldsymbol{X} \;=\; \boldsymbol{b}^\top \boldsymbol{\mu} + (\boldsymbol{F}^\top \boldsymbol{b})^\top \boldsymbol{Z}$$

has Gaussian distribution with mean $\boldsymbol{b}^\top \boldsymbol{\mu}$ and variance $\|\boldsymbol{F}^\top \boldsymbol{b}\|^2 = \boldsymbol{b}^\top \boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{b} = \boldsymbol{b}^\top \boldsymbol{\Sigma} \boldsymbol{b}$.

Since $\mathbb{E}(\boldsymbol{Z}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{Z}) = \boldsymbol{I}_k$, the general rules for expectations and covariances imply that $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

Remark 3.3 (Density functions). The k-variate standard Gaussian distribution has the following Lebesgue density $\boldsymbol{\phi}$:

$$\boldsymbol{\phi}(\boldsymbol{x}) \;=\; \prod_{i=1}^k \phi(x_i) \;=\; (2\pi)^{-k/2} \exp\Big(-\frac{\|\boldsymbol{x}\|^2}{2}\Big).$$

For any vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and any symmetric, positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$, the Gaussian distribution $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has Lebesgue density function $\boldsymbol{\phi}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ given by

$$\begin{aligned}\boldsymbol{\phi}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) \;&=\; \det(\boldsymbol{\Sigma})^{-1/2} \boldsymbol{\phi}\big(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})\big) \\ &=\; (2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\Big(-\frac{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}{2}\Big).\end{aligned}$$

This follows from the transformation formula for Lebesgue measure under diffeomorphisms.

An important fact is that the image of a Gaussian random vector under an affine mapping is again Gaussian. With similar arguments as in Remark 3.2 one can prove the following result:

Lemma 3.4. *Let $\boldsymbol{X}$ be a random vector with distribution $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For any vector $\boldsymbol{a} \in \mathbb{R}^\ell$ and any matrix $\boldsymbol{B} \in \mathbb{R}^{\ell\times k}$,*

$$\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X} \;\sim\; \mathrm{N}_\ell\big(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top\big).$$

This lemma implies an essential property of standard Gaussian distributions, their so-called *rotational invariance.*

Corollary 3.5 (Rotational invariance of standard Gaussian distributions). *Let $\boldsymbol{X}$ be a random vector with k-variate standard Gaussian distribution, and let $\boldsymbol{T} \in \mathbb{R}^{k\times k}$ be an orthogonal matrix. Then, $\boldsymbol{TX}$ follows a standard Gaussian distribution too.*

Exercise 3.6. Prove Lemma 3.4.

Exercise 3.7 (Stochastic independence and normal distributions). Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top]^\top$ be a random vector with components $\boldsymbol{X}_i \in \mathbb{R}^{k(i)}$ such that

$$\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{bmatrix} \sim \mathcal{N}_{k(1)+k(2)}\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right).$$

Show that the following assertions are equivalent:

(i) $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are stochastically independent;
(ii) $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top = \boldsymbol{0}$.

Exercise 3.8. Let $\boldsymbol{X}$ be a random vector with distribution $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, and let $\boldsymbol{A}$ be a symmetric matrix in $\mathbb{R}^{k\times k}$. Show that the distribution of $\boldsymbol{X}^\top \boldsymbol{AX}$ depends only on the eigenvalues of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\boldsymbol{\Sigma}^{1/2}$.

3.2 Special Univariate Distributions

In the subsequent sections, the following special distributions will recur frequently:

Definition 3.9 (Chi-squared, student and F distributions). Let $Z_0, Z_1, Z_2, \ldots$ be stochastically independent, standard Gaussian random variables.

(a) The *Chi-squared distribution (χ^2 distribution) with k degrees of freedom* is defined as the distribution of

$$\sum_{i=1}^{k} Z_i^2.$$

It is denoted by the symbol χ_k^2, and its β-quantile is written as $\chi_{k;\beta}^2$.

(b) *Student's t distribution (student distribution, t distribution) with k degrees of freedom* is defined as the distribution of

$$\frac{Z_0}{S} \quad \text{with} \quad S := \sqrt{\frac{1}{k}\sum_{i=1}^{k} Z_i^2}.$$

It is denoted by the symbol t_k, and its β-quantile is written as $t_{k;\beta}$.

(c) *Fisher's F distribution (F distribution) with k and ℓ degrees of freedom* is defined as the distribution of

$$\frac{S^2}{T^2} \quad \text{with} \quad S^2 := \frac{1}{k}\sum_{i=1}^{k} Z_i^2, \quad T^2 := \frac{1}{\ell}\sum_{i=k+1}^{k+\ell} Z_i^2.$$

It is denoted by the symbol $F_{k,\ell}$, and its β-quantile is written as $F_{k,\ell;\beta}$.

The next exercises imply that all distributions in Definition 3.9 have continuous distribution functions.

Exercise 3.10. Let k and ℓ be positive integers.

(a) Show by means of Fubini's theorem that χ^2_k, t_k and $F_{k,\ell}$ have continuous distribution function.

(b) Show that for $0 < \beta < 1$,

$$t_{k;1-\beta} \;=\; -t_{k;\beta} \quad \text{and} \quad F_{k,\ell;1-\beta} \;=\; \frac{1}{F_{\ell,k;\beta}}.$$

Remark 3.11 (Moment-generating functions). The next exercise uses the *moment-generating function* of random variables $Y \in \mathbb{R}$. It is defined as the mapping $\boldsymbol{m}_Y : \mathbb{R} \to (0,\infty]$, $\boldsymbol{m}_Y(t) := \mathbb{E}\exp(tY)$. If $\boldsymbol{m}_Y < \infty$ on a nondegenerate interval, then the distribution of Y is uniquely determined by $\boldsymbol{m}_Y$. That is, if Y' is another random variable such that $\boldsymbol{m}_{Y'} \equiv \boldsymbol{m}_Y$, then the distributions of Y and Y' coincide. Moreover, for stochastically independent random variables Y and Z,

$$\boldsymbol{m}_{Y+Z} \;=\; \boldsymbol{m}_Y \boldsymbol{m}_Z.$$

Exercise 3.12 (Chi-squared and gamma distributions). For $a, b > 0$ let $\mathrm{Gamma}(a,b)$ be the gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. That means, $\mathrm{Gamma}(a,b)$ is a distribution on $(0,\infty)$ with Lebesgue density

$$f_{a,b}(x) \;:=\; b^{-1} f_a(b^{-1}x), \quad f_a(x) \;:=\; \Gamma(a)^{-1} x^{a-1} e^{-x},$$

where $\Gamma(a) := \int_0^\infty x^{a-1}e^{-x}\,dx$. In particular, if $Y \sim \mathrm{Gamma}(a,1)$, then $bY \sim \mathrm{Gamma}(a,b)$.

(i) Show that for $Y \sim \mathrm{Gamma}(a,b)$ and $t \in \mathbb{R}$,

$$\mathbb{E}\exp(tY) \;=\; \begin{cases} (1-bt)^{-a} & \text{if } t < b^{-1}, \\ \infty & \text{if } t \geq b^{-1}. \end{cases}$$

(ii) Show that for $Z \sim \mathrm{N}(0,1)$,

$$\mathbb{E}\exp(tZ^2) \;=\; \begin{cases} (1-2t)^{-1/2} & \text{if } t < 2^{-1}, \\ \infty & \text{if } t \geq 2^{-1}. \end{cases}$$

(iii) Deduce from parts (i–ii) and Remark 3.11 that for any integer $k > 0$,

$$\chi^2_k \;=\; \mathrm{Gamma}(k/2, 2).$$

Exercise 3.13 (Student type distributions). Let Z and $S > 0$ be stochastically independent random variables, where $Z \sim \mathrm{N}(0,1)$ and $\mathbb{E}(S^2) = 1$.

(a) Show by means of Fubini's theorem that $T := Z/S$ has distribution function

$$F(t) \;=\; \mathbb{E}\,\Phi(tS)$$

and Lebesgue density

$$f(t) \;=\; \mathbb{E}(S\phi(tS)).$$

(b) Suppose that $\mathbb{P}(S \neq 1) > 0$. Show by means of Jensen's inequality that

$$F(t) \begin{cases} > \Phi(t) & \text{if } t < 0, \\ < \Phi(t) & \text{if } t > 0. \end{cases}$$

Deduce from that the inequalities

$$F^{-1}(\beta) \begin{cases} < \Phi^{-1}(\beta) & \text{if } \beta < 1/2, \\ > \Phi^{-1}(\beta) & \text{if } \beta > 1/2. \end{cases}$$

Exercise 3.14 (Densities of student distributions). Show by means of Exercises 3.13 and 3.12 that student's t distribution t_k has Lebesgue density

$$f_k(t) \;=\; C_k(1 + t^2/k)^{-(k+1)/2} \quad \text{with} \quad C_k \;=\; \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\,\Gamma(k/2)}.$$

Chi-squared and F distributions play an important role in connection with confidence regions. Behind that is the following basic result:

Lemma 3.15. *Let $\boldsymbol{X}$ be a Gaussian random vector with mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and nonsingular covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$. Then*

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \;\sim\; \chi^2_k.$$

Proof of Lemma 3.15. We may write $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ with a standard Gaussian random vector $\boldsymbol{Z} \in \mathbb{R}^k$. But then

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \;=\; \big\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})\big\|^2 \;=\; \|\boldsymbol{Z}\|^2 \;=\; \sum_{i=1}^k Z_i^2 \;\sim\; \chi^2_k.$$

□

3.3 The Joint Distribution of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$

To simplify the subsequent formulae, we set

$$\boldsymbol{\Gamma} \;:=\; \boldsymbol{D}^\top \boldsymbol{D}.$$

Recall that $\boldsymbol{D}$ is assumed to have full rank, i.e. $\boldsymbol{\Gamma}$ is symmetric and positive definite.

Theorem 3.16. *The estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$ are stochastically independent. Moreover,*

$$\widehat{\boldsymbol{\theta}} \;\sim\; \mathrm{N}_p\big(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Gamma}^{-1}\big), \quad \textit{and} \quad (n-p)\,\frac{\widehat{\sigma}^2}{\sigma^2} \;\sim\; \chi^2_{n-p}.$$

Remark 3.17. This theorem generalizes the well-known result of Gosset–Fisher about the sample mean of a Gaussian sample: Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with distribution $\mathrm{N}(\mu, \sigma^2)$. Then, the sample mean $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and the sample variance $S^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$ are stochastically independent, where

$$\bar{Y} \;\sim\; \mathrm{N}(\mu, \sigma^2/n) \quad \text{and} \quad (n-1)S^2/\sigma^2 \;\sim\; \chi^2_{n-1}.$$

This follows from Theorem 3.16 when $\boldsymbol{D} = \mathbf{1}$ and $\boldsymbol{\theta} = \mu$, leading to $\widehat{\boldsymbol{\theta}} = \bar{Y}$.

Proof of Theorem 3.16. As to the distribution of the LSE, recall that $\widehat{\boldsymbol{\theta}}$ is an affine function of the Gaussian random vector $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. Hence, by Lemma 3.4 and Lemma 2.15,

$$\widehat{\boldsymbol{\theta}} \;\sim\; \mathrm{N}_p\big(\mathbb{E}(\widehat{\boldsymbol{\theta}}), \mathrm{Var}(\widehat{\boldsymbol{\theta}})\big) \;=\; \mathrm{N}_p\big(\boldsymbol{\theta}, \sigma^2 \boldsymbol{\Gamma}^{-1}\big).$$

As to the remaining assertions, let $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n$ be an orthonormal basis of $\mathbb{R}^n$ such that

$$\boldsymbol{M} = \boldsymbol{D}\mathbb{R}^p \;=\; \mathrm{span}(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p),$$

and thus $\boldsymbol{M}^\perp = \mathrm{span}(\boldsymbol{t}_{p+1}, \ldots, \boldsymbol{t}_n)$. Then, the matrix $\boldsymbol{T} := [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n] \in \mathbb{R}^{n \times n}$ is orthogonal, and we may write

$$\boldsymbol{\varepsilon} \;=\; \sigma \boldsymbol{T} \boldsymbol{Z} \;=\; \sigma \sum_{i=1}^n Z_i \boldsymbol{t}_i$$

with the random vector $\boldsymbol{Z} := \sigma^{-1} \boldsymbol{T}^\top \boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{I}_n)$; see Lemma 3.4. On the one hand, it follows from $\boldsymbol{D}^\top \boldsymbol{t}_i = 0$ for $i > p$ that

$$\widehat{\boldsymbol{\theta}} \;=\; \boldsymbol{\Gamma}^{-1} \boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \;=\; \boldsymbol{\theta} + \sigma \boldsymbol{\Gamma}^{-1} \boldsymbol{D}^\top \sum_{i=1}^n Z_i \boldsymbol{t}_i \;=\; \boldsymbol{\theta} + \sigma \boldsymbol{\Gamma}^{-1} \boldsymbol{D}^\top \sum_{i=1}^p Z_i \boldsymbol{t}_i,$$

that means, $\widehat{\boldsymbol{\theta}}$ is a function of $(Z_i)_{i=1}^p$. On the other hand, with $\bar{\boldsymbol{H}} = \boldsymbol{I}_n - \boldsymbol{H}$, the orthogonal projection of $\boldsymbol{Y}$ onto $\boldsymbol{M}^\perp$ is given by

$$\bar{\boldsymbol{H}}(\boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \;=\; \bar{\boldsymbol{H}}\boldsymbol{\varepsilon} \;=\; \sigma \sum_{i=p+1}^n Z_i \boldsymbol{t}_i,$$

because $\bar{\boldsymbol{H}} \boldsymbol{D} = \mathbf{0}$ and $\bar{\boldsymbol{H}} \boldsymbol{t}_i = 0$ for $i \le p$. Hence

$$\widehat{\sigma}^2 \;=\; \frac{\|\bar{\boldsymbol{H}}\boldsymbol{Y}\|^2}{n-p} \;=\; \frac{\sigma^2}{n-p} \Big\| \sum_{i=p+1}^n Z_i \boldsymbol{t}_i \Big\|^2 \;=\; \frac{\sigma^2}{n-p} \sum_{i=p+1}^n Z_i^2.$$

This shows that $\widehat{\sigma}$ is a function of $(Z_i)_{i=p+1}^n$, whence $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$ are stochastically independent. Moreover, $(n-p)\widehat{\sigma}^2/\sigma^2 = \sum_{i=p+1}^n Z_i^2$ follows a chi-squared distribution with $n-p$ degrees of freedom. □

A first application: confidence intervals for σ. It follows from Theorem 3.16 that for arbitrary $0 \le \beta_1 < \beta_2 \le 1$,

$$\mathbb{P}\Big(\chi^2_{n-p;\beta_1} \le (n-p)\,\frac{\widehat{\sigma}^2}{\sigma^2} \le \chi^2_{n-p;\beta_2}\Big) \;=\; \beta_2 - \beta_1,$$

where we set $\chi^2_{n-p;0} := 0$ and $\chi^2_{n-p;1} := \infty$. That means, with probability $\beta_2 - \beta_1$, the unknown standard deviation σ is contained in the interval

$$\Bigg[\widehat{\sigma}\sqrt{\frac{n-p}{\chi^2_{n-p;\beta_2}}},\, \widehat{\sigma}\sqrt{\frac{n-p}{\chi^2_{n-p;\beta_1}}}\Bigg].$$

If we fix a test level $\alpha \in (0,1)$ and set $\beta_1 = 0, \beta_2 = 1-\alpha$, this leads to the lower $(1-\alpha)$-confidence bound

$$\widehat{\sigma}\sqrt{\frac{n-p}{\chi^2_{n-p;1-\alpha}}}$$

for σ. The choice $\beta_1 = \alpha$ and $\beta_2 = 1$ leads to the upper $(1-\alpha)$-confidence bound

$$\widehat{\sigma}\sqrt{\frac{n-p}{\chi^2_{n-p;\alpha}}}$$

for σ. If we choose $\beta_1 \in (0, \alpha)$ and $\beta_2 = 1 - \alpha + \beta_1$, we obtain a bounded $(1-\alpha)$-confidence interval for σ. A standard choice for β_1 would be $\alpha/2$, but this is not necessarily the best choice.

Exercise 3.18. Write a computer program that returns for a given integer $k > 0$ and test level $\alpha \in (0,1)$ a number $\beta = \beta(k,\alpha) \in (0,\alpha)$ and the pair

$$(L, U) \;=\; \left(\sqrt{\frac{k}{\chi^2_{k;1-\alpha+\beta}}},\, \sqrt{\frac{k}{\chi^2_{k;\beta}}}\right)$$

such that the ratio U/L is approximately minimal.

3.4 Student Confidence Intervals and Tests

This section is about inference for a single linear function $\boldsymbol{\psi}^\top\boldsymbol{\theta}$ of $\boldsymbol{\theta}$ with a given vector $\boldsymbol{\psi} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$. Particularly useful are confidence regions for $\boldsymbol{\psi}^\top\boldsymbol{\theta}$. As a by-product one obtains a p-value for the null hypothesis "$\boldsymbol{\psi}^\top\boldsymbol{\theta} = 0$", as explained later.

3.4.1 Student Confidence Regions

For the construction of confidence regions we consider the *student statistic*

$$T_{\boldsymbol{\psi}} \;:=\; \frac{\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}}{\widehat{\sigma}_{\boldsymbol{\psi}}}$$

with the standard error

$$\widehat{\sigma}_{\boldsymbol{\psi}} \;=\; \widehat{\sigma}\sqrt{\boldsymbol{\psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\psi}} \;=\; \widehat{\sigma}\|\boldsymbol{a}_{\boldsymbol{\psi}}\|,$$

where $\boldsymbol{a}_{\boldsymbol{\psi}} := \boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\psi}$. This standard error is our substitute for the true standard deviation

$$\sigma_{\boldsymbol{\psi}} \;=\; \sigma\sqrt{\boldsymbol{\psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\psi}} \;=\; \sigma\|\boldsymbol{a}_{\boldsymbol{\psi}}\|$$

of the GME $\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}}$.

Note that the random variable $T_{\boldsymbol{\psi}}$ involves the data (via $\widehat{\theta}$ and $\widehat{\sigma}$) as well as the unknown parameter $\boldsymbol{\theta}$. It is a *pivotal statistic* in the sense that its distribution does not depend on any unknown parameters.

Corollary 3.19. *For any vector $\boldsymbol{\psi} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$, the random variable $T_{\boldsymbol{\psi}}$ follows a student distribution with $n-p$ degrees of freedom.*

Proof of Corollary 3.19. It follows from Lemma 3.4 and Theorem 3.16 that

$$Z \;:=\; \frac{\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}}{\sigma_{\boldsymbol{\psi}}}$$

has a standard Gaussian distribution and is stochastically independent from $\widehat{\sigma}$. Moreover, $S^2 := (n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$, so

$$T_{\boldsymbol{\psi}} \;=\; \frac{Z}{\sqrt{\widehat{\sigma}^2/\sigma^2}} \;=\; \frac{Z}{\sqrt{(n-p)^{-1}S^2}}$$

follows the asserted student distribution. □

A particular consequence of Corollary 3.19 is that for any test level $\alpha \in (0,1)$,

$$\left.\begin{array}{c} \mathbb{P}\big(T_{\boldsymbol{\psi}} \le t_{n-p;1-\alpha}\big) \\ \mathbb{P}\big(T_{\boldsymbol{\psi}} \ge -t_{n-p;1-\alpha}\big) \\ \mathbb{P}\big(|T_{\boldsymbol{\psi}}| \le t_{n-p;1-\alpha/2}\big) \end{array}\right\} \;=\; 1-\alpha.$$

Converting the inequalities for $T_{\boldsymbol{\psi}}$ into inequalities for $\boldsymbol{\psi}^\top\boldsymbol{\theta}$ leads to the following confidence regions: The

- lower $(1-\alpha)$-confidence bound $\quad \boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \widehat{\sigma}_{\boldsymbol{\psi}} t_{n-p;1-\alpha}$,
- upper $(1-\alpha)$-confidence bound $\quad \boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} + \widehat{\sigma}_{\boldsymbol{\psi}} t_{n-p;1-\alpha}$,
- $(1-\alpha)$-confidence interval $\quad \big[\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}_{\boldsymbol{\psi}} t_{n-p;1-\alpha/2}\big]$.

Which of these confidence regions is to be used, has to be agreed on prior to the data analysis.

Since student distributions are continuous, one could even interpret the one-sided bounds as strict bounds and replace the closed with the open confidence interval $\big(\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}_{\boldsymbol{\psi}} t_{n-p;1-\alpha/2}\big)$ without changing the exact confidence level $1-\alpha$.

Example 3.20 (Simple linear regression, Example 1.2). Here $X \in \mathbb{R}$ and $f(x) = a + bx$. With the centered vector $\tilde{\boldsymbol{X}} := (X_i - \bar{X})_{i=1}^n$,

$$\widehat{b} \;=\; \frac{\tilde{\boldsymbol{X}}^\top \boldsymbol{Y}}{\|\tilde{\boldsymbol{X}}\|^2} \;\sim\; \mathrm{N}\Big(b, \frac{\sigma^2}{\|\tilde{\boldsymbol{X}}\|^2}\Big).$$

Hence, a $(1-\alpha)$-confidence interval for the slope parameter b is given by

$$\Big[\widehat{b} \pm \frac{\widehat{\sigma}}{\|\tilde{\boldsymbol{X}}\|}\, t_{n-2;1-\alpha/2}\Big].$$

Next we consider the regression function at an arbitrary fixed point $x \in \mathbb{R}$. Recall that

$$\widehat{f}(x) \;=\; \widehat{a} + \widehat{b}x \;=\; \bar{Y} + \widehat{b}(x - \bar{X}) \;\sim\; \mathrm{N}\big(f(x), \sigma(x)^2\big)$$

with

$$\sigma(x) \;:=\; \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{\|\widetilde{\boldsymbol{X}}\|^2}}.$$

If we replace the factor σ in $\sigma(x)$ by $\widehat{\sigma}$, then we obtain the standard error $\widehat{\sigma}(x)$ of $\widehat{f}(x)$. A $(1-\alpha)$-confidence interval for $f(x)$ is given by

$$\big[\widehat{f}(x) \pm \widehat{\sigma}(x) t_{n-2;1-\alpha/2}\big].$$

As a function of x, the upper and lower bound are hyperbolas with asymptotes

$$x \;\mapsto\; \widehat{f}(x) \,\pm\, \widehat{\sigma}\,\frac{|x-\bar{X}|}{\|\tilde{\boldsymbol{X}}\|}\,t_{n-2;1-\alpha/2}.$$

Note that

$$\liminf_{|x|\to\infty}\; \min\Big\{|y| \,:\, y\in\big[\widehat{f}(x) \,\pm\, \widehat{\sigma}(x)t_{n-2;1-\alpha/2}\big]\Big\} \;=\; \infty$$

if and only if

$$|\widehat{b}| \;>\; \frac{\widehat{\sigma}}{\|\tilde{\boldsymbol{X}}\|}\,t_{n-2;1-\alpha/2},$$

that means, the closed confidence interval for b does not contain 0.

Figure 3.1 shows a data set with $n = 51$ observations, together with the regression line $\widehat{f}$ and the 95%-confidence bounds plus asymptotes. The corresponding student quantile is $t_{n-2;1-\alpha/2} = t_{49;0.975} = 2.010$ (rounded to three decimals).

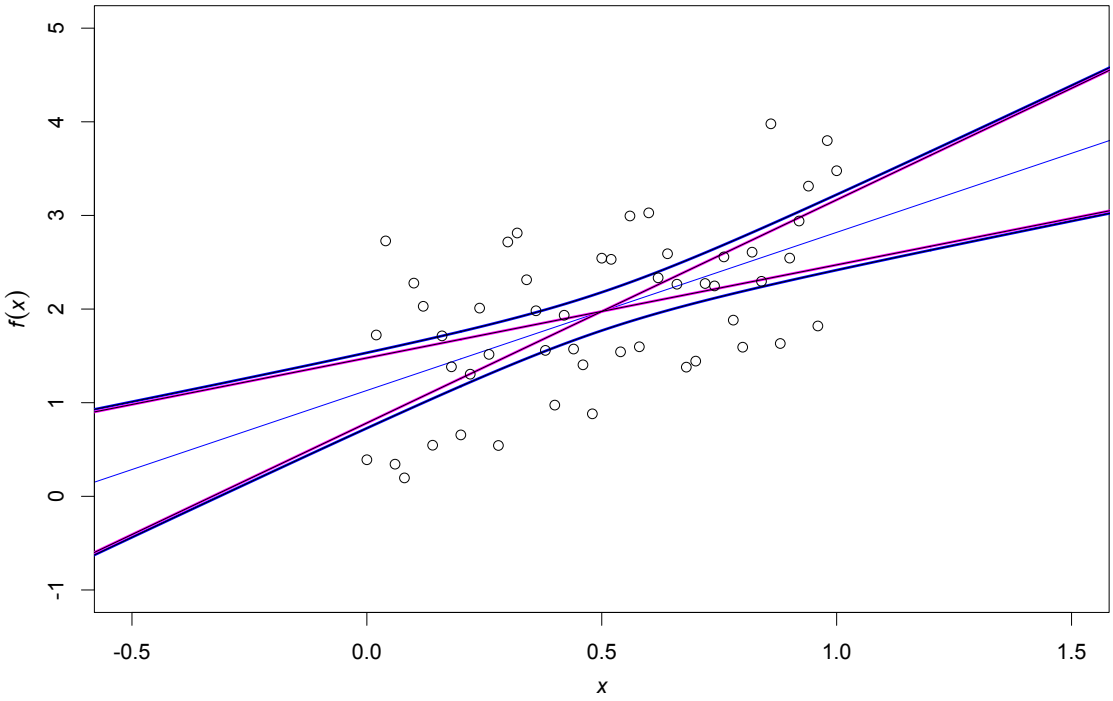

Figure 3.1: Regression line plus pointwise 95%-confidence intervals for $f(x)$.

Example 3.21 (One-way ANCOVA, Example 1.4). We consider covariates $C \in \{1, 2, \ldots, L\}$ and $W \in \mathbb{R}$, and the model equation reads $Y = a(C) + bW + \varepsilon$. For $1 \le j < k \le L$,

$$\widehat{a}(k) - \widehat{a}(j) \;=\; \bar{Y}(k) - \bar{Y}(j) - \widehat{b}\big(\bar{W}(k) - \bar{W}(j)\big) \;\sim\; \mathrm{N}\big(a(k) - a(j), \sigma(j,k)^2\big)$$

with the standard deviation

$$\sigma(j,k) \;:=\; \sigma\sqrt{\frac{1}{n(j)} + \frac{1}{n(k)} + \frac{\big(\bar{W}(k) - \bar{W}(j)\big)^2}{\|\tilde{\boldsymbol{W}}\|^2}},$$

where $\tilde{\boldsymbol{W}} := \big(W_i - \bar{W}(C_i)\big)_{i=1}^n$. Again we replace σ with $\widehat{\sigma}$ and obtain the standard error $\widehat{\sigma}(j,k)$ of $\widehat{a}(k) - \widehat{a}(j)$. Then

$$\big[\widehat{a}(k) - \widehat{a}(j) \,\pm\, \widehat{\sigma}(j,k) t_{n-L-1;1-\alpha/2}\big]$$

is a $(1-\alpha)$-confidence interval for $a(k) - a(j)$.

Exercise 3.22. Consider the data set 'Trees.txt' and the variables $Y := \log(\text{volume})$, $X(1) := \log(\text{height})$ and $X(2) := \log(\text{diameter})$. Compute a 95%-confidence interval for each of the parameters a, b_1, b_2 in the model equation

$$Y = a + b_1 X(1) + b_2 X(2) + \varepsilon.$$

Which conclusion can you draw about the model equation $Y = a + X(1) + 2X(2) + \varepsilon$?

3.4.2 Student Tests

Instead of confidence bounds or intervals for the parameter $\boldsymbol{\psi}^\top \boldsymbol{\theta}$ one may test hypotheses about it. For any fixed constant c_o, we introduce the test statistic

$$T_{\boldsymbol{\psi}}(c_o) \;:=\; \frac{\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} - c_o}{\widehat{\sigma}_{\boldsymbol{\psi}}}.$$

Moreover, let $\mathrm{tcdf}_k(\cdot)$ be the distribution function of t_k.

One-sided tests. Suppose we want to test

$$H_o : \boldsymbol{\psi}^\top \boldsymbol{\theta} \le c_o \quad \text{versus} \quad H_A : \boldsymbol{\psi}^\top \boldsymbol{\theta} > c_o.$$

Then, the null hypothesis H_o may be rejected at level $\alpha \in (0,1)$ if

$$T_{\boldsymbol{\psi}}(c_o) \;\ge\; t_{n-p;1-\alpha}.$$

This is equivalent to the right-sided p-value

$$1 - \mathrm{tcdf}_{n-p}(T_{\boldsymbol{\psi}}(c_o)) \;=\; \mathrm{tcdf}_{n-p}(-T_{\boldsymbol{\psi}}(c_o))$$

being less than or equal to α.

The level of this test can be checked as follows:

$$\begin{aligned} \mathbb{P}(H_o \text{ is rejected}) \;&=\; \mathbb{P}\big(T_{\boldsymbol{\psi}}(c_o) \ge\; t_{n-p;1-\alpha}\big) \\ &=\; \mathbb{P}\Big(T_{\boldsymbol{\psi}} + \frac{\boldsymbol{\psi}^\top \boldsymbol{\theta} - c_o}{\widehat{\sigma}_{\boldsymbol{\psi}}} \ge\; t_{n-p;1-\alpha}\Big). \end{aligned}$$

Under the null hypothesis H_o, the right hand side is less than or equal to

$$\mathbb{P}\big(T_{\boldsymbol{\psi}} \ge\; t_{n-p;1-\alpha}\big) \;=\; \alpha,$$

with equality in case of $\boldsymbol{\psi}^\top \boldsymbol{\theta} = c_o$.

Analogously, if we want to test

$$H_o : \boldsymbol{\psi}^\top \boldsymbol{\theta} \geq c_o \quad \text{versus} \quad H_A : \boldsymbol{\psi}^\top \boldsymbol{\theta} < c_o,$$

then we may reject the null hypothesis H_o at level $\alpha \in (0,1)$ if

$$T_{\boldsymbol{\psi}}(c_o) \;\leq\; -t_{n-p;1-\alpha},$$

which is equivalent to the left-sided p-value

$$\mathrm{tcdf}_{n-p}(T_{\boldsymbol{\psi}}(c_o))$$

being less than or equal to α.

Two-sided test. Now we consider the testing problem

$$H_o : \boldsymbol{\psi}^\top \boldsymbol{\theta} = c_o \quad \text{versus} \quad H_A : \boldsymbol{\psi}^\top \boldsymbol{\theta} \neq c_o.$$

Here one may reject the null hypothesis at level α if

$$|T_{\boldsymbol{\psi}}(c_o)| \;\geq\; t_{n-p;1-\alpha/2}.$$

The corresponding p-value is equal to

$$2 \cdot \big(1 - \mathrm{tcdf}_{n-p}\big(|T_{\boldsymbol{\psi}}(c_o)|\big)\big) \;=\; 2 \cdot \mathrm{tcdf}_{n-p}\big(-|T_{\boldsymbol{\psi}}(c_o)|\big).$$

In case of rejection, one may even claim with confidence $1-\alpha$ that

$$\boldsymbol{\psi}^\top \boldsymbol{\theta} \begin{cases} < \; c_o & \text{if } \boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} < c_o, \\ > \; c_o & \text{if } \boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} > c_o. \end{cases}$$

The reason for that is that the two-sided test rejects the null hypothesis if and only if c_o is not contained in the open $(1-\alpha)$-confidence interval $\big(\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}_{\boldsymbol{\psi}} t_{n-p;1-\alpha/2}\big)$ for $\boldsymbol{\psi}^\top \boldsymbol{\theta}$.

Example 3.23 (Confidence region for a cusp). We consider the linear model of quadratic regression, that is, $X \in \mathbb{R}$ and $Y = f(X) + \varepsilon$ with $f(x) = a_0 + a_1 x + a_2 x^2/2$. (The factor $1/2$ for x^2 will turn out to be convenient later.) Under the additional assumption that $a_2 < 0$, we want to determine a confidence region for the maximizer

$$x_* \;=\; \frac{-a_1}{a_2}$$

of f, which is the unique point $x \in \mathbb{R}$ such that $f'(x) = a_1 + a_2 x$ equals 0.

A naive approach would be to compute both for a_1 and a_2 a $(1-\alpha/2)$-confidence interval, and to deduce from that bounds for the ratio $x_* = -a_1/a_2$. A more elegant method is to test the null hypothesis

$$H_o(x) : f'(x) = 0$$

for each point $x \in \mathbb{R}$ at level α. Note that the GME of $f'(x)$ is given by

$$\widehat{f}'(x) \;=\; \widehat{a}_1 + \widehat{a}_2 x$$

Its standard error equals

$$\widehat{\sigma}(x) \;=\; \sqrt{\widehat{\Sigma}_{11} + 2\widehat{\Sigma}_{12}x + \widehat{\Sigma}_{22}x^2},$$

where

$$\begin{bmatrix} \widehat{\Sigma}_{00} & \widehat{\Sigma}_{01} & \widehat{\Sigma}_{02} \\ \widehat{\Sigma}_{10} & \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{20} & \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{bmatrix} \;:=\; \widehat{\sigma}^2 \boldsymbol{\Gamma}^{-1}$$

is the estimated covariance matrix of $\widehat{\boldsymbol{\theta}} = (\widehat{a}_0, \widehat{a}_1, \widehat{a}_2)^\top$. Now we set

$$\begin{aligned} C_\alpha = C_\alpha(\boldsymbol{X}, \boldsymbol{Y}) \;&:=\; \big\{x \in \mathbb{R} : H_o(x) \text{ is not rejected at level } \alpha\big\} \\ &=\; \big\{x \in \mathbb{R} : |\widehat{f}'(x)| < \widehat{\sigma}(x) t_{n-3;1-\alpha/2}\big\}. \end{aligned}$$

This defines an exact $(1-\alpha)$-confidence region for x_*, because

$$\begin{aligned} \mathbb{P}\big(C_\alpha \ni x_*\big) \;&=\; \mathbb{P}\big(|\widehat{f}'(x_*)| < \widehat{\sigma}(x) t_{n-3;1-\alpha/2}\big) \\ &=\; \mathbb{P}\big(\big|\widehat{f}'(x_*) - f'(x_*)\big| < \widehat{\sigma}(x_*) t_{n-3;1-\alpha/2}\big) \\ &=\; 1-\alpha. \end{aligned}$$

The question is, whether C_α is really useful. With $\tau := t_{n-p;1-\alpha/2}$ we may write

$$\begin{aligned} C_\alpha \;&=\; \big\{x \in \mathbb{R} : \widehat{a}_1^2 + 2\widehat{a}_1\widehat{a}_2 x + \widehat{a}_2^2 x^2 < \widehat{\Sigma}_{11}\tau^2 + 2\widehat{\Sigma}_{12}\tau^2 x + \widehat{\Sigma}_{22}\tau^2 x^2\big\} \\ &=\; \big\{x \in \mathbb{R} : (\widehat{a}_2^2 - \widehat{\Sigma}_{22}\tau^2)x^2 + 2(\widehat{a}_1\widehat{a}_2 - \widehat{\Sigma}_{12}\tau^2)x < \widehat{\Sigma}_{11}\tau^2 - \widehat{a}_1^2\big\}. \end{aligned}$$

In case of $\widehat{a}_2^2 > \widehat{\Sigma}_{22}\tau^2$, this is a bounded open interval with midpoint

$$\widehat{x}_* \;=\; \frac{\widehat{\Sigma}_{12}\tau^2 - \widehat{a}_1\widehat{a}_2}{\widehat{a}_2^2 - \widehat{\Sigma}_{22}\tau^2}.$$

Interestingly, the condition $\widehat{a}_2^2 > \widehat{\Sigma}_{22}\tau^2$ is equivalent to

$$|\boldsymbol{\psi}^\top \boldsymbol{\theta}| \;>\; \widehat{\sigma}_{\boldsymbol{\psi}} t_{n-3;1-\alpha/2} \quad \text{with } \boldsymbol{\psi} = (0,0,1)^\top,$$

that means, the p-value for the null hypothesis "$a_2 = 0$" is strictly smaller than α, and the (closed) confidence interval for a_2 does not contain the value 0.

Exercise 3.24. Complement and implement the procedure in Example 3.23.

Exercise 3.25. Consider simple linear regression, that means, $X \in \mathbb{R}$ and $Y = f(X) + \varepsilon$ with an unknown regression function $f(x) = a + bx$. Assuming that $b \neq 0$, construct a $(1-\alpha)$-confidence region for the unique point

$$x_* \;:=\; f^{-1}(0) \;=\; \frac{-a}{b}.$$

3.5 F Confidence Regions and Tests

So far we constructed confidence intervals for a single linear function $\boldsymbol{\psi}^\top \boldsymbol{\theta}$ of $\boldsymbol{\theta}$. Now our goal is to construct a confidence region for the full vector $\boldsymbol{\theta}$ or for a tuple $\big(\boldsymbol{\psi}^\top \boldsymbol{\theta}\big)_{\boldsymbol{\psi} \in \mathcal{P}}$ with an arbitrary subset $\mathcal{P}$ of $\mathbb{R}^p$, finite or infinite.

3.5.1 F Tests and Confidence Ellipsoids

A possible measure of the distance between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ is given by the *F statistic*

$$F \;:=\; \frac{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^\top \boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})}{p\,\widehat{\sigma}^2}.$$

Note that this random variable F involves the data as well as the unknown parameter $\boldsymbol{\theta}$. As explained in the next corollary, it is a *pivotal statistic* in the sense that its distribution does not depend on any unknown parameters.

Corollary 3.26. *The random variable F has distribution $F_{p,n-p}$.*

Proof of Corollary 3.26. Recall from Theorem 3.16 that $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}$ are stochastically independent, where $\widehat{\boldsymbol{\theta}} \sim \mathrm{N}_p(\boldsymbol{\theta}, \sigma^2\boldsymbol{\Gamma}^{-1})$ and $S^2 := (n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$. This implies that

$$S_o^2 \;:=\; \frac{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^\top \boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})}{\sigma^2} \;=\; \left\|\sigma^{-1}\boldsymbol{\Gamma}^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})\right\|^2$$

has distribution χ^2_p, because $\sigma^{-1}\boldsymbol{\Gamma}^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is a standard Gaussian random vector. Moreover, S_o^2 and S^2 are stochastically independent. Consequently,

$$F \;=\; \frac{S_o^2}{p\,\widehat{\sigma}^2/\sigma^2} \;=\; \frac{S_o^2/p}{S^2/(n-p)}$$

follows an F distribution with p and $n-p$ degrees of freedom. □

Corollary 3.26 leads to two statistical procedures. On the one hand, let $\boldsymbol{\theta}_o$ be a given candidate $\boldsymbol{\theta}_o$ for the unknown parameter vector $\boldsymbol{\theta}$. When testing

$$H_o : \boldsymbol{\theta} = \boldsymbol{\theta}_o \quad \text{versus} \quad H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_o,$$

we may reject the null hypothesis H_o at level α if the F test statistic

$$F(\boldsymbol{\theta}_o) \;:=\; \frac{(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)^\top \boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)}{p\,\widehat{\sigma}^2}$$

exceeds $F_{p,n-p;1-\alpha}$. This is equivalent to the (right-sided) p-value

$$1 - \mathrm{Fcdf}_{p,n-p}(F(\boldsymbol{\theta}_o))$$

being less than or equal to α, where $\mathrm{Fcdf}_{k,\ell}$ denotes the distribution function of $F_{k,\ell}$.

This test has exact level α, because under H_o, the F test statistic $F(\boldsymbol{\theta}_o)$ coincides with the pivotal statistic F and has distribution $F_{p,n-p}$.

On the other hand,

$$\begin{aligned} C_\alpha = C_\alpha(\boldsymbol{X},\boldsymbol{Y}) \;&:=\; \left\{\boldsymbol{\eta} \in \mathbb{R}^p : F(\boldsymbol{\eta}) \le F_{p,n-p;1-\alpha}\right\} \\ &=\; \left\{\boldsymbol{\eta} \in \mathbb{R}^p : (\widehat{\boldsymbol{\theta}}-\boldsymbol{\eta})^\top \boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\eta}) \le \widehat{\sigma}^2 p F_{p,n-p;1-\alpha}\right\} \end{aligned}$$

defines an exact $(1-\alpha)$-confidence region for $\boldsymbol{\theta}$. That means,

$$\mathbb{P}(C_\alpha \ni \boldsymbol{\theta}) \;=\; 1-\alpha,$$

because $C_\alpha \ni \boldsymbol{\theta}$ if and only if $F(\boldsymbol{\theta}) = F$ is not greater than $F_{p,n-p;1-\alpha}$, and the latter event has probability $1-\alpha$.

Concerning the shape of C_α, if $\boldsymbol{\Gamma}$ is the identity matrix $\boldsymbol{I}_p$, then C_α is a closed ball in $\mathbb{R}^p$ with center $\widehat{\boldsymbol{\theta}}$ and radius $\widehat{c} := \widehat{\sigma}\sqrt{pF_{p,n-p;1-\alpha}}$. In general, let $\boldsymbol{\Gamma} = \sum_{j=1}^p \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^\top$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p > 0$ and orthonormal eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p \in \mathbb{R}^p$. Then, $\boldsymbol{\Gamma}^{\pm 1/2} = \sum_{j=1}^p \lambda_j^{\pm 1/2} \boldsymbol{u}_j \boldsymbol{u}_j^\top$, and

$$\begin{aligned}
C_\alpha \;&=\; \left\{\widehat{\boldsymbol{\theta}} + \widehat{c}\boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^p, \boldsymbol{v}^\top \boldsymbol{\Gamma} \boldsymbol{v} \leq 1\right\} \\
&=\; \left\{\widehat{\boldsymbol{\theta}} + \widehat{c}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^p, \|\boldsymbol{w}\| \leq 1\right\} \qquad \big(\text{with } \boldsymbol{w} = \boldsymbol{\Gamma}^{1/2}\boldsymbol{v}\big) \\
&=\; \left\{\widehat{\boldsymbol{\theta}} + \widehat{c}\sum_{j=1}^p x_j \, \lambda_j^{-1/2} \boldsymbol{u}_j : \boldsymbol{x} \in \mathbb{R}^p, \|\boldsymbol{x}\| \leq 1\right\} \qquad \big(\text{with } x_j = \boldsymbol{u}_j^\top \boldsymbol{w}\big).
\end{aligned}$$

That means, the closed unit ball in $\mathbb{R}^p$ with center $\mathbf{0}$ is rescaled in direction $\boldsymbol{u}_j$ by the factor $\widehat{c}\lambda_j^{-1/2}$ for $j = 1, \ldots, p$. This leads to an ellipsoid centered at $\mathbf{0}$, and finally, the latter set is shifted by $\widehat{\boldsymbol{\theta}}$.

3.5.2 Simultaneous Confidence Intervals via F Tests

The confidence ellipsoid C_α for $\boldsymbol{\theta}$ can be visualized in case of $p \leq 3$, but it is not so clear what to do with it in dimension $p > 3$. The following lemma and corollary provide a very useful connection between the F and student test statistics.

Lemma 3.27 (Henry Scheffé). *For any vector $\boldsymbol{v} \in \mathbb{R}^p$,*

$$\sqrt{\boldsymbol{v}^\top \boldsymbol{\Gamma} \boldsymbol{v}} \;=\; \max_{\boldsymbol{\psi} \in \mathbb{R}^p \setminus \{\mathbf{0}\}} \frac{|\boldsymbol{\psi}^\top \boldsymbol{v}|}{\sqrt{\boldsymbol{\psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\psi}}}.$$

In particular, for any constant $c \geq 0$,

$$\boldsymbol{v}^\top \boldsymbol{\Gamma} \boldsymbol{v} \;\leq\; c^2;$$

if and only if

$$|\boldsymbol{\psi}^\top \boldsymbol{v}| \;\leq\; c\sqrt{\boldsymbol{\psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\psi}} \quad \textit{for all } \boldsymbol{\psi} \in \mathbb{R}^p.$$

This lemma is true for any symmetric and positive definite matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times p}$. In our special case of $\boldsymbol{\Gamma} = \boldsymbol{D}^\top \boldsymbol{D}$, recall that

$$\sqrt{\boldsymbol{\psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\psi}} \;=\; \frac{\widehat{\sigma}_{\boldsymbol{\psi}}}{\widehat{\sigma}} \quad \text{and} \quad (\widehat{\boldsymbol{\theta}} - \boldsymbol{\eta})^\top \boldsymbol{\Gamma} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\eta}) \;=\; p\widehat{\sigma}^2 F(\boldsymbol{\eta}).$$

Hence, applying Lemma 3.27 to $\boldsymbol{v} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\eta}$ and $c = \widehat{\sigma}\sqrt{pF_{p,n-p;1-\alpha}}$ yields the aforementioned connection between F and student statistics:

Corollary 3.28 (Henry Scheffé). *For any vector $\boldsymbol{\eta} \in \mathbb{R}^p$,*

$$\sqrt{pF(\boldsymbol{\eta})} \;=\; \max_{\boldsymbol{\psi} \in \mathbb{R}^p \setminus \{\mathbf{0}\}} |T_{\boldsymbol{\psi}}(\boldsymbol{\psi}^\top \boldsymbol{\eta})|,$$

and thus,

$$\sqrt{pF} \;=\; \max_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{\mathbf{0}\}} |T_{\boldsymbol{\psi}}|.$$

In particular, $\boldsymbol{\theta}$ lies in the confidence ellipsoid C_α if and only if

$$\left|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}\right| \;\le\; \widehat{\sigma}_{\boldsymbol{\psi}}\sqrt{pF_{p,n-p;1-\alpha}} \quad \textit{for all } \boldsymbol{\psi}\in\mathbb{R}^p.$$

This corollary shows that the confidence region C_α yields *simultaneous $(1-\alpha)$-confidence intervals* for arbitrary linear functions $\boldsymbol{\psi}^\top\boldsymbol{\theta}$ of $\boldsymbol{\theta}$. Namely, we replace the student confidence intervals

$$\left[\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}_{\boldsymbol{\psi}} t_{n-p;1-\alpha/2}\right]$$

with the confidence intervals

$$\left[\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}_{\boldsymbol{\psi}}\sqrt{pF_{p,n-p;1-\alpha}}\right],$$

i.e. we replace the student quantile $t_{n-p;1-\alpha/2}$ with the quantity $\sqrt{pF_{p,n-p;1-\alpha}}$. Since $C_\alpha \ni \boldsymbol{\theta}$ with probability $1-\alpha$, we may claim with confidence $1-\alpha$ that

$$\boldsymbol{\psi}^\top\boldsymbol{\theta} \;\in\; \left[\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}_{\boldsymbol{\psi}}\sqrt{pF_{p,n-p;1-\alpha}}\right] \quad \text{for all } \boldsymbol{\psi}\in\mathbb{R}^p.$$

Proof of Lemma 3.27. Note first that $\boldsymbol{v}^\top\boldsymbol{\Gamma}\boldsymbol{v} = \|\boldsymbol{\Gamma}^{1/2}\boldsymbol{v}\|^2$. It follows from the Cauchy-Schwarz inequality that for any $\boldsymbol{w}\in\mathbb{R}^p$,

$$\left|\boldsymbol{w}^\top\boldsymbol{\Gamma}^{1/2}\boldsymbol{v}\right| \;\le\; \|\boldsymbol{w}\|\|\boldsymbol{\Gamma}^{1/2}\boldsymbol{v}\|$$

with equality if $\boldsymbol{w}$ and $\boldsymbol{\Gamma}^{1/2}\boldsymbol{v}$ are collinear. Hence,

$$\begin{aligned}
\|\boldsymbol{\Gamma}^{1/2}\boldsymbol{v}\| \;&=\; \max_{\boldsymbol{w}\in\mathbb{R}^p\setminus\{\mathbf{0}\}} \frac{\left|\boldsymbol{w}^\top\boldsymbol{\Gamma}^{1/2}\boldsymbol{v}\right|}{\|\boldsymbol{w}\|} \\
&=\; \max_{\boldsymbol{w}\in\mathbb{R}^p\setminus\{\mathbf{0}\}} \frac{\left|(\boldsymbol{\Gamma}^{1/2}\boldsymbol{w})^\top\boldsymbol{v}\right|}{\left\|\boldsymbol{\Gamma}^{-1/2}(\boldsymbol{\Gamma}^{1/2}\boldsymbol{w})\right\|} \\
&=\; \max_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{\mathbf{0}\}} \frac{|\boldsymbol{\psi}^\top\boldsymbol{v}|}{\|\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi}\|} \\
&=\; \max_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{\mathbf{0}\}} \frac{|\boldsymbol{\psi}^\top\boldsymbol{v}|}{\sqrt{\boldsymbol{\psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\psi}}}.
\end{aligned}$$

□

Example 3.20 (Simple linear regression, continued). As seen before, for any fixed $x\in\mathbb{R}$, a $(1-\alpha)$-confidence interval for the value $f(x)$ is given by $\left[\widehat{f}(x) \pm \widehat{\sigma}(x)t_{n-2;1-\alpha/2}\right]$. Now we replace the student quantile $t_{n-2;1-\alpha/2}$ with $\sqrt{2F_{2,n-2;1-\alpha}}$ and obtain the interval

$$\left[\widehat{f}(x) \;\pm\; \widehat{\sigma}(x)\sqrt{2F_{2,n-2;1-\alpha}}\right].$$

We may claim with confidence $1-\alpha$ that $f(x)$ lies within this interval, *simultaneously for all* $x\in\mathbb{R}$. This is a direct consequence of Corollary 3.28. One can even say that

$$\mathbb{P}\Big(f(x) \;\in\; \left[\widehat{f}(x) \;\pm\; \widehat{\sigma}(x)\sqrt{2F_{2,n-2;1-\alpha}}\right] \text{ for all } x\in\mathbb{R}\Big) \;=\; 1-\alpha.$$

This follows from the fact that the set of all vectors $\lambda(1, x)^\top$, $x \in \mathbb{R}$, $\lambda \in \mathbb{R}$, is a dense subset of $\mathbb{R}^2$; see also Subsection 3.5.3.

Figure 3.2 shows again the data from Figure 3.1 and the regression line $\widehat{f}$, plus the pointwise and simultaneous 95%-confidence bounds for $f(x)$. For the pointwise bounds one needs $t_{n-2;1-\alpha/2} = t_{49;0.975} = 2.010$, while for the simultaneous bounds one needs $\sqrt{2F_{2,n-2;1-\alpha}} = \sqrt{2F_{2,49;0.95}} = 2.525$ (rounded to three decimals).

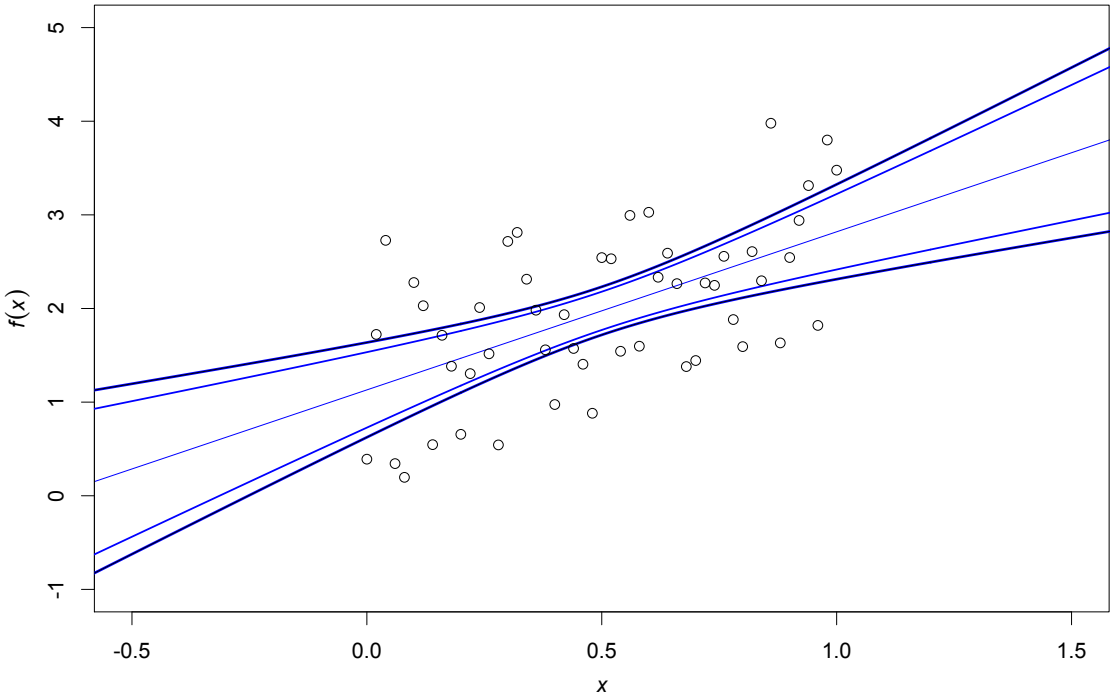

Figure 3.2: Regression line, pointwise and simultaneous 95%-confidence intervals for $f(x)$.

Example 3.29 (Polynomial regression). Let $X \in \mathbb{R}$, and consider the model equation $Y = f(X) + \varepsilon$, where the regression function $f : \mathbb{R} \to \mathbb{R}$ is an unknown polynomial of given order $d \geq 1$. For our purposes, it is convenient to work with orthonormal basis polynomials. That means, for a given data vector $\boldsymbol{X} \in \mathbb{R}^n$ with at least $d + 1$ different components, we construct polynomials $f_0(x), f_1(x), \ldots, f_d(x)$ of degree $0, 1, \ldots, d$, respectively, such that

$$f_j(\boldsymbol{X})^\top f_k(\boldsymbol{X}) \;=\; 1_{[j=k]} \quad \text{for } j, k \in \{0, 1, \ldots, d\}.$$

In other words, the resulting design matrix $\boldsymbol{D} = \big[f_0(\boldsymbol{X}), f_1(\boldsymbol{X}), \ldots, f_d(\boldsymbol{X})\big]$ satisfies $\boldsymbol{\Gamma} = \boldsymbol{D}^\top \boldsymbol{D} = \boldsymbol{I}_{d+1}$.

Suppose that

$$f(x) \;=\; \sum_{j=0}^{d} \theta_{j+1} f_j(x)$$

with an unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$. The LSE for $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{D}^\top \boldsymbol{Y} \sim \mathrm{N}_{d+1}(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_{d+1}).$$

Hence, the GME for $f(x)$ is given by

$$\widehat{f}(x) \;=\; \boldsymbol{\psi}(x)^\top \widehat{\boldsymbol{\theta}} \;\sim\; \mathrm{N}\big(f(x), \sigma^2 \|\boldsymbol{\psi}(x)\|^2\big),$$

where

$$\boldsymbol{\psi}(x) \;:=\; \big(f_0(x), f_1(x), \ldots, f_d(x)\big)^\top.$$

Moreover, for $1 \le \ell \le d$, the GME for the ℓ-th derivative $f^{(\ell)}(x)$ of f at x is given by

$$\widehat{f}^{(\ell)}(x) \;=\; \boldsymbol{\psi}^{(\ell)}(x)^\top \widehat{\boldsymbol{\theta}} \;\sim\; \mathrm{N}\big(f^{(\ell)}(x), \sigma^2 \|\boldsymbol{\psi}^{(\ell)}(x)\|^2\big)$$

with

$$\boldsymbol{\psi}^{(\ell)}(x) \;:=\; \big(f_0^{(\ell)}(x), f_1^{(\ell)}(x), \ldots, f_d^{(\ell)}(x)\big).$$

We may claim with confidence $1 - \alpha$, that

$$f(x) \;\in\; \Big[\widehat{f}(x) \,\pm\, \widehat{\sigma}\|\boldsymbol{\psi}(x)\|\sqrt{(d+1)F_{d+1,n-d-1;1-\alpha}}\Big]$$

and

$$f^{(\ell)}(x) \in \Big[\widehat{f}^{(\ell)}(x) \,\pm\, \widehat{\sigma}\|\boldsymbol{\psi}^{(\ell)}(x)\|\sqrt{(d+1)F_{d+1,n-d-1;1-\alpha}}\Big]$$

for any $x \in \mathbb{R}$ and $\ell \in \{1, \ldots, d\}$.

The upper panel of Figure 3.3 shows a scatter plot of simulated data vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^n$ with sample size $n = 201$. In addition, one sees the polynomial fit $\widehat{f}$ of order $d = 2$, the pointwise 95%-confidence bounds

$$\widehat{f}(x) \,\pm\, \widehat{\sigma}\|\boldsymbol{\psi}(x)\| t_{n-3;0.975}$$

and the simultaneous 95%-confidence bounds

$$\widehat{f}(x) \,\pm\, \widehat{\sigma}\|\boldsymbol{\psi}(x)\|\sqrt{3F_{3,n-3,0.95}}.$$

This scatter plot shows some systematic differences between the observed vector $\boldsymbol{Y}$ and the fitted vector $\widehat{\boldsymbol{Y}} = \widehat{f}(\boldsymbol{X})$: There are certain regions for the values X_i in which the differences $Y_i - \widehat{Y}_i$ tend to be systematically less than 0 or systematically greater than 0. In the lower panel one sees the data vectors and the simultaneous bounds for f, together with the true function f and its best approximation $\check{f}$ by a polynomial of the given order d. That means,

$$\check{f}(x) \;=\; \sum_{j=0}^{d} \theta_{j+1} f_j(x) \quad \text{with} \quad \boldsymbol{\theta} \;:=\; \boldsymbol{D}^\top f(\boldsymbol{X}).$$

Note that for many values x, the true value $f(x)$ is not contained within the simultaneous bounds, but the latter do enclose $\check{f}(x)$. Here the estimated standard deviation turned out to be $\widehat{\sigma} \approx 0.7952$, whereas the true standard deviation was $\sigma = 0.5$.

For the same data, Figure 3.4 shows analogous plots for polynomial regression of order $d = 8$. Now the difference between $f(x)$ and $\check{f}(x)$ is rather small, as long as x is in the range $\big[\min(\boldsymbol{X}), \max(\boldsymbol{X})\big]$. For values x outside the latter range, the confidence intervals become very wide. For the same data, Figure 3.5 shows the true and estimated first and second derivatives of the regression function(s), together with simultaneous 95%-confidence bounds.

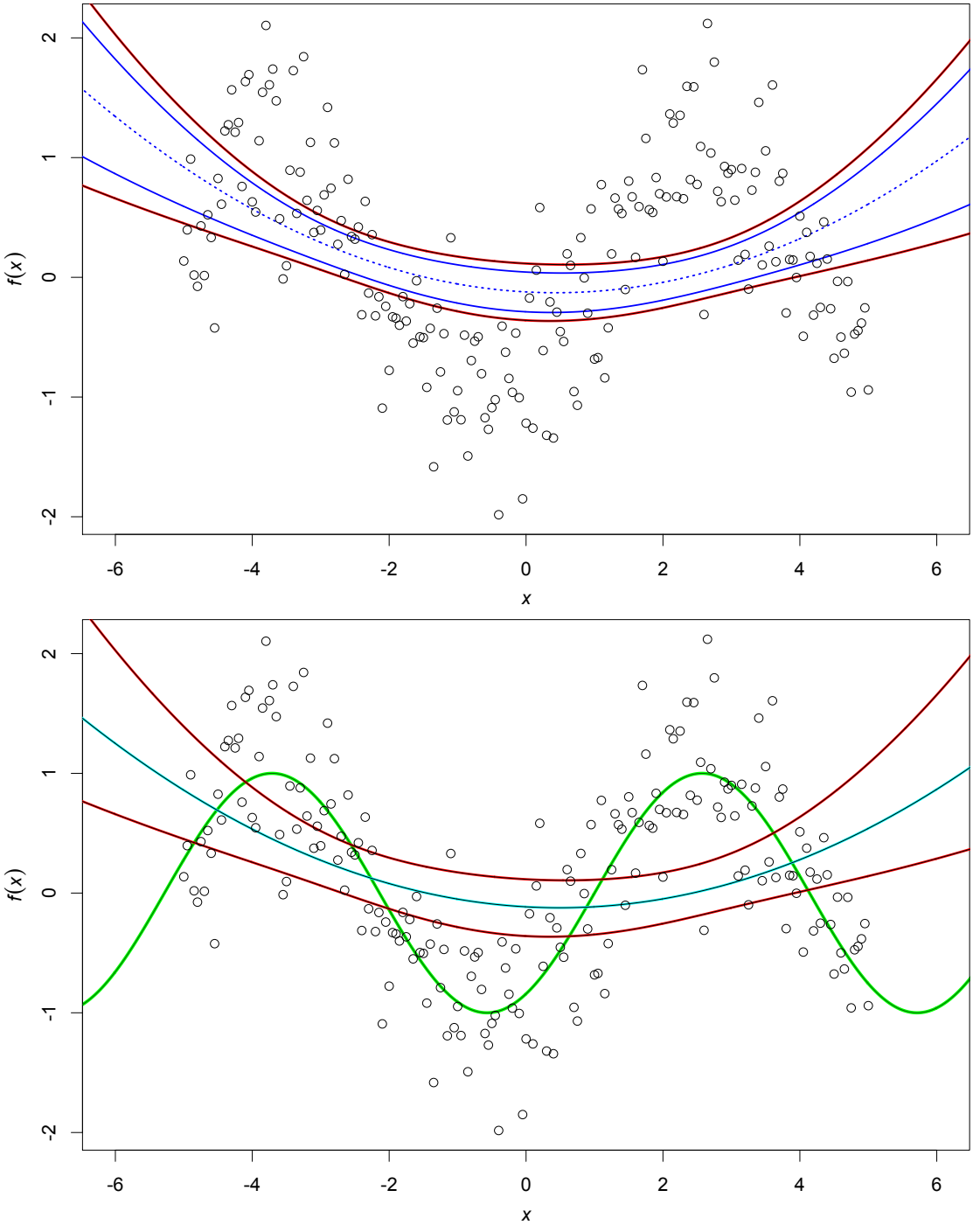

Figure 3.3: Simulated data example for quadratic regression. Upper panel: Estimated regression function (dashed blue), pointwise (blue) and simultaneous (red) 95%-confidence bounds. Lower panel: Simultaneous 95%-confidence bounds (red), true function f (green) and its approximation $\check{f}$ (cyan).

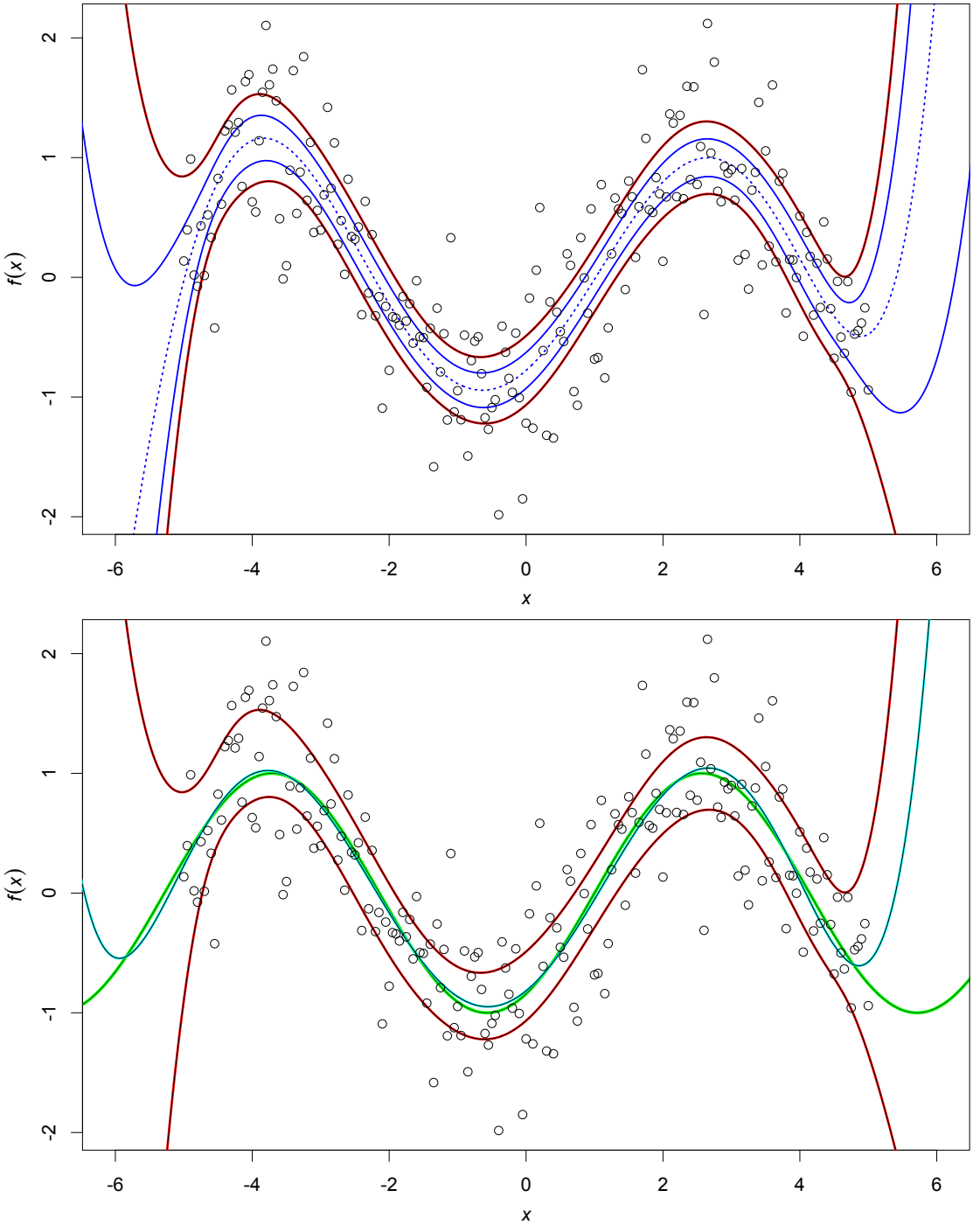

Figure 3.4: Simulated data example for polynomial regression of order $d = 6$. Upper panel: Estimated regression function (dashed blue), pointwise (blue) and simultaneous (red) 95%-confidence bounds. Lower panel: Simultaneous 95%-confidence bounds (red), true function f (green) and its approximation $\tilde{f}$ (cyan).

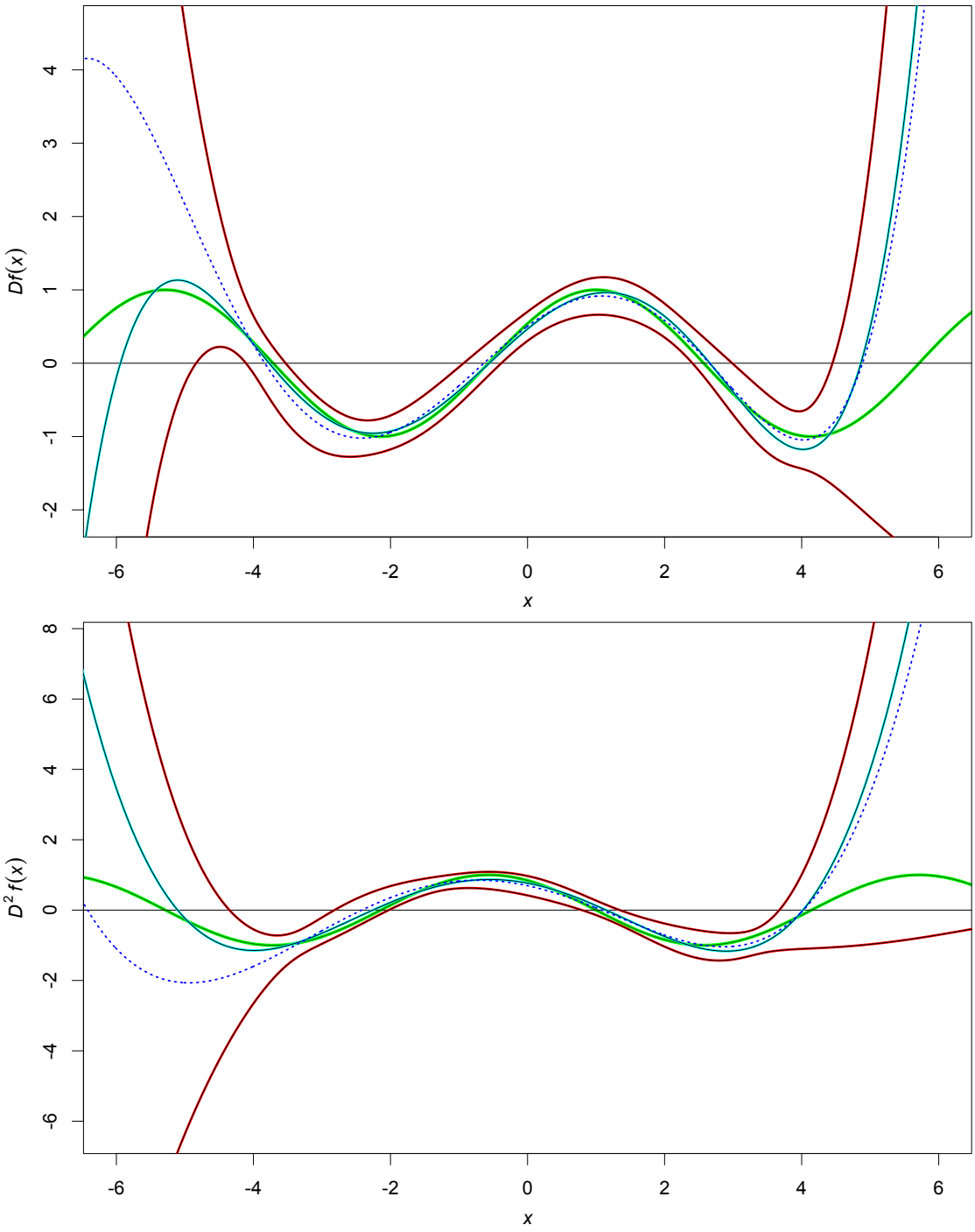

Figure 3.5: Simulated data example for polynomial regression of order $d = 6$. Estimated derivative $\widehat{f}^{(\ell)}$ (dashed blue) with simultaneous 95%-confidence bounds (red), and the true derivaties $f^{(\ell)}$ (green), $\check{f}^{(\ell)}$ (cyan). Here $\ell = 1$ (upper panel) and $\ell = 2$ (lower panel).

Exercise 3.30. Consider polynomials given by $p_0(x) = 1$, $p_1(x) = x - b_0$ and

$$p_{k+1}(x) \;=\; xp_k(x) - b_k p_k(x) - c_k p_{k-1}(x),$$

for $k \geq 1$ with given parameters $b_0, b_1, b_2, \ldots$ and $c_1, c_2, c_3, \ldots$. Show that

$$p_{k+1}^{(\ell)}(x) \;=\; \ell p_k^{(\ell-1)}(x) + xp_k^{(\ell)}(x) - b_k p_k^{(\ell)}(x) - c_k p_{k-1}^{(\ell)}(x)$$

for $k, \ell \geq 1$.

Exercise 3.31. Consider the model of polynomial regression. Write a computer program which computes and visualizes pointwise and simultaneous $(1-\alpha)$-confidence intervals for $f(x)$, $x \in \mathbb{R}$. The input arguments should be the data vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^n$, the maximal degree $d > 0$ of the polynomials, the test level $\alpha \in (0,1)$ and a vector $\boldsymbol{x}$ of points at which f should be estimated and narrowed down. Optionally, you could also compute and visualize pointwise and simultaneous $(1-\alpha)$-confidence intervals for $f^{(\ell)}(x)$, $x \in \mathbb{R}$ and $\ell \in \{1, \ldots, d\}$.

3.5.3 Generalizations

Inference about a vector-valued linear function of $\boldsymbol{\theta}$. Suppose we are interested in $\boldsymbol{\psi}_j^\top \boldsymbol{\theta}$, $1 \leq j \leq d$, for given linearly independent vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_d \in \mathbb{R}^p$. In other words, we are interested in the vector $\boldsymbol{\Psi}^\top \boldsymbol{\theta} \in \mathbb{R}^d$, where $\boldsymbol{\Psi} := [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_d] \in \mathbb{R}^{p \times d}$ has rank $d \leq p$. (The subsequent considerations are valid for any $d \in \{1, \ldots, p\}$, but our main interest is in the case $1 < d < p$.)

The F tests and confidence ellipsoid for $\boldsymbol{\theta}$ can be modified as follows:

$$\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} \;=\; \big(\boldsymbol{\psi}_j^\top \widehat{\boldsymbol{\theta}}\big)_{j=1}^d$$

is an estimator of $\boldsymbol{\Psi}^\top \boldsymbol{\theta}$ with distribution

$$\mathrm{N}_d\big(\boldsymbol{\Psi}^\top \boldsymbol{\theta},\, \sigma^2\, \boldsymbol{\Psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\Psi}\big) \;=\; \mathrm{N}_d\big(\boldsymbol{\Psi}^\top \boldsymbol{\theta}, \sigma^2 \boldsymbol{\Gamma}_{\boldsymbol{\Psi}}^{-1}\big),$$

where

$$\boldsymbol{\Gamma}_{\boldsymbol{\Psi}} \;:=\; (\boldsymbol{\Psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\Psi})^{-1}.$$

Moreover, it is stochastically independent from $\widehat{\sigma}$. This implies that

$$F_{\boldsymbol{\Psi}} \;:=\; \frac{(\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\Psi}^\top \boldsymbol{\theta})^\top \boldsymbol{\Gamma}_{\boldsymbol{\Psi}} (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\Psi}^\top \boldsymbol{\theta})}{d\, \widehat{\sigma}^2}$$

is a pivotal statistic with distribution $F_{d,n-p}$. Again, this leads to three statistical procedures.

Tests of one-point hypotheses for $\boldsymbol{\Psi}^\top \boldsymbol{\theta}$. For any given vector $\boldsymbol{w}_o \in \mathbb{R}^d$, we can test

$$H_o : \boldsymbol{\Psi}^\top \boldsymbol{\theta} = \boldsymbol{w}_o \quad \text{versus} \quad H_A : \boldsymbol{\Psi}^\top \boldsymbol{\theta} \neq \boldsymbol{w}_o,$$

by means of the test statistic

$$F_{\boldsymbol{\Psi}}(\boldsymbol{w}_o) \;:=\; \frac{(\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{w}_o)^\top \boldsymbol{\Gamma}_{\boldsymbol{\Psi}} (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{w}_o)}{d\, \widehat{\sigma}^2}.$$

We may reject the null hypothesis H_o at level α if $F_{\boldsymbol{\Psi}}(\boldsymbol{w}_o)$ is greater than or equal to $F_{d,n-p;1-\alpha}$. The latter condition is equivalent to the (right-sided) p-value

$$1 - \mathrm{Fcdf}_{d,n-p}(F_{\boldsymbol{\Psi}}(\boldsymbol{w}_o))$$

being less than or equal to α.

Exercise 3.32. Show that for any matrix $\boldsymbol{\Psi} \in \mathbb{R}^{p\times d}$ with rank d and any vector $\boldsymbol{w}_o \in \mathbb{R}^d$, testing the null hypothesis "$\boldsymbol{\Psi}^\top\boldsymbol{\theta} = \boldsymbol{w}_o$" with the data $(\boldsymbol{X}, \boldsymbol{Y})$ is equivalent to testing the null hypothesis "$\boldsymbol{\Psi}^\top\boldsymbol{\theta} = \boldsymbol{0}$" with the modified data $(\boldsymbol{X}, \boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\theta}_o)$ for some suitable $\boldsymbol{\theta}_o \in \mathbb{R}^p$. (Hint: One may use $\boldsymbol{\theta}_o = \boldsymbol{\Psi}\boldsymbol{v}_o$ for some $\boldsymbol{v}_o \in \mathbb{R}^d$.)

Exercise 3.33. Consider once more the data set 'Trees.txt' and the modified variables $Y := \log(\text{volume})$, $X(1) := \log(\text{height})$ and $X(2) := \log(\text{diameter})$. Assuming the model equation $Y = a + b_1X(1) + b_2X(2) + \varepsilon$, test the null hypothesis that $b_1 = 1$ and $b_2 = 2$ at level $\alpha = 5\%$.

Exercise 3.34. The oxygen saturation of blood (X, in percent) is an important physiological parameter. In particular, during surgery or in intensive care, this parameter has to be monitored permanently. It can be measured with high precision by analyzing blood samples chemically. Alternatively, one can measure the absorption of light when it passes through the skin of fingertips. Pulse oxymeters are technical devices which produce a proxy Y for X, based on such non-invasive, optical measurements. Ideally,

$$Y \;=\; X + \varepsilon$$

with a random measurement error ε such that $\mathbb{E}(\varepsilon) = 0$ and $\sigma := \mathrm{Std}(\varepsilon)$ is reasonably small. From a doctor's point of view, one should rather consider the undersaturations $\tilde{X} := 100 - X$ and $\tilde{Y} := 100 - Y$, because saturations below 70 are highly critical and have to be avoided. Under normal circumstances, the value X is way above 90. Hence, a pulse oxymeter should work with high precision if X lies between 70 and 100, i.e. if $\tilde{X}$ lies between 0 and 30.

Suppose we want to test whether a particular device works properly for a given person. To this end, one determines the actual saturation X_i and the device's measurement Y_i at n different time points, where the ambient air is manipulated such that the true saturations vary to some extent.

(a) Suppose that

$$\tilde{Y} \;=\; a + b\tilde{X} + \varepsilon$$

with unknown real constants a and b. Specify a suitable test for the null hypothesis that the device works correctly, that means,

$$H_o \,:\, a = 0, b = 1.$$

(b) Suppose that

$$\tilde{Y} \;=\; a + b\tilde{X} + c\tilde{X}^2 + \varepsilon$$

with unknown real constants a, b and c. Specify a suitable test for the null hypothesis that the device works correctly, that means,

$$H_o \,:\, a = 0, b = 1, c = 0.$$

(c) Assuming the model of quadratic or cubic regression, specify a test of the null hypothesis that Y is an affine function of X plus measurement error.

(d) Implement your tests and apply them to one or two subsets of the data set 'Pulsoxymeter.txt'.

The latter data set consists of several columns. The test person (prob), the measurements of the chemical analysis of blood samples (X) and the optical measurements by several pulse oxymeters for each device ($Y(1), Y(2), \ldots, Y(11)$). Now we restrict our attention to one particular test person and compare the values of X with the values $Y(k)$ of one particular pulse oxymeter.

Confidence ellipsoids for $\boldsymbol{\Psi}^\top\boldsymbol{\theta}$. A $(1-\alpha)$-confidence region for $\boldsymbol{\Psi}^\top\boldsymbol{\theta}$ is given by the d-dimensional ellipsoid

$$\begin{aligned} C_{\boldsymbol{\Psi},\alpha} = C_{\boldsymbol{\Psi},\alpha}(\boldsymbol{X},\boldsymbol{Y}) &:= \big\{\boldsymbol{w}\in\mathbb{R}^d : F_{\boldsymbol{\Psi}}(\boldsymbol{w}) \le F_{d,n-p;1-\alpha}\big\} \\ &= \big\{\boldsymbol{w}\in\mathbb{R}^d : (\boldsymbol{\Psi}^\top\widehat{\boldsymbol{\theta}}-\boldsymbol{w})^\top\boldsymbol{\Gamma}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}^\top\widehat{\boldsymbol{\theta}}-\boldsymbol{w}) \le \widehat{\sigma}^2 d F_{d,n-p;1-\alpha}\big\}. \end{aligned}$$

Simultaneous confidence intervals. Suppose we are interested in the values $\boldsymbol{\psi}^\top\boldsymbol{\theta}$ for all vectors $\boldsymbol{\psi}$ in a given finite or infinite family $\mathcal{P}\subset\mathbb{R}^p\setminus\{\mathbf{0}\}$. The next result provides simultaneous $(1-\alpha)$-confidence intervals for $\boldsymbol{\psi}^\top\boldsymbol{\theta}$, $\boldsymbol{\psi}\in\mathcal{P}$.

Theorem 3.35 (Henry Scheffé). *Let $\mathcal{P}$ be an arbitrary subset of $\mathbb{R}^p\setminus\{\mathbf{0}\}$, and let*

$$d := \dim(\operatorname{span}(\mathcal{P})).$$

Then

$$\sup_{\boldsymbol{\psi}\in\mathcal{P}} |T_{\boldsymbol{\psi}}| \le \max_{\boldsymbol{\psi}\in\operatorname{span}(\mathcal{P})\setminus\{\mathbf{0}\}} |T_{\boldsymbol{\psi}}| = \sqrt{dF_{\mathcal{P}}}$$

with a random variable

$$F_{\mathcal{P}} \sim F_{d,n-p}.$$

In particular,

$$\mathbb{P}\left(|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}}-\boldsymbol{\psi}^\top\boldsymbol{\theta}| \le \widehat{\sigma}_{\boldsymbol{\psi}}\sqrt{dF_{d,n-p;1-\alpha}} \text{ for all } \boldsymbol{\psi}\in\mathcal{P}\right) \ge 1-\alpha.$$

Equality holds true if the set $\{\lambda\boldsymbol{\psi} : \boldsymbol{\psi}\in\mathcal{P}, \lambda\in\mathbb{R}\}$ is a dense subset of $\operatorname{span}(\mathcal{P})$.

Proof of Theorem 3.35. By definition of d, there exist linearly independent vectors $\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_d\in\mathcal{P}$ such that

$$\operatorname{span}(\mathcal{P}) = \operatorname{span}(\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_d) = \big\{\boldsymbol{\Psi}\boldsymbol{\lambda} : \boldsymbol{\lambda}\in\mathbb{R}^d\big\}$$

with $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_d]$. Then, it follows from Lemma 3.27 that

$$\begin{aligned}
\sqrt{dF_{\boldsymbol{\Psi}}} &= \max_{\boldsymbol{\lambda} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{\left|\boldsymbol{\lambda}^\top (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\Psi}^\top \boldsymbol{\theta})\right|}{\widehat{\sigma} \sqrt{\boldsymbol{\lambda}^\top \boldsymbol{\Gamma}_{\boldsymbol{\Psi}}^{-1} \boldsymbol{\lambda}}} \\
&= \max_{\boldsymbol{\lambda} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{\left|(\boldsymbol{\Psi}\boldsymbol{\lambda})^\top \widehat{\boldsymbol{\theta}} - (\boldsymbol{\Psi}\boldsymbol{\lambda})^\top \boldsymbol{\theta}\right|}{\widehat{\sigma} \sqrt{(\boldsymbol{\Psi}\boldsymbol{\lambda})^\top \boldsymbol{\Gamma}^{-1} (\boldsymbol{\Psi}\boldsymbol{\lambda})}} \\
&= \max_{\boldsymbol{\psi} \in \mathrm{span}(\mathcal{P}) \setminus \{\mathbf{0}\}} \frac{\left|\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top \boldsymbol{\theta}\right|}{\widehat{\sigma} \sqrt{\boldsymbol{\psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\psi}}} \\
&= \max_{\boldsymbol{\psi} \in \mathrm{span}(\mathcal{P}) \setminus \{\mathbf{0}\}} |T_{\boldsymbol{\psi}}| \\
&\geq \sup_{\boldsymbol{\psi} \in \mathcal{P}} |T_{\boldsymbol{\psi}}|.
\end{aligned}$$

Equality holds true, if the set $\{\lambda\boldsymbol{\psi} : \lambda \in \mathbb{R}, \boldsymbol{\psi} \in \mathcal{P}\}$ is dense in $\mathrm{span}(\mathcal{P})$, because $|T_{\boldsymbol{\psi}}|$ is continuous in $\boldsymbol{\psi} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$, and $|T_{\lambda\boldsymbol{\psi}}| = |T_{\boldsymbol{\psi}}|$ for arbitrary $\lambda \neq 0$ and $\boldsymbol{\psi} \neq \mathbf{0}$. Since $F_{\boldsymbol{\Psi}} \leq F_{d,n-p;1-\alpha}$ with probability $1 - \alpha$, these considerations lead to the asserted (in)equality for

$$\mathbb{P}\left(\left|\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top \boldsymbol{\theta}\right| \leq \widehat{\sigma}_{\boldsymbol{\psi}} \sqrt{dF_{d,n-p;1-\alpha}} \text{ for all } \boldsymbol{\psi} \in \mathcal{P}\right).$$

□

Exercise 3.36. Consider polynomial regression of order $d \geq 2$. That means, $X \in \mathbb{R}$, and $Y = f(X) + \varepsilon$ with an unknown polynomial $f(x) = \sum_{j=0}^{d} \theta_{j+1} x^j$.

(a) Define an appropriate family $\mathcal{P}$ of vectors $\boldsymbol{\psi} \in \mathbb{R}^{d+1} \setminus \{0\}$, and determine $\mathrm{span}(\mathcal{P})$ as well as $\dim(\mathrm{span}(\mathcal{P}))$, in the following three situations:

(a.1) We are interested in $f'(x)$, for any $x \in \mathbb{R}$.
(a.2) We are interested in $f''(x)$, for any $x \in \mathbb{R}$.
(a.3) We are interested in $f(b) - f(a) - f'(a)(b-a)$, for arbitrary different $a, b \in \mathbb{R}$.

(b) Check in situations (a.1–3) whether

$$\{\lambda\psi : \psi \in \mathcal{P}, \lambda \in \mathbb{R}\}$$

is dense in $\mathrm{span}(\mathcal{P})$.

3.5.4 A Geometrical Approach to F Tests

In the previous sections, F tests appeared as a building block of confidence ellipsoids for linear functions of $\boldsymbol{\theta}$. In the present section we shall develop a seemingly different, purely geometrical approach.

Setting and testing problem. Our starting point is a linear model with n-dimensional observation vector

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}.$$

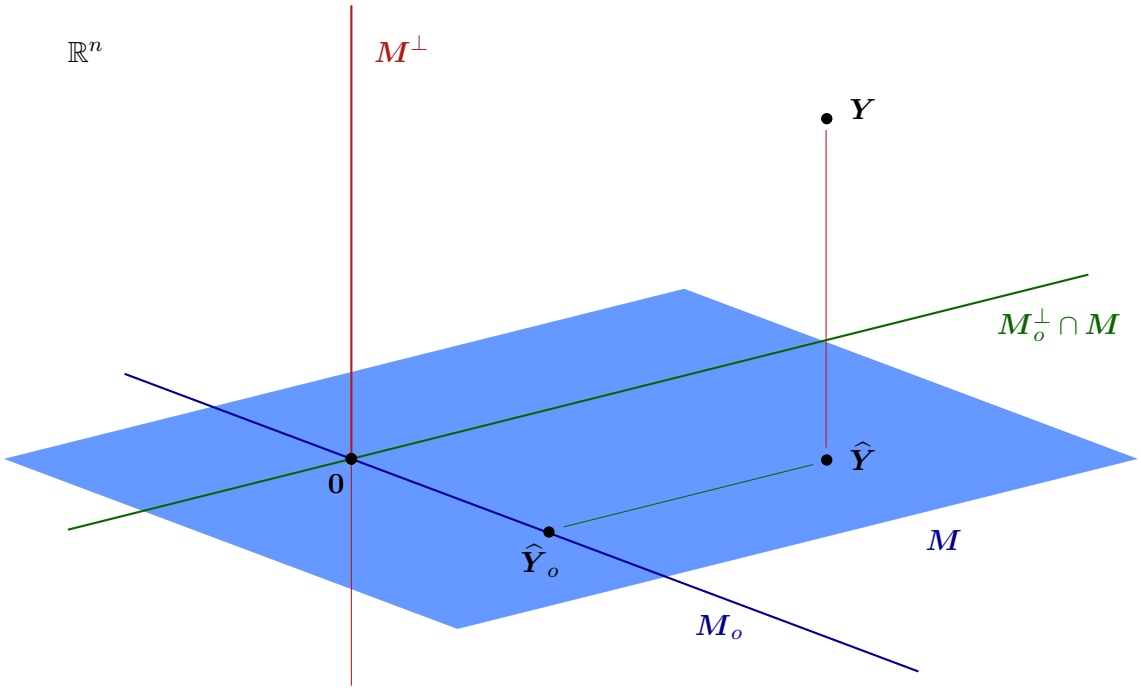

Figure 3.6: Geometry of an F test.

Here $\boldsymbol{\mu}$ is an unknown parameter vector which is assumed to lie in a given p-dimensional model space $\boldsymbol{M} \subset \mathbb{R}^n$, and $\boldsymbol{\varepsilon}$ is a random vector with distribution $\mathrm{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\sigma > 0$ is unknown as well. Now let $\boldsymbol{M}_o$ be a linear subspace of $\boldsymbol{M}$ with dimension $p_o < p$. We would like to test

$$H_o : \boldsymbol{\mu} \in \boldsymbol{M}_o \quad \text{versus} \quad H_A : \boldsymbol{\mu} \notin \boldsymbol{M}_o.$$

Construction of a test. In addition to the hat matrix $\boldsymbol{H}$, describing the orthogonal projection onto $\boldsymbol{M}$, there is a second matrix $\boldsymbol{H}_o$ for the orthogonal projection onto $\boldsymbol{M}_o$. With $\widehat{\boldsymbol{Y}} := \boldsymbol{H}\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}_o := \boldsymbol{H}_o\boldsymbol{Y}$, the observation vector $\boldsymbol{Y}$ can be decomposed into three orthogonal components:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{H}_o\boldsymbol{Y} + (\boldsymbol{H} - \boldsymbol{H}_o)\boldsymbol{Y} + (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} \\ &= \widehat{\boldsymbol{Y}}_o + (\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o) + (\boldsymbol{Y} - \widehat{\boldsymbol{Y}}). \end{aligned}$$

This corresponds to the representation of $\mathbb{R}^n$ as a direct sum of three orthogonal subspaces:

$$\mathbb{R}^n = \boldsymbol{M}_o \oplus (\boldsymbol{M} \cap \boldsymbol{M}_o^\perp) \oplus \boldsymbol{M}^\perp.$$

Figure 3.6 illustrates these decompositions of $\boldsymbol{Y}$ and $\mathbb{R}^n$.

Note that

$$\begin{aligned} \boldsymbol{Y} - \widehat{\boldsymbol{Y}} &= (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\varepsilon} \qquad \text{because } \boldsymbol{\mu} \in \boldsymbol{M}, \\ \widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o &= \begin{cases} (\boldsymbol{H} - \boldsymbol{H}_o)\boldsymbol{\mu} + (\boldsymbol{H} - \boldsymbol{H}_o)\boldsymbol{\varepsilon} & \text{in general,} \\ (\boldsymbol{H} - \boldsymbol{H}_o)\boldsymbol{\varepsilon} & \text{if } \boldsymbol{\mu} \in \boldsymbol{M}_o, \end{cases} \\ \widehat{\boldsymbol{Y}}_o &= \begin{cases} \boldsymbol{H}_o\boldsymbol{\mu} + \boldsymbol{H}_o\boldsymbol{\varepsilon} & \text{in general,} \\ \boldsymbol{\mu} + \boldsymbol{H}_o\boldsymbol{\varepsilon} & \text{if } \boldsymbol{\mu} \in \boldsymbol{M}_o. \end{cases} \end{aligned}$$

Consequently, under the null hypothesis H_o, both vectors $\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o$ and $\boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ contain only parts of the *noise vector* $\boldsymbol{\varepsilon}$, whereas the *signal vector* $\boldsymbol{\mu}$ is hidden entirely in $\widehat{\boldsymbol{Y}}_o$.

If the null hypothesis H_o is violated, the distance $\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o\|$ tends to be larger than $\|(\boldsymbol{H} - \boldsymbol{H}_o)\varepsilon\|$. This will be investigated in more detail in a later section, but a simple calculation shows that

$$\mathbb{E}\big(\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o\|^2\big) \;=\; (p - p_o)\sigma^2 + \|\boldsymbol{\mu} - \boldsymbol{H}_o\boldsymbol{\mu}\|^2,$$

whereas $\mathbb{E}\big(\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2\big) = (n - p)\sigma^2$. Hence, we introduce the test statistic

$$F \;:=\; \frac{\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o\|^2/(p - p_o)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2/(n - p)} \;=\; \frac{\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o\|^2/\dim(\boldsymbol{M} \cap \boldsymbol{M}_o^\perp)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2/\dim(\boldsymbol{M}^\perp)}. \tag{3.1}$$

The enumerator of F could also be written as

$$\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o\|^2/(p - p_o) \;=\; \big(\|\widehat{\boldsymbol{Y}}\|^2 - \|\widehat{\boldsymbol{Y}}_o\|^2\big)/(p - p_o),$$

and its denominator is equal to

$$\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2/(n - p) \;=\; \big(\|\boldsymbol{Y}\|^2 - \|\widehat{\boldsymbol{Y}}\|^2\big)/(n - p) \;=\; \widehat{\sigma}^2.$$

Theorem 3.37. *Under the null hypothesis that $\boldsymbol{\mu} \in \boldsymbol{M}_o$, the test statistic F in* (3.1) *has distribution $F_{p-p_o,n-p}$.*

Consequently, we may reject the null hypothesis H_o at level α if

$$F \;\geq\; F_{p-p_o,n-p;1-\alpha}.$$

The corresponding (right-sided) p-value is

$$1 - \mathrm{Fcdf}_{p-p_o,n-p}(F).$$

Proof of Theorem 3.37. Similarly as in the proof of Theorem 3.16, we consider a suitable orthonormal basis $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n$ of $\mathbb{R}^n$. This time we require that

$$\boldsymbol{M}_o \;=\; \mathrm{span}(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{p_o}) \quad \text{and} \quad \boldsymbol{M} \;=\; \mathrm{span}(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p).$$

Moreover, we write

$$\boldsymbol{\varepsilon} \;=\; \sigma \sum_{i=1}^{n} Z_i \boldsymbol{t}_i$$

with the standard Gaussian random vector $\boldsymbol{Z} := \sigma^{-1}(\boldsymbol{t}_i^\top \boldsymbol{\varepsilon})_{i=1}^n \sim \mathrm{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Under the null hypothesis H_o,

$$\begin{aligned} F \;&=\; \frac{(p - p_o)^{-1}\|(\boldsymbol{H} - \boldsymbol{H}_o)\boldsymbol{\varepsilon}\|^2}{(n - p)^{-1}\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}\|^2} \\ &=\; \frac{(p - p_o)^{-1}\big\|\sigma \sum_{i=p_o+1}^{p} Z_i \boldsymbol{t}_i\big\|^2}{(n - p)^{-1}\big\|\sigma \sum_{i=p+1}^{n} Z_i \boldsymbol{t}_i\big\|^2} \\ &=\; \frac{(p - p_o)^{-1} \sum_{i=p_o+1}^{p} Z_i^2}{(n - p)^{-1} \sum_{i=p+1}^{n} Z_i^2}. \end{aligned}$$

By definition of F distributions, the latter ratio has the asserted distribution $F_{p-p_o,n-p}$. □

Exercise 3.38 (R-squared and F test). If $\mathcal{F} \supset \mathcal{F}_o = \{\text{constant functions}\}$ with $\dim(\mathcal{F}) > 1$, then the standard output of statistical software for the linear model $\mathcal{F}$ includes the value R^2 as well as the test statistic F and the corresponding p-value for the null hypothesis "$f \in \mathcal{F}_o$" versus "$f \in \mathcal{F} \setminus \mathcal{F}_o$". What is the relationship between R^2 and F?

Exercise 3.39. Suppose you want to test whether a linear subspace $\boldsymbol{M}_o \subset \boldsymbol{M}$ (or a smaller linear model $\mathcal{F}_o \subset \mathcal{F}$) fits a data vector $\boldsymbol{Y}$ sufficiently well, where $p_o := \dim(\boldsymbol{M}_o) < p := \dim(\boldsymbol{M}) < n$. Suppose you only know n, p_o, p and the variance estimators $\widehat{\sigma}_o^2$ for the simple as well as $\widehat{\sigma}^2$ for the full model. Write the test statistic F as a function of these quantities n, p_o, p, $\widehat{\sigma}_o^2$ and $\widehat{\sigma}^2$.

Example 3.40 (One-way ANOVA). Starting from a covariate $X \in \{1, 2, \ldots, L\}$, we use the paper and blackboard notation and identify the response vector $\boldsymbol{Y}$ with an array $(Y_{js})_{j,s}$, where

$$Y_{js} \;=\; f_j + \varepsilon_{js}, \quad 1 \le j \le L,\ 1 \le s \le n(j),$$

with unknown parameters $f_1, f_2, \ldots, f_L$ and independent random variables $\varepsilon_{js} \sim \mathcal{N}(0, \sigma^2)$.

To verify a true association between X and Y, one could test the following null hypothesis:

$$H_o : f_1 = f_2 = \cdots = f_L.$$

This null hypothesis corresponds to the space $\boldsymbol{M}_o$ of all constant arrays, and its dimension is $p_o = 1$. The full model space $\boldsymbol{M}$ consists of all arrays $(g_j)_{j,s}$ with $g_1, \ldots, g_L \in \mathbb{R}$, and its dimension is L.

Here $\widehat{\boldsymbol{Y}} = \big(\bar{Y}_{j\cdot}\big)_{j,s}$ and $\widehat{\boldsymbol{Y}}_o = \big(\bar{Y}\big)_{j,s}$. Hence, the *total sum of squares*,

$$\mathrm{SS}_{\mathrm{total}} \;:=\; \|\boldsymbol{Y} - \bar{Y}\mathbf{1}\|^2 \;=\; \sum_{j,s}(Y_{js} - \bar{Y})^2$$

may be written as the *sum of squares within groups*,

$$\mathrm{SS}_{\mathrm{within}} \;:=\; \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2 \;=\; \sum_{j,s}(Y_{js} - \bar{Y}_{j\cdot})^2,$$

plus the *sum of squares between groups*,

$$\mathrm{SS}_{\mathrm{between}} \;:=\; \|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_o\|^2 \;=\; \sum_{j,s}(\bar{Y}_{j\cdot} - \bar{Y})^2 \;=\; \sum_{j} n(j)(\bar{Y}_{j\cdot} - \bar{Y})^2.$$

That means,

$$\mathrm{SS}_{\mathrm{total}} \;=\; \mathrm{SS}_{\mathrm{within}} + \mathrm{SS}_{\mathrm{between}}.$$

The null hypothesis H_o that all parameters f_j are identical is rejected at level α if

$$F \;=\; \frac{\mathrm{SS}_{\mathrm{between}}/(L-1)}{\mathrm{SS}_{\mathrm{within}}/(n-L)}$$

is greater than or equal to $F_{L-1,n-L;1-\alpha}$.

The phrase "analysis of variance" refers to such decompositions of the total sum of squares into different components.

Example 3.41 (Michelson's and Morley's measurements of the speed of light). A famous data example from physics are measurements of the speed of light (in vacuum) by Albert A. Michelson and Edward Morley at the end of the nineteenth century. One particular data set, available in R as 'morley', comprises $n = 100$ measurements. These measurements are reported in kilometers per second, and from each raw value the number $299'000$ has been subtracted. The observations correspond to $L = 5$ different experiments, and in each experiment they did $n_o = 20$ measurements. Ideally,

$$Y_{js} \;=\; \mu + \varepsilon_{js}, \quad 1 \le j \le 5,\ 1 \le s \le 20,$$

where μ is the true value of the speed of light. Alternatively, one could think about the model

$$Y_{js} \;=\; f_j + \varepsilon_{js}, \quad 1 \le j \le 5,\ 1 \le s \le 20,$$

with certain values $f_1, f_2, \ldots, f_5$ reflecting different circumstances of the experiment. In fact, Michelson's original goal was to prove the existence of 'ether', i.e. a medium for electromagnetic waves. So he assumed that f_j reflects the speed of the lab relative to the surrounding ether. But there exist other, unintended, reasons for the f_j to be different.

To test the null hypothesis that $f_1 = f_2 = \cdots = f_5$, we need the following sums of squares:

$$\begin{aligned} \mathrm{SS}_{\mathrm{within}} &= 523'510, \\ \mathrm{SS}_{\mathrm{between}} &= 94'514. \end{aligned}$$

This leads to the F test statistic

$$F \;=\; \frac{94'514/4}{523'510/95} \;=\; \frac{23'628.5}{5'510.632} \;\approx\; 4.2878,$$

and the p-value for the null hypothesis equals

$$1 - \mathrm{Fcdf}_{4,95}(F) \;\approx\; 0.003114.$$

Hence, one may conclude with high confidence that there have been systematic differences between the five different experiments.

Eventually, Michelson and his colleagues concluded from numerous high-precision measurements that presumably there is no ether. Michelson got a Nobel award for his experimental methods, and his findings inspired Albert Einstein to develop the special theory of relativity.

Example 3.42 (Nested function spaces). We consider a linear model $\mathcal{F}$ with basis functions $f_1, \ldots, f_p : \mathcal{X} \to \mathbb{R}$. Now we want to test the null hypothesis that the unknown regression function belongs to the smaller model $\mathcal{F}_o := \mathrm{span}(f_1, \ldots, f_{p_o})$, where $0 \le p_o < p$. (In case of $p_o = 0$, we just set $\mathcal{F}_o = \{0\}$.) Under the assumption that the vectors $f_1(\boldsymbol{X}), f_2(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})$ are linearly independent, we are talking about the model space $\boldsymbol{M} := \mathrm{span}\big(f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})\big)$ and its p_o-dimensional subspace $\boldsymbol{M}_o := \mathrm{span}\big(f_1(\boldsymbol{X}), \ldots, f_{p_o}(\boldsymbol{X})\big)$.

If we write $f(x) = \sum_{j=1}^{p} \theta_j f_j(x)$, then the null hypothesis reads

$$H_o : (\theta_j)_{j > p_o} \;=\; \mathbf{0}.$$

The corresponding F test statistic may be written in two ways. On the one hand,

$$F \;=\; \frac{(\widehat{\theta}_j)_{j>p_o}^\top \boldsymbol{\Gamma}_o (\widehat{\theta}_j)_{j>p_o}}{(p-p_o)\widehat{\sigma}^2},$$

where

$$\boldsymbol{\Gamma}_o \;:=\; \big((\boldsymbol{\Gamma}^{-1})_{j,k>p_o}\big)^{-1}.$$

On the other hand, one may write

$$F \;=\; \frac{(p-p_o)^{-1}\big\|\widehat{f}(\boldsymbol{X})-\widehat{f}_o(\boldsymbol{X})\big\|^2}{(n-p)^{-1}\big\|\boldsymbol{Y}-\widehat{f}(\boldsymbol{X})\big\|^2} \;=\; \frac{\big\|\widehat{f}(\boldsymbol{X})-\widehat{f}_o(\boldsymbol{X})\big\|^2}{(p-p_o)\,\widehat{\sigma}^2},$$

where $\widehat{f}_o$ and $\widehat{f}$ are the LSE of f in the smaller and the full model, respectively.

Warning: With the LSE $\widehat{\boldsymbol{\theta}}$ for the full model,

$$\widehat{f}_o \;\not\equiv\; \sum_{j=1}^{p_o} \widehat{\theta}_j f_j \quad \text{in general!}$$

This is only guaranteed, if

$$\big\{f_1(\boldsymbol{X}),\ldots,f_{p_o}(\boldsymbol{X})\big\} \;\perp\; \big\{f_{p_o+1}(\boldsymbol{X}),\ldots,f_p(\boldsymbol{X})\big\}.$$

Exercise 3.43. In Exercise 2.7, theoretical problems of linear regression as a means to obtain "adjusted data" have been discussed. In this exercise, these reflections are complemented by analysing the specific data set "SozialdiensteBE.txt". Each observation (row) corresponds to one social service in the Canton of Bern.

(a) Try out different linear models for these data and search for a set of covariates having a significant effect on the response

$$Y \;:=\; \text{expenditures for social welfare per year and inhabitant.}$$

(b) The Canton of Bern used a standard multiple regression model with the covariates

$$\begin{aligned}
X(1) &= \text{percentage of foreigners},\\
X(2) &= \text{percentage of refugees},\\
X(3) &= \text{percentage of people getting subsidies (Ergänzungsleistungen)},\\
X(4) &= \text{percentage of vacant apartments}.
\end{aligned}$$

Assuming that

$$Y = \theta_0 + \theta_1 X(1) + \theta_2 X(2) + \theta_3 X(3) + \theta_4 X(4) + \varepsilon$$

(with ε being interpreted as noise rather than misperformance), calculate the LSE $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} = (\theta_0,\theta_1,\ldots,\theta_4)^\top \in \mathbb{R}^5$ and interpret your results.

(c) Now consider the more complex model

$$Y = \theta_0 + \sum_{j=1}^{4} \theta_j X(j) + \sum_{1 \le \ell < k \le 4} \gamma_{\ell k} X(\ell) X(k) + \varepsilon$$

with $\theta_j \in \mathbb{R}$, $0 \le j \le 4$, and $\gamma_{\ell k} \in \mathbb{R}$, $1 \le \ell < k \le 4$. Do you reach the same conclusions as for the previous model? Compare the two models. Does the inclusion of interactions yield a significantly better fit?

Example 3.44 (Polynomial regression). As in Example 3.29, let $X \in \mathbb{R}$, and consider the model equation $Y = f(X) + \varepsilon$, where the unknown regression function $f : \mathbb{R} \to \mathbb{R}$ is assumed to be a polynomial of given order $d \ge 1$. Again, suppose that $\boldsymbol{X}$ has at least $d+1$ different components, and let $f_0(x), f_1(x), \ldots, f_d(x)$ be polynomials of degree $0, 1, \ldots, d$, respectively, such that the vectors $f_0(\boldsymbol{X}), f_1(\boldsymbol{X}), \ldots, f_d(\boldsymbol{X})$ are orthonormal. With $f(x) = \sum_{j=0}^{d} \theta_{j+1} f_j(x)$, the LSE of $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}} \;=\; \big(f_j(\boldsymbol{X})^\top \boldsymbol{Y}\big)_{j=0}^{d} \;\sim\; \mathrm{N}_{d+1}(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_{d+1}).$$

For any order $d_o \in \{0, \ldots, d\}$, the orthogonal projection of $\boldsymbol{Y}$ onto $\operatorname{span}\big(f_0(\boldsymbol{X}), \ldots, f_{d_o}(\boldsymbol{X})\big)$ is given by

$$\sum_{j=0}^{d_o} \widehat{\theta}_{j+1} f_j(\boldsymbol{X}).$$

In case of $d_o < d$, the null hypothesis that f has order d_o is equivalent to the null hypothesis

$$H_o : f(\boldsymbol{X}) \in \operatorname{span}\big(f_0(\boldsymbol{X}), \ldots, f_{d_o}(\boldsymbol{X})\big).$$

The corresponding F test statistic is given by

$$F_{d_o} \;:=\; \frac{\sum_{j=d_o+1}^{d} \widehat{\theta}_{j+1}^2}{(d-d_o)\widehat{\sigma}^2}$$

with distribution $F_{d-d_o, n-d-1}$ under H_o, and the corresponding p-value equals

$$\pi_{d_o} \;:=\; 1 - F_{d-d_o, n-d-1}(F_{d_o}).$$

For the particular data we simulated in Example 3.29, let us start with the assumption that f is a polynomial of order $d = 10$. This yields an estimated standard deviation $\widehat{\sigma} \approx 0.4921$. The following table shows for $d_o = 0, 1, \ldots, 9$, the value of the test statistic F_{d_o} and the corresponding p-value π_{d_o} (rounded to three decimals).

d_o	0	1	2	3	4	5	6	7	8	9
F_{d_o}	41.753	43.501	48.104	46.491	41.060	11.550	2.992	0.900	1.220	2.286
π_{d_o}	0.000	0.000	0.000	0.000	0.000	0.001	0.021	0.443	0.298	0.133

This table shows that degrees smaller than 6 are inappropriate for the data, even $d = 6$ seems to be too small.

Exercise 3.45. In Exercise 3.34 we tested for a single test person and a single pulse oxymeter, whether the device is working properly. Now consider all data for one pulse oxymeter and specify one or several potential model equations for the measurement Y, the true oxygen saturation W and the test person C. Fit this model or these models to the data. Specifically, how would you test the null hypothesis that the device is working identically with each test person?

Example 3.46 (Additive model for two-way ANOVA). We consider $X = (C, D)$ with factors $C \in \{1, \dots, L\}$ and $D \in \{1, \dots, M\}$. Now we switch to paper and blackboard notation and consider the observation array $\boldsymbol{Y} = (Y_{jks})_{j,k,s}$ with

$$Y_{jks} \;=\; \mu + a_j + b_k + \varepsilon_{jks}, \quad 1 \le j \le L,\ 1 \le k \le M,\ 1 \le s \le n(j,k).$$

Here $\mu \in \mathbb{R}$, $\boldsymbol{a} \in \mathbb{R}^L$ and $\boldsymbol{b} \in \mathbb{R}^M$ are unknown parameters such that $a_+ := \sum_{j=1}^L a_j = 0$ and $b_+ := \sum_{k=1}^M b_k = 0$, while the ε_{jks} are independent random variables with distribution $\mathrm{N}(0, \sigma^2)$.

This model equation means that the influence of the two factors is purely additive. The corresponding model space $\boldsymbol{M}$ consists of all arrays

$$\big(g_j + h_k\big)_{j,k,s}$$

with arbitrary numbers g_j and h_k. It can be written as the direct sum of three subspaces:

$$\begin{aligned}
\boldsymbol{M}_0 &:= \big\{(\mu)_{j,k,s} : \mu \in \mathbb{R}\big\} \quad \text{with } \dim(\boldsymbol{M}_0) = 1,\\
\boldsymbol{M}_1 &:= \big\{(a_j)_{j,k,s} : \boldsymbol{a} \in \mathbb{R}^L, a_+ = 0\big\} \quad \text{with } \dim(\boldsymbol{M}_1) = L - 1,\\
\boldsymbol{M}_2 &:= \big\{(b_k)_{j,k,s} : \boldsymbol{b} \in \mathbb{R}^M, b_+ = 0\big\} \quad \text{with } \dim(\boldsymbol{M}_2) = M - 1.
\end{aligned}$$

With the orthogonal projections $\widehat{\boldsymbol{Y}}_\ell$ of $\boldsymbol{Y}$ onto $\boldsymbol{M}_\ell$ ($0 \le \ell \le 2$) and $\widehat{\boldsymbol{Y}}_{0,\ell}$ of $\boldsymbol{Y}$ onto $\boldsymbol{M}_0 + \boldsymbol{M}_\ell$ ($\ell = 1, 2$) we can perform the following F tests:

(a) The null hypothesis that neither C nor D have any association with Y corresponds to the assumption that

$$\boldsymbol{a} \;=\; \boldsymbol{0} \quad \text{and} \quad \boldsymbol{b} \;=\; \boldsymbol{0}.$$

The corresponding null model space is $\boldsymbol{M}_0$, and the appropriate test statistic is

$$F \;=\; \frac{\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_0\|^2}{(M + L - 2)\,\widehat{\sigma}^2} \;=\; \frac{\|\widehat{\boldsymbol{Y}}\|^2 - \|\widehat{\boldsymbol{Y}}_0\|^2}{(M + L - 2)\,\widehat{\sigma}^2}$$

with distribution $F_{M+L-2,n-M-L+1}$ under the null hypothesis.

(b) The null hypothesis that the factor C has no association with Y corresponds to the assumption that

$$\boldsymbol{a} \;=\; \boldsymbol{0}.$$

The corresponding null model space is $\boldsymbol{M}_0 + \boldsymbol{M}_2$, and the appropriate test statistic is

$$F \;=\; \frac{\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{0,2}\|^2}{(L - 1)\,\widehat{\sigma}^2} \;=\; \frac{\|\widehat{\boldsymbol{Y}}\|^2 - \|\widehat{\boldsymbol{Y}}_{0,2}\|^2}{(L - 1)\,\widehat{\sigma}^2}$$

with distribution $F_{L-1,n-L-M+1}$ under the null hypothesis.

(c) Analogously, one may test the null hypothesis that the factor D has no association with Y by means of the test statistic

$$F \;=\; \frac{\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{0,1}\|^2}{(M-1)\,\widehat{\sigma}^2} \;=\; \frac{\|\widehat{\boldsymbol{Y}}\|^2 - \|\widehat{\boldsymbol{Y}}_{0,1}\|^2}{(M-1)\,\widehat{\sigma}^2}$$

with distribution $F_{M-1,n-L-M+1}$ under the null hypothesis.

An important special case is a *balanced design*, that means, all group sizes $n(j,k)$ are identical,

$$n(j,k) \;=\; n_o \quad \text{for } 1 \le j \le L, 1 \le k \le M.$$

This implies that the three spaces $\boldsymbol{M}_0, \boldsymbol{M}_1, \boldsymbol{M}_2$ are pairwise orthogonal: For arbitrary vectors $\boldsymbol{v} = (a_j)_{j,k,s} \in \boldsymbol{M}_1$ and $\boldsymbol{w} = (b_k)_{j,k,s} \in \boldsymbol{M}_2$,

$$\begin{aligned}
\mathbf{1}^\top \boldsymbol{v} \;&=\; \sum_{j,k,s} a_j \;=\; a_+ M n_o \;=\; 0, \\
\mathbf{1}^\top \boldsymbol{w} \;&=\; \sum_{j,k,s} b_k \;=\; L b_+ n_o \;=\; 0, \\
\boldsymbol{v}^\top \boldsymbol{w} \;&=\; \sum_{j,k,s} a_j b_k \;=\; a_+ b_+ n_o \;=\; 0.
\end{aligned}$$

In case of a balanced design, the GMEs for the parameters μ, a_j and b_k are obtained easily. To this end, we consider the partial means

$$\begin{aligned}
\bar{Y}_{jk\cdot} \;&:=\; \frac{1}{n_o} \sum_s Y_{jks}, \\
\bar{Y}_{j\cdot\cdot} \;&:=\; \frac{1}{M n_o} \sum_{k,s} Y_{jks} \;=\; \frac{1}{M} \sum_k \bar{Y}_{jk}, \\
\bar{Y}_{\cdot k\cdot} \;&:=\; \frac{1}{L n_o} \sum_{j,s} Y_{jks} \;=\; \frac{1}{L} \sum_j \bar{Y}_{jk}
\end{aligned}$$

and the total mean $\bar{Y} = L^{-1} \sum_j \bar{Y}_{j\cdot\cdot} = M^{-1} \sum_k \bar{Y}_{\cdot k\cdot}$. Now one can write

$$\boldsymbol{Y} \;=\; \widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{\varepsilon}} \quad \text{and} \quad \widehat{\boldsymbol{Y}} \;=\; \widehat{\boldsymbol{Y}}_0 + \widehat{\boldsymbol{Y}}_1 + \widehat{\boldsymbol{Y}}_2,$$

where

$$\begin{aligned}
\widehat{\boldsymbol{Y}}_0 \;&:=\; (\bar{Y})_{j,k,s} \;\in\; \boldsymbol{M}_0, \\
\widehat{\boldsymbol{Y}}_1 \;&:=\; (\bar{Y}_{j\cdot\cdot} - \bar{Y})_{j,k,s} \;\in\; \boldsymbol{M}_1, \\
\widehat{\boldsymbol{Y}}_2 \;&:=\; (\bar{Y}_{\cdot k\cdot} - \bar{Y})_{j,k,s} \;\in\; \boldsymbol{M}_2.
\end{aligned}$$

In particular,

$$\widehat{\mu} \;=\; \bar{Y}, \quad \widehat{a}_j \;=\; \bar{Y}_{j\cdot\cdot} - \bar{Y} \quad \text{and} \quad \widehat{b}_k \;=\; \bar{Y}_{\cdot k\cdot} - \bar{Y}.$$

The F test statistics introduced before have now a simplified representation:

(a) For the null hypotheses that $\boldsymbol{a} = \mathbf{0}$ and $\boldsymbol{b} = \mathbf{0}$, we get

$$F \;=\; \frac{\|\widehat{\boldsymbol{Y}}_1\|^2 + \|\widehat{\boldsymbol{Y}}_2\|^2}{(M+L-2)\,\widehat{\sigma}^2}.$$

(b) For the null hypothesis that $\boldsymbol{a} = \mathbf{0}$, we get

$$F \;=\; \frac{\|\widehat{\boldsymbol{Y}}_1\|^2}{(L-1)\,\widehat{\sigma}^2}.$$

(c) For the null hypothesis that $\boldsymbol{b} = \mathbf{0}$, we get

$$F \;=\; \frac{\|\widehat{\boldsymbol{Y}}_2\|^2}{(M-1)\,\widehat{\sigma}^2}.$$

Example 3.47 (Hearing tests). As a data example for the testing problems in Example 3.46, we consider the data set 'Hearing.txt'. Twenty-four test persons listened to recordings of four different lists of words with some background noise. The measurements Y are the percentages of correctly identified words, and the two factors were the person ($C \in \{1, 2, \ldots, 24\}$) and the word list ($D \in \{1, 2, 3, 4\}$). Here we have a balanced design with $n_o = 1$, and it turned out that

$$\|\widehat{\boldsymbol{Y}}_1\|^2 \;=\; 3231.62, \quad \|\widehat{\boldsymbol{Y}}_2\|^2 \;=\; 920.458 \quad \text{and} \quad \|\widehat{\boldsymbol{\varepsilon}}\|^2 \;=\; 2506.54.$$

Moreover, $n - L - M + 1 = LM - L - M + 1 = (L-1)(M-1) = 23 \cdot 3 = 69$, whence

$$\widehat{\sigma}^2 \;=\; 2506.54/69 \;=\; 36.327.$$

For the null hypotheses that neither C nor D has an association with Y, the test statistic turns out to be

$$F \;=\; \frac{3231.62 + 920.458}{(24 + 4 - 2)\,36.327} \;=\; 4.940,$$

and the resulting p-value equals $1 - \mathrm{Fcdf}_{26,69}(4.940) < 0.0001$.

For the null hypothesis that the factor C (test person) is superfluous, the test statistic turns out to be

$$F \;=\; \frac{3231.62}{(24-1)\,36.327} \;=\; 3.868,$$

and the corresponding p-value equals $1 - \mathrm{Fcdf}_{23,69}(3.868) < 0.0001$.

Finally, for the null hypothesis that the factor D (word list) is superfluous, the test statistic has the value

$$F \;=\; \frac{920.458}{(4-1)\,36.327} \;=\; 8.446$$

and the p-value equals $1 - \mathrm{Fcdf}_{3,69}(8.446) < 0.0001$.

Exercise 3.48. Consider once more the data set 'Exam.txt'. Convert this into a data set with three variables, namely, exam score (Y), student ($C \in \{1, 2, \ldots, 88\}$) and subject ($D \in \{1, 2, 3, 4, 5\}$). Analyze your data with a suitable linear model.

Exercise 3.49 (Incomplete designs). In Example 3.46 we tacitly assumed that all group sizes $n(j,k)$ are strictly positive, a so-called *complete design*. However, in case of an additive model for the contributions of the two factors, one may also consider data sets in which some combinations of C and D are not represented, a so-called *incomplete design*. The dimension of the corresponding model space $\boldsymbol{M} \subset \mathbb{R}^n$ depends on the *incidence matrix*

$$\begin{bmatrix} 1_{[n(1,1)>0]} & 1_{[n(1,2)>0]} & \cdots & 1_{[n(1,M)>0]} \\ 1_{[n(2,1)>0]} & 1_{[n(2,2)>0]} & \cdots & 1_{[n(2,M)>0]} \\ \vdots & \vdots & & \vdots \\ 1_{[n(L,1)>0]} & 1_{[n(L,2)>0]} & \cdots & 1_{[n(L,M)>0]} \end{bmatrix}.$$

(a) Determine the dimension of the model space $\boldsymbol{M}$ for the following three incidence matrices:

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

(b) Show that the dimension of the model space $\boldsymbol{M}$ equals $M+L-1$, provided that the following conditions are satisfied:

- Each row and each column of the incidence matrix contains at least one entry 1.
- Let (j_0, k_0) and $(\widetilde{j}, \widetilde{k})$ be two different elements of

$$\mathcal{D} \;:=\; \big\{(j,k) : n(j,k) > 0\big\}.$$

Then, there exists a sequence $(j_1, k_1), (j_2, k_2), \ldots, (j_m, k_m)$ in $\mathcal{D}$ such that $(j_m, k_m) = (\widetilde{j}, \widetilde{k})$ and for $1 \le \ell \le m$, either $j_{\ell-1} = j_\ell$ or $k_{\ell-1} = k_\ell$.

Example 3.50 (Two-way ANOVA). We consider the same setting as in Example 3.46, assuming a complete, balanced design with group size $n_o > 1$. The general model equation is

$$Y_{jks} \;=\; f_{jk} + \varepsilon_{jks} \;=\; \mu + a_j + b_k + h_{jk} + \varepsilon_{jks}, \quad 1 \le j \le L,\ 1 \le k \le M,\ 1 \le s \le n_o,$$

with parameters $\mu \in \mathbb{R}$, $\boldsymbol{a} \in \mathbb{R}^L$ and $\boldsymbol{b} \in \mathbb{R}^M$ as in Example 3.46 and an additional matrix $\boldsymbol{h} \in \mathbb{R}^{L \times M}$ of interactions such that

$$h_{j+} \;=\; 0 \quad \text{for } 1 \le j \le L, \quad h_{+k} \;=\; 0 \quad \text{for } 1 \le k \le M.$$

The corresponding model space is the direct sum of pairwise orthogonal linear subspaces $\boldsymbol{M}_0$, $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$. Here the first three spaces are defined as in Example 3.46, and $\boldsymbol{M}_3$ consists of all arrays $(h_{jk})_{j,k,s}$ with a matrix $\boldsymbol{h} \in \mathbb{R}^{L \times M}$ whose row sums and column sums are equal to 0. The dimension of $\boldsymbol{M}_3$ is equal to $LM - L - M + 1 = (L-1)(M-1)$.

For the array $\boldsymbol{Y}$, this means that the residual array from Example 3.46 is split once more into two parts. This yields the representation

$$\boldsymbol{Y} \;=\; \widehat{\boldsymbol{Y}}_0 + \widehat{\boldsymbol{Y}}_1 + \widehat{\boldsymbol{Y}}_2 + \widehat{\boldsymbol{Y}}_3 + \widehat{\boldsymbol{\varepsilon}}$$

with $\widehat{\boldsymbol{Y}}_0, \widehat{\boldsymbol{Y}}_1, \widehat{\boldsymbol{Y}}_2$ as before, and

$$\begin{aligned} \widehat{\boldsymbol{Y}}_3 \;&=\; \big(\bar{Y}_{jk\cdot} - \bar{Y}_{j\cdot\cdot} - \bar{Y}_{\cdot k\cdot} + \bar{Y}\big)_{j,k,s}, \\ \widehat{\boldsymbol{\varepsilon}} \;&=\; \big(Y_{jks} - \bar{Y}_{ks\cdot}\big)_{j,k,s}. \end{aligned}$$

Suppose we want to test the null hypothesis H_o that the influence of the factors C and D on Y is purely additive. This is equivalent to all interactions being 0, that means,

$$H_o : \boldsymbol{h} \;=\; \boldsymbol{0}.$$

The corresponding test statistic is

$$F \;=\; \frac{\|\widehat{\boldsymbol{Y}}_3\|^2}{(L-1)(M-1)\,\widehat{\sigma}^2}$$

with distribution $F_{(L-1)(M-1),n-LM}$ under H_o.

The connection between the geometrical and the former description of F tests. Recall that in Section 3.5.3, we considered the null hypothesis

$$H_o : \boldsymbol{\Psi}^\top \boldsymbol{\theta} \;=\; \boldsymbol{w}_o$$

for a given matrix $\boldsymbol{\Psi} \in \mathbb{R}^{p\times d}$ with rank d and a given vector $\boldsymbol{w}_o \in \mathbb{R}^d$. At first, we would like to get rid of the vector $\boldsymbol{w}_o$. For that purpose, we use a trick treated in Exercise 3.32: We choose an arbitrary vector $\boldsymbol{\theta}_o \in \mathbb{R}^p$ such that $\boldsymbol{\Psi}^\top \boldsymbol{\theta}_o = \boldsymbol{w}_o$, for instance, $\boldsymbol{\theta}_o = \boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Psi})^{-1}\boldsymbol{w}_o$. Then $\tilde{\boldsymbol{Y}} := \boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\theta}_o$ satisfies the model equation

$$\tilde{\boldsymbol{Y}} \;=\; \boldsymbol{D}\tilde{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}$$

with $\tilde{\boldsymbol{\theta}} := \boldsymbol{\theta} - \boldsymbol{\theta}_o$, and $\boldsymbol{\Psi}^\top\boldsymbol{\theta} = \boldsymbol{w}_o$ if and only if $\boldsymbol{\Psi}^\top\tilde{\boldsymbol{\theta}} = \boldsymbol{0}$. Hence, we may assume without loss of generality that $\boldsymbol{w}_o = \boldsymbol{0}$ and consider the null hypothesis

$$H_o : \boldsymbol{\Psi}^\top \boldsymbol{\theta} \;=\; \boldsymbol{0}.$$

With $\boldsymbol{\mu} = \boldsymbol{D}\boldsymbol{\theta}$ one can write

$$\boldsymbol{\Psi}^\top\boldsymbol{\theta} \;=\; \boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{\theta} \;=\; \boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{D}^\top\boldsymbol{\mu} \;=\; \boldsymbol{A}^\top\boldsymbol{\mu}$$

with the matrix

$$\boldsymbol{A} \;=\; \boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi} \;\in\; \mathbb{R}^{n\times d}.$$

That means, the null hypothesis H_o is equivalent to

$$H_o : \boldsymbol{\mu} \in \boldsymbol{M}_o$$

with the linear subspace

$$\boldsymbol{M}_o \;:=\; \big\{\boldsymbol{y} \in \boldsymbol{M} : \boldsymbol{A}^\top\boldsymbol{y} = \boldsymbol{0}\big\} \;=\; \big\{\boldsymbol{y} \in \boldsymbol{M} : \boldsymbol{y} \perp \boldsymbol{A}\mathbb{R}^d\big\}$$

of $\boldsymbol{M}$. Note that the d columns of $\boldsymbol{A}$ are linearly independent vectors in $\boldsymbol{M}$. Consequently, its column space $\boldsymbol{A}\mathbb{R}^d$ is a d-dimensional subspace of $\boldsymbol{M}$. Hence, $\boldsymbol{M}_o$ is the orthogonal complement of $\boldsymbol{A}\mathbb{R}^d$ within $\boldsymbol{M}$ and has dimension $p_o = p - d$. Likewise, $\boldsymbol{A}\mathbb{R}^d$ is the orthogonal complement of $\boldsymbol{M}_o$ within $\boldsymbol{M}$, that means, $\boldsymbol{A}\mathbb{R}^d = \boldsymbol{M} \cap \boldsymbol{M}_o^\perp$. In particular, the orthogonal projection onto that space is given by the matrix[1]

$$\begin{aligned}
\boldsymbol{H}_1 \;&:=\; \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top \\
&=\; \boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}\,(\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi})^{-1}\,\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{D}^\top \\
&=\; \boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}\,(\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi})^{-1}\,\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{D}^\top \\
&=\; \boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}\,\boldsymbol{\Gamma}_{\boldsymbol{\Psi}}\,\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{D}^\top.
\end{aligned}$$

[1] Note the analogous formula for the hat matrix $\boldsymbol{H} = \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top$ describing the orthogonal projection onto $\boldsymbol{D}\mathbb{R}^p$.

Consequently, the geometrical approach leads to the F test statistic

$$\begin{aligned} F &= \frac{\|\boldsymbol{H}_1\boldsymbol{Y}\|^2/d}{\widehat{\sigma}^2} \\ &= \frac{\boldsymbol{Y}^\top\boldsymbol{H}_1\boldsymbol{Y}}{d\widehat{\sigma}^2} \\ &= \frac{\boldsymbol{Y}^\top\boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}\,\boldsymbol{\Gamma}_{\boldsymbol{\Psi}}\,\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{D}^\top\boldsymbol{Y}}{d\widehat{\sigma}^2} \\ &= \frac{(\boldsymbol{\Psi}^\top\widehat{\boldsymbol{\theta}})^\top\boldsymbol{\Gamma}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}^\top\widehat{\boldsymbol{\theta}})}{d\widehat{\sigma}^2} \\ &= F_{\boldsymbol{\Psi}}(\boldsymbol{0}). \end{aligned}$$

3.6 Alternative Simultaneous Confidence Intervals

Scheffé's method is just one of several possibilities to construct simultaneous confidence intervals for linear functions $\boldsymbol{\psi}^\top\boldsymbol{\theta}$ of the parameter, where $\boldsymbol{\psi}$ is running through a given set $\mathcal{P} \subset \mathbb{R}^p \setminus \{\boldsymbol{0}\}$. The general goal is to find a critical value

$$c_\alpha \;=\; c_{\alpha,\mathcal{P}}$$

such that

$$\mathbb{P}\big(|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}| \le \widehat{\sigma}_{\boldsymbol{\psi}}c_\alpha \text{ for all } \boldsymbol{\psi} \in \mathcal{P}\big) \;\ge\; 1-\alpha.$$

In other words, we may claim with confidence $1-\alpha$ that

$$\boldsymbol{\psi}^\top\boldsymbol{\theta} \;\in\; \big[\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} \pm \widehat{\sigma}_{\boldsymbol{\psi}}c_\alpha\big] \quad \text{for all } \boldsymbol{\psi} \in \mathcal{P}.$$

Scheffé's method yields the critical value

$$c_\alpha \;=\; \sqrt{dF_{d,n-p;1-\alpha}} \quad \text{with } d := \dim(\operatorname{span}(\mathcal{P})).$$

In what follows, we shall discuss two alternative methods.

3.6.1 The Bonferroni Method

Suppose that $\mathcal{P}$ is a finite set, that is,

$$q \;:=\; \#\mathcal{P} \;<\; \infty.$$

We know already that

$$P\big(|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}| > \widehat{\sigma}_{\boldsymbol{\psi}}t_{n-p;1-\gamma/2}\big) \;=\; \gamma$$

for any vector $\boldsymbol{\psi} \in \mathcal{P}$ and arbitrary numbers $\gamma \in (0,1)$. Consequently, if we define

$$c_\alpha \;:=\; t_{n-p;1-(\alpha/q)/2},$$

then the simple Bonferroni inequality implies that

$$\begin{aligned}
&\mathbb{P}\big(|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}| \le \widehat{\sigma}_{\boldsymbol{\psi}} c_\alpha \text{ for all } \boldsymbol{\psi} \in \mathcal{P}\big) \\
&\quad = \; 1 - \mathbb{P}\big(|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}| > \widehat{\sigma}_{\boldsymbol{\psi}} c_{\boldsymbol{\psi}} c_\alpha \text{ for at least one } \boldsymbol{\psi} \in \mathcal{P}\big) \\
&\quad \ge \; 1 - \sum_{\boldsymbol{\psi}\in\mathcal{P}} \mathbb{P}\big(|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}| > \widehat{\sigma}_{\boldsymbol{\psi}} c_\alpha\big) \\
&\quad = \; 1 - \sum_{\boldsymbol{\psi}\in\mathcal{P}} \alpha/q \\
&\quad = \; 1 - \alpha.
\end{aligned}$$

3.6.2 Tukey's Method

The Bonferroni method is certainly conservative. But numerical examples and asymptotic considerations show that in special cases it is often not too far from an exact method which has been introduced by John W. Tukey in special ANOVA settings. We present here a generalization of Tukey's approach. For an arbitrary set $\mathcal{P} \subset \mathbb{R}^p \setminus \{\mathbf{0}\}$, one considers the random variable

$$T_{\mathcal{P}} \;:=\; \sup_{\boldsymbol{\psi}\in\mathcal{P}} |T_{\boldsymbol{\psi}}| \;=\; \sup_{\boldsymbol{\psi}\in\mathcal{P}} \frac{|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}|}{\widehat{\sigma}_{\boldsymbol{\psi}}}.$$

The next lemma shows that $T_{\mathcal{P}}$ is a pivotal statistic with continuous distribution.

Lemma 3.51. *The distribution of $T_{\mathcal{P}}$ is continuous and does not depend on $\boldsymbol{\theta}$ or σ. Rather, $T_{\mathcal{P}}$ has the same distribution as*

$$\frac{\sup_{\boldsymbol{\psi}\in\mathcal{P}} |\boldsymbol{b}_{\boldsymbol{\psi}}^\top \boldsymbol{Z}|}{\sqrt{S^2/(n-p)}}$$

with stochastically independent random variables $\boldsymbol{Z} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{I})$, $S^2 \sim \chi^2_{n-p}$, and the unit vectors

$$\boldsymbol{b}_{\boldsymbol{\psi}} \;:=\; \|\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi}\|^{-1}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi} \;\in\; \mathbb{R}^p, \quad \boldsymbol{\psi} \in \mathcal{P}.$$

In general, the distribution of $T_{\mathcal{P}}$ is not a standard distribution. But it can be simulated easily by refined Monte Carlo methods as explained in the appendix. In some special cases it can be determined exactly or numerically. The critical value

$$c_\alpha \;:=\; (1-\alpha)\text{-quantile of } T_{\mathcal{P}}$$

satisfies the equation

$$\mathbb{P}\Big(|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta}| \le \widehat{\sigma}_{\boldsymbol{\psi}} c_\alpha \text{ for all } \boldsymbol{\psi} \in \mathcal{P}\Big) \;=\; 1-\alpha.$$

Proof of Lemma 3.51. It follows from Theorem 3.16 that $\boldsymbol{Z} := \sigma^{-1}\boldsymbol{\Gamma}^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ and $S^2 := (n-p)\widehat{\sigma}^2/\sigma^2$ are stochastically independent with $\boldsymbol{Z} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{I})$ and $S^2 \sim \chi^2_{n-p}$. Moreover,

$$\begin{aligned}
\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top\boldsymbol{\theta} \;&=\; \sigma\,(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi})^\top\boldsymbol{Z} \;=\; \sigma\|\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi}\|\,\boldsymbol{b}_{\boldsymbol{\psi}}^\top\boldsymbol{Z}, \\
\widehat{\sigma}_{\boldsymbol{\psi}} \;&=\; \widehat{\sigma}\sqrt{\boldsymbol{\psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\psi}} \;=\; \sigma\|\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi}\|\,\sqrt{S^2/(n-p)},
\end{aligned}$$

whence

$$T_{\mathcal{P}} \;=\; \frac{\sup_{\boldsymbol{\psi}\in\mathcal{P}} |\boldsymbol{b}_{\boldsymbol{\psi}}^{\top}\boldsymbol{Z}|}{\sqrt{S^2/(n-p)}}.$$

The enumerator $W := \sup_{\boldsymbol{\psi}\in\mathcal{P}} |\boldsymbol{b}_{\boldsymbol{\psi}}^{\top}\boldsymbol{Z}|$ of $T_{\mathcal{P}}$ is no larger than $\|\boldsymbol{Z}\|$, and for any fixed $\boldsymbol{\psi}_o \in \mathcal{P}$, it is no smaller than $|\boldsymbol{b}_{\boldsymbol{\psi}_o}^{\top}\boldsymbol{Z}|$, the modulus of a standard Gaussian random variable. This shows that $T_{\mathcal{P}} > 0$ almost surely. But for any fixed $x > 0$,

$$\mathbb{P}\big(T_{\mathcal{P}} = x\big) \;=\; \mathbb{E}\,\mathbb{P}\big(S^2 \;=\; (n-p)W^2/x^2 \,\big|\, \boldsymbol{Z}\big) \;=\; 0,$$

because S^2 has a continuous distribution function. Consequently, the distribution of $T_{\mathcal{P}}$ is continuous. □

3.6.3 Examples of Simultaneous Confidence Regions

Let us start with a general remark about the Bonferroni method and Tukey's method. Both are advisable if one is interested primarily in simultaneous confidence bounds for given linear functions $\boldsymbol{\psi}^{\top}\boldsymbol{\theta}$, $\boldsymbol{\psi} \in \mathcal{P}$, of the parameter $\boldsymbol{\theta}$. As illustrated subsequently, these simultanous bounds also imply simultaneous bounds for arbitrary $\boldsymbol{\psi}^{\top}\boldsymbol{\theta}$ with $\boldsymbol{\psi} \in \operatorname{span}(\mathcal{P})$, although Scheffé's method may provide more accurate results for $\boldsymbol{\psi} \notin \mathcal{P}$.

Example 3.52 (Simple linear regression). Recall that

$$\widehat{f}(x) \;=\; \bar{Y} + \widehat{b}(x - \bar{X}), \quad \widehat{b} \;=\; \|\tilde{\boldsymbol{X}}\|^{-2}\tilde{\boldsymbol{X}}^{\top}\boldsymbol{Y}$$

with $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \bar{X}\mathbf{1}$, and for arbitrary $x_1, x_2 \in \mathbb{R}$,

$$\operatorname{Cov}\big(\widehat{f}(x_1), \widehat{f}(x_2)\big) \;=\; \sigma^2\gamma(x_1, x_2) \quad \text{with} \quad \gamma(x_1, x_2) \;:=\; \frac{1}{n} + \frac{(x_1 - \bar{X})(x_2 - \bar{X})}{\|\tilde{\boldsymbol{X}}\|^2}.$$

We first focus on two different potential values $x_1 < x_2$ of X. We know that

$$P\big(f(x_j) \in \big[\widehat{f}(x_j) \pm d_\alpha(x_j)\big] \text{ for } j = 1, 2\big) \;\geq\; 1 - \alpha,$$

provided that

$$d_\alpha(x_j) \;:=\; \widehat{\sigma}(x_j)t_{n-2;1-\alpha/4}$$

with

$$\widehat{\sigma}(x_j) \;:=\; \widehat{\sigma}\sqrt{\gamma(x_j, x_j)}.$$

Alternatively,

$$P\big(f(x_j) \in \big[\widehat{f}(x_j) \pm d_\alpha(x_j)\big] \text{ for } j = 1, 2\big) \;=\; 1 - \alpha,$$

provided that

$$d_\alpha(x_j) \;:=\; \widehat{\sigma}(x_j)c_\alpha$$

with the $(1-\alpha)$-quantile c_α of the following random variable:

$$\frac{\max\{|W_1|, |W_2|\}}{(n-2)^{-1/2}S} \tag{3.2}$$

with independent random variables $S^2 \sim \chi^2_{n-2}$ and

$$\boldsymbol{W} \sim \mathrm{N}_2\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad \rho = \frac{\gamma(x_1, x_2)}{\sqrt{\gamma(x_1, x_1)\gamma(x_2, x_2)}}.$$

In both cases, the two inequalities $|\widehat{f}(x_j) - f(x_j)| \leq d_\alpha(x_j)$ lead to inequalities for $f(x)$ at *arbitrary* positions $x \in \mathbb{R}$: We write

$$\widehat{f}(x) - f(x) = \lambda_1(x)\big(\widehat{f}(x_1) - f(x_1)\big) + \lambda_2(x)\big(\widehat{f}(x_2) - f(x_1)\big)$$

with

$$\lambda_1(x) := \frac{x_2 - x}{x_2 - x_1} \quad \text{and} \quad \lambda_2(x) := \frac{x - x_1}{x_2 - x_1}.$$

Then

$$\big|\widehat{f}(x) - f(x)\big| \leq |\lambda_1(x)| d_\alpha(x_1) + |\lambda_2(x)| d_\alpha(x_2).$$

Numerical example. We consider once more the $n = 51$ observation pairs from Example 3.20. Figure 3.7 shows the data and the estimator $\widehat{f}$, together with simultaneous 95%-confidence intervals for f via Scheffé's method and via Tukey's method. The bounds from Scheffé's method are given by

$$\widehat{f}(x) \pm \widehat{\sigma}(x)\sqrt{2F_{2,49;0.95}} \approx \widehat{f}(x) \pm \widehat{\sigma}(x)\sqrt{2 \cdot 3.1866} \approx \widehat{f}(x) \pm \widehat{\sigma}(x) \cdot 2.5245.$$

Tukey's method yields piecewise linear bounds, namely,

$$\widehat{f}(x) \pm (|\lambda_1(x)|\widehat{\sigma}(x_1) + |\lambda_2(x)|\widehat{\sigma}(x_2)) \cdot \widehat{\kappa}_{0.95},$$

where $\widehat{\kappa}_{0.95}$ is a Monte Carlo estimator of the 0.95-quantile of the random variable (3.2) with $n = 51$. In the upper panel of Figure 3.7,

$$x_1 = -0.1, \ x_2 = 1.1, \ \rho \approx -0.6119, \quad \text{and} \quad \widehat{\kappa}_{0.95} = 2.259.$$

In the lower panel,

$$x_1 = 0.333, \ x_2 = 0.667, \ \rho \approx 0.5131, \quad \text{and} \quad \widehat{\kappa}_{0.95} = 2.276.$$

Note that the Bonferroni method would use $t_{n-2;1-\alpha/4} = t_{49;0.9875} \approx 2.3124$ in place of $\widehat{\kappa}_{0.95}$.

Interestingly, Scheffé's method performs rather well in comparison with Tukey's method. As expected, for x close to x_1 or x_2, the Tukey intervals are a bit shorter than the Scheffé intervals, but at other locations x, the Scheffé intervals are clearly more narrow.

Example 3.53 (Extending finitely many bounds). The confidence bands in the previous example can be generalized as follows: Let $\mathcal{F} = \mathrm{span}(f_1, \ldots, f_p)$ with basis functions $f_j : \mathcal{X} \to \mathbb{R}$. Let $\boldsymbol{x} = (x_j)_{j=1}^p$ be a fixed tuple in $\mathcal{X}^p$ such that the matrix

$$\boldsymbol{B} := \big[f_1(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x})\big] \in \mathbb{R}^{p \times p}$$

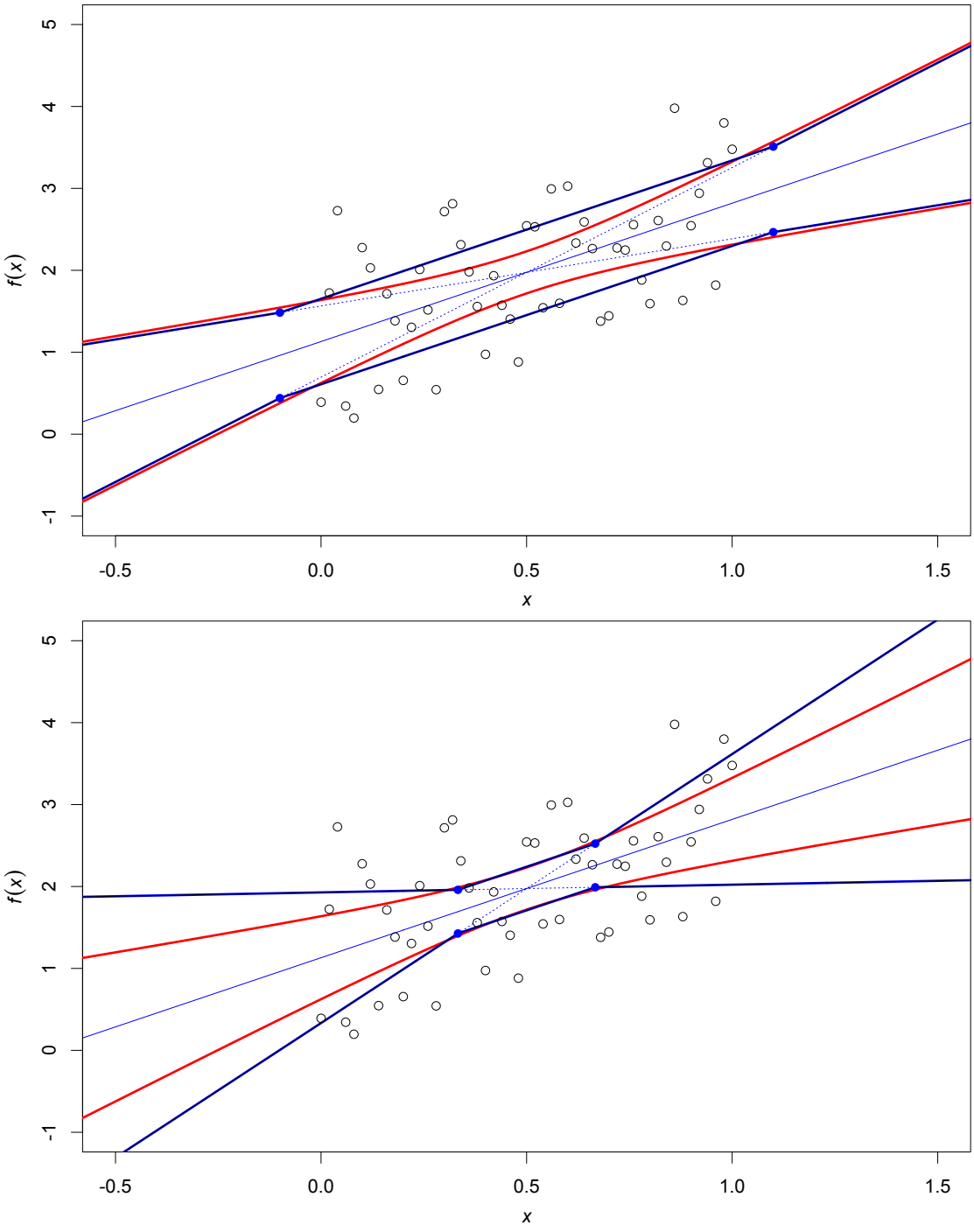

Figure 3.7: Simultaneous 95%-confidence bounds for $f(x)$ in simple linear regression: Scheffé's method (red) and Tukey's method with two starting points (blue).

has full rank. The Bonferroni method or Tukey's method yield constants $\widehat{d}(x_j) = \widehat{\sigma}(x_j)c_\alpha$, $1 \leq j \leq p$, such that with confidence $1-\alpha$ we may assume that

$$\big|\widehat{f}(x_j) - f(x_j)\big| \;\leq\; \widehat{d}(x_j) \quad \text{for } 1 \leq j \leq p.$$

Now, any function $g \in \mathcal{F}$ may be written as $g = \sum_{k=1}^p \eta_k f_k$ for some $\boldsymbol{\eta} \in \mathbb{R}^p$. In particular, $g(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\eta}$, so $\boldsymbol{\eta} = \boldsymbol{B}^{-1}g(\boldsymbol{x})$. Consequently, for any $x \in \mathcal{X}$,

$$g(x) \;=\; \sum_{k=1}^p \eta_k f_k(x) \;=\; \sum_{k=1}^p \sum_{j=1}^p (\boldsymbol{B}^{-1})_{kj} g(x_j) f_k(x) \;=\; \sum_{j=1}^p \lambda_j(x) g(x_j)$$

with

$$\lambda_j(x) \;:=\; \sum_{k=1}^p (\boldsymbol{B}^{-1})_{kj} f_k(x).$$

Applying this to $g := \widehat{f} - f$ shows that for arbitrary $x \in \mathcal{X}$,

$$\big|\widehat{f}(x) - f(x)\big| \;\leq\; \Big|\sum_{j=1}^p \lambda_j(x)\big(\widehat{f}(x_j) - f(x_j)\big)\Big| \;\leq\; \sum_{j=1}^p |\lambda_j(x)|\widehat{d}(x_j).$$

In some settings, Tukey's method involves the following new type of distributions.

Definition 3.54 (Studentized range). Let $\mathbb{Z} \sim \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{I})$ and $S^2 \sim \chi^2_\ell$ be stochastically independent. The distribution of

$$\frac{\max(Z_1, \ldots, Z_k) - \min(Z_1, \ldots, Z_k)}{\sqrt{S^2/\ell}}$$

is called the *studentized range distribution with parameters k and ℓ*. It is denoted with $Q_{k,\ell}$, and its β-quantile is denoted with $Q_{k,\ell;\beta}$.

In R, the distribution and quantile function of $Q_{k,\ell}$ are available as

$$\mathsf{ptukey}(\cdot, \mathsf{nmeans} = k, \mathsf{df} = \ell) \quad \text{and} \quad \mathsf{qtukey}(\cdot, \mathsf{nmeans} = k, \mathsf{df} = \ell),$$

respectively.

Example 3.55 (One-way ANOVA, Example 1.1). We use the blackboard-and-paper notation from Section 2.6.4. That means, we observe

$$Y_{js} \;=\; f_j + \varepsilon_{js}, \quad 1 \leq j \leq L,\ 1 \leq s \leq n(j),$$

with unknown parameters $f_1, \ldots, f_L \in \mathbb{R}$ and independent random errors $\varepsilon_{js} \sim \mathrm{N}(0, \sigma^2)$. Often, one is interested mainly in the differences $f_j - f_k$, where $1 \leq j < k \leq L$. With $\boldsymbol{\theta} = (f_j)_{j=1}^L$ and the standard basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_L$ of $\mathbb{R}^L$, this corresponds to the family

$$\mathcal{P} \;:=\; \big\{\boldsymbol{e}_j - \boldsymbol{e}_k : 1 \leq j < k \leq L\big\}.$$

It consists of $L(L-1)/2$ vectors spanning the $(L-1)$-dimensional space of all vectors $\boldsymbol{v} \in \mathbb{R}^L$ such that $v_+ = 0$. It follows from Scheffé's method that with probability $1-\alpha$,

$$f_j - f_k \;\in\; \Big[\bar{Y}_{j\cdot} - \bar{Y}_{k\cdot} \pm \widehat{\sigma}_{jk}\sqrt{(L-1)F_{L-1,n-L;1-\alpha}}\Big]$$

for $1 \le j < k \le L$, where $\widehat{\sigma}_{jk} := \widehat{\sigma}\sqrt{n(j)^{-1} + n(k)^{-1}}$.

The Bonferroni method leads to the simultaneous confidence intervals

$$\big[\bar{Y}_{j\cdot} - \bar{Y}_{k\cdot} \pm \widehat{\sigma}_{jk} t_{n-L;1-\alpha/(L(L-1))}\big].$$

Tukey's method is particularly easy to apply in the special case of a balanced design with group sizes

$$n(1) = n(2) = \cdots = n(L) = n_o.$$

If we define $Z_j := n_o^{1/2}(\bar{Y}_{j\cdot} - f_j)/\sigma$ and $S^2 := (n-L)\widehat{\sigma}^2/\sigma^2$, then the random variables $Z_1, \ldots, Z_L, S^2$ are stochastically independent, where $Z_j \sim \mathrm{N}(0,1)$ and $S^2 \sim \chi^2_{n-L}$. Moreover, $\widehat{\sigma}_{jk} = \widehat{\sigma}\sqrt{2/n_o}$, and thus

$$\begin{aligned} T_{\mathcal{P}} &= \max_{1\le j<k\le L} \frac{\big|\bar{Y}_{j\cdot} - \bar{Y}_{k\cdot} - f_j + f_k\big|}{\widehat{\sigma}_{jk}} \\ &= 2^{-1/2} \max_{1\le j<k\le L} \frac{|Z_j - Z_k|}{\sqrt{S^2/(n-L)}} \\ &= 2^{-1/2} \frac{\max(Z_1,\ldots,Z_L) - \min(Z_1,\ldots,Z_L)}{\sqrt{S^2/(n-L)}}. \end{aligned}$$

Hence $\sqrt{2}\,T_{\mathcal{P}}$ has the distribution $Q_{L,n-L}$ of a studentized range with parameters L and $n-L$. Hence, one may claim with confidence $1-\alpha$ that

$$\begin{aligned} f_j - f_k &\in \big[\bar{Y}_{j\cdot} - \bar{Y}_{k\cdot} \pm \widehat{\sigma}_{jk} 2^{-1/2} Q_{L,n-L;1-\alpha}\big] \\ &= \big[\bar{Y}_{j\cdot} - \bar{Y}_{k\cdot} \pm \widehat{\sigma}\sqrt{1/n_o}\, Q_{L,(n_o-1)L;1-\alpha}\big] \end{aligned}$$

for $1 \le j < k \le L$, where $Q_{L,n-L;1-\alpha}$ is the $(1-\alpha)$-quantile of $Q_{L,n-L}$.

Table 3.1 contains for $\alpha = 0.05$ and various combinations of L, n_o the three critical values

$$\begin{aligned} c_\alpha^{\text{Scheffe}} &:= \sqrt{(L-1)F_{L-1,(n_o-1)L;1-\alpha}}, \\ c_\alpha^{\text{Bonf}} &:= t_{(n_o-1)L;1-\alpha/(L(L-1))}, \\ c_\alpha^{\text{Tukey}} &:= 2^{-1/2} Q_{L-1,(n_o-1)L;1-\alpha}. \end{aligned}$$

Exercise 3.56. Let $\boldsymbol{v} \in \mathbb{R}^L$ be an unknown vector in $\mathbb{R}^L$. Suppose we only know that for some constant $d > 0$,

$$|v_k - v_j| \le d \quad \text{for } 1 \le j < k \le L.$$

Which bound can be deduced for

$$\Big|\sum_{j=1}^{L} \psi_j v_j\Big|,$$

if $\boldsymbol{\psi}$ is a given vector in $\mathbb{R}^L$ such that $\sum_{j=1}^{L} \psi_j = 0$?

Exercise 3.57. Consider a one-way ANOVA with balanced design, that is, we observe $Y_{js} = f_j + \varepsilon_{js}$, $1 \le j \le L$ and $1 \le s \le n_o$, with unknown paramaeters $f_1, \ldots, f_L$ and independent errors

L	n_o	$c_{0.05}^{\text{Scheffe}}$	$c_{0.05}^{\text{Bonf}}$	$c_{0.05}^{\text{Tukey}}$
3	10	2.5901	2.5525	2.4795
5	10	3.2117	2.9521	2.8415
10	10	4.2274	3.3693	3.2445
3	50	2.4730	2.4217	2.3677
5	50	3.1039	2.8327	2.7483
10	50	4.1342	3.2803	3.1784
3	100	2.4602	2.4077	2.3556
5	100	3.0919	2.8197	2.7379
10	100	4.1236	3.2704	3.1710

Table 3.1: Some critical values for group comparisons in balanced one-way ANOVA.

$\varepsilon_{js} \sim \mathrm{N}(0, \sigma^2)$. With Tukey's or the Bonferroni method we find a constant $\widehat{d} = \widehat{\sigma}\sqrt{2/n_o}c_\alpha > 0$ such that with confidence $1-\alpha$, we may claim that

$$f_j - f_k \;\in\; \big[\widehat{f_j} - \widehat{f_k} \pm \widehat{d}\big] \quad \text{for } 1 \le j < k \le L.$$

What can you claim about the quantities

$$f_j - \frac{1}{L-1}\sum_{k:\,k\neq j} f_k, \quad f_j - \max_{k:\,k\neq j} f_k, \quad \text{rank of } f_j \text{ within } (f_k)_{k=1}^L$$

for $1 \le j \le L$?

Exercise 3.58. Suppose that we observe

$$Y_{jks} \;=\; f_j + b_k + \varepsilon_{ijs}, \quad 1 \le i \le L, 1 \le k \le M, 1 \le s \le n_o,$$

with unknown real parameters $f_1, \ldots, f_L$ and $b_1, \ldots, b_M$ such that $\sum_{k=1}^M b_k = 0$, and with independent random errors $\varepsilon_{jks} \sim \mathrm{N}(0, \sigma^2)$, where $\sigma > 0$ is unknown. Show how to apply Tukey's method and studentized range distributions to obtain simultaneous confidence intervals for the so-called main contrasts $f_j - f_{j'}$, $1 \le j < j' \le L$.

3.6.4 Comparison of the Methods

A precise and general comparison of the previous methods is impossible. But at least in asymptotic frameworks with $n - p$ tending to infinity, some comparisons are possible.

Lemma 3.59 (Asymptotics of various quantiles). *The student quantiles admit the following approximation:*

$$\frac{t_{k;1-\delta}}{\sqrt{2\log(1/\delta)}} \;\to\; 1 \quad \textit{as } \delta \downarrow 0,\ k \to \infty \textit{ and } \frac{\log(1/\delta)}{k} \to 0.$$

For fixed $\gamma \in (0,1)$, the γ-quantiles of the F distribution and the studentized range with parameters k and ℓ satisfy

$$F_{k,\ell;\gamma} \;\to\; 1 \quad \textit{and} \quad \frac{Q_{k,\ell;\gamma}}{2\sqrt{2\log(k)}} \;\to\; 1 \quad \textit{as } k, \ell \to \infty.$$

Before proving this lemma, let us show how to apply it. In some special settings such as balanced one-way ANOVA or balanced two-way ANOVA one encounters the following situation: The parameter vector $\boldsymbol{\theta}$ contains $L \geq 3$ parameters $f_1, \ldots, f_L$ such that the corresponding Gauss-Markov estimators $\widehat{f}_1, \ldots, \widehat{f}_L$ are stochastically independent with $\widehat{f}_j \sim \mathrm{N}(f_j, \sigma^2\gamma)$ for some known constant $\gamma > 0$. Now we are interested in simultaneous $(1-\alpha)$-confidence intervals for the so-called main contrasts $f_j - f_k$, $1 \leq j < k \leq L$. These confidence intervals have the form

$$\big[\widehat{f}_j - \widehat{f}_k \pm \widehat{\sigma}\sqrt{2\gamma}\, c_\alpha\big]$$

with

$$c_\alpha \;=\; \begin{cases} \sqrt{(L-1)F_{L-1,n-p;1-\alpha}} & \text{(Scheffé)}, \\ t_{n-p;1-\alpha/[L(L-1)]} & \text{(Bonferroni)}, \\ 2^{-1/2}Q_{L,n-p;1-\alpha} & \text{(Tukey)}. \end{cases}$$

Now, the expansions in Lemma 3.59 imply that as $L \to \infty$ and $n - p \to \infty$,

$$c_\alpha \;=\; \begin{cases} \sqrt{L}\,(1+o(1)) & \text{(Scheffé)} \\ 2\sqrt{\log L}\,(1+o(1)) & \text{(Bonferroni)} \\ 2\sqrt{\log L}\,(1+o(1)) & \text{(Tukey)} \end{cases}$$

(where the expansion for the Bonferroni method requires that $\log(L)/(n-p) \to 0$). This shows that Scheffé's method is rather conservative in comparison with the Bonferroni or Tukey's method, while the latter two yield similar results.

Proof of Lemma 3.59. We consider stochastically independent random variables $Z \sim \mathrm{N}(0,1)$, $\boldsymbol{Z} \sim \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{I}_k)$ and $T_\ell^2 \sim \chi_\ell^2$. Since $\mathbb{E}(T_\ell^2/\ell) = 1$ and $\mathrm{Var}(T_\ell^2/\ell) = 2/\ell$, it follows from the Tsheyshev inequality that for arbitrary fixed $\varepsilon > 0$,

$$\mathbb{P}\big(T_\ell^2/\ell \notin [1 \pm \varepsilon]\big) \;\leq\; 2/(\ell\varepsilon^2).$$

Hence T_ℓ^2/ℓ converges to 1 in probability as $\ell \to \infty$. The same consideration applies to $S_k^2 := \|\boldsymbol{Z}\|^2 \sim \chi_k^2$, so the random variable

$$F \;:=\; \frac{S_k^2/k}{T_\ell^2/\ell} \;\sim\; F_{k,\ell}$$

satisfies

$$F \;\to_p\; 1 \quad \text{as } k, \ell \to \infty.$$

But this is equivalent to the statement that for any fixed $\gamma \in (0,1)$,

$$F_{k,\ell;\gamma} \;\to\; 1 \quad \text{as } k, \ell \to \infty.$$

It follows from Exercise 3.61 that

$$\frac{\max(\boldsymbol{Z})}{\sqrt{2\log(k)}} \;\to_p\; 1 \quad \text{as } k \to \infty.$$

For symmetry reasons, $\min(\boldsymbol{Z})$ has the same distribution as $-\max(\boldsymbol{Z})$. Consequently, the range $\max(\boldsymbol{Z}) - \min(\boldsymbol{Z})$ is equal to $2\sqrt{2\log(k)}(1 + o_p(1))$ as $k \to \infty$. As a consequence, the random variable

$$Q \;:=\; \frac{\max(\boldsymbol{Z}) - \min(\boldsymbol{Z})}{\sqrt{T_\ell^2/\ell}} \;\sim\; Q_{k,\ell}$$

satisfies

$$\frac{Q}{2\sqrt{2\log(k)}} \;\to_p\; 1 \quad \text{as } k, \ell \to \infty.$$

In particular, for any fixed $\gamma \in (0, 1)$,

$$\frac{Q_{k,\ell;\gamma}}{2\sqrt{2\log(k)}} \;\to\; 1 \quad \text{as } k, \ell \to \infty.$$

For the student quantiles we have to work a bit harder. With $Y_k := \sqrt{S_k^2/k}$, the ratio Z/Y_k has distribution t_k. As shown in Exercise 3.13,

$$t_{k;1-\delta} \;>\; \Phi^{-1}(1 - \delta) \quad \text{for arbitrary } \delta \in (0, 1/2),$$

and in Exercise 3.61 it is shown that

$$\frac{\Phi^{-1}(1 - \delta)}{\sqrt{2\log(1/\delta)}} \;\to\; 1 \quad \text{as } \delta \downarrow 0.$$

On the other hand, Exercise 3.60 implies that $\Phi(-x) \;\leq\; \exp(-x^2/2)/2$ for arbitrary $x \;\geq\; 0$. Hence,

$$\mathbb{P}(Z/Y_k > t) \;=\; \mathbb{E}\,\mathbb{P}(Z > tY_k \,|\, Y_k) \;=\; \mathbb{E}\,\Phi(-tY_k) \;\leq\; \mathbb{E}\exp(-t^2Y_k^2/2)/2.$$

It follows from Exercise 3.12 that

$$\mathbb{E}\exp(\lambda S_k^2) \;=\; (1 - 2\lambda)^{-k/2} \quad \text{for arbitrary } \lambda < 1/2.$$

Setting $\lambda = -t^2/(2k)$, this leads to the inequality

$$\mathbb{P}(Z/Y_k > t) \;\leq\; \Big(1 + \frac{t^2}{k}\Big)^{-k/2}/2.$$

Now we choose $t > 0$ such that the right hand side equals δ. This yields the inequality

$$\begin{aligned} t_{k;1-\delta} \;&\leq\; \sqrt{k\big((2\delta)^{-2/k} - 1\big)} \\ &=\; \sqrt{k\bigg(\exp\Big(\frac{2\log(1/\delta) - \log 4}{k}\Big) - 1\bigg)} \\ &=\; \sqrt{2\log(1/\delta)(1 + o(1))} \quad \text{as } \delta \downarrow 0, \frac{\log(1/\delta)}{k} \to 0. \end{aligned}$$

□

Exercise 3.60. Show that the standard Gaussian distribution function Φ satisfies

$$\frac{\phi(x)}{\sqrt{1+x^2/4}+x/2} \;\le\; 1-\Phi(x) \;\le\; \frac{\phi(x)}{\sqrt{2/\pi+x^2/4}+x/2} \quad \text{for } x \ge 0.$$

In particular, $1-\Phi(x) \le \exp(-x^2/2)/2$ for all $x \ge 0$.

Hint: Consider the function $\Delta := 1-\Phi-\phi/h$, where $h(x) := x/2+\sqrt{c+x^2/4}$ for some constant $c>0$. Show that

$$\Delta' \;=\; \frac{\phi}{h^2}(h'-c).$$

Now, verify and use the fact that $\Delta(0) = 1/2 - 1/\sqrt{2\pi c}$ whereas $\lim_{x\to\infty}\Delta(x) = 0$.

Exercise 3.61. Show by means of Exercise 3.60 that

$$\Phi^{-1}(1-\delta) \begin{cases} \le \sqrt{2\log(1/\delta)-\log 4} & \text{for } 0<\delta\le 1/2, \\ = \sqrt{2\log(1/\delta)}\,(1+o(1)) & \text{for } \delta \downarrow 0. \end{cases}$$

Further, show that for independent, standard Gaussian random variables $Z_1, Z_2, \ldots, Z_n$ and fixed constants $c \in [0,2]$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\max_{i=1,2,\ldots,n} Z_i \le \sqrt{c\log n}\Big) \;=\; \begin{cases} 1 & \text{if } c=2, \\ 0 & \text{if } c<2. \end{cases}$$

3.7 Non-Central F Distributions and Approximation Errors

To compute the power of an F test, i.e. the probability that it rejects the null hypothesis, one needs *non-central F distributions*.

Definition 3.62 (Non-central chi-squared and F distributions). Let $Z_1, Z_2, Z_3, \ldots$ be a sequence of stochastically independent, standard Gaussian random variables, and let $\delta, \tilde\delta \ge 0$ be fixed numbers.

(a) The *non-central chi-squared distribution with k degrees of freedom and non-centrality parameter (NCP) δ^2* is defined as the distribution of

$$(Z_1+\delta)^2 + Z_2^2 + \cdots + Z_k^2.$$

It is denoted with $\chi^2_k(\delta^2)$.

(b) The *non-central F distribution with k and ℓ degrees of freedom and non-centrality parameters (NCPs) δ^2 and $\tilde\delta^2$* is defined as the distribution of

$$\frac{k^{-1}\big((Z_1+\delta)^2 + Z_2^2 + \cdots + Z_k^2\big)}{\ell^{-1}\big((Z_{k+1}+\tilde\delta)^2 + Z_{k+2}^2 + \cdots + Z_{k+\ell}^2\big)}.$$

It is denoted with the symbol $F_{k,\ell}(\delta^2, \tilde\delta^2)$.

Remark 3.63 (Moments of $\chi^2_k(\delta^2)$). For a random variable $S^2 \sim \chi^2_k(\delta^2)$,

$$\mathbb{E}(S^2) \;=\; k+\delta^2 \quad \text{and} \quad \mathrm{Std}(S^2) \;=\; \sqrt{2k+4\delta^2}.$$

This follows from writing

$$(Z_1+\delta)^2+Z_2^2+\cdots+Z_k^2 \;=\; k+\delta^2+\sum_{i=1}^{k}(Z_i^2-1)+2\delta Z_1$$

and noting that the $k+1$ random variables $Z_1^2-1,\ldots,Z_k^2-1$ and $2\delta Z_1$ are centered and uncorrelated with variance $\mathrm{Var}(Z_i^2-1)=2$, $\mathrm{Var}(2\delta Z_1)=4\delta^2$. In particular,

$$\frac{S^2}{k+\delta^2} \;\to_p\; 1 \quad \text{as } k+\delta^2\to\infty.$$

Remark 3.64 (Stochastic orders with respect to NCPs). The distributions $\chi^2_k(\delta^2)$ and $F_{k,\ell}(\delta^2,\tilde\delta^2)$ are continuous and strictly increasing (in the sense of stochastic order) in $\delta\geq 0$. Furthermore, $F_{k,\ell}(\delta^2,\tilde\delta^2)$ is continuous and strictly decreasing (in the sense of stochastic order) in $\tilde\delta\geq 0$. More precisely, let T_δ and $U_{\delta,\tilde\delta}$ be the random variables described in Definition 3.62 (a) and (b), respectively. For an arbitrary threshold $c>0$, both probabilities $\mathbb{P}(T_\delta\leq c)$ and $\mathbb{P}(U_{\delta,\tilde\delta}\leq c)$ are a continuous and strictly decreasing function of $\delta\geq 0$ with limit 0 as $\delta\to\infty$. Furthermore, $\mathbb{P}(U_{\delta,\tilde\delta}\leq c)$ is a continuous and strictly increasing function of $\tilde\delta\geq 0$ with limit 1 as $\tilde\delta\to\infty$.

Exercise 3.65. Let Z be a standard Gaussian random variable. For $\delta\in\mathbb{R}$ and $r\geq 0$ let

$$h(\delta,r) \;:=\; \mathbb{P}\left((Z+\delta)^2\leq r\right).$$

Show that $h:\mathbb{R}\times[0,\infty)$ is a continuous function with $h(\cdot,0)\equiv 0$, where $h(-\delta,r)=h(\delta,r)>0$ for $r>0$. Further, show that for fixed $r>0$, $h(\delta,r)$ is strictly decreasing in $\delta\geq 0$ with limit 0 as $\delta\to\infty$.

Prove Remark 3.64 by conditioning on all random variables Z_i, except Z_1 or Z_{k+1}.

Remark 3.66 (Representing $\chi^2_k(\delta^2)$ and $F_{k,\ell}(\delta^2,\tilde\delta^2)$ as a Poisson mixture). As shown in Exercise 3.67 below, non-central chi-squared distributions may be represented as mixtures of ordinary chi-squared distributions. Precisely,

$$\chi^2_k(\delta^2) \;=\; \sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!}\cdot\chi^2_{k+2j} \quad \text{with } \lambda:=\delta^2/2.$$

The probability weights $e^{-\lambda}\lambda^j/j!$ on the right hand side are the weights of the Poisson distribution $\mathrm{Poiss}(\lambda)$.

In other words, let $N,Z_1,Z_2,Z_3,\ldots$ be stochastically independent random variables, where $N\sim\mathrm{Poiss}(\delta^2/2)$ and $Z_i\sim\mathrm{N}(0,1)$. Then $\chi^2_k(\delta^2)$ is the distribution of

$$\sum_{i=1}^{k+2N} Z_i^2.$$

Similarly, let $N,Z_1,Z_2,Z_3,\ldots$ and $\tilde N,\tilde Z_1,\tilde Z_2,\tilde Z_3,\ldots$ be stochastically independent random variables such that $N\sim\mathrm{Poiss}(\delta^2/2)$, $\tilde N\sim\mathrm{Poiss}(\tilde\delta^2/2)$ and $Z_i,\tilde Z_i\sim\mathrm{N}(0,1)$. Then $F_{k,\ell}(\delta^2,\tilde\delta^2)$ is the distribution of

$$\frac{k^{-1}\sum_{i=1}^{k+2N}Z_i^2}{\ell^{-1}\sum_{j=1}^{\ell+2\tilde N}\tilde Z_i^2}.$$

Exercise 3.67 (Non-central χ^2 distributions). **(a)** We have verified in Exercise 3.12 that for $Z \sim \mathrm{N}(0,1)$ and $t < 1/2$,

$$\mathbb{E}\exp(tZ^2) \;=\; (1-2t)^{-1/2}.$$

Show that for arbitrary $\delta \in \mathbb{R}$ and $t < 1/2$,

$$\mathbb{E}\exp\bigl(t(Z+\delta)^2\bigr) \;=\; (1-2t)^{-1/2}\exp(-\delta^2/2)\exp\Bigl(\frac{\delta^2/2}{1-2t}\Bigr)$$

(b) Deduce from part (a) and Remark 3.11 the Poisson representation in Remark 3.66.

A first implication. Let $\boldsymbol{Y}$ be a random vector with distribution $\mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Then,

$$\|\boldsymbol{Y}\|^2 \;\sim\; \chi^2_n\bigl(\|\boldsymbol{\mu}\|^2\bigr).$$

To verify this claim, let $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n$ be an orthonormal basis of $\mathbb{R}^n$ such that $\boldsymbol{\mu} = \|\boldsymbol{\mu}\|\boldsymbol{t}_1$. Then $\boldsymbol{Y}$ has the same distribution as

$$\boldsymbol{\mu} + \sum_{i=1}^n Z_i \boldsymbol{t}_i \;=\; (Z_1 + \|\boldsymbol{\mu}\|)\boldsymbol{t}_1 + \sum_{i=2}^n Z_i \boldsymbol{t}_i$$

with independent, standard Gaussian random variables $Z_1, Z_2, \ldots, Z_n$. Consequently, $\|\boldsymbol{Y}\|^2 = \sum_{i=1}^n (\boldsymbol{t}_i^\top \boldsymbol{Y})^2$ has the same distribution as

$$(Z_1 + \|\boldsymbol{\mu}\|)^2 + \sum_{i=2}^n Z_i^2 \;\sim\; \chi^2_n\bigl(\|\boldsymbol{\mu}\|^2\bigr).$$

Application to F tests. Now, we consider a random vector

$$\boldsymbol{Y} \;=\; \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

with an unknown mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ an an unobserved random error $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. For given linear spaces $\boldsymbol{M}_o \subset \boldsymbol{M} \subset \mathbb{R}^n$ with dimensions $p_o < p < n$, we want to test

$$H_o : \boldsymbol{\mu} \in \boldsymbol{M}_o \quad \text{versus} \quad H_A : \boldsymbol{\mu} \in \boldsymbol{M} \setminus \boldsymbol{M}_o.$$

With the corresponding projection matrices $\boldsymbol{H}_o$ and $\boldsymbol{H}$, recall the F test statistic

$$F \;=\; \frac{\|\boldsymbol{H}\boldsymbol{Y} - \boldsymbol{H}_o\boldsymbol{Y}\|^2/(p-p_o)}{\|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}\|^2/(n-p)}.$$

Under the null hypothesis H_o, this test statistic has an F distribution with $p - p_o$ and $n - p$ degress of freedom. The following theorem specifies its distribution in the general case, without any assumptions on $\boldsymbol{\mu}$.

Theorem 3.68. *The test statistic F has distribution $F_{p-p_o,n-p}(\delta^2, \tilde{\delta}^2)$, where*

$$\delta \;:=\; \frac{\|\boldsymbol{H}\boldsymbol{\mu} - \boldsymbol{H}_o\boldsymbol{\mu}\|}{\sigma} \quad \textit{and} \quad \tilde{\delta} \;:=\; \frac{\|\boldsymbol{\mu} - \boldsymbol{H}\boldsymbol{\mu}\|}{\sigma}.$$

Remark 3.69. Note that Theorem 3.68 does *not* assume that $\boldsymbol{\mu} \in \boldsymbol{M}$. In fact, one may re-interpret the F test as a test of

$$H_o : \boldsymbol{H}\boldsymbol{\mu} \in \boldsymbol{M}_o \quad \text{versus} \quad H_A : \boldsymbol{H}\boldsymbol{\mu} \in \boldsymbol{M} \setminus \boldsymbol{M}_o.$$

Under this null hypothesis, F has distribution $F_{p-p_o,n-p}(0, \tilde{\delta}^2)$ with $\tilde{\delta} = \|\boldsymbol{\mu} - \boldsymbol{H}\boldsymbol{\mu}\|/\sigma$. This fact and Remark 3.64 imply that under H_o,

$$\mathbb{P}(H_o \text{ is rejected}) \;=\; \mathbb{P}(F \geq F_{p-p_o,n-p;1-\alpha}) \;\leq\; \alpha,$$

with equality if and only if $\boldsymbol{\mu} = \boldsymbol{H}\boldsymbol{\mu}$, i.e. $\boldsymbol{\mu} \in \boldsymbol{M}$.

Remark 3.70 (Power of the F test). Under the usual assumption that $\boldsymbol{\mu} \in \boldsymbol{M}$ (i.e. $\boldsymbol{\mu} = \boldsymbol{H}\boldsymbol{\mu}$), the test statistic F follows $F_{p-p_o,n-p}(\delta^2, 0)$, where $\delta = \|\boldsymbol{\mu} - \boldsymbol{H}_o\boldsymbol{\mu}\|/\sigma$. If we denote the distribution function of $F_{p-p_o,n-p}(\delta^2, 0)$ temporarily with $G_\delta(\cdot)$, then

$$\begin{aligned}
\mathbb{P}(H_o \text{ is rejected}) \;&= \mathbb{P}(F \geq F_{p-p_o,n-p;1-\alpha}) \\
&= \; 1 - G_\delta(F_{p-p_o,n-p;1-\alpha}) \begin{cases} = \; \alpha & \text{if } \delta = 0, \\ > \; \alpha & \text{if } \delta > 0, \\ \to \; 1 & \text{as } \delta \to \infty. \end{cases}
\end{aligned}$$

Proof of Theorem 3.68. We just refine the proof of Theorem 3.37. We choose an orthonormal basis $\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n$ of $\mathbb{R}^n$ such that

$$\begin{aligned}
\boldsymbol{M}_o \;&=\; \operatorname{span}(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{p_o}), \\
\boldsymbol{M} \;&=\; \operatorname{span}(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p), \\
\boldsymbol{H}\boldsymbol{\mu} - \boldsymbol{H}_o\boldsymbol{\mu} \;&=\; \|\boldsymbol{H}\boldsymbol{\mu} - \boldsymbol{H}_o\boldsymbol{\mu}\|\, \boldsymbol{t}_{p_o+1} \;=\; \sigma\delta\, \boldsymbol{t}_{p_o+1}, \\
\boldsymbol{\mu} - \boldsymbol{H}\boldsymbol{\mu} \;&=\; \|\boldsymbol{\mu} - \boldsymbol{H}\boldsymbol{\mu}\|\, \boldsymbol{t}_{p+1} \;=\; \sigma\tilde{\delta}\, \boldsymbol{t}_{p+1}.
\end{aligned}$$

With the random vector $\boldsymbol{Z} := \sigma^{-1}\big(\boldsymbol{t}_i^\top \boldsymbol{\varepsilon}\big)_{i=1}^n \sim \mathrm{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$ we can write

$$\boldsymbol{Y} \;=\; \boldsymbol{\mu} + \sigma \sum_{i=1}^{n} Z_i \boldsymbol{t}_i$$

and may conclude that

$$\begin{aligned}
\sigma^{-1}(\boldsymbol{H}\boldsymbol{Y} - \boldsymbol{H}_o\boldsymbol{Y}) \;&=\; (Z_{p_o+1} + \delta)\boldsymbol{t}_{p_o+1} + Z_{p_o+2}\boldsymbol{t}_{p_o+2} + \cdots + Z_p\boldsymbol{t}_p, \\
\sigma^{-1}(\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}) \;&=\; (Z_{p+1} + \tilde{\delta})\boldsymbol{t}_{p+1} + Z_{p+2}\boldsymbol{t}_{p+2} + \cdots + Z_n\boldsymbol{t}_n.
\end{aligned}$$

Hence,

$$F \;=\; \frac{\big((Z_{p_o+1} + \delta)^2 + Z_{p_o+2}^2 + \cdots + Z_p^2\big)/(p - p_o)}{\big((Z_{p+1} + \tilde{\delta})^2 + Z_{p+2}^2 + \cdots + Z_n^2\big)/(n - p)} \;\sim\; F_{p-p_o,n-p}(\delta^2, \tilde{\delta}^2),$$

by definition of the non-central F distributions. □

Confidence bounds for approximation errors. In many applications, the conclusion that a simplified model ($\boldsymbol{M}_o$) is not entirely correct is not very surprising. Nevertheless it is possible that the distance between $\boldsymbol{\mu}$ and $\boldsymbol{M}_o$ is irrelevant for practical purposes. Moreover, even if the F test does not reject the null hypothesis that $\boldsymbol{\mu} \in \boldsymbol{M}_o$, there is no evidence for the null hypothesis to be true. A potential remedy for both problems are confidence bounds for the standardized distance

$$\delta \;=\; \frac{\|\boldsymbol{\mu} - \boldsymbol{H}_o\boldsymbol{\mu}\|}{\sigma}.$$

Let us assume that $\boldsymbol{\mu} \in \boldsymbol{M}$. Then our test statistic F has distribution $F_{p-p_o,n-p}(\delta^2, 0)$, and its distribution function is denoted by G_δ. That means,

$$G_\delta(r) \;=\; \mathbb{P}(F \le r).$$

According to Remark 3.64, for any fixed $r > 0$, $G_\eta(r)$ is a continuous and strictly decreasing function of $\eta \ge 0$ with limit 0 as $\eta \to \infty$. Moreover, $G_\delta(F)$ is uniformly distributed on $[0, 1]$. Consequently, for $\alpha \in (0, 1)$,

$$1 - \alpha \;=\; \begin{cases} \mathbb{P}(G_\delta(F) \le 1 - \alpha), \\ \mathbb{P}(G_\delta(F) \ge \alpha), \\ \mathbb{P}\big(G_\delta(F) \in [\alpha/2, 1 - \alpha/2]\big). \end{cases}$$

Solving the inequalities for $G_\delta(F)$ on the right hand side for $\delta \ge 0$, we obtain the following confidence regions for δ:

- The lower $(1 - \alpha)$-confidence bound

$$\begin{aligned} a_\alpha(F) \;&:=\; \min\big\{\eta \ge 0 : G_\eta(F) \le 1 - \alpha\big\} \\ &\begin{cases} = \; 0 & \text{if } G_0(F) \le 1 - \alpha, \\ = \; \text{unique } \eta > 0 \text{ such that } G_\eta(F) = 1 - \alpha & \text{if } G_0(F) > 1 - \alpha, \end{cases} \end{aligned}$$

- the upper $(1 - \alpha)$-confidence bound

$$\begin{aligned} b_\alpha(F) \;&:=\; \min\big\{\eta \ge 0 : G_\eta(F) \le \alpha\big\} \\ &\begin{cases} = \; 0 & \text{if } G_0(F) \le \alpha, \\ = \; \text{unique } \eta > 0 \text{ such that } G_\eta(F) = \alpha & \text{if } G_0(F) > \alpha, \end{cases} \end{aligned}$$

- the $(1 - \alpha)$-confidence interval

$$\big[a_{\alpha/2}(F), b_{\alpha/2}(F)\big].$$

Note that $1 - G_0(F)$ is our p-value for the null hypothesis "$\boldsymbol{\mu} \in \boldsymbol{M}_o$". Thus, the lower $(1 - \alpha)$-confidence bound $a_\alpha(F)$ is strictly positive if and only if this p-value is strictly smaller than α.

Exercise 3.71. Consider a one-way ANOVA with balanced design. That is, we observe

$$Y_{js} \;=\; f_j + \varepsilon_{js}, \quad 1 \le j \le L,\ 1 \le s \le n_o,$$

with unknown parameters $f_1, f_2, \ldots, f_L$ and independent random errors $\varepsilon_{js} \sim \mathrm{N}(0, \sigma^2)$.

(a) Specify the distribution of the test statistic F for the null hypothesis

$$H_o : f_1 = f_2 = \cdots = f_L.$$

In particular, provide an explicit expression for the NCP δ^2.

(b) Sketch the distribution function $G_\delta(\cdot) = \mathbb{P}(F \leq \cdot)$ for $\delta \in \{0, 1, 2, 3\}$, $L = 5$ and $n_o = 20$.

(c) Sketch for $L = 5$, $n_o = 20$ and $\alpha = 0.05$ the power function

$$\delta \;\mapsto\; \mathbb{P}(F > F_{L-1,L(n_o-1);1-\alpha}).$$

(d) Suppose you obtain $F = 1.2$ in the previous setting. Determine an upper 95%-confidence bound for the parameter $\sigma^{-1}\sqrt{\sum_{j=1}^{L}(f_j - \bar{f})^2}$.

3.8 Calibration

For the sake of simplicity, we discuss the calibration problem only in the context of simple linear regression. That means, we consider generic observations $(X, Y) \in \mathbb{R} \times \mathbb{R}$ such that

$$Y \;=\; a + bX + \varepsilon$$

with unknown parameters a and $b \neq 0$ and a random measruement error ε. Here, we think about X being a quantity of interest which may be measured exactly with an expensive method, while Y is an indirect measurement which is relatively easy to obtain.

Data generation and analysis consists of two separate parts:

Calibration phase. In the first phase, *calibration (or training) data* $\mathcal{D}$, consisting of (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) are obtained. We assume that these n pairs are stochastically independent and satisfy

$$Y_i \;=\; a + X_i + \varepsilon_i, \quad 1 \leq i \leq n,$$

with fixed values X_i and independent errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim \mathrm{N}(0, \sigma^2)$. The calibration data are used to estimate the unknown parameters a, b and $\sigma > 0$.

Prediction phase. Later on, we are dealing with one or several pairs (X_o, Y_o), where only Y_o is observed. By means of $\mathcal{D}$ and Y_o, we would like to estimate the value of X_o or compute a confidence region for X_o, assuming that

$$Y_o \;=\; a + bX_o + \varepsilon_o \quad \text{with} \quad \varepsilon_o \;\sim\; \mathrm{N}(0, \lambda\sigma^2).$$

Here, $\lambda \in (0, 1]$ is a given scaling factor. For instance, $\lambda = 1/m$, if Y_o is the average of m independent measurements for the same X_o.

Note that ε_1, ε_2, …, ε_n and ε_o are considered as independent random variables, while X_1, X_2, …, X_n and X_o are treated as fixed numbers.

Point estimation. A natural estimator for X_o would be

$$\widehat{X}_o \;=\; \frac{Y_o - \widehat{a}}{\widehat{b}} \;=\; \bar{X} + \frac{Y_o - \bar{Y}}{\widehat{b}}.$$

It results from solving the approximate equation $Y_o \approx \widehat{a} + \widehat{b}X_o = \bar{Y} + \widehat{b}(X_o - \bar{X})$ for X_o.

Instead of the point estimator $\widehat{X}_o$ we would like to compute a confidence interval $C(\mathcal{D}, Y_o)$ for X_o. A first and commonly used confidence interval is

$$C(\mathcal{D}, Y_o) \;:=\; \Big[\frac{Y_o \,\pm\, \widehat{\sigma}\sqrt{\lambda}\Phi^{-1}(1-\alpha/2) - \widehat{a}}{\widehat{b}}\Big] \;=\; \Big[\bar{X} + \frac{Y_o - \bar{Y} \,\pm\, \widehat{\sigma}\sqrt{\lambda}\Phi^{-1}(1-\alpha/2)}{\widehat{b}}\Big].$$

It is motivated by the fact that with probability $1-\alpha$, the unknown value X_o satisfies the inequalities

$$|Y_o - a - bX_o| \;=\; |\varepsilon_o| \;\leq\; \sqrt{\lambda}\sigma\Phi^{-1}(1-\alpha/2).$$

Hence, we rely on $(\widehat{a}, \widehat{b}, \widehat{\sigma})$ being sufficiently close to the true triple (a, b, σ), so the uncertainty about X_o is mainly due to the variability of Y_o. However, if we want to take into account the uncertainty of $(\widehat{a}, \widehat{b}, \widehat{\sigma})$, there are two different points of view.

Single-use confidence regions. Let us first consider only *one* pair (X_o, Y_o). Our requirement is that

$$\mathbb{P}\big(X_o \in C(\mathcal{D}, Y_o)\big) \;\geq\; 1-\alpha$$

for arbitrary values of a, b, σ and X_o. Here, the randomness in $\mathcal{D}$ as well as in Y_o is taken into account.

The standard recipe, computing a point estimator for X_o plus or minus a certain constant times a standard error, would only work approximately. Instead we resort to the *inversion of tests*, similarly as in Example 3.23. To this end, we consider the random variable

$$\begin{aligned} Y_o - \widehat{a} - \widehat{b}X_o \;&=\; Y_o - \bar{Y} - \widehat{b}(X_o - \bar{X}) \\ &=\; \varepsilon_o - \bar{\varepsilon} - (\widehat{b} - b)(X_o - \bar{X}). \end{aligned}$$

With $Q := \sum_{i=1}^n (X_i - \bar{X})^2$, the random variables ε_o, $\bar{\varepsilon}$ and $\widehat{b} - b = \sum_{i=1}^n \varepsilon_i (X_i - \bar{X})/Q$ are stochastically independent and Gaussian with mean 0 and variances $\lambda\sigma^2$, σ^2/n and σ^2/Q, respectively. Consequently,

$$Y_o - \widehat{a} - \widehat{b}X_o \;\sim\; \mathrm{N}\Big(0,\, \sigma^2\Big(\lambda + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{Q}\Big)\Big).$$

Moreover, this random variable and $\widehat{\sigma}$ are stochastically independent. Hence for $x \in \mathbb{R}$,

$$T(x) \;:=\; \frac{Y_o - \widehat{a} - \widehat{b}x}{\widehat{\sigma}\sqrt{\lambda + 1/n + (x - \bar{X})^2/Q}} \;=\; \frac{Y_o - \bar{Y} - \widehat{b}(x - \bar{X})}{\widehat{\sigma}\sqrt{\lambda + 1/n + (x - \bar{X})^2/Q}}$$

defines a test statistic for the null hypothesis that $X_o = x$, and

$$T(X_o) \;\sim\; t_{n-1}.$$

Hence, a $(1-\alpha)$-confidence region for X_o is given by

$$C_\alpha(\mathcal{D}, Y_o) \;:=\; \big\{x \in \mathbb{R} : |T(x)| \le t_{n-2;1-\alpha/2}\big\}.$$

Elementary algebra reveals that the inequality $|T(x)| \le t_{n-2;1-\alpha/2}$ is equivalent to the quadratic inequality

$$(\widehat{b}^2 - c^2/Q)(x-\bar{X})^2 - 2\widehat{b}(Y_o-\bar{Y})(x-\bar{X}) \;\le\; \widehat{c}_\alpha^2(\lambda+1/n) - (Y_o-\bar{Y})^2, \tag{3.3}$$

where

$$\widehat{c}_\alpha \;:=\; \widehat{\sigma}\, t_{n-2;1-\alpha/2}.$$

Typically, $\widehat{b}^2 > \widehat{c}_\alpha^2/Q$, which is equivalent to the student test of "$b=0$" giving a p-value strictly smaller than α. Then, the inequality (3.3) may be solved for x, and we obtain the following $(1-\alpha)$-confidence interval for X_o:

$$C_\alpha(\mathcal{D}, Y_o) \;=\; \left[\bar{X} + \frac{\widehat{b}(Y_o-\bar{Y})}{\widehat{b}^2 - \widehat{c}_\alpha^2/Q} \pm \frac{\widehat{c}_\alpha\sqrt{(\lambda+1/n)(\widehat{b}^2 - c^2/Q) + (Y_o-\bar{Y})^2/Q}}{\widehat{b}^2 - c^2/Q}\right].$$

Multiple-use confidence intervals. Often the calibration data $\mathcal{D}$ are used for inference about *many* future pairs (X_o, Y_o). In this case it is appropriate to consider the minimal *conditional* coverage probability of a confidence region $C(\mathcal{D}, Y_o)$ for X_o, given the calibration data $\mathcal{D}$. That means, we are wondering about the value of

$$\inf_{X_o \in \mathbb{R}} \mathbb{P}\big(X_o \in C(\mathcal{D}, Y_o) \,\big|\, \mathcal{D}\big). \tag{3.4}$$

If the latter quantity is at least $1-\alpha$, one can guarantee that, in the long run, for at most $\alpha \cdot 100$ percent of future pairs (X_o, Y_o), the resulting confidence region $C(\mathcal{D}, Y_o)$ fails to cover X_o. Unfortunately, the infimum (3.4) is a random variable, and we cannot exclude calibration data $\mathcal{D}$ of low quality, resulting in poor estimators of a, b and σ.

However, it is feasible to construct confidence intervals $C(\mathcal{D}, Y_o)$ such that with given confidence $1-\beta$, the minimal conditional coverage probability (3.4) is at least $1-\alpha$. Here is an explicit recipe to construct $C(\mathcal{D}, Y_o)$:

Step 1: We start with an upper confidence bound for σ, namely,

$$\overline{\sigma} \;:=\; \widehat{\sigma}\sqrt{\frac{n-2}{\chi^2_{n-2;\gamma_1}}}$$

for some $\gamma_1 \in (0,1)$. That means,

$$\mathbb{P}\left(\sigma \le \overline{\sigma}\right) \;=\; 1-\gamma_1,$$

because the random variable $(n-2)\widehat{\sigma}^2/\sigma^2$ has distribution χ^2_{n-2}.

Step 2: Let us assume temporarily that $\sigma > 0$ was known. Then we could adapt Scheffé' s method and claim with confidence $1-\gamma_2$ that

$$\big|\widehat{f}(x) - f(x)\big| \;\le\; \sigma\sqrt{1/n + (x-\bar{X})^2/Q}\sqrt{2\chi^2_{n-2;1-\gamma_2}} \quad \text{for all } x \in \mathbb{R}.$$

Since $\widehat{f}(\cdot)$ and $\widehat{\sigma}$ are stochastically independent, we may conclude that

$$\mathbb{P}\Big(\sigma \le \overline{\sigma} \text{ and } f(x) \in \big[\widehat{\ell}(x), \widehat{u}(x)\big] \text{ for all } x \in \mathbb{R}\Big) \;\ge\; (1-\gamma_1)(1-\gamma_2),$$

where

$$\big[\widehat{\ell}(x), \widehat{u}(x)\big] \;:=\; \Big[\widehat{f}(x) \,\pm\, \overline{\sigma}\sqrt{1/n + (x-\bar{X})^2/Q}\sqrt{2\chi^2_{n-2;1-\gamma_2}}\Big].$$

For suitable parameters γ_1, γ_2, the bound $(1-\gamma_1)(1-\gamma_2)$ equals $1-\beta$, for instance, if $\gamma_1 = \gamma_2 = 1 - \sqrt{1-\beta}$.

Step 3: Note that

$$\mathbb{P}\Big(Y_o \in \big[f(X_o) \pm \lambda^{1/2}\sigma\,\Phi^{-1}(1-\alpha/2)\big]\Big) \;=\; 1-\alpha,$$

and with probability at least $1-\beta$,

$$\begin{aligned}&\big[f(x) \pm \lambda^{1/2}\sigma\,\Phi^{-1}(1-\alpha/2)\big]\\ &\quad\subset\; \big[\widehat{\ell}(x) - \lambda^{1/2}\overline{\sigma}\,\Phi^{-1}(1-\alpha/2),\; \widehat{u}(x) + \lambda^{1/2}\overline{\sigma}\,\Phi^{-1}(1-\alpha/2)\big]\end{aligned}$$

for any $x \in \mathbb{R}$. Hence the confidence region

$$C(\mathcal{D}, Y_o) \;:=\; \Big\{x \in \mathbb{R} : Y_o \in \big[\widehat{\ell}(x) - \lambda^{1/2}\overline{\sigma}\,\Phi^{-1}(1-\alpha/2), \widehat{u}(x) + \lambda^{1/2}\overline{\sigma}\,\Phi^{-1}(1-\alpha/2)\big]\Big\}$$

has the desired properties. Its explicit computation amounts to solving two quadratic inequalities, and it is a compact interval, provided that $\widehat{b}^2 > 2\overline{\sigma}^2\chi^2_{n-2;1-\gamma_2}/Q$.

3.9 Random Effects

In certain applications, it is appropriate to view some components of $\boldsymbol{\theta}$ as random variables. In the present section, we describe two particular examples of such models with *random effects*.

One-way ANOVA. Starting from $X \in \{1, 2, \ldots, L\}$, our model equation (in blackboard-and-paper notation) was:

$$Y_{js} \;=\; \mu + a_j + \varepsilon_{js}, \quad 1 \le j \le L,\; 1 \le s \le n(j),$$

where $a_+ = 0$, say.

Specifically, one could think about L different test persons who participate in a certain performance test once or several times. In the model equation above, Y_{js} would be the measure of performance of the j-th test person in its s-th trial. The parameter a_j describes the the average performance of person j in comparison with the other $L-1$ persons in this study. However, if we view the L test persons as a random sample from a larger population of individuals, one should rather consider the following model:

$$Y_{js} \;=\; \mu + A_j + \varepsilon_{js}, \quad 1 \le j \le L,\; 1 \le s \le n(j),$$

with $L + n$ stochastically independent random variables A_j and ε_{js}, where

$$A_j \sim \mathrm{N}(0, \sigma_A^2) \quad \text{and} \quad \varepsilon_{js} \sim \mathrm{N}(0, \sigma^2).$$

Now, μ stands for the average performance of an individual from the population, σ_A quantifies the person-to-person variation of performance, and σ quantifies the random fluctuations of one person's performance over time.

The null hypothesis that $\boldsymbol{a} = \mathbf{0}$, corresponds now to the null hypothesis that $\sigma_A = 0$. But often, the insight that $\sigma_A > 0$ is not very surprising. More interesting would be confidence bounds for σ_A or σ_A/σ. At least in case of a balanced design, such bounds can be computed quite easily. From now on we assume that

$$n(1) = n(2) = \cdots = n(L) = n_o.$$

In this case,

$$\bar{Y}_{j\cdot} = \mu + A_j + \bar{\varepsilon}_{j\cdot} \quad \text{and} \quad \bar{Y} = \mu + \bar{A} + \bar{\varepsilon}$$

with $\bar{A} = L^{-1} \sum_{j=1}^{L} A_j$. Here,

$$\mathrm{SS}_{\mathrm{within}} = \sum_{j,s} (Y_{js} - \bar{Y}_{j\cdot})^2 = \sum_{j,s} (\varepsilon_{js} - \bar{\varepsilon}_{j\cdot})^2$$

and

$$\mathrm{SS}_{\mathrm{between}} = \sum_{j,s} (\bar{Y}_{j\cdot} - \bar{Y})^2 = n_o \sum_{j} (V_j - \bar{V})^2$$

with

$$V_j := A_j + \bar{\varepsilon}_{j\cdot} \sim \mathrm{N}\big(0, \sigma_A^2 + \sigma^2/n_o\big).$$

The random variables $V_1, V_2, \ldots, V_L$ and $(\varepsilon_{js} - \bar{\varepsilon}_{j\cdot})_{j,s}$ are centered, uncorrelated, and their joint distribution is Gaussian. Thus they are independent, and this implies that $\mathrm{SS}_{\mathrm{between}}$ and $\mathrm{SS}_{\mathrm{within}}$ are stochastically independent, where

$$\frac{\mathrm{SS}_{\mathrm{between}}}{n_o \sigma_A^2 + \sigma^2} \sim \chi^2_{L-1} \quad \text{and} \quad \frac{\mathrm{SS}_{\mathrm{within}}}{\sigma^2} \sim \chi^2_{n-L}.$$

This implies that the standard F test statistic

$$F = \frac{\mathrm{SS}_{\mathrm{between}}/(L-1)}{\mathrm{SS}_{\mathrm{within}}/(n-L)}$$

satisfies the relation

$$\frac{F}{n_o(\sigma_A/\sigma)^2 + 1} \sim F_{L-1,n-L}.$$

This implies confidence bounds for the ratio σ_A/σ. On the one hand, with probability $1 - \alpha$,

$$F \le [n_o(\sigma_A/\sigma)^2 + 1]\, F_{L-1,n-L;1-\alpha},$$

and the latter inequality is equivalent to

$$\sigma_A/\sigma \ge \sqrt{\Big(\frac{F}{F_{L-1,n-L;1-\alpha}} - 1\Big)^+ \Big/ n_o}.$$

On the other hand, with probability $1-\alpha$,

$$F \;\geq\; [n_o(\sigma_A/\sigma)^2 + 1]\, F_{L-1,n-L;\alpha},$$

and this leads to the upper $(1-\alpha)$-confidence bound

$$\sqrt{\Big(\frac{F}{F_{L-1,n-L;\alpha}} - 1\Big)^+ \Big/ n_o}$$

for σ_A/σ.

Two-way ANOVA: Cross classification without interactions, balanced design. Starting from covariables $C \in \{1,\ldots,L\}$ and $D \in \{1,\ldots,M\}$, we considered the model equation

$$Y_{jks} \;=\; \mu + a_j + b_k + \varepsilon_{jks}, \quad 1 \leq j \leq L,\ 1 \leq k \leq M,\ 1 \leq s \leq n_o,$$

where $a_+ = b_+ = 0$.

In the specific Example 3.47, it might be appropriate to view the test persons as random sample from a population. This leads to a *mixed model* with the random effect "test person" and the fixed effect "word list".

Generally, let

$$Y_{jks} \;=\; \mu + A_j + b_k + \varepsilon_{jks}, \quad 1 \leq j \leq L,\ 1 \leq k \leq M,\ 1 \leq s \leq n_o,$$

with independent random variables $A_j \sim \mathrm{N}(0,\sigma_A^2)$, $\varepsilon_{jks} \sim \mathrm{N}(0,\sigma^2)$, and unknown parameters $\mu \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^M$, where $b_+ = 0$. Here,

$$\begin{aligned} \bar{Y}_{j\cdot\cdot} \;&=\; \mu + A_j + \bar{\varepsilon}_{j\cdot\cdot}, \\ \bar{Y}_{\cdot k\cdot} \;&=\; \mu + \bar{A} + b_k + \bar{\varepsilon}_{\cdot k\cdot}, \\ \bar{Y} \;&=\; \mu + \bar{A} + \bar{\varepsilon}_{\cdot}. \end{aligned}$$

This implies that the F test of the null hypothesis "$\boldsymbol{b} = \boldsymbol{0}$" remains the same, because under the null hypothesis, the residual array $\widehat{\boldsymbol{\varepsilon}}$ as well as the array $\widehat{\boldsymbol{Y}}_2 = \big(\bar{Y}_{\cdot k\cdot} - \bar{Y}\big)_{j,k,s}$ depend only on $\boldsymbol{\varepsilon}$.

Now, let us investigate the F test statistic for the null hypothesis "$\boldsymbol{a} = \boldsymbol{0}$" or "$\sigma_A = 0$":

$$\begin{aligned} F \;&=\; \frac{(L-1)^{-1}\sum_{j,k,s}(\bar{Y}_{j\cdot\cdot} - \bar{Y})^2}{(n-L-M+1)^{-1}\sum_{j,k,s}(Y_{jks} - \bar{Y}_{j\cdot\cdot} - \bar{Y}_{\cdot k\cdot} + \bar{Y})^2} \\ &=\; \frac{(L-1)^{-1} M n_o \sum_j (V_j - \bar{V})^2}{(n-L-M+1)^{-1}\sum_{j,k,s}(\varepsilon_{jks} - \bar{\varepsilon}_{j\cdot\cdot} - \bar{\varepsilon}_{\cdot k\cdot} + \bar{\varepsilon})^2} \end{aligned}$$

with $V_j := A_j + \bar{\varepsilon}_{j\cdot\cdot} \sim \mathrm{N}(0, \sigma_A^2 + \sigma^2/(Mn_o))$. Here one can verify that

$$\frac{F}{Mn_o(\sigma_A/\sigma)^2 + 1} \;\sim\; F_{L-1,n-L-M+1}.$$

Again, this fact leads to confidence bounds for the ratio σ_A/σ with given confidence level.

Exercise 3.72. Compute for Example 3.47 a lower 95%-confidence bound for σ_A/σ in the following mixed model:

$$Y_{jk} \;=\; A_j + g_k + \varepsilon_{jk}, \quad 1 \le j \le 24,\ 1 \le k \le 4,$$

with independent random variables $A_j \sim \mathrm{N}(0, \sigma_A^2)$, $\varepsilon_{jk} \sim \mathrm{N}(0, \sigma^2)$, and unknown parameters $\boldsymbol{g} \in \mathbb{R}^4$ and $\sigma_A \ge 0$, $\sigma > 0$.

Chapter 4

Regression Diagnostics

In the previous chapter, we derived various statistical procedures under the strong assumption that the errors ε_i are independent with distribution $\mathrm{N}(0, \sigma^2)$. An obvious question is what happens if the errors are independent and homoscedastic, but not necessarily Gaussian. As we shall see, this question is linked to another one: Are there single design points X_i which are "outliers" in the sense that the corresponding observation (X_i, Y_i) has a strong influence on the estimator $\widehat{f}$ or the fit $\widehat{\boldsymbol{Y}}$? A second important question is how to check various model assumptions graphically.

4.1 Leverage

The results of a linear model fit are to be taken with a pinch of salt, if single observations have a strong influence. We are not talking about outliers in $\boldsymbol{Y}$ but about special design matrices, i.e. special configurations of the points $X_i \in \mathcal{X}$. To identify potentially problematic observations, we consider the vector $\widehat{\boldsymbol{Y}} = \boldsymbol{D}\widehat{\boldsymbol{\theta}}$ and the residual vector $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.

In case of $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$,

$$\mathbb{E}(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^\top) \;=\; \sigma^2(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})^\top \;=\; \sigma^2(\boldsymbol{I} - \boldsymbol{H}).$$

In particular,

$$\mathbb{E}\big((Y_i - \widehat{Y}_i)^2\big) \;=\; \sigma^2(1 - H_{ii}).$$

The number H_{ii} is the so-called *leverage* of the i-th observation. It is a number between 0 and 1. The larger it is, the stronger the influence of observation (X_i, Y_i) on the fitted vector $\widehat{\boldsymbol{Y}}$. As mentioned before,

$$\sum_{i=1}^{n} H_{ii} \;=\; p.$$

Hence,

$$\max_{i=1,\ldots,n} H_{ii} \;\geq\; \frac{p}{n}.$$

Consequently, a necessary condition for the maximal leverage being small is that the number p of parameters is small compared to the number n of observations.

Example 4.1 (Simple linear regression, Example 1.2). With the sum $Q := \sum_{i=1}^n (X_i - \bar{X})^2$ of squared centered X-values, $\widehat{Y}_i$ is given by

$$\bar{Y} + \frac{\sum_{j=1}^n (X_j - \bar{X}) Y_j}{Q} (X_i - \bar{X}) \;=\; \sum_{j=1}^n H_{ij} Y_j$$

with

$$H_{ij} \;=\; \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{Q}.$$

Hence, the leverage of the i-th observation is given by

$$H_{ii} \;=\; \frac{1}{n} + \frac{(X_i - \bar{X})^2}{Q}.$$

Figure 4.1 shows the regression line for a simulated data vector $\boldsymbol{Y} \in \mathbb{R}^{20}$ and two different vectors $\boldsymbol{X} \in \mathbb{R}^{20}$. Below a bar plot of the respective leverages is displayed. To illustrate the influence of the observation X_i with the largest X-value, we also show the resulting regression lines after replacing the corresponding value Y_i with $Y_i \pm 10$.

Exercise 4.2. Deduce a general formula for the leverages in the model of one-way ANCOVA (Example 1.4).

4.2 An Application of the Central Limit Theorem

The essential message of Lindeberg's Central Limit Theorem is that the sum of stochastically independent random variables follows approximately a Gaussian distribution if each summand has only little influence on the total sum. Theorem A.14 in Section A.6 provides precise formulations of this statement.

Now, we apply this result to linear models. We consider stochastically independent random errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, assuming only that for some constant $K \geq 1$,

$$\mathrm{Std}(\varepsilon_i) \;=\; \sigma \quad \text{and} \quad \mathbb{E}(\varepsilon_i^4) \;\leq\; K\sigma^4 \quad \text{for } 1 \leq i \leq n. \tag{4.1}$$

The next theorem implies that our student tests and confidence regions are still approximately valid, if the maximal leverage gets small. Precisely, as the maximal leverage converges to 0, all standardized Gauss–Markov estimators

$$Z_{\boldsymbol{\psi}} \;=\; \frac{\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top \boldsymbol{\theta}}{\sigma_{\boldsymbol{\psi}}}$$

and the corresponding student pivotal statistics

$$T_{\boldsymbol{\psi}} \;=\; \frac{\boldsymbol{\psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\psi}^\top \boldsymbol{\theta}}{\widehat{\sigma}_{\boldsymbol{\psi}}}$$

are approximately standard Gaussian, even if the single errors are non-Gaussian.

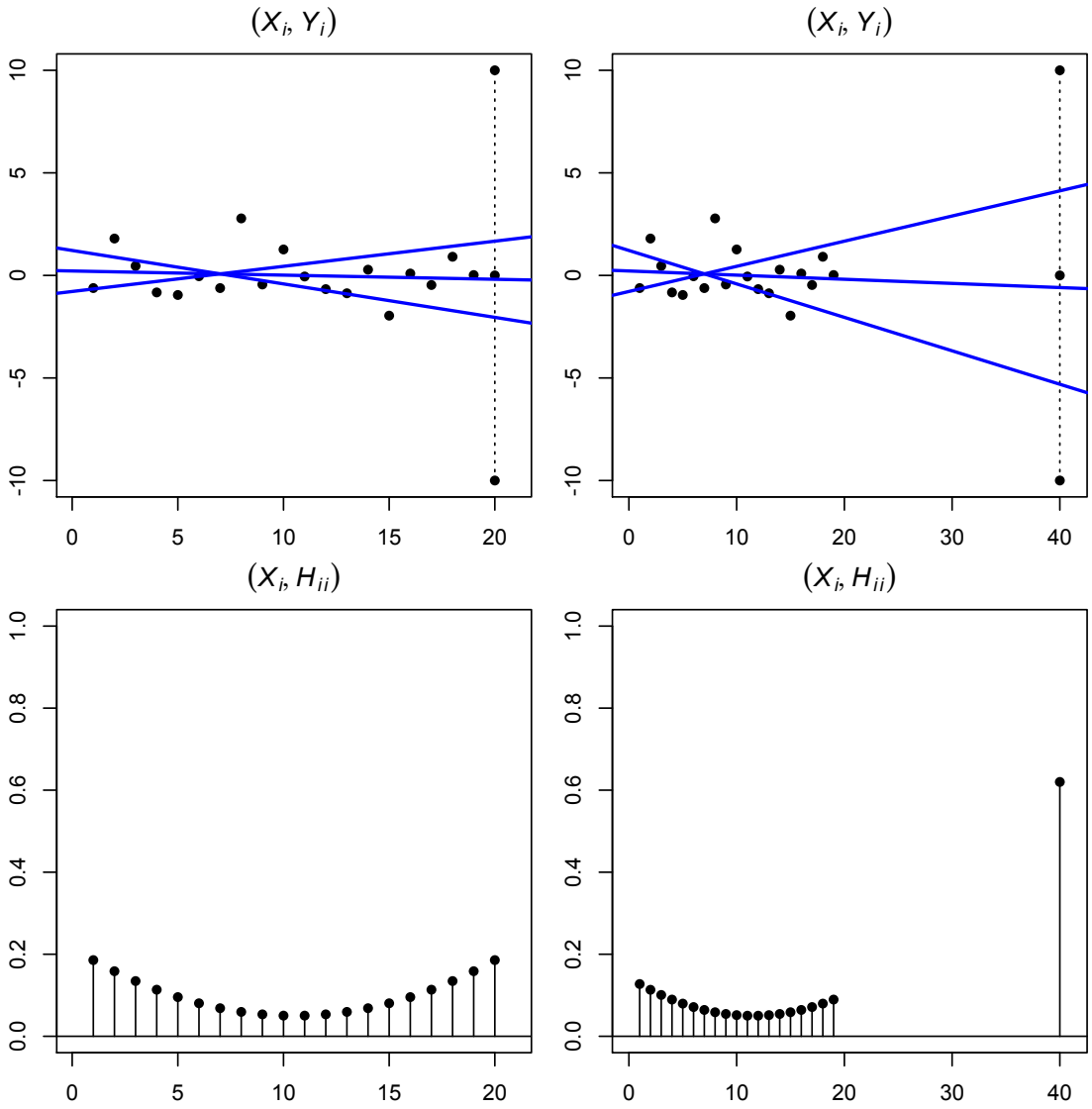

Figure 4.1: Illustrating leverage for simple linear regression.

Theorem 4.3. *Under condition* (4.1),

$$\sup_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{0\},\, r\in\mathbb{R}} \left\{ \begin{array}{l} \big|\mathbb{P}(Z_{\boldsymbol{\psi}} \le r) - \Phi(r)\big| \\ \big|\mathbb{P}(T_{\boldsymbol{\psi}} \le r) - \Phi(r)\big| \end{array} \right\} \to 0 \quad \textit{as} \quad \max_{i=1,\ldots,n} H_{ii} \to 0.$$

Proof of Theorem 4.3. For $\boldsymbol{\psi} \in \mathbb{R}^p \setminus \{0\}$ we define the unit vector

$$\boldsymbol{b} = \boldsymbol{b}(\boldsymbol{\psi}) := (\boldsymbol{\psi}^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{\psi})^{-1/2} \boldsymbol{D}\boldsymbol{\Gamma}^{-1}\boldsymbol{\psi} \in \mathbb{R}^n$$

with $\boldsymbol{\Gamma} = \boldsymbol{D}^\top \boldsymbol{D}$. Then,

$$Z_{\boldsymbol{\psi}} = \boldsymbol{b}^\top \boldsymbol{\varepsilon}/\sigma \quad \text{and} \quad T_{\boldsymbol{\psi}} = (\sigma/\widehat{\sigma}) Z_{\boldsymbol{\psi}}.$$

We first focus on the random variables $Z_{\boldsymbol{\psi}}$. With $Y_i := b_i \varepsilon_i/\sigma$, the assumptions of Theorem A.14

are satisfied, and

$$\Lambda = \Lambda(\boldsymbol{\psi}) \;\le\; \sum_{i=1}^{n} |b_i|^3\, \mathbb{E}(|\varepsilon_i|^3)/\sigma^3 \;\le\; \sum_{i=1}^{n} |b_i|^3 (K\sigma^4)^{3/4}/\sigma^3 \;\le\; \max_{1\le i\le n} |b_i|\, K^{3/4},$$

because of $\sum_i b_i^2 = 1$. Hence, it suffices to show that

$$\max_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{0\},\, 1\le i\le n} |b_i(\boldsymbol{\psi})| \;\to\; 0.$$

But with the standard basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ of $\mathbb{R}^n$, the left hand side may be written as

$$\begin{aligned}
\max_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{\boldsymbol{0}\},\, 1\le i\le n} \frac{|\boldsymbol{\psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{D}^\top \boldsymbol{e}_i|}{\sqrt{\boldsymbol{\psi}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\psi}}} &= \max_{\boldsymbol{v}\in\mathbb{R}^p\setminus\{\boldsymbol{0}\},\, 1\le i\le n} \frac{|\boldsymbol{v}^\top \boldsymbol{\Gamma}^{-1/2} \boldsymbol{D}^\top \boldsymbol{e}_i|}{\|\boldsymbol{v}\|} \\
&= \max_{1\le i\le n} \|\boldsymbol{\Gamma}^{-1/2} \boldsymbol{D}^\top \boldsymbol{e}_i\| \\
&= \max_{1\le i\le n} \sqrt{\boldsymbol{e}_i^\top \boldsymbol{D} \boldsymbol{\Gamma}^{-1} \boldsymbol{D}^\top \boldsymbol{e}_i} \\
&= \sqrt{\max_{1\le i\le n} H_{ii}}.
\end{aligned}$$

In the first step, we considered the vector $\boldsymbol{v} := \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi}$. In the second step, we deduced from the Cauchy–Schwarz inequality that $|\boldsymbol{v}^\top \boldsymbol{\Gamma}^{-1/2} \boldsymbol{D}^\top \boldsymbol{e}_i|/\|\boldsymbol{v}\|$ is not greater than $\|\boldsymbol{\Gamma}^{-1/2} \boldsymbol{D}^\top \boldsymbol{e}_i\|$, with equality in case of $\boldsymbol{v} = \boldsymbol{\Gamma}^{-1/2} \boldsymbol{D}^\top \boldsymbol{e}_i$. Consequently,

$$\Delta_Z := \sup_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{0\},\, r\in\mathbb{R}} \big|\mathbb{P}(Z_{\boldsymbol{\psi}} \le r) - \Phi(r)\big| \;\to\; 0 \quad \text{as } \max_{1\le i\le n} H_{ii} \to 0.$$

It remains to prove the same conclusion for

$$\Delta_T \;:=\; \sup_{\boldsymbol{\psi}\in\mathbb{R}^p\setminus\{0\},\, r\in\mathbb{R}} \big|\mathbb{P}(T_{\boldsymbol{\psi}} \le r) - \Phi(r)\big|.$$

But $T_{\boldsymbol{\psi}} = Z_{\boldsymbol{\psi}}/S$ with the ratio $S := \widehat{\sigma}/\sigma$, and it follows from Theorem 2.17 that

$$\begin{aligned}
\mathbb{E}((S-1)^2) \;&=\; \mathbb{E}\Big(\Big(\frac{S^2-1}{S+1}\Big)^2\Big) \\
&\le\; \mathbb{E}\big((S^2-1)^2\big) \;\le\; \Delta_S := \frac{(K-3)^+ + 2}{n-p} \;\to\; 0 \quad \text{as } \max_{1\le i\le n} H_{ii} \to 0.
\end{aligned}$$

Hence, for arbitrary $\delta \in (0, 1]$, Markov's inequality implies that

$$\mathbb{P}(|S-1| \ge \delta) \;\le\; \mathbb{E}((S-1)^2)/\delta^2 \;\le\; \Delta_S/\delta^2,$$

Moreover, for any $r \in \mathbb{R}$,

$$[T_{\boldsymbol{\psi}} \le r] \;=\; [Z_{\boldsymbol{\psi}} \le Sr] \begin{cases} \subset\; [Z_{\boldsymbol{\psi}} \le r_{2\delta}] \cup [|S-1| > \delta], \\ \supset\; [Z_{\boldsymbol{\psi}} \le r_{1\delta}] \setminus [|S-1| > \delta], \end{cases}$$

where $r_{1\delta}$ and $r_{2\delta}$ are the minimum and maximum of $\big\{(1-\delta)r, (1+\delta)r\big\}$, respectively. Consequently,

$$\begin{aligned}
&\mathbb{P}(T_{\boldsymbol{\psi}} \le r) - \Phi(r) \\
&\quad \begin{cases} \le\; \mathbb{P}(Z_{\boldsymbol{\psi}} \le r_{2\delta}) + \mathbb{P}(|S-1| > \delta) - \Phi(r) \;\le\; \Phi(r_{2\delta}) - \Phi(r) + \Delta_Z + \Delta_S/\delta^2, \\ \ge\; \mathbb{P}(Z_{\boldsymbol{\psi}} \le r_{1\delta}) - \mathbb{P}(|S-1| > \delta) - \Phi(r) \;\ge\; \Phi(r_{1\delta}) - \Phi(r) - \Delta_Z - \Delta_S/\delta^2. \end{cases}
\end{aligned}$$

These considerations show that

$$\Delta_T \;\le\; \sup_{r\in\mathbb{R},\,\xi\in\{-1,1\}} \bigl|\Phi((1+\xi\delta)r) - \Phi(r)\bigr| + \Delta_Z + \Delta_S/\delta^2.$$

If we set $\delta := \Delta_S^{1/3}$, say, the right hand side converges to 0 as $\max_{i=1,\dots,n} H_{ii} \to 0$. □

4.3 Residual Analysis

4.3.1 Q-Q-Plots for Normality

To check the plausibility of our assumption that the errors ε_i have a Gaussian distribution, one could employ P-P-plots or Q-Q-plots of the residuals. At first, we introduce these methods for samples of independent, identically distributed (i.i.d.) observations and general continuous distributions. Then, we modify the methods for our regression setting.

Plots for i.i.d. observations. Suppose that $X_1, X_2, \dots, X_n$ are stochastically independent with continuous distribution function F. Then, $F(X_1), F(X_2), \dots, F(X_n)$ are stochastically independent with uniform distribution on $[0,1]$. For the order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ this implies the following formulae:

$$\mathbb{E}\,F(X_{(k)}) \;=\; \frac{k}{n+1} \quad\text{and}\quad \mathrm{Var}\bigl(F(X_{(k)})\bigr) \;=\; \frac{\mathbb{E}\,F(X_{(k)})\bigl(1-\mathbb{E}\,F(X_{(k)})\bigr)}{n+2} \;\le\; \frac{1}{4(n+2)},$$

see Exercise 4.4. Hence we may expect that $F(X_{(k)})$ is rather close to $k/(n+1)$.

Exercise 4.4 (Uniform order statistics). Let $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ be the order statistics of independent random variables $U_1, U_2, \dots, U_n$ with distribution $\mathrm{Unif}[0,1]$.

(a) Show that $U_{(k)}$ has density function f_k on $[0,1]$, where

$$f_k(x) \;:=\; n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k}.$$

Hint: For $x \in [0,1]$, $U_{(k)} \le x$ if and only if $\sum_{i=1}^n 1_{[U_i \le x]} \ge k$, and the latter random sum follows a binomial distribution.

(b) Show that part (a) implies the formula

$$\int_0^1 x^\ell (1-x)^m \, dx \;=\; (\ell+m+1)^{-1}\binom{\ell+m}{\ell}^{-1}$$

for arbitrary integers $\ell, m \ge 0$. Then, show that

$$\mathbb{E}(U_{(k)}) \;=\; \frac{k}{n+1} \quad\text{and}\quad \mathrm{Var}(U_{(k)}) \;=\; \frac{\mathbb{E}(U_{(k)})\bigl(1-\mathbb{E}(U_{(k)})\bigr)}{n+2} \;\le\; \frac{1}{4(n+2)}.$$

P-P-Plots. To check the assumption that the X_i follow the distribution (function) F, one can generate a scatter plot of the pairs

$$\Bigl(\frac{k}{n+1}, F(X_{(k)})\Bigr) \;\in\; [0,1]\times[0,1]$$

and check whether they are close to the line $y = x$.

Q-Q-Plots. Alternatively, one may generate a scatter plot of the pairs

$$\Big(F^{-1}\Big(\frac{k}{n+1}\Big), X_{(k)}\Big) \;\in\; \mathbb{R}\times\mathbb{R}$$

and check whether they are close to the line $y = x$.

Location and scale families. Suppose that

$$\varepsilon_i \;\sim\; F_o\Big(\frac{\cdot-\mu}{\sigma}\Big)$$

for a given continuous distribution function F_o and unknown parameters $\mu \in \mathbb{R}$, $\sigma > 0$. With suitable estimators $\widehat{\mu}$ and $\widehat{\sigma}$, a modified P-P-plot would depict the pairs

$$\Big(\frac{k}{n+1}, F_o\Big(\frac{X_{(k)}-\widehat{\mu}}{\widehat{\sigma}}\Big)\Big).$$

For a modified Q-Q-plot there are two possible variants: One could generate a scatter plot of the pairs

$$\Big(F_o^{-1}\Big(\frac{k}{n+1}\Big), X_{(k)}\Big)$$

and check whether they are close to *some* straight line. Or one generates a scatter plot of the pairs

$$\Big(F_o^{-1}\Big(\frac{k}{n+1}\Big), \frac{X_{(k)}-\widehat{\mu}}{\widehat{\sigma}}\Big)$$

and checks whether they are close to the line $y = x$.

Exercise 4.5 (Q-Q-Plots for t distributions). **(a)** Write a program which generates a Q-Q-plot for an arbitrary data vector $\boldsymbol{X}$ and the distribution function $F_o = F_\nu$ of t_ν for any given value $\nu > 0$. Use the sample median and a suitably standardized interquartile range as a location and scale estimator.

(b) Find a data set with daily stock prices for some company or the daily values of a stock index over a longer period of time. Convert these values $K_1, K_2, K_3, \ldots$ into *log-returns*

$$X_t \;:=\; \log_{10}(K_{t+1}/K_t).$$

Now use our program from part (a) to check the assumption that the log-returns follows a t distribution up to an affine transformation.

Additional question: Is it a plausible assumption that the log-returns are stochastically independent?

Linear models. When fitting a linear model, one sorts the residuals $\widehat{\varepsilon}_i$ in ascending order and obtains $\widehat{\varepsilon}_{(1)} \le \widehat{\varepsilon}_{(2)} \le \cdots \le \widehat{\varepsilon}_{(n)}$. Then one generates a scatter plot of the pairs

$$\Big(\Phi^{-1}\Big(\frac{i}{n+1}\Big), \widehat{\varepsilon}_{(i)}\Big)$$

or of the pairs

$$\Big(\Phi^{-1}\Big(\frac{i}{n+1}\Big), \frac{\widehat{\varepsilon}_{(i)}}{\widehat{\sigma}}\Big). \tag{4.2}$$

Under the assumption that the errors are homoscedastic with centered Gaussian distribution, these points should be close to the straight line through $(0,0)$ with slope σ or 1, respectively, at least if the ratio p/n is small. This is explained in Exercise 4.6.

Obviously, the points being close to a straight line is rather vague. To get a feeling for the typical appearance of a Q-Q-plot in case of homoscedastic Gaussian errors, one should compare the Q-Q-plot for the original data with a scatter plot of the points

$$\Big(\Phi^{-1}\Big(\frac{i}{n+1}\Big), \widehat{Z}_{(i)}\Big)$$

or

$$\Big(\Phi^{-1}\Big(\frac{i}{n+1}\Big), \frac{\widehat{Z}_{(i)}}{\|\widehat{\boldsymbol{Z}}\|/\sqrt{n-p}}\Big), \tag{4.3}$$

where $\boldsymbol{Z} \in \mathbb{R}^n$ is a simulated random vector with distribution $\mathrm{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, $\widehat{\boldsymbol{Z}} := (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$, and $\widehat{Z}_{(1)} \le \widehat{Z}_{(2)} \le \cdots \le \widehat{Z}_{(n)}$ are the ordered components of $\widehat{\boldsymbol{Z}}$. Under the null hypothesis that $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, the scatter plot of the pairs (4.2) and the scatter plot of the pairs (4.3) have the same distribution. Precisely, the two random vectors

$$\Big(\frac{\widehat{\varepsilon}_{(i)}}{\widehat{\sigma}}\Big)_{i=1}^n \quad \text{and} \quad \Big(\frac{\widehat{Z}_{(i)}}{\|\widehat{\boldsymbol{Z}}\|/\sqrt{n-p}}\Big)_{i=1}^n$$

have the same distribution.

Exercise 4.6 (Estimation of the error distribution). Suppose that the errors ε_i are stochastically independent with distribution Q, where $\int x\,Q(dx) = 0$ and $\sigma^2 = \int x^2\,Q(dx) < \infty$. The empirical distribution $\widehat{Q}$ of the residuals $\widehat{\varepsilon}_i$ is given by

$$\widehat{Q}(B) := \frac{1}{n}\sum_{i=1}^n 1_{[\widehat{\varepsilon}_i \in B]} \quad \text{for } B \subset \mathbb{R}.$$

Show that $\widehat{Q}$ is a consistent estimator of Q in the following sense: For any bounded, Lipschitz-continuous function $h : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\Big|\int h(x)\,\widehat{Q}(dx) - \int h(x)\,Q(dx)\Big| \to 0 \quad \text{as } p/n \to 0.$$

(Hint: Consider first the empirical distribution $\check{Q}$ of the errors ε_i.)

Exercise 4.7 (Approximation by Lipschitz-continuous functions). Let $(\mathcal{X}, d)$ be a metric space, and let $h : \mathcal{X} \to \mathbb{R}$ be bounded. For $L > 0$ define

$$\begin{aligned} h_{L,1}(x) &:= \inf_{y\in\mathcal{X}} \big(h(y) + Ld(x,y)\big), \\ h_{L,2}(x) &:= \sup_{y\in\mathcal{X}} \big(h(y) - Ld(x,y)\big). \end{aligned}$$

Prove the following claims:

(a) $h_{L,1}$ and $h_{L,2}$ are Lipschitz-continuous with constant L, and

$$\inf_{y\in\mathcal{X}} h(y) \le h_{L,1} \le h \le h_{L,2} \le \sup_{y\in\mathcal{X}} h(y).$$

(b) For any fixed $x \in \mathcal{X}$, $h_{L,1}$ and $h_{L,2}$ are non-decreasing and non-increasing in L, respectively. Moreover, if h is continuous at x, then

$$\lim_{L\to\infty} h_{L,j}(x) \;=\; h(x) \quad \text{for } j = 1, 2.$$

(By means of this exercise, one can replace the assumption of Lipschitz-continuity in Exercise 4.6 with continuity.)

4.3.2 Plots of Residuals versus Functions of the Covariates or the Fit

Besides the assumption of normality, one should also check the linear model itself and the assumption of homoscedastic errors. To this end, one searches for certain features of the residual vector $\widehat{\boldsymbol{\varepsilon}}$ which indicate a violation of those assumptions.

One possibility is the graphical display of the pairs

$$(V_i, \widehat{\varepsilon}_i),$$

where $\boldsymbol{V} = (V_i)_{i=1}^n \in \mathbb{R}^n$ is a vector which may depend on $\boldsymbol{X}$ or $\widehat{\boldsymbol{Y}}$. Often one chooses $V_i = \widehat{Y}_i$ or the values of a certain numerical covariate. When looking at such a scatter plot, one should check it for two types of trends:

- Trends in the local mean: Suppose that the residuals tend to be negative in certain regions determined by $\boldsymbol{V}$ and positive in other regions. This indicates that our linear model is incorrect, i.e. $\mathbb{E}(\boldsymbol{Y}) \notin \boldsymbol{M}$.

- Trends in the local variability: Suppose that the local means of the residuals are approximately 0, but their moduli show some association with components of $\boldsymbol{V}$. This indicates that the errors are heteroscedastic.

Judging such a scatter plot may be difficult if the distribution of the components of $\boldsymbol{V}$ is very non-uniform. This can often be ameliorated by replacing $\boldsymbol{V}$ with the vector of its *ranks*. If numerous components of $\boldsymbol{V}$ are tied, one should not use the usual averaged ranks but ranks in $\{1, 2, \ldots, n\}$ with random allocation amoung observations with equal values of V_i.

Example 4.8 (Baseball). We illustrate these methods with a data set about $n = 263$ professional baseball players. The response Y is the annual income (in $1'000$ USD), and we consider just one numerical covariate X, the number of seasons (including the current) the player was on a professional team. Since only a few values X_i are larger than 20, we replace X with $\min(X, 20)$.

At first, we assume the model of simple linear regression, i.e. $Y = a + bX + \varepsilon$. The corresponding LSEs are $\widehat{a} = 256.39$ and $\widehat{b} = 38.322$ with estimated standard deviation $\widehat{\sigma} = 413.634$ and $R^2_{\text{adj}} = 0.159$.

Figure 4.2 shows the data and the regression line. This plot indicates already that the model assumptions are implausible. Two residual plots are shown in Figure 4.3. The upper left panel depicts a Q-Q-plot for normality of the standardized residuals, and the upper right panel shows a scatter plot of the pairs $(X_i, \widehat{\varepsilon}_i)$. The Q-Q-plot exhibits a strong deviation from normality towards

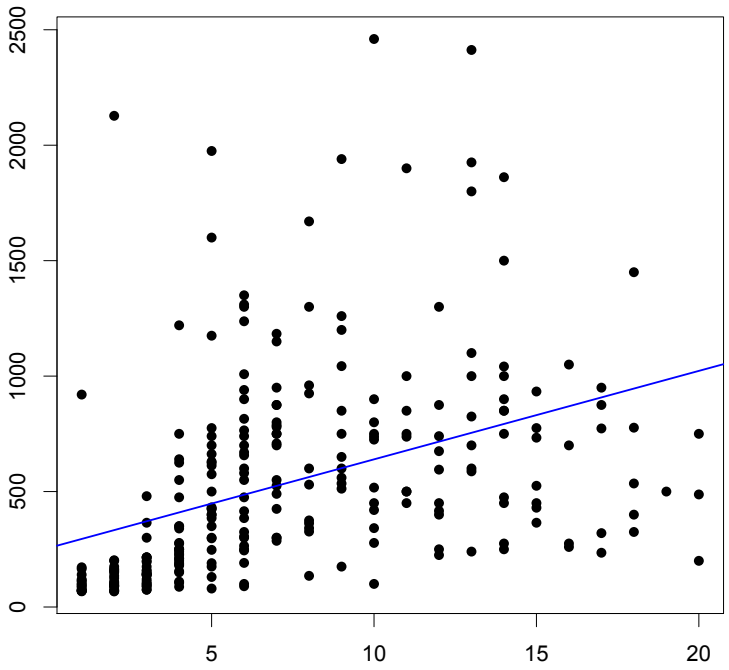

Figure 4.2: Linear fit for raw data in Example 4.8.

a right-skewed distribution. This becomes even more obvious when comparing it with two Q-Q-plots from simulated Gaussian vectors as in (4.3).

Since the Q-Q-plot for normality indicates a right-skewed distribution, we try to replace the raw response with its log-transform $\tilde{Y} := \log_{10}(Y)$, see also the next section. Figure 4.4 shows the transformed data and the resulting regression line. Now $\widehat{a} = 2.253$, $\widehat{b} = 0.0439$, $\widehat{\sigma} = 0.326$ and $R^2_{\text{adj}} = 0.292$. The new residual plots are shown in Figure 4.5. Now, the Q-Q-plot for normality looks very nice. But the residuals show a clear trend, tending to be negative if X is relatively small or large and positive if X is close to its center. This indicates that the simple linear regression model is too simplistic.

The scatter plot of the pairs $(X_i, \tilde{Y}_i)$ indicates that the strongest changes occur for smaller values of X. Hence, we also tried a log-transformed covariate $\tilde{X} := \log_{10}(X)$. Despite this double transformation, one sees clear trends in the resulting residual plots for linear or quadratic regression. This is also supported by F tests with cubic or higher order polynomial regression. On the other hand, the latter analyses indicate that cubic regression could be appropriate. Hence, we tried the model

$$\tilde{Y} \;=\; f(\tilde{X}) + \varepsilon \quad \text{with} \quad f(x) \;=\; \sum_{j=0}^{3} \tilde{a}_0 x^j.$$

The upper left panel of Figure 4.6 shows a scatter plot of the observations $(\tilde{X}_i, \tilde{Y}_i)$, together with the estimated regression function $\widehat{f}$ and pointwise as well as simultaneous 95%-confidence intervals for $f(x)$. In the upper right panel one sees the resulting Q-Q-plot (4.2) for normality. This plot looks okay, except for two large outliers. The lower left panel shows the pairs $(X_i, \widehat{\varepsilon}_i)$. It is difficult to assess this plot, because of the non-uniform distribution of the $\tilde{X}_i$ with many ties. Hence the lower right panel shows a scatter plot of the pairs $(R_i, \widehat{\varepsilon}_i)$, where $(R_i)_{i=1}^n$ is a

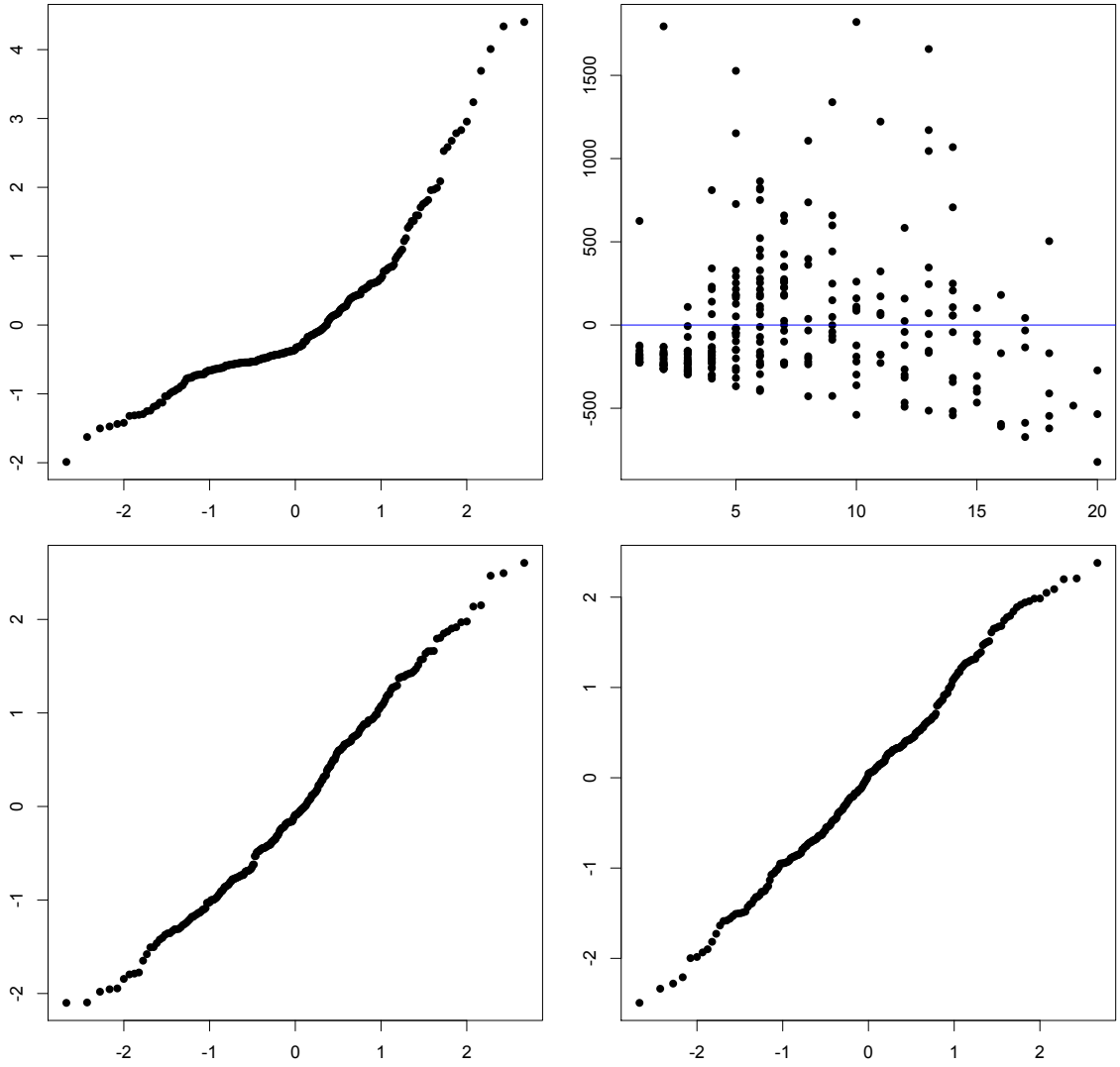

Figure 4.3: Residual plots for original data (upper panels) and simulated data (lower panels) in Example 4.8.

rank vector of $\boldsymbol{X}$ with random allocation of ranks within tied observations. Now one sees that the model assumptions are rather plausible; only for very small values of $\tilde{X}$ one sees a slightly smaller variability and two large outliers.

Exercise 4.9. Perform a residual analysis for the data set "Goats.txt". Discuss your results.

Exercise 4.10 (Tukey's F test). The scatter plots of pairs $\big(V_i(\widehat{\boldsymbol{Y}}), \widehat{\varepsilon}_i\big)$ correspond to a variant of F tests which has been proposed in some special settings by J. Tukey: Suppose we observe $\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$, and let $\boldsymbol{M} \subset \mathbb{R}^n$ be a given p-dimensional model space. Now we would like to test the null hypothesis that $\boldsymbol{\mu} \in \boldsymbol{M}$. To this end, one could embed $\boldsymbol{M}$ into a larger linear subspace $\boldsymbol{M}_*$ with $p < p_* = \dim(\boldsymbol{M}_*) < n$ and perform an F test of "$\boldsymbol{\mu} \in \boldsymbol{M}$" versus "$\boldsymbol{\mu} \in \boldsymbol{M}_* \setminus \boldsymbol{M}$". Interestingly, we do not have to specify the space $\boldsymbol{M}_*$ beforehand, but we may choose $\boldsymbol{M}_*$ by means of $\widehat{\boldsymbol{Y}}$!

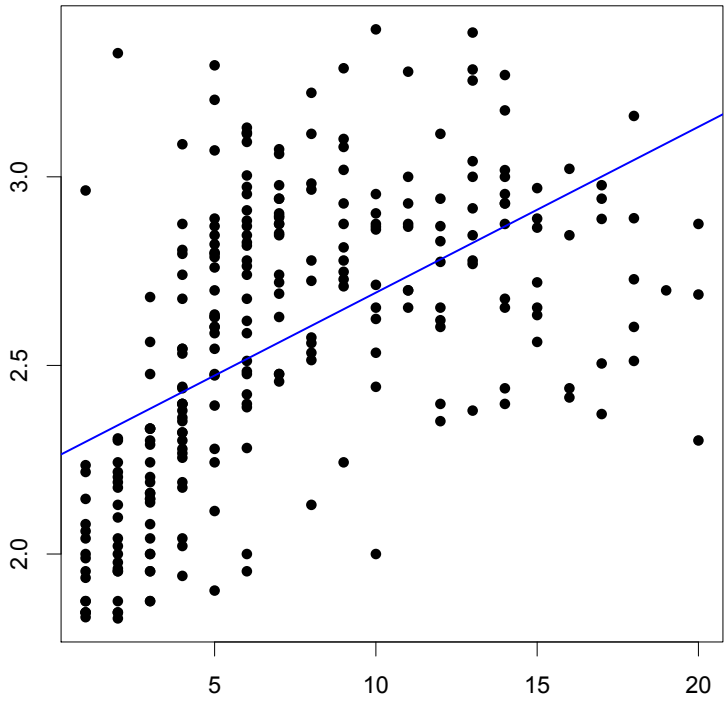

Figure 4.4: Linear fit for log-transformed responses in Example 4.8.

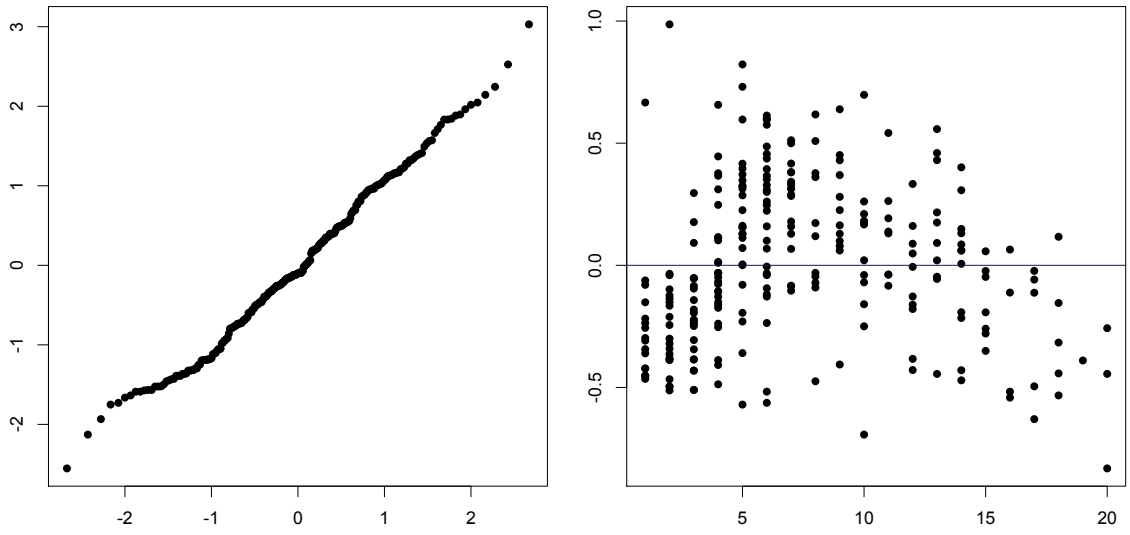

Figure 4.5: Residual plots for log-transformed responses in Example 4.8.

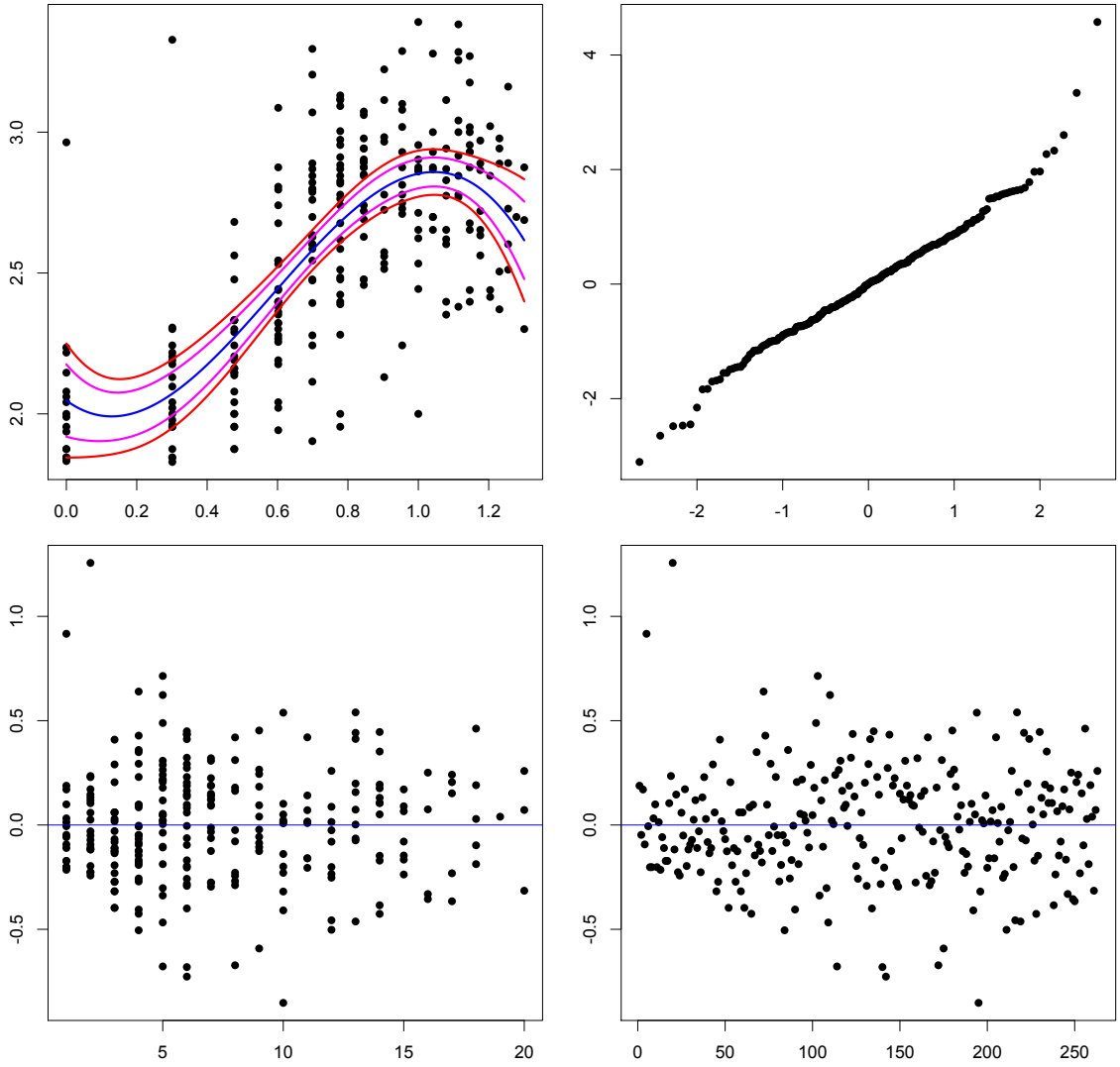

Figure 4.6: Cubic fit and residual plots for doubly log-transformed data in Example 4.8.

For $\boldsymbol{x} \in \boldsymbol{M}$ and $p < j \leq p_*$ let $\boldsymbol{b}_j(\boldsymbol{x}) \in \boldsymbol{M}^\perp$ such that

$$\boldsymbol{b}_j(\boldsymbol{x})^\top \boldsymbol{b}_k(\boldsymbol{x}) \;=\; 1_{[j=k]} \quad \text{for } p < j, k \leq p_*.$$

Show that

$$\frac{\sum_{j=p+1}^{p_*} (\boldsymbol{b}_j(\widehat{\boldsymbol{Y}})^\top \widehat{\boldsymbol{\varepsilon}})^2/(p_* - p)}{\big(\|\widehat{\boldsymbol{\varepsilon}}\|^2 - \sum_{j=p+1}^{p_*} (\boldsymbol{b}_j(\widehat{\boldsymbol{Y}})^\top \widehat{\boldsymbol{\varepsilon}})^2\big)/(n - p_*)} \;\sim\; F_{p_*-p, n-p_*}$$

in case of $\boldsymbol{\mu} \in \boldsymbol{M}$.

Exercise 4.11 (Tukey's F test for non-additivity). Consider independent observations

$$Y_{jk} \;\sim\; \mathrm{N}(\mu_{jk}, \sigma^2), \quad 1 \leq j \leq L, 1 \leq k \leq M.$$

We assume that

$$\mu_{jk} \;=\; \mu + a_j + b_k$$

with unknown parameters $\mu \in \mathbb{R}$, $\boldsymbol{a} \in \mathbb{R}^L$ and $\boldsymbol{b} \in \mathbb{R}^M$, where $a_+ = 0 = b_+$. The corresponding Gauss–Markov estimators are given by

$$\widehat{\mu} := \overline{Y}, \quad \widehat{a}_j := \overline{Y}_{j\cdot} - \overline{Y} \quad \text{and} \quad \widehat{b}_k := \overline{Y}_{\cdot k} - \overline{Y}.$$

Now, we would like to test whether $(\mu_{jk})_{j,k}$ has indeed the additive structure above.

(a) Show by means of Exercise 4.10, that under the null hypothesis of an additive structure,

$$\frac{W^2}{\big(\sum_{j,k} \widehat{\varepsilon}_{jk}^2 - W^2\big)/(LM - L - M)} \sim F_{1,LM-L-M},$$

where $\widehat{\varepsilon}_{jk} := Y_{jk} - \overline{Y}_{j\cdot} - \overline{Y}_{\cdot k} + \overline{Y}$, and

$$W := \sum_{j,k} \widehat{a}_j \widehat{b}_k \widehat{\varepsilon}_{jk} \Big/ \sqrt{\sum_j \widehat{a}_j^2 \sum_k \widehat{b}_k^2}.$$

(b) Apply this test to the data set "Hearing.txt".

Exercise 4.12. For some $d \in \mathbb{N}$, let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable in an open neighborhood of $\mathbf{0}$. With the standard basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$ of $\mathbb{R}^d$ we define

$$h_j(x) := f(x\boldsymbol{e}_j) - f(\mathbf{0}) \quad \text{for } 1 \le j \le d \text{ and } x \in \mathbb{R}.$$

Suppose that $h_j'(0) \neq 0$ for $1 \le j \le d$. Show that

$$f(\boldsymbol{x}) = f(\mathbf{0}) + \sum_{1 \le j \le d} h_j(x_j) + \sum_{1 \le j < k \le d} c_{jk} h_j(x_j) h_k(x_k) + o(\|\boldsymbol{x}\|^2) \quad \text{as } \boldsymbol{x} \to \mathbf{0}$$

with suitable constants $c_{jk} \in \mathbb{R}$.

4.4 Transformations

If the residual plots indicate heteroscedastic errors, an obvious question is how to proceed. Often it is possible to enforce homoscedasticity by means of a simple non-linear transformation of the raw response. In many applications with non-negative response Y, the standard deviation of Y_i seems to be proportional to $(\mathbb{E}\, Y_i)^\gamma$ for some $\gamma \in (0, 1]$. In case of Poisson distributed variables, this is true with $\gamma = 1/2$, see Exercise 4.13. This suggests to replace Y_i with $T_\gamma(Y_i)$, where

$$T_\gamma(y) := \begin{cases} \dfrac{y^{1-\gamma} - 1}{1 - \gamma} & \text{if } 0 < \gamma < 1, \\ \log(y) & \text{if } \gamma = 1. \end{cases}$$

Indeed, suppose that Y may be written as $Y = \mu + \mu^\gamma Z$ with a real constant $\mu > 0$ and a random variable Z such that $\mathbb{E}(Z) = 0$ and $\mu^\gamma \operatorname{Std}(Z) << \mu$. In case of $0 < \gamma < 1$,

$$\begin{aligned} T_\gamma(Y) &= \frac{\mu^{1-\gamma}\big(1 + \mu^{\gamma-1} Z\big)^{1-\gamma} - 1}{1 - \gamma} \\ &\approx \frac{\mu^{1-\gamma}\big(1 + (1-\gamma)\mu^{\gamma-1} Z\big) - 1}{1 - \gamma} \\ &= T_\gamma(\mu) + Z. \end{aligned}$$

Here, we utlilized the Taylor expansion $(1+x)^{1-\gamma} = 1 + (1-\gamma)x + O(x^2)$ as $x \to 0$. In case of $\gamma = 1$, the Taylor expansion $\log(1+x) = x + O(x^2)$ as $x \to 0$ implies that

$$T_1(Y) \;=\; \log(\mu) + \log(1+Z) \;\approx\; \log(\mu) + Z \;=\; T_1(\mu) + Z.$$

Analysis of variance or regression analysis with Poisson distributed response are quite frequent in biology or medicine, for instance, when counting cells under the microscope. Another example is the analysis of low-dose X-ray images. Here one often takes the transform

$$\tilde{T}(y) \;:=\; 2\sqrt{1+y}$$

in place of $T_{1/2}(y) = 2\sqrt{y}$, which improves the Gaussian approximation.

Exercise 4.13. Let Y be a random variable with distribution $\mathrm{Poiss}(\lambda)$.

(a) Show by means of the CLT and Slutsky's lemma or with characteristic functions that the standardized random variable $(Y-\lambda)/\sqrt{\lambda}$ converges in distribution to $\mathrm{N}(0,1)$.

(b) Show by means of part (a) and Slutsky's lemma that for any fixed $a \geq 0$, the distribution of $\sqrt{a+Y} - \sqrt{a+\lambda}$ converges weakly to a Gaussian distribution with mean 0 and standard deviation $1/2$ as $\lambda \to \infty$.

Chapter 5

Nonparametric Regression

In this chapter, we consider the special case of a numerical covariate X. We have already seen the models of simple linear and polynomial regression. But with increasing order of the polynomials, the latter method becomes numerically and statistically unstable. For the general situation that

$$Y \;=\; f(X) + \varepsilon$$

with a sufficiently *smooth*, but unknown function $f : \mathbb{R} \to \mathbb{R}$, there are numerous alternative approaches under the general name *nonparametric regression*. We shall touch on three of them. There is a rich literature about nonparametric regression. For the second method in this chapter, local polynomials, we refer to the monograph of Fan and Gijbels (1996) and the literature cited therein.

5.1 Spline Regression

5.1.1 Definition of Splines

A function $f : [a, b] \to \mathbb{R}$ is called *spline of order* d with *knots* $a = t_0 < t_1 < \cdots < t_m = b$ if it satisfies the following conditions:

(a) On each interval $[t_{k-1}, t_k]$, f is a polynomial of order d.

(b) f is $d - 1$ times continuously differentiable.

(In case of $d = 1$, the latter requirement means that f is continuous.) In particular, we talk about

linear splines, if $d = 1$,

quadratic splines, if $d = 2$,

cubic splines, if $d = 3$.

In what follows we write

$$\mathcal{S}_d(t_0, t_1, \ldots, t_m) \;:=\; \big\{\text{splines of order } d \text{ with knots } t_0, t_1, \ldots, t_m\big\}.$$

Remark. Some authors would talk about splines of order $d+1$ here, because on each interval $[t_{k-1}, t_k]$ the polynomial function f is given by $d+1$ parameters.

5.1.2 Polynomial Representation and a First Basis

One can easily verify that the set $\mathcal{S}_d(t_0, t_1, \ldots, t_m)$ is a linear space of functions on $[a, b]$. But what is the dimension, and what would be possible basis functions?

By assumption, for $k = 1, \ldots, m$, there exist coefficients $b_{k,0}, b_{k,1}, \ldots, b_{k,d}$ such that

$$f(x) \;=\; P_k(x) := \sum_{j=0}^{d} b_{k,j}(x - t_{k-1})^j \quad \text{for } x \in [t_{k-1}, t_k].$$

With $\Delta_k := t_k - t_{k-1}$, one could also write

$$\begin{aligned} P_k(x) \;&=\; \sum_{j=0}^{d} b_{k,j}(x - t_k + \Delta_k)^j \\ &=\; \sum_{j=0}^{d} b_{k,j} \sum_{i=0}^{j} \binom{j}{i} \Delta_k^{j-i}(x - t_k)^i \\ &=\; \sum_{i=0}^{d} \Big(\sum_{j=i}^{d} \binom{j}{i} \Delta_k^{j-i} b_{k,j} \Big)(x - t_k)^i. \end{aligned}$$

Hence,

$$\begin{aligned} b_{k+1,i} \;&=\; \frac{P_{k+1}^{(i)}(t_k)}{i!} \;=\; \frac{f^{(i)}(t_k+)}{i!} \;=\; \frac{f^{(i)}(t_k-)}{i!} \;=\; \frac{P_k^{(i)}(t_k)}{i!} \\ &=\; \sum_{j=i}^{d} \binom{j}{i} \Delta_k^{j-i} b_{k,j} \quad \text{for } 1 \le k < m, 0 \le i < d. \end{aligned} \tag{5.1}$$

Consequently, if we specify the polynomials $P_1, P_2, \ldots, P_k$ one after another, we have $d+1$ free parameters $b_{1,0}, \ldots, b_{1,d}$ for P_1. Having specified $P_1, \ldots, P_k$ for some $k < m$, the coefficients $b_{k+1,0}, \ldots, b_{k+1,d-1}$ are given by (5.1), and only the coefficient $b_{k+1,d}$ of P_{k+1} can be chosen arbitrarily. Precisely, $b_{k+1,d} - b_{k,d}$ specifies the change of the d-th derivative of f at the point t_k, divided by $d!$,

$$b_{k+1,d} - b_{k,d} \;=\; \frac{f^{(d)}(t_k+) - f^{(d)}(t_k-)}{d!}.$$

These considerations show that

$$\dim\left(\mathcal{S}_d(t_0, t_1, \ldots, t_m)\right) \;=\; (d+1) + (m-1) \;=\; d + m.$$

They also suggest a first basis for $\mathcal{S}_d(t_0, t_1, \ldots, t_m)$:

$$\begin{aligned} f_i(x) \;&:=\; (x - t_0)^{i-1} \quad \text{for } i = 1, \ldots, d+1, \\ f_{d+1+k}(x) \;&:=\; (x - t_k)_+^d \quad \text{for } k = 1, \ldots, m-1, \end{aligned}$$

where $y_+ := \max(y, 0)$ denotes the positive part of a real number y. Writing an arbitrary function $f \in \mathcal{S}_d(t_0, t_1, \ldots, t_m)$ as a linear combination

$$f \;=\; \sum_{i=1}^{d+m} \theta_i f_i$$

of these basis functions, the connection between the coefficients θ_i and $b_{k,j}$ is:

$$\begin{aligned} \theta_i \;&=\; b_{1,i-1} \quad \text{for } i = 1, \ldots, d+1, \\ \theta_{d+1+k} \;&=\; b_{k+1,d} - b_{k,d} \quad \text{for } k = 1, \ldots, m-1. \end{aligned}$$

That is, for $1 \le k < m$, the value $d!\,\theta_{d+1+k}$ is the change of the d-th derivative of f at the point t_k.

In the special case of $m + 1 = 4$ knots, one could also choose any basis $f_1, f_2, \ldots, f_{d+1}$ for the polynomials of order d, and then augment these by the two functions

$$f_{d+2}(x) \;:=\; (t_1 - x)_+^d \quad \text{and} \quad f_{d+3}(x) \;:=\; (x - t_2)_+^d.$$

The advantage is that the latter two functions have disjoint support.

In the special case of $m + 1 = 5$ knots, one could use, for instance, the following basis:

$$\begin{aligned} f_i(x) \;&:=\; (x - t_2)^{i-1} \quad \text{for } 1 \le i \le d, \\ f_{d+1}(x) \;&:=\; (t_1 - x)_+^d, \\ f_{d+2}(x) \;&:=\; (t_2 - x)_+^d, \\ f_{d+3}(x) \;&:=\; (x - t_2)_+^d, \\ f_{d+4}(x) \;&:=\; (x - t_3)_+^d. \end{aligned}$$

5.1.3 B Splines

A disadvantage of the basis functions defined in Section 5.1.2 is that the resulting design matrix is often ill-conditioned. Indeed, two columns $\big((X_i - t_{k-1})_+^d\big)_{i=1}^n$ can be nearly collinear. Hence, we would like to construct basis functions $f_1, f_2, \ldots, f_{m+d}$ such that the angles between the vectors $f_j(\boldsymbol{X})$ are sufficiently large.

Precisely, we want to specify nonnegative basis functions $B_1, \ldots, B_{d+m}$ such that

$$\{x \in [a, b] : B_j(x) > 0\} \;=\; (t_{j-1-d}, t_j) \cap [a, b].$$

Here, we specify arbitrary additional knots $t_{-d} \le t_{1-d} \le \cdots < t_0 = a$ and $b = t_m \le t_{m+1} \le \cdots \le t_{m+d}$. For the general theory, we refer to de Boor (2002) or Schumaker (1981); see also Section A.4 in the appendix for more details.

Special case 1: Linear Splines. A function $f \in \mathcal{S}_1(t_0, t_1, \ldots, t_m)$ is uniquely defined by its values at the $m + 1$ knots. In particular, let $B_{i,1} \in \mathcal{S}_1(t_0, t_1, \ldots, t_m)$ be such that

$$B_{i,1}(t_j) \;:=\; \begin{cases} 1 & \text{if } j = i - 1, \\ 0 & \text{if } j \ne i - 1. \end{cases}$$

Figure 5.1 shows these basis functions in case of $m = 5$ and $(t_0, t_1, \ldots, t_m) = (0, 1, 2, 4, 5, 6)$. The basis function B_3 is emphasized.

For this particular basis,

$$\theta_i \;=\; f(t_{i-1})$$

for any $f \in \mathcal{S}_1(t_0, t_1, \ldots, t_m)$ and $i = 1, \ldots, m+1$.

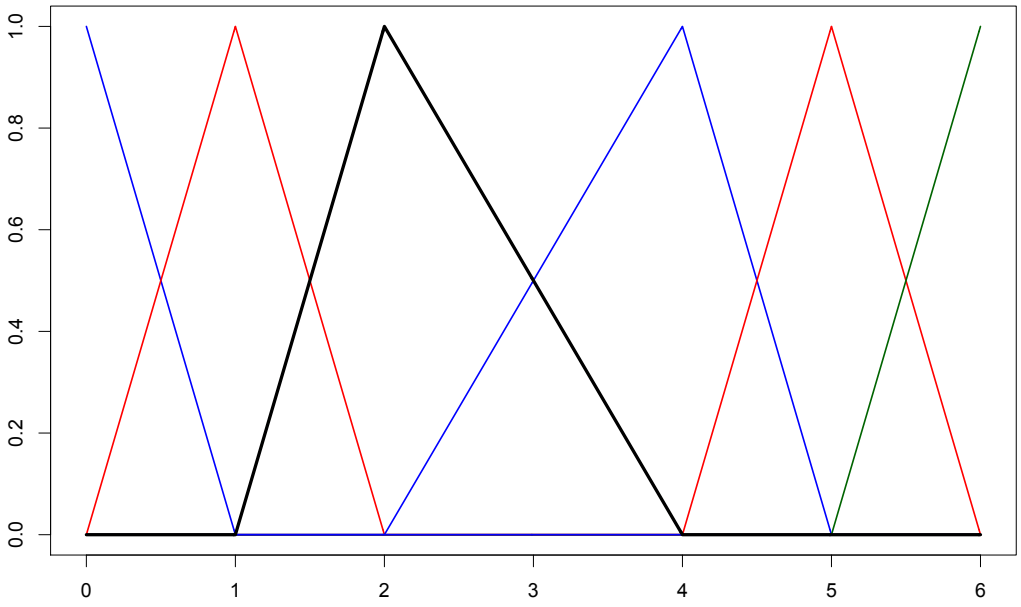

Figure 5.1: Basis functions $B_{j,1}$ for $\mathcal{S}_1(0, 1, 2, 4, 5, 6)$.

B splines of arbitrary order. The basis functions for linear splines appear within a general recursive construction. For a given integer $d_o \geq 1$, we start with the functions

$$B_{z,0}(x) \;:=\; 1_{[t_{z-1} \leq x < t_z]}, \quad 1 - d_o \leq z \leq m + d_o.$$

Then, for $d = 1, 2, \ldots, d_o$, we define the auxiliary quantities $\Delta_{z,d} := t_z - t_{z-d}$ and the functions

$$B_{z,d}(x) \;=\; \frac{x - t_{z-1-d}}{\Delta_{z-1,d}} B_{z-1,d-1}(x) + \frac{t_z - x}{\Delta_{z,d}} B_{z,d-1}(x)$$

for $1 - d_o + d \leq z \leq m + d_o$. As shown in Section A.4, for each $d \in \{1, 2, \ldots, d_o\}$, the functions $B_{j,d}$, $1 \leq j \leq m + d$, constitute a basis for $\mathcal{S}_d(t_0, t_1, \ldots, t_m)$ with the desired property that

$$B_{j,d} \begin{cases} = 0 & \text{on } \mathbb{R} \setminus (t_{j-d}, t_j), \\ > 0 & \text{on } (t_{j-d}, t_j). \end{cases} \tag{5.2}$$

Moreover,

$$\sum_{j=1}^{m+d} B_{j,d} \;\equiv\; 1 \;\text{ on } [a, b]. \tag{5.3}$$

Figures 5.2 and 5.3 show these B spline basis functions for $\mathcal{S}_d(0, 1, 2, 4, 5, 6)$, $d = 2, 3$. In both cases, we used $t_z := 0.1 \cdot z$ for $z < 0$ and $t_{5+z} := 6 + 0.1 \cdot z$ for $z > 0$.

Figure 5.2: Basis functions $B_{j,2}$ for $\mathcal{S}_2(0,1,2,4,5,6)$.

Special case: Cubic splines with equidistant knots. Suppose that the knots t_k are equidistant, that is, $t_k - t_{k-1} = \Delta > 0$ for $1 \le k \le m$. Augmenting these knots by $t_z = a + z\Delta$ for $z \in \mathbb{Z}$, a basis of B splines for $\mathcal{S}_3(t_0, \ldots, t_m)$ is given by

$$B_{k,3}(x) \;:=\; b_o\big((x - t_{k-2})/\Delta\big),$$

where

$$b_o(s) \;:=\; \begin{cases} 2/3 - s^2 + |s|^3/2 & \text{if } |s| \le 1, \\ (2 - |s|)^3/6 & \text{if } 1 \le |s| \le 2, \\ 0 & \text{if } |s| \ge 2. \end{cases}$$

That these functions $B_{k,3}$, $1 \le k \le m+3$, are cubic splines can be deduced from

$$b_o'(s) \;=\; \begin{cases} -2s + (3/2)\,\mathrm{sign}(s)s^2 & \text{if } |s| \le 1, \\ -\,\mathrm{sign}(s)(2 - |s|)^2/2 & \text{if } 1 \le |s| \le 2, \\ 0 & \text{if } |s| \ge 2, \end{cases}$$

$$b_o''(s) \;=\; \begin{cases} -2 + 3|s| & \text{if } |s| \le 1, \\ 2 - |s| & \text{if } 1 \le |s| \le 2, \\ 0 & \text{if } |s| \ge 2, \end{cases}$$

$$b_o'''(s) \;=\; \begin{cases} 3\,\mathrm{sign}(s) & \text{if } 0 < |s| < 1, \\ -\,\mathrm{sign}(s) & \text{if } 1 < |s| < 2, \\ 0 & \text{if } |s| > 2. \end{cases}$$

That they satisfy the properties (5.2) and (5.3) can be verified with elementary calculations.

5.1.4 Precision in Case of Linear Splines

Suppose that all values X_i are in a compact interval $[a, b]$, and we work with linear splines with equidistant knots $t_{m,j} = a + (j/m)(b - a)$, $0 \le j \le m$. This means, the estimated regression

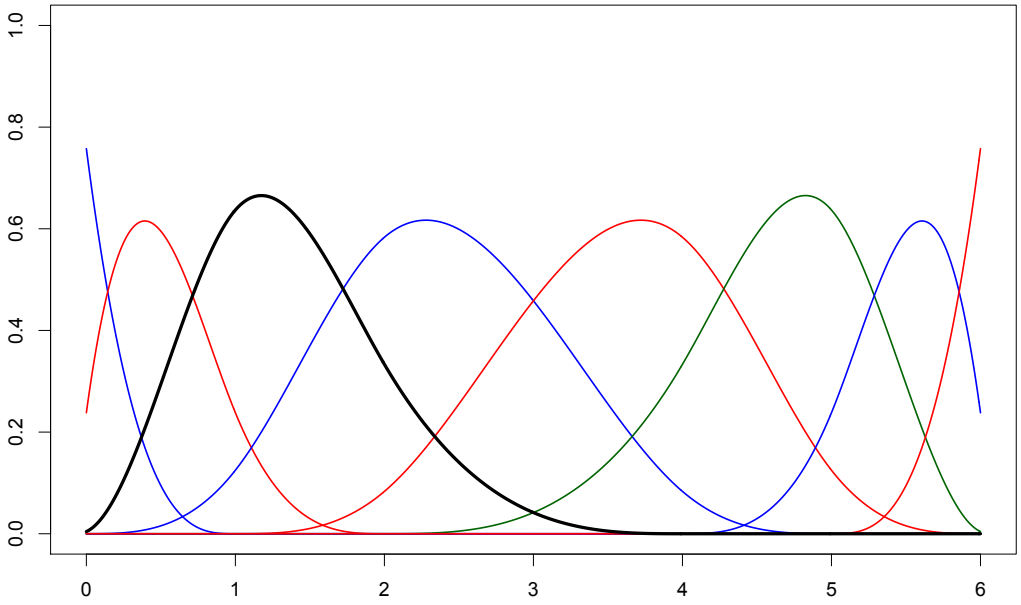

Figure 5.3: Basis functions $B_{j,3}$ for $\mathcal{S}_3(0,1,2,4,5,6)$.

function $\widehat{f}$ may be viewed as an estimator of the function

$$f_{n,m} := \operatorname*{arg\,min}_{g \in \mathcal{S}_1(t_{m,0},\ldots,t_{m,m})} \|f - g\|_n,$$

where

$$\|h\|_n := \sqrt{n^{-1} \sum_{i=1}^{n} h(X_i)^2};$$

see also Section 2.5.

Theorem 5.1. *In case of homoscedastic errors ε_i with variance σ^2,*

$$\mathbb{E}\big(\|\widehat{f} - f\|_n^2\big) \le \|f - f_{n,m}\|_n^2 + \frac{(m+1)\sigma^2}{n}.$$

If f is twice differentiable with $|f''| \le L$, then

$$\|f - f_{n,m}\|_n^2 \le \frac{L^2(b-a)^4}{64m^4}.$$

If we choose $m = m(n) = (C + o(1))n^{1/5}$ for some $C > 0$, then

$$\mathbb{E}\big(\|\widehat{f} - f\|_n^2\big) = O(n^{-4/5}).$$

This theorem shows that by means of linear splines, one can estimate a twice differentiable regression function with bounded second derivative up to an estimation error of order $O_p(n^{-2/5})$. In fact, one can show that under these assumptions, no other estimator would exhibit a faster rate of convergence. But proving the latter result is beyond the scope of the present course.

Proof of Theorem 5.1. With the hat matrix $\boldsymbol{H}$, we may write

$$\begin{aligned}\mathbb{E}\big(\|\widehat{f} - f\|_n^2\big) &= \mathbb{E}\big(n^{-1}\|\boldsymbol{H}\boldsymbol{Y} - f(\boldsymbol{X})\|^2\big)\\ &= \mathbb{E}\big(n^{-1}\|\boldsymbol{H}f(\boldsymbol{X}) - f(\boldsymbol{X}) + \boldsymbol{H}\boldsymbol{\varepsilon}\|^2\big)\\ &= \|f - f_{n,m}\|_n^2 + \mathbb{E}\big(n^{-1}\|\boldsymbol{H}\boldsymbol{\varepsilon}\|^2\big)\\ &= \|f - f_{n,m}\|_n^2 + \frac{p(n,m)\sigma^2}{n},\end{aligned}$$

where $p(n,m)$ is the dimension of the model space

$$\big\{g(\boldsymbol{X}) : g \in \mathcal{S}_1(t_{m,0},\ldots,t_{m,m})\big\}.$$

Now, the first part of Theorem 5.1 follows from the fact that $p(n,m) \le m+1$.

As a surrogate for $f_{n,m}$, we consider the interpolation spline $f_m \in \mathcal{S}_1(t_{m,0},\ldots,t_{m,m})$ with $f_m(t_{m,i}) = f(t_{m,i})$ for $i = 0,\ldots,m$. Then

$$\begin{aligned}\|f - f_{n,m}\|_n^2 &\le \|f - f_m\|_n^2\\ &\le \max_{x\in[a,b]} |f(x) - f_m(x)|^2\\ &\le \frac{L^2(b-a)^4}{64m^4}.\end{aligned}$$

The latter inequality is a consequence of Exercise 5.2. This proves the second part of Theorem 5.1, and the last part follows from an elementary calculation. □

Exercise 5.2 (Linear interpolation and extrapolation). Let f be twice differentiable on $[a,b]$ with $|f''| \le L$. For fixed points $a \le x_0 < x_1 \le b$, set

$$g(x) := f(x_0) + \frac{x - x_0}{x_1 - x_0}\,(f(x_1) - f(x_0)).$$

That means, g is the unique affine function such that $g(x_0) = f(x_0)$ and $g(x_1) = f(x_1)$. Show that

$$|f(x) - g(x)| \le \frac{L}{2}\,|x - x_0||x - x_1|$$

for arbitrary $x \in [a,b]$. Deduce from that inequality that

$$|f - g| \le L(x_1 - x_0)^2/8 \quad \text{on } [x_0, x_1].$$

5.2 Local Polynomials

If the regression function f is d times continuously differentiable, it follows from Taylor's formula that

$$f(x+s) = \sum_{k=0}^{d} f^{(k)}(x)\frac{s^k}{k!} + o(s^d) \quad \text{as } s \to 0.$$

Thus, f may be approximated locally by a polynomial or order d. To estimate $f(x)$, one could choose a neighborhood $U(x)$ of x and work with the model of polynomial regression for the subsample of observations (X_i, Y_i) such that $X_i \in U(x)$.

Here is a more general description: For a fixed point x, we choose nonnegative weights $w_i(x)$, $1 \le i \le n$, such that

$$\#\{X_i : w_i(x) > 0\} \;\ge\; d+1.$$

Then we minimize the weighted sum of squares

$$\sum_{i=1}^{n} w_i(x)\Big(Y_i - \sum_{k=0}^{d} a_k \frac{(X_i - x)^k}{k!}\Big)^2$$

as a function of $\boldsymbol{a} = (a_k)_{k=0}^d \in \mathbb{R}^{d+1}$. Let $\widehat{\boldsymbol{a}}(x) := (\widehat{a}_k(x))_{k=0}^d$ be the unique minimizer. Then $\widehat{a}_k(x)$ may be viewed as an estimator of $f^{(k)}(x)$.

5.2.1 Examples for the Weights $w_i(x)$

Nearest neighbor method. We choose an integer $k = k(n)$ between $d+1$ and n and define

$$w_i(x) \;:=\; \begin{cases} 1 & \text{if } |x - X_i| \le R_k(x), \\ 0 & \text{if } |x - X_i| > R_k(x). \end{cases}$$

Here $R_1(x) \le R_2(x) \le \cdots \le R_n(x)$ are the distances $|x - X_j|$ between the point x and the observed X-values in nondecreasing order. The numbers $k(n)$ should satisfy $\lim_{n\to\infty} k(n) = \infty$ and $\lim_{n\to\infty} k(n)/n = 0$.

Kernel functions. Let $K : \mathbb{R} \to \mathbb{R}$ be a nonnegative function such that $0 < \int K(x)\,dx < \infty$. Then, we define

$$w_i(x) \;:=\; K\Big(\frac{x - X_i}{h}\Big)$$

with a suitable *bandwidth* $h = h(x, \boldsymbol{X}) > 0$. In our explicit examples, we use the *Epanechnikov kernel*

$$K(x) \;:=\; \max(1 - x^2, 0).$$

Alternatively, one could use $K(z) = \exp(-z^2/2)$.

5.2.2 Explicit Computation

We may rewrite the weighted sum of squares as

$$\sum_{i=1}^{n} w_i(x)\Big(Y_i - \sum_{k=0}^{d} a_k \frac{(X_i - x)^k}{k!}\Big)^2 \;=\; \big\|\boldsymbol{Y}(x) - \boldsymbol{D}(x)\boldsymbol{a}\big\|^2$$

with

$$\boldsymbol{Y}(x) \;:=\; \big(\sqrt{w_i(x)}\,Y_i\big)_{i=1}^n \;\in\; \mathbb{R}^n$$

and

$$\boldsymbol{D}(x) \;:=\; \Big(\sqrt{w_i(x)}\,\frac{(X_i - x)^{j-1}}{(j-1)!}\Big)_{i\le n, j\le d+1} \;\in\; \mathbb{R}^{n\times(d+1)}.$$

Consequently,

$$\widehat{\boldsymbol{a}}(x) \;=\; \operatorname*{arg\,min}_{\boldsymbol{a}\in\mathbb{R}^{d+1}} \big\|\boldsymbol{Y}(x) - \boldsymbol{D}(x)\boldsymbol{a}\big\|^2.$$

The special cases $d = 0$ and $d = 1$. For an arbitrary vector $\boldsymbol{v} \in \mathbb{R}^n$, we define its weighted mean

$$\bar{v}(x) \;:=\; \sum_{i=1}^{n} \pi_i(x) v_i \quad \text{with} \quad \pi_i(x) \;:=\; w_i(x) \Big/ \sum_{j=1}^{n} w_j(x).$$

In the special case $d = 0$, we end up with the estimator

$$\widehat{f}(x) \;=\; \bar{Y}(x).$$

In the special case $d = 1$, we may write

$$\widehat{f}(x) \;=\; \bar{Y}(x) + \widehat{a}_1(x)(x - \bar{X}(x))$$

with

$$\begin{aligned} \widehat{a}_1(x) \;&:=\; \sum_{i=1}^{n} \pi_i(x) \frac{(X_i - \bar{X}(x))Y_i}{S(x)^2}, \\ S(x)^2 \;&:=\; \sum_{i=1}^{n} \pi_i(x)(X_i - \bar{X}(x))^2. \end{aligned}$$

These formulae result from elementary calculations of the subsequent exercise.

Exercise 5.3. Let X_o and Y_o be random variables on a common probability space such that $\mathbb{E}(X_o^2), \mathbb{E}(Y_o^2) < \infty$ and $\mathrm{Var}(X_o) > 0$. Show that

$$\mathbb{E}\big((Y_o - a - bX_o)^2\big)$$

is minimal in $a, b \in \mathbb{R}$ if and only if

$$\begin{aligned} a \;&=\; \mathbb{E}(Y_o) - b\,\mathbb{E}(X_o), \\ b \;&=\; \frac{\mathrm{Cov}(X_o, Y_o)}{\mathrm{Var}(X_o)} \;=\; \frac{\mathbb{E}\big((X_o - \mathbb{E}(X_o))Y_o\big)}{\mathrm{Var}(X_o)}. \end{aligned}$$

Apply this result to two fixed vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and the random variables $X_o := x_J$, $Y_o = y_J$, where J is a random variable with values in $\{1, 2, \ldots, n\}$ and given weights $\pi_i = \mathbb{P}(J = i)$.

Example 5.4 (Baseball data). For the data in Example 4.8, Figure 5.4 shows a scatter plot of the decimal logarithms of annual incomes (Y) versus the number of years, truncated at 20 (X). In addition, we show the the LSE $\widehat{f}$ for f, assuming that f belongs to $\mathcal{S}_3(t_0, t_1, t_2, t_3, t_4)$ with $t_j = 0.5 + 5j$. This is the curve in the center. The other four curves are pointwise and simultaneous (via Scheffé) 95%-confidence bounds for f. Precisely, the estimator and confidence bounds are for the approximating function

$$f_o \;:=\; \operatorname*{arg\,min}_{g \in \mathcal{S}_3(t_0,t_1,t_2,t_3,t_4)} \|f - g\|_n.$$

Figure 5.5 shows the same data and the locally linear estimators $\widehat{f}$, based on the Epanechnikov kernel and the global bandwidths $h = 2$ and $h = 4$.

Exercise 5.5. Implement the locally linear and locally quadratic estimators. Compare the methods by means of simulated data.

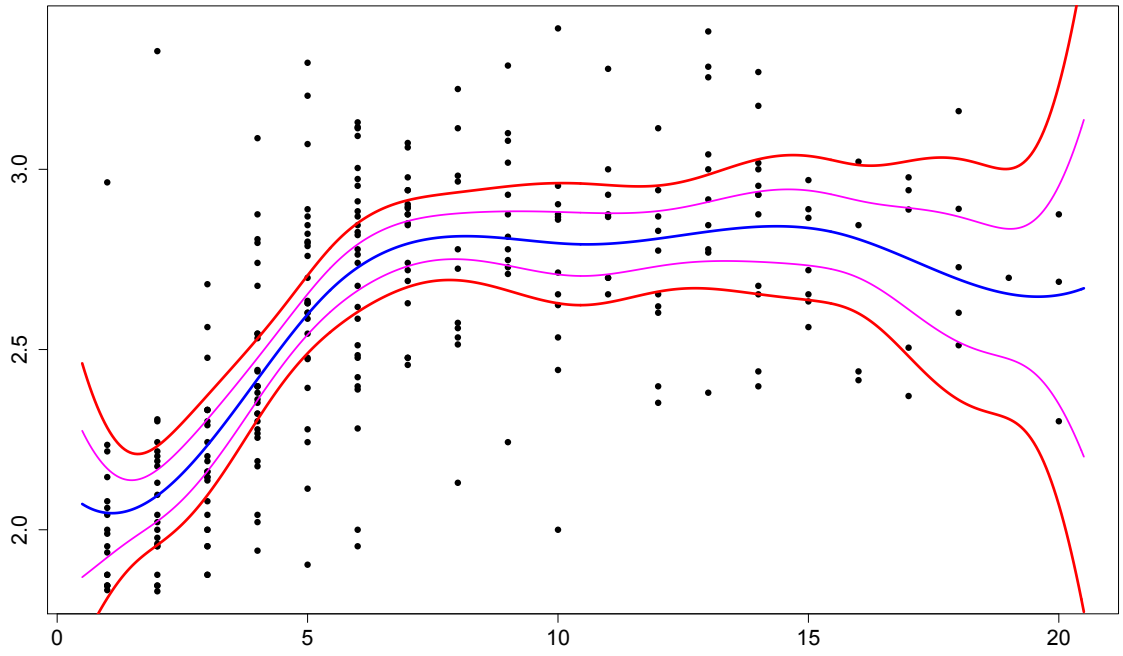

Figure 5.4: Spline estimator for f in Example 5.4.

5.2.3 Precision of Locally Linear Estimators

In this section, we derive a rough bound for the precision of locally linear estimators. For more precise and general results we refer to Fan and Gijbels (1996).

The locally linear estimator $\widehat{f}(x)$ can be written as

$$\widehat{f}(x) \;=\; \sum_{i=1}^{n} \pi_i(x)\kappa(x, X_i)Y_i,$$

where

$$\kappa(x,t) \;:=\; 1 + \frac{(x - \bar{X}(x))(t - \bar{X}(x))}{S(x)^2}.$$

One can interpret $\bar{X}(x)$ and $S(x)$ as mean and standard deviation of a discrete probability distribution Q_x,

$$Q_x(B) \;:=\; \sum_{i=1}^{n} \pi_i(x) 1_{[X_i \in B]}.$$

Then, the expected value of $\widehat{f}(x)$ may be represented as

$$\sum_{i=1}^{n} \pi_i(x)\kappa(x, X_i) f(X_i) \;=\; \int f(t)\kappa(x,t)\, Q_x(dt).$$

One can easily verify that

$$\int \kappa(x,t)\, Q_x(dt) \;=\; 1 \quad \text{and} \quad \int t\kappa(x,t)\, Q_x(dt) \;=\; x.$$

Hence, $\mathbb{E}\,\widehat{f}(x) = f(x)$, whenever f is an affine function. But now we want to assume only that f is twice continuously differentiable with derivatives f' and f'', where

$$|f''| \;\leq\; L.$$

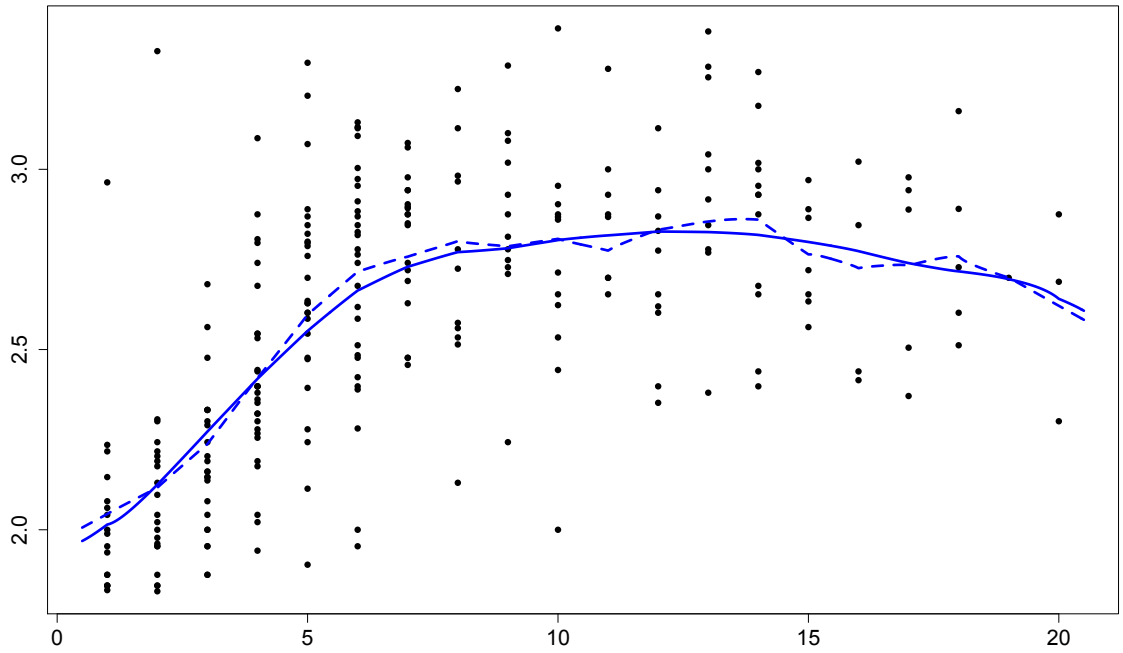

Figure 5.5: Locally linear estimator for f in Example 5.4, with bandwidths $h = 2$ (dashed line) and $h = 4$ (continuous line).

In this case,

$$f(t) \;=\; f(x) + f'(x)(t-x) + r(x,t) \quad \text{with} \quad |r(x,t)| \;\leq\; \frac{L(t-x)^2}{2},$$

so

$$\mathbb{E}\,\widehat{f}(x) \;=\; f(x) + \int r(x,t)\kappa(x,t)\,Q(dt).$$

Introducing the absolute moments

$$M_r(x) \;:=\; \int |t-x|^r\,Q(dt), \quad r \geq 1,$$

of Q_x, we may conclude that

$$\begin{aligned}
\big|\mathbb{E}\,\widehat{f}(x) - f(x)\big| \;&\leq\; \int |r(x,t)||\kappa(x,t)|\,Q_x(dt) \\
&\leq\; \frac{L}{2}\Big(M_2(x) + \frac{|x-\bar{X}(x)|}{S(x)} \int \frac{(t-x)^2|t-\bar{X}(x)|}{S(x)}\,Q_x(dt)\Big) \\
&\leq\; \frac{L}{2}\Big(M_2(x) + \frac{|x-\bar{X}(x)|}{S(x)}\Big(M_4(x) \int \frac{|t-\bar{X}(x)|^2}{S(x)^2}\,Q_x(dt)\Big)^{1/2}\Big) \\
&\leq\; \frac{L}{2}M_4(x)^{1/2}\Big(1 + \frac{|x-\bar{X}(x)|}{S(x)}\Big),
\end{aligned}$$

where the last two steps follows from the Cauchy-Schwarz inequality and the definition of $S(x)$. On the other hand, with the maximal variance

$$\sigma^2 \;:=\; \max_{i=1,\ldots,n} \mathrm{Var}(\varepsilon_i),$$

the variance of $\widehat{f}(x)$ may be bounded by

$$\begin{aligned}
\operatorname{Var}\big(\widehat{f}(x)\big) &= \sum_{i=1}^{n} \pi_i(x)^2 \kappa(x, X_i)^2 \operatorname{Var}(\varepsilon_i) \\
&\le \sigma^2 \max_{i=1,\dots,n} \pi_i(x) \int \kappa(x,t)^2 \, Q_x(dt) \\
&= \sigma^2 \max_{i=1,\dots,n} \pi_i(x) \Big(1 + \frac{(x-\bar{X}(x))^2}{S(x)^2} \int \frac{(t-\bar{X}(x))^2}{S(x)^2} \, Q_x(dt)\Big) \\
&= \sigma^2 \max_{i=1,\dots,n} \pi_i(x) \Big(1 + \frac{(x-\bar{X}(x))^2}{S(x)^2}\Big),
\end{aligned}$$

where the second last step follows from $\int (t - \bar{X}(x)) \, Q(dt) = 0$.

All in all, we obtain the following inequality:

Lemma 5.6 (Error bound for locally linear estimators). *Under the previous conditions,*

$$\begin{aligned}
\mathbb{E}\big(\big(\widehat{f}(x) - f(x)\big)^2\big) &= \big(\mathbb{E}\,\widehat{f}(x) - f(x)\big)^2 + \operatorname{Var}\big(\widehat{f}(x)\big) \\
&\le \Big(\frac{L^2 M_4(x)}{2} + \sigma^2 \max_{i=1,\dots,n} \pi_i(x)\Big)\Big(1 + \frac{(x-\bar{X}(x))^2}{S(x)^2}\Big). \qquad (5.4)
\end{aligned}$$

This lemma shows that, both, $M_4(x)$ and $\max_i \pi_i(x)$ should be as small as possible, but these goals can not be achieved simultaneously.

Example 5.7. Specifically, let $X_i = i/n$ for $1 \le i \le n$, and suppose that the weights have been defined via the nearest neighbor method with some number $k(n) \ge 2$. Then one can show that for any $x \in [0,1]$,

$$k(n) \;\le\; \#\{i\colon w_i(x) = 1\} \;\le\; k(n) + 1 \quad \text{and} \quad \max_{1 \le i \le n} w_i(x) |X_i - x| \;\le\; k(n)/n.$$

Thus,

$$\max_{1 \le i \le n} \pi_i(x) \;\le\; k(n)^{-1} \quad \text{and} \quad M_4(x) \;\le\; (k(n)/n)^4.$$

Moreover, Q_x is the uniform distribution on a set of the form $\{j/n : a(x) < j \le b(x)\}$ with integers $0 \le a(x) < b(x) \le n$ such that $b(x) - a(x) \ge k(n)$, and this distribution has mean $n^{-1}(a(x) + b(x) + 1)$ and variance $n^{-2}((b(x) - a(x))^2 - 1)/12$. Moreover, the distance between x and $\bar{X}(x)$ is at most $n^{-1}(b(x) - a(x) + 1)/2$, so

$$\frac{(x-\bar{X}(x))^2}{S(x)^2} \;\le\; \frac{3(b(x) - a(x) + 1)}{b(x) - a(x) - 1} \;\le\; 3 + 6/(k(n) - 1) \;\le\; 9.$$

Consequently, for fixed L and σ^2, we see that

$$\mathbb{E}\big((\widehat{f}(x) - f(x))^2\big) \;=\; O\big(n^{-4} k(n)^4 + k(n)^{-1}\big)$$

uniformly in $x \in [0,1]$. To obtain a minimal order of magnitude for this bound, $n^{-4}k(n)^4$ and $k(n)^{-1}$ should be of the same order. This is the case if $k(n)$ is precisely of order $O(n^{4/5})$, and then,

$$\mathbb{E}\big((\widehat{f}(x) - f(x))^2\big) \;=\; O(n^{-4/5})$$

uniformly in $x \in [0,1]$. In particular,

$$\widehat{f}(x) - f(x) \;=\; O_p(n^{-2/5}).$$

The same orders of magnitude result with kernel-based weights, provided that the global bandwidth $h = h(n)$ is of order $O(n^{-1/5})$.

5.3 Regularization

Finally, we describe a special case of smoothing methods which are known under the general names *regularization* or *penalization*. Generally, we estimate the regression function f by minimizing

$$\sum_{i=1}^{n}(Y_i - g(X_i))^2 + \lambda \operatorname{Pen}(g)$$

over all functions $g : \mathbb{R} \to \mathbb{R}$. Here, $\lambda > 0$ is a given penalty parameter, and $\operatorname{Pen}(\cdot)$ is a *penalty function*. The penalty $\operatorname{Pen}(g) \in [0,\infty]$ measures "irregularity" of g. By means of this penalty we prevent overfitting in the sense of choosing functions g which essentially interpolate the observations (X_i, Y_i). The parameter λ plays a similar role as $k(n)$ and $h(n)^{-1}$ in connection with local polynomials.

5.3.1 Smoothing Splines

For a function $g : \mathbb{R} \to \mathbb{R}$, let

$$\operatorname{Pen}_1(g) \;:=\; \int_{\mathbb{R}} g'(x)^2\,dx,$$

provided that g is absolutely continuous; otherwise we set $\operatorname{Pen}_1(g) := \infty$. Further, let

$$\operatorname{Pen}_2(g) \;:=\; \int_{\mathbb{R}} g''(x)^2\,dx,$$

if g is differentiable with absolutely continuous derivative g'; otherwise we set $\operatorname{Pen}_2(g) := \infty$. When minimizing $\operatorname{Pen}_1(g)$ or $\operatorname{Pen}_2(g)$ under certain constraints, functions of the following type appear:

Definition 5.8 (Natural linear and cubic splines). Consider $m \in \mathbb{N}$ and real numbers $t_0 < t_1 < \cdots < t_m$.

A continuous function $f : \mathbb{R} \to \mathbb{R}$ is called *natural linear spline* with knots $t_0, t_1, \ldots, t_m$, if f is constant on $(-\infty, t_0]$, constant on $[t_m, \infty)$, and affine on each interval $[t_{j-1}, t_j]$, $1 \le j \le m$. The set of all such functions is denoted with $\mathcal{S}_1^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$.

A function $f : \mathbb{R} \to \mathbb{R}$ is called *natural cubic spline* with knots $t_0, t_1, \ldots, t_m$, if f is twice continuously differentiable, and if the second derivative f'' has the following properties: $f'' = 0$ on $(-\infty, t_0] \cup [t_m, \infty)$, and f'' is affine on each interval $[t_{j-1}, t_j]$, $1 \le j \le m$. The set of all such functions is denoted with $\mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$.

Remark 5.9. Both function classes $\mathcal{S}_1^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$ and $\mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$ are linear spaces with dimension $m + 1$.

In case of $\mathcal{S}_1^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$, one can simply extend any function f in the usual spline space $\mathcal{S}_1(t_0, t_1, \ldots, t_m)$ to a constant function on $(-\infty, t_0]$ and on $[t_m, \infty)$. In particular, any basis f_1, f_2, …, f_{m+1} of $\mathcal{S}_1(t_0, t_1, \ldots, t_m)$ becomes thus a basis of $\mathcal{S}_1^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$.

In case of $\mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$, the restriction of any function $f \in \mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$ to $[t_0, t_m]$ defines a function in $\mathcal{S}_3(t_0, t_1, \ldots, t_m)$ with the additional property that

$$f''(t_0) \;=\; f''(t_m) \;=\; 0. \tag{5.5}$$

On the other hand, any function $f \in \mathcal{S}_3(t_0, t_1, \ldots, t_m)$ satisfying (5.5) (with $f''(t_0)$ and $f''(t_m)$ interpreted as one-sided derivatives) may be extended to a function in $\mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$ via

$$f(x) \;:=\; \begin{cases} f(t_0) + f'(t_0)(x - t_0) & \text{for } x \le t_0, \\ f(t_m) + f'(t_m)(x - t_m) & \text{for } x \ge t_m. \end{cases} \tag{5.6}$$

To construct a specific basis of $\mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$, we start with an arbitrary basis f_1, f_2, …, f_{m+3} of $\mathcal{S}_3(t_0, t_1, \ldots, t_m)$ such that $f_{m+2}(x)$ and $f_{m+3}(x)$ are proportional to $(t_1 - x)_+^3$ and $(x - t_{m-1})_+^3$, respectively. In particular, $f_{m+2} = f'_{m+2} = f''_{m+2} = 0$ on $[t_1, t_m]$, and $f_{m+3} = f'_{m+3} = f''_{m+3} = 0$ on $[t_0, t_{m-1}]$. Consequently,

$$\begin{aligned} \tilde{f}_j(x) \;&:= f_j(x) - \frac{f_j''(t_0)}{f_{m+2}''(t_0)} f_{m+2}(x) - \frac{f_j''(t_m)}{f_{m+3}''(t_m)} f_{m+3}(x) \\ &= f_j(x) - \frac{f_j''(t_0)(t_1 - x)_+^3}{6(t_1 - t_0)} - \frac{f_j''(t_m)(x - t_{m-1})_+^3}{6(t_m - t_{m-1})}, \quad x \in [t_0, t_m], \end{aligned}$$

defines linearly independent functions $\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_{m+1}$ in $\mathcal{S}_3(t_0, t_1, \ldots, t_m)$ with the additional property (5.5). By means of the extension (5.6), we obtain a basis of $\mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$.

The following lemma about smooth interpolation will be the key ingredient for our main result in this section.

Lemma 5.10. *Consider $m \in \mathbb{N}$ and real numbers $t_0 < t_1 < \cdots < t_m$ and $z_0, z_1, \ldots, z_m$. Further, let $g : \mathbb{R} \to \mathbb{R}$ be an arbitrary function with the property that*

$$g(t_j) \;=\; z_j \quad \text{for } j = 0, 1, \ldots, m. \tag{5.7}$$

(a) *There exists a unique function $g_o \in \mathcal{S}_1^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$ which satisfes* (5.7)*. This function minimizes $\mathrm{Pen}_1(g)$ among all functions g satisfying* (5.7)*.*

(b) *There exists a unique function $g_o \in \mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$ which satisfes* (5.7)*. This function minimizes $\mathrm{Pen}_2(g)$ among all functions g satisfying* (5.7)*.*

Proof of Lemma 5.10. The proof of part (a) is essentially a consequence of Exercise 5.11. Hence, we focus on part (b). For $x \in \mathbb{R}$ define

$$\lambda_0(x) \;:=\; \frac{t_m - x}{t_m - t_0} \quad \text{and} \quad \lambda_m(x) \;:=\; \frac{x - t_0}{t_m - t_0}.$$

Obviously, $\lambda_0(x) + \lambda_m(x) = 1$, and in case of $x \in [t_0, t_m]$, $\lambda_0(x), \lambda_m(x) \in [0,1]$. Now, for an arbitrary function g with $\mathrm{Pen}_2(g) < \infty$ and any point $x \in [t_0, t_m]$, we investigate the difference

$$\Delta(x) \;:=\; \lambda_0(x)g(t_0) + \lambda_m(x)g(t_m) - g(x)$$

in more detail:

$$\begin{aligned}
\Delta(x) \;&=\; \lambda_0(x)\big(g(t_0) - g(x)\big) + \lambda_m(x)\big(g(t_m) - g(x)\big) \\
&=\; -\lambda_0(x)\int_{t_0}^{x} g'(s)\,ds + \lambda_m(x)\int_{x}^{t_m} g'(s)\,ds \\
&=\; -\lambda_0(x)\int_{t_0}^{x} \big(g'(s) - g'(x)\big)\,ds + \lambda_m(x)\int_{x}^{t_m} \big(g'(s) - g'(x)\big)\,ds \\
&\qquad \big(\text{because } \lambda_0(x)(x - t_0) = \lambda_m(x)(t_m - x)\big) \\
&=\; \lambda_0(x)\int_{t_0}^{x}\int_{s}^{x} g''(t)\,dt\,ds + \lambda_m(x)\int_{x}^{t_m}\int_{x}^{s} g''(t)\,dt\,ds \\
&=\; \lambda_0(x)\int_{t_0}^{x} (t - t_0)g''(t)\,dt + \lambda_m(x)\int_{x}^{t_m} (t_m - t)g''(t)\,dt.
\end{aligned}$$

Consequently, for $x \in [t_0, t_m]$,

$$\Delta(x) \;=\; \int_{\mathbb{R}} K(x,t)g''(t)\,dt$$

with

$$K(x,t) \;:=\; \begin{cases} \dfrac{(t - t_0)(t_m - x)}{t_m - t_0} & \text{for } t_0 \le t \le x, \\[2ex] \dfrac{(x - t_0)(t_m - t)}{t_m - t_0} & \text{for } x \le t \le t_m, \\[2ex] 0 & \text{else.} \end{cases}$$

The constraints (5.7) on g are equivalent to

$$g(t_0) \;=\; z_0,\;\; g(t_m) \;=\; z_m \quad \text{and} \quad \int_{\mathbb{R}} K(t_j,t)g''(t)\,dt \;=\; c_j \quad \text{for } 1 \le j < m, \tag{5.8}$$

where $c_j := \lambda_0(t_j)z_0 + \lambda_m(t_j)z_m - z_j$. The functions $K(t_j,\cdot)$, $1 \le j < m$, are in the space $\mathcal{S}_1^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$, have compact support and are linearly independent. Hence, there exists a unique linear combination $h = \sum_{j=1}^{m-1} a_j K(t_j,\cdot)$ saytisfying the equations

$$\int_{\mathbb{R}} K(t_j,t)h(t)\,dt \;=\; c_j, \quad 1 \le j < m.$$

In particular, there exists a unique function $g_o \in \mathcal{S}_3^{\mathrm{nat}}(t_0, t_1, \ldots, t_m)$ solving (5.7), and $g_o'' = h$. For any other function g satisfying (5.7) and $\mathrm{Pen}_2(g) < \infty$, we can write $g'' = h + r$ with a function $r \in L^2(\mathbb{R})$ satisfying the equations

$$\int_{\mathbb{R}} K(t_j,t)r(t)\,dt \;=\; 0, \quad 1 \le j < m.$$

In particular, $\int_{\mathbb{R}} h(t)r(t)\,dt = 0$, whence

$$\int_{\mathbb{R}} g''(t)^2\,dt \;=\; \int_{\mathbb{R}} h(t)^2\,dt + \int_{\mathbb{R}} r(t)^2\,dt \;\ge\; \int_{\mathbb{R}} h(t)^2\,dt.$$

Equality holds if and only if $r = 0$ almost everywhere, which is equivalent to $g \equiv g_o$. □

Exercise 5.11. Let $g : [a,b] \to \mathbb{R}$ be absolutely continuous, that means, there exists an integrable function $g' : [a,b] \to \mathbb{R}$ such that $g(x) = g(a) + \int_a^x g'(t)\,dt$ for all $x \in [a,b]$. Show that

$$\int_a^b g'(t)^2\,dt \;\geq\; \frac{\big(g(b)-g(a)\big)^2}{b-a}$$

with equality if and only if $g' = \big(g(b)-g(a)\big)/(b-a)$ almost everywhere on $[a,b]$.

Now we come back to our regularization estimator:

Theorem 5.12. *Consider arbitrary data vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^n$ such that $\#\{X_1,\ldots,X_n\} \geq 2$. For arbitrary numbers $k \in \{1,2\}$ and $\lambda > 0$, there exists a unique function $\widehat{f}_\lambda : \mathbb{R} \to \mathbb{R}$ minimizing*

$$H_\lambda(g) \;:=\; \sum_{i=1}^n (Y_i - g(X_i))^2 + \lambda\,\mathrm{Pen}_k(g)$$

among all functions $g : \mathbb{R} \to \mathbb{R}$. Denoting the different elements of $\{X_1,\ldots,X_n\}$ with $t_0 < t_1 < \ldots < t_m$, the function $\widehat{f}_\lambda$ belongs to $\mathcal{S}^{\mathrm{nat}}_{2k-1}(t_0,t_1,\ldots,t_m)$.

Proof of Theorem 5.12 and construction of $\widehat{f}_\lambda$. The target functional $H_\lambda(g)$ may be rewritten as $S_0^2(\boldsymbol{X},\boldsymbol{Y}) + \tilde{H}_\lambda(g)$ with

$$\begin{aligned}
\tilde{H}_\lambda(g) \;&:=\; \sum_{i=0}^m w_i (y_i - g(t_i))^2 + \lambda\,\mathrm{Pen}_k(g),\\
w_i \;&:=\; \#\{\ell : X_\ell = t_i\},\\
y_i \;&:=\; w_i^{-1} \sum_{\ell:\,X_\ell = t_i} Y_\ell
\end{aligned}$$

and $S_0^2(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=0}^m \sum_{\ell:\,X_\ell=t_i} (Y_\ell - y_i)^2$. Now let

$$\mathcal{F} \;:=\; \begin{cases} \mathcal{S}^{\mathrm{nat}}_1(t_0,t_1,\ldots,t_m) & \text{if } k=1,\\ \mathcal{S}^{\mathrm{nat}}_3(t_0,t_1,\ldots,t_m) & \text{if } k=2. \end{cases}$$

According to Lemma 5.10, for each function $g : \mathbb{R} \to \mathbb{R}$ there exists a unique function $g_o \in \mathcal{F}$ such that $g_o = g$ on $\{t_0,t_1,\ldots,t_m\}$. Moreover, $\mathrm{Pen}_k(g_o) \leq \mathrm{Pen}_k(g)$ with equality if and only if $g = g_o$. Hence, it suffices to consider functions $g \in \mathcal{F}$. Now, we choose a basis $f_1, f_2, \ldots, f_{m+1}$ of $\mathcal{F}$ and write $g = \sum_{j=1}^{m+1} \theta_j f_j$ for some $\boldsymbol{\theta} \in \mathbb{R}^{m+1}$. In this case,

$$\tilde{H}_\lambda(g) \;=\; c - 2\boldsymbol{b}^\top\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{A}_\lambda \boldsymbol{\theta}$$

with $c = \sum_{i=0}^m w_i y_i^2$ and

$$\begin{aligned}
\boldsymbol{b} \;&:=\; \Big(\sum_{i=0}^m w_i y_i f_j(t_i)\Big)_{j=1}^{m+1},\\
\boldsymbol{A}_\lambda \;&:=\; \Big(\sum_{i=0}^m w_i f_j(t_i) f_\ell(t_i) + \lambda \int_{t_0}^{t_m} f_j^{(k)}(t) f_\ell^{(k)}(t)\,dt\Big)_{j,\ell=1}^{m+1}.
\end{aligned}$$

The latter matrix $\boldsymbol{A}_\lambda$ is symmetric and positive definite, because $\boldsymbol{\theta}^\top \boldsymbol{A}_\lambda \boldsymbol{\theta} = 0$ would imply that $g^{(k)} \equiv 0$ and $g = 0$ on $\{t_0,\ldots,t_m\}$, whence $g \equiv 0$ and $\boldsymbol{\theta} = \boldsymbol{0}$. Consequently,

$$\widehat{\boldsymbol{\theta}}_\lambda = \big(\widehat{\theta}_{\lambda,j}\big)_{j=1}^{m+1} \;:=\; \boldsymbol{A}_\lambda^{-1}\boldsymbol{b}$$

is the unique minimizer of $\tilde{H}_\lambda$ and yields the unique minimizer $\widehat{f}_\lambda := \sum_{j=1}^{m+1} \widehat{\theta}_{\lambda,j} B_j$ of H_λ. $\square$

5.3.2 A Related Approach

As in the previous section, let $t_0 < t_1 < \cdots < t_m$ be the different elements of $\{X_1, X_2, \ldots, X_n\}$, and write $\boldsymbol{w} = (w_i)_{i=0}^m$, $\boldsymbol{y} = (y_i)_{i=0}^m$ with

$$w_i \;=\; \#\{\ell : X_\ell = t_i\}, \quad \text{and} \quad y_i \;=\; w_i^{-1} \sum_{\ell\,:\,X_\ell = t_i} Y_\ell.$$

If we are only interested in estimating f on the set $\{t_0, t_1, \ldots, t_m\}$ of observed X-values, there exists a rather simple and general approach to regularization: One minimizes

$$\begin{aligned} H_\lambda(\boldsymbol{g}) \;&:=\; \sum_{i=0}^m w_i (y_i - g_i)^2 + \lambda \boldsymbol{g}^\top \boldsymbol{A} \boldsymbol{g} \\ &= \boldsymbol{y}^\top \operatorname{diag}(\boldsymbol{w}) \boldsymbol{y} - 2 \boldsymbol{y}^\top \operatorname{diag}(\boldsymbol{w}) \boldsymbol{g} + \boldsymbol{g}^\top (\operatorname{diag}(\boldsymbol{w}) + \lambda \boldsymbol{A}) \boldsymbol{g} \end{aligned}$$

with respect to $\boldsymbol{g} = (g_i)_{i=0}^m \in \mathbb{R}^{m+1}$, where $\boldsymbol{A}$ is a symmetric and positive semidefinite matrix in $\mathbb{R}^{(m+1)\times(m+1)}$. The vector $\boldsymbol{g}$ corresponds to $\big(g(t_i)\big)_{i=0}^m$ with $g : \mathbb{R} \to \mathbb{R}$. Then, the unique minimizer

$$\widehat{\boldsymbol{f}}_\lambda \;:=\; (\operatorname{diag}(\boldsymbol{w}) + \lambda \boldsymbol{A})^{-1} \operatorname{diag}(\boldsymbol{w}) \boldsymbol{y}$$

of H_λ is our estimator of $\big(f(t_i)\big)_{i=0}^m$.

Concerning $\boldsymbol{A}$, there are several possibilities. Suppose we want to approximate the functional $\int_{t_0}^{t_m} g^{(k)}(t)^2\, dt$ for some integer $k \geq 1$ by $\boldsymbol{g}^\top \boldsymbol{A} \boldsymbol{g}$, where we assume that $m \geq k$. To this end, we construct for $j \in \{0, 1, \ldots, m-k\}$ a vector $\boldsymbol{v}_j = (v_{ij})_{i=0}^m \in \mathbb{R}^{m+1}$ such that

$$v_{ij} \;=\; 0 \;\text{ if } i \notin \{j, \ldots, j+k\}$$

and

$$\boldsymbol{v}_j^\top (t_i^s)_{i=0}^m \;=\; 1_{[s=k]} \;\text{ for } s = 0, 1, \ldots, k.$$

For $k \geq 3$, this may be achieved, for instance, with orthogonal polynomials. If $g^{(k)}$ is approximately constant on $[t_j, t_{j+k}]$, then

$$(t_{j+k} - t_j) \big(\boldsymbol{v}_j^\top \big(g(t_i)\big)_{i=0}^m\big)^2 \;\approx\; (k!)^2 \int_{t_j}^{t_{j+k}} g^{(k)}(t)^2\, dt.$$

Hence, if we define

$$\boldsymbol{A} \;:=\; \frac{1}{(k!)^2 \min(k, m+1-k)} \sum_{j=0}^{m-k} (t_{j+k} - t_j) \boldsymbol{v}_j \boldsymbol{v}_j^\top,$$

then $\boldsymbol{g}^\top \boldsymbol{A} \boldsymbol{g}$ may be viewed as a surrogate for $\int_{t_0}^{t_m} g^{(k)}(t)^2\, dt$. The factor $k_o = \min(k, m+1-k)$ stems from the fact that each interval $[t_{\ell-1}, t_\ell]$ is contained in $\min(\ell, m+1-\ell, k) \leq k_o$ intervals $[t_j, t_{j+k}]$, $0 \leq j \leq m-k$.

Specifically, for $k = 1$ we obtain the vectors

$$\boldsymbol{v}_j \;=\; \frac{1}{t_{j+1} - t_j} \big(1_{[i=j+1]} - 1_{[i=j]}\big)_{i=0}^m, \quad 0 \leq j \leq m-1,$$

and for $k = 2$ we end up with

$$\boldsymbol{v}_j \;=\; \frac{1}{t_{j+2} - 2t_{j+1} - t_j} \Big(\frac{1_{[i=j+2]} - 1_{[i=j+1]}}{t_{j+2} - t_{j+1}} - \frac{1_{[i=j+1]} - 1_{[i=j]}}{t_{j+1} - t_j} \Big)_{i=0}^m, \quad 0 \leq j \leq m-2.$$

5.3.3 Choosing the Regularization Parameter

Regularization estimators as well as locally polynomial estimators and other nonparametric procedures depend often on certain tuning parameters. In what follows, we describe two possible strategies for choosing the regularization parameter $\lambda > 0$ automatically. Similar ideas are applicable to other nonparametric techniques.

Cross validation. For $i = 1, 2, \ldots, n$, let $[\boldsymbol{X}_{-i}, \boldsymbol{Y}_{-i}]$ be the data matrix $[\boldsymbol{X}, \boldsymbol{Y}]$ without the row (X_i, Y_i). Now one estimates for given $\lambda > 0$ the regression function f from the reduced data $[\boldsymbol{X}_{-i}, \boldsymbol{Y}_{-i}]$ by $\widehat{f}_{\lambda,-i}$. The overall quality of these n estimators is measured by the sum of squares $Q(\lambda)$,

$$Q(\lambda) \ := \ \sum_{i=1}^{n} (Y_i - \widehat{f}_{\lambda,-i}(X_i))^2.$$

Now, one minimizes $Q(\lambda)$ over all $\lambda > 0$ in a given set.

Comparing two variance estimators. For $\lambda > 0$, let $\widehat{f}_\lambda$ be the corresponding regularization estimator of the regression function f. This yields an estimator of the standard deviation $\sigma > 0$ of the (homoscedastic) errors, namely,

$$\widehat{\sigma}_\lambda \ := \ \left(\frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{f}_\lambda(X_i))^2 \right)^{1/2}.$$

In addition, let $\widehat{\sigma}_*$ be an estimator for σ which does not depend on λ or $\widehat{f}_\lambda$ and is reliable for a large class of regression functions f; explicit examples will follow soon. As explained in Exercise 5.13, the estimator $\widehat{\sigma}_\lambda$ is monotone increasing in $\lambda > 0$. Now, we compare $\widehat{\sigma}_\lambda$ with $\widehat{\sigma}_*$ and choose $\lambda > 0$ such that both are essentially identical.

Exercise 5.13. For an arbitrary set $\mathcal{X}$, let $Q : \mathcal{X} \to \mathbb{R}$ and $P : \mathcal{X} \to [0, \infty]$ be two functions. Suppose that for real parameters $0 \leq \lambda < \mu$ there exist minimizers x_λ of $Q + \lambda P$ and x_μ of $Q + \mu P$ over $\mathcal{X}$, where $P(x_\lambda), P(x_\mu) < \infty$. Show that

$$Q(x_\lambda) \ \leq \ Q(x_\mu) \quad \text{and} \quad P(x_\lambda) \ \geq \ P(x_\mu).$$

A first estimator of the noise level. In case of $X_1 \leq X_2 \leq \cdots \leq X_n$ one could estimate σ by

$$\widehat{\sigma}_* \ := \ \left(\frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 \right)^{1/2}, \tag{5.9}$$

as proposed by Rice (1981). The rationale behind this estimator is that $Y_{i+1} - Y_i \approx \varepsilon_{i+1} - \varepsilon_i$, provided that $f(X_i) \approx f(X_{i+1})$, and $\mathbb{E}\big((\varepsilon_{i+1} - \varepsilon_i)^2\big) = 2\sigma^2$ in case of homoscedastic errors.

Exercise 5.14. Suppose that $X_1 \leq X_2 \leq \cdots \leq X_n$, and let $\widehat{\sigma}_*$ be given by (5.9). Suppose further that $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ and $\mathbb{E}(\varepsilon_i^4) \leq K\sigma^4$ for $1 \leq i \leq n$ and some constant $K \geq 1$.

(a) Show that

$$\mathbb{E}(\widehat{\sigma}_*^2) \;=\; \sigma^2 + \rho^2 \quad \text{with } \rho^2 := \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (f(X_{i+1}) - f(X_i))^2.$$

Show also that

$$\rho^2 \;\le\; \frac{\max_{1\le i<n}(X_{i+1} - X_i)}{2(n-1)} \int_{X_1}^{X_n} f'(t)^2 \, dt.$$

(b) Show that the "estimator"

$$\check{\sigma}_*^2 \;:=\; \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (\varepsilon_{i+1} - \varepsilon_i)^2$$

satisfies the following inequalities:

$$\mathbb{E}(\check{\sigma}_*^2) \;=\; \sigma^2, \quad \mathrm{Var}(\check{\sigma}_*^2) \;\le\; \frac{K\sigma^4}{n-1} \quad \text{and} \quad \mathbb{E}\big((\widehat{\sigma}_*^2 - \check{\sigma}_*^2 - \rho^2)^2\big) \;\le\; \frac{8\sigma^2}{n-1}\rho^2.$$

(c) Deduce the inequality

$$\mathbb{E}\Big(\Big(\frac{\widehat{\sigma}^2}{\sigma^2 + \rho^2} - 1\Big)^2\Big) \;\le\; \frac{2K+4}{n-1}.$$

A second estimator of the noise level. Let us assume that the errors have a variance depending continuously on X_i, that is, $\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma(X_i)^2$ for some continuous function $\sigma(\cdot)$. The quantity $\widehat{\sigma}_\lambda^2 = n^{-1}\sum_{i=1}^n (Y_i - \widehat{f}_\lambda(X_i))^2$ may be viewed as a proxy for $n^{-1}\sum_{i=1}^n \varepsilon_i^2$ and thus as an estimator of

$$\sigma_*^2 \;:=\; \frac{1}{n}\sum_{i=1}^n \sigma(X_i)^2.$$

Now, with $t_0 < t_1 < \cdots < t_m$, $w_0, w_1, \ldots, w_m \ge 1$ and $y_0, y_1, \ldots, y_m \in \mathbb{R}$ as before, we estimate $\sigma(t_j)^2$ as follows: If $w_j > 1$, let

$$\widehat{\sigma}(t_j)^2 \;:=\; \frac{1}{w_j - 1} \sum_{i:\; X_i = t_j} (Y_i - y_j)^2.$$

If $w_0 = 1$, we set $\widehat{\sigma}(t_0)^2 := \widehat{\sigma}(t_1)^2$, and in case of $w_m = 1$ we set $\widehat{\sigma}(t_m)^2 := \widehat{\sigma}(t_{m-1})^2$. For $0 < j < m$ with $w_j = 1$, we set

$$\widehat{\sigma}(t_j)^2 \;:=\; \frac{(y_j - (1-\lambda_j)y_{j-1} - \lambda_j y_{j+1})^2}{1 + (1-\lambda_j)^2/w_{j-1} + \lambda_j^2/w_{j+1}} \quad \text{with} \quad \lambda_j \;:=\; \frac{t_j - t_{j-1}}{t_{j+1} - t_{j-1}}.$$

Then we set

$$\widehat{\sigma}_*^2 \;:=\; \frac{1}{n}\sum_{i=1}^n \widehat{\sigma}(X_i)^2.$$

Exercise 5.15. Show that $\mathbb{E}(\widehat{\sigma}(t_j)^2) = \sigma(t_j)^2$ in the following two situations:

(i) $w_j > 1$.

(ii) $w_j = 1$, $0 < j < m$, $\sigma(t_{j-1}) = \sigma(t_j) = \sigma(t_{j+1})$, and f is affine on $[t_{j-1}, t_{j+1}]$.

Chapter 6

General Considerations about Estimation

In this chapter, the estimation of regression functions is embedded into a rather general framework. To do so, we are using concepts of statistical decision theory which are treated in more detail within courses on mathematical statistics. In particular, we introduce so-called (log-)likelihood functions.

6.1 Means and Quantiles as Optimal Predictors

Suppose we want to predict the value of a yet unobserved random variable $Y \in \mathbb{R}$ by a fixed real number v with maximal precision. For the time being, the distribution of Y is assumed to be known. Depending on how we define "precision", we obtain different solutions. In general, we quantify the size of the *prediction error* $v - Y$ by its *prediction loss*

$$\rho(v - Y)$$

for a given convex *loss function* $\rho : \mathbb{R} \to \mathbb{R}$. We also assume that ρ is coercive, that is,

$$\rho(t) \ \to \ \infty \quad \text{as } |t| \to \infty.$$

Our goal is to determine a prediction $v \in \mathbb{R}$ such that the *mean prediction loss*

$$\mathbb{E}\,\rho(v - Y)$$

is minimal. The next lemma provides explicit solutions for two and a half particular cases.

Lemma 6.1 (Optimal prediction).

(a) Mean squared prediction error: *Suppose that* $\mathbb{E}(Y^2) < \infty$. *In case of* $\rho(t) = t^2$,

$$\mathbb{E}\,\rho(v - Y) \ = \ \mathrm{Var}(Y) + (v - \mathbb{E}(Y))^2.$$

Hence, the optimal prediction of Y *is given by*

$$v \ = \ \mathbb{E}(Y).$$

(b) Mean absolute prediction error: *Suppose that* $\mathbb{E}\,|Y| < \infty$. *Let* $\rho(t) = |t|$. *Then* $\mathbb{E}\,\rho(v-Y)$ *is minimal in* v *if and only if* v *is a median of* $\mathcal{L}(Y)$, *that is,*

$$\mathbb{P}(Y < v) \;\le\; 1/2 \;\le\; \mathbb{P}(Y \le v).$$

(c) Quantiles: *Suppose that* $\mathbb{E}\,|Y| < \infty$. *For a given number* $\gamma \in (0,1)$ *let*

$$\rho(t) \;:=\; (1-2\gamma)t + |t| \;=\; \begin{cases} 2(1-\gamma)t & \text{if } t \ge 0, \\ 2\gamma|t| & \text{if } t \le 0. \end{cases}$$

Then $\mathbb{E}\,\rho(v-Y)$ *is minimal in* v *if and only if* v *is a* γ*-quantile of* $\mathcal{L}(Y)$, *that is,*

$$\mathbb{P}(Y < v) \;\le\; \gamma \;\le\; \mathbb{P}(Y \le v).$$

Remark 6.2 (Existence of moments). The assumptions in Lemma 6.1 that $\mathbb{E}(Y^2)$ or $\mathbb{E}\,|Y|$ are finite, may be weakened as follows. For any convex loss function ρ, we could replace the prediction loss $\rho(v-Y)$ by the difference

$$\rho(v-Y) - \rho(v_o - Y)$$

with an arbitrary reference value v_o. If $\mathbb{E}\,\rho(v-Y)$ is finite for all $v \in \mathbb{R}$, then minimizing $v \mapsto \mathbb{E}\,\rho(v-Y)$ is equivalent to minimizing $v \mapsto \mathbb{E}\big[\rho(v-Y) - \rho(v_o-Y)\big]$. The modified prediction loss has the advantage that in case of $\rho(t) = t^2$, it suffices to assume that $\mathbb{E}\,|Y| < \infty$, and in case of $\rho(t) = (1-2\gamma)t + |t|$, we do not need any further moment assumption. The main conclusions that $\mathbb{E}(Y)$ or any γ-quantile of $\mathcal{L}(Y)$ are optimal, respectively, remain valid and may be viewed as special cases of the more general Theorem 6.3 below.

In what follows, the left- and right-sided derivatives of a function $h : \mathbb{R} \to \mathbb{R}$ are denoted by

$$h'(x-) \;:=\; \lim_{y \nearrow x} \frac{h(y)-h(x)}{y-x} \quad \text{and} \quad h'(x+) \;:=\; \lim_{y \searrow x} \frac{h(y)-h(x)}{y-x},$$

respectively. In what follows, we assume some basic knowledge about convex functions on the real line as presented, for instance, in Section 3.1 of Dümbgen (2021). In particular, if $\rho : \mathbb{R} \to \mathbb{R}$ is convex, then for arbitrary points $s < t < u$,

$$\frac{\rho(t)-\rho(s)}{t-s} \;\le\; \rho'(t-) \;\le\; \rho'(t+) \;\le\; \frac{\rho(u)-\rho(t)}{u-t}. \tag{6.1}$$

Theorem 6.3. *Let* $\rho : \mathbb{R} \to \mathbb{R}$ *be convex, and suppose that for arbitrary* $v \in \mathbb{R}$, *the expectations* $\mathbb{E}\,\rho'((v-Y)\pm)$ *exist in* $\mathbb{R}$. *For arbitrary* $v_o \in \mathbb{R}$,

$$R(v) \;:=\; \mathbb{E}\big[\rho(v-Y) - \rho(v_o-Y)\big]$$

defines a convex function $R : \mathbb{R} \to \mathbb{R}$ *such that*

$$R'(v\pm) \;=\; \mathbb{E}\,\rho'((v-Y)\pm).$$

If, in addition, ρ *is coercive, then* R *is coercive too, that is,* $R(v) \to \infty$ *as* $|v| \to \infty$. *In this case, the set* $V_* := \arg\min_{v \in \mathbb{R}} R(v)$ *is a compact real interval. It consists of all points* $v \in \mathbb{R}$ *such that* $R'(v-) \le 0 \le R'(v+)$.

The inequalities (6.1) imply that $\mathbb{E}\,\rho'((v-Y)\pm)$ exists in $\mathbb{R}$ for arbitrary $v \in \mathbb{R}$ whenever $\mathbb{E}\,\rho(v-Y)$ exists in $\mathbb{R}$ for arbitrary $v \in \mathbb{R}$.

Proof of Lemma 6.1. The main statement in part (a) is a well-known identity from probability theory. If we only assume that $\mathbb{E}\,|Y| < \infty$, optimality of $v = \mathbb{E}(Y)$ follows from the following calculations:

$$\begin{aligned}\mathbb{E}\big[\rho(v-Y) - \rho(v_o-Y)\big] &= \mathbb{E}\big[-2vY + v^2 + 2v_oY - v_o^2\big]\\ &= -2v\,\mathbb{E}(Y) + v^2 + 2v_o\,\mathbb{E}(Y) - v_o^2\\ &= (v-\mathbb{E}(Y))^2 - (v_o-\mathbb{E}(Y))^2.\end{aligned}$$

Part (b) is a special case of part (c) with $\gamma = 1/2$, and part (c) could be derived with elementary calculations. But the arguing would be similar to the proof of Theorem 6.3. Thus we derive part (c) from Theorem 6.3: The function $\mathbb{R} \ni t \mapsto \rho(t) := (1-2\gamma)t + |t|$ is convex and coercive, and its one-sided derivatives at t are given by

$$\begin{aligned}\rho'(t-) &= 1-2\gamma+1_{[t>0]}-1_{[t\le 0]} = 2(1_{[t>0]}-\gamma),\\ \rho'(t+) &= 1-2\gamma+1_{[t\ge 0]}-1_{[t<0]} = 2(1_{[t\ge 0]}-\gamma).\end{aligned}$$

Since the functions $\rho'(\cdot\pm)$ are bounded, $R(v) := \mathbb{E}\big(\rho(v-Y)-\rho(v_o-Y)\big)$ defines a convex function $R:\mathbb{R}\to\mathbb{R}$ with one-sided derivatives

$$\begin{aligned}R'(v-) &= 2(\mathbb{P}(v-Y>0)-\gamma) = 2(\mathbb{P}(Y<v)-\gamma),\\ R'(v+) &= 2(\mathbb{P}(v-Y\ge 0)-\gamma) = 2(\mathbb{P}(Y\le v)-\gamma).\end{aligned}$$

As stated in Theorem 6.3, a point v minimizes R if and only if $R'(v-) \le 0 \le R'(v+)$, and the latter conditions are equivalent to the inequalities $\mathbb{P}(Y<v) \le \gamma \le \mathbb{P}(Y\le v)$. □

Proof of Theorem 6.3. For arbitrary real numbers $a<b$ and two different points $v,w \in [a,b]$, it follows from (6.1) that

$$\rho'((a-Y)+) \le \frac{\rho(w-Y)-\rho(v-Y)}{w-v} \le \rho'((b-Y)-).$$

In particular, if $[a,b] \ni v_o$, we obtain the inequality

$$\big|\rho(v-Y)-\rho(v_o-Y)\big| \le |v-v_o|\max\big\{-\rho'((a-Y)+),\rho'((b-Y)-)\big\}$$

for any $v \in [a,b]$, and the random variable on the right hand side has finite expectation. Hence $R(v)$ is well-defined in $\mathbb{R}$. One can easily deduce from convexity of ρ and linearity of expectations that R is convex, too.

As to the one-sided derivatives, for different points $v,w \in (a,b)$,

$$\frac{R(w)-R(v)}{w-v} = \mathbb{E}\,\frac{\rho(w-Y)-\rho(v-Y)}{w-v},$$

and

$$\Big|\frac{\rho(w-Y)-\rho(v-Y)}{w-v}\Big| \;\le\; \max\big\{-\rho'((a-Y)+),\rho'((b-Y)-)\big\},$$
$$\frac{\rho(w-Y)-\rho(v-Y)}{w-v} \;\to\; \begin{cases}\rho'((v-Y)+) & \text{as } w\searrow v,\\ \rho'((v-Y)-) & \text{as } w\nearrow v.\end{cases}$$

Hence, by dominated convergence,

$$\frac{R(w)-R(v)}{w-v} \;\to\; \begin{cases}\mathbb{E}\,\rho'((v-Y)+) & \text{as } w\searrow v,\\ \mathbb{E}\,\rho'((v-Y)-) & \text{as } w\nearrow v.\end{cases}$$

Now suppose that ρ is convex and coercive. Coercivity of any convex function $h:\mathbb{R}\to\mathbb{R}$ is equivalent to $\lim_{t\to\infty} h'(t\pm)\in(0,\infty]$ and $\lim_{t\to-\infty} h'(t\pm)\in[-\infty,0)$. Convexity of ρ implies that $\rho'((v-Y)\pm)$ is monotone increasing in $v\in\mathbb{R}$. Hence, by monotone convergence, $R'(v\pm)=\mathbb{E}\,\rho'((v-Y)\pm)\to\lim_{t\to\infty}\rho'(t\pm)>0$ as $v\to\infty$, while $R'(v\pm)=\mathbb{E}\,\rho'((v-Y)\pm)\to\lim_{t\to-\infty}\rho'(t\pm)<0$ as $v\to-\infty$. This proves coercivity of R.

If R is convex and coercive, the set V_* of its minimizers has to be closed and nonempty (by continuity and coercivity) and an interval (by convexity). In other words, V_* is a compact interval. That $v\in V_*$ is equivalent to $R'(v-)\le 0\le R'(v+)$ is a standard result of convex analysis: On the one hand, if $R'(v-)>0$, then $R(w)<R(v)$ for $w<v$ sufficiently close to v, and $R(v+)<0$ would imply that $R(w)<R(v)$ for $w>v$ sufficiently close to v. On the other hand, if $R'(v-)\le 0\le R'(v+)$, then (6.1) implies that $R(w)\ge R(v)+R'(v-)(w-v)\ge R(v)$ for all $w<v$, and $R(w)\ge R(v)+R'(v+)(w-v)\ge R(v)$ for all $w>v$. □

Empirical mean prediction error. In most applications, the distribution of Y is unknown and has to be estimated from empirical data. Suppose we observe stochastically independent copies $Y_1,Y_2,\ldots,Y_n$ of Y. Then the *empirical mean prediction error*

$$\widehat{R}(v) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\rho(v-Y_i)$$

suggests itself as a surrogate for $R(v):=\mathbb{E}\,\rho(v-Y)$. For the explicit examples in Lemma 6.1, minimizing $\widehat{R}(\cdot)$ leads to the sample mean, a sample median or a sample γ-quantile, respectively.

Here is a general result showing that for large sample sizes n, any minimizer $\widehat{v}$ of the empirical risk function $\widehat{R}$ has to be close to the set V_* of minimizers of the theoretical risk function R.

Theorem 6.4. *Let $\rho:\mathbb{R}\to\mathbb{R}$ be convex and coercive such that $\mathbb{E}\,\rho'((v-Y)\pm)$ exists in $\mathbb{R}$ for any $v\in\mathbb{R}$. Let $\widehat{v}$ be any minimizer of the empirical risk function $\widehat{R}$, and let V_* be the set of minimizers v of $R(v)=\mathbb{E}\big[\rho(v-Y)-\rho(v_o-Y)\big]$. Then for any fixed $\delta>0$,*

$$\lim_{n\to\infty}\mathbb{P}\big(\widehat{v}\le\min(V_*)-\delta \;\text{ or }\; \widehat{v}\ge\max(V_*)+\delta\big) \;=\; 0.$$

Proof of Theorem 6.4. Note that $R'(v\pm)$ and $\widehat{R}'(v\pm)$ are monotone increasing in $v\in\mathbb{R}$. By definition of V_*, the points $v_1:=\min(V_*)-\delta$ and $v_2:=\max(V_*)+\delta$ satisfy the inequalities

$R'(v_1+) < 0 < R'(v_2-)$. If $\widehat{v} \le v_1$, then

$$0 \;\le\; \widehat{R}(\widehat{v}+) \;\le\; \widehat{R}(v_1+) \;=\; Z_{n1} + R'(v_1+)$$

with

$$Z_{n1} \;:=\; \widehat{R}'(v_1+) - R'(v_1+) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big[\rho'((v_1 - Y_i)+) - \mathbb{E}\,\rho'((v_1 - Y)+)\big].$$

Likewise, $\widehat{v} \ge v_2$ implies that

$$0 \;\ge\; \widehat{R}(\widehat{v}-) \;\ge\; \widehat{R}(v_2-) \;=\; Z_{n2} + R'(v_2-)$$

with

$$Z_{n2} \;:=\; \widehat{R}'(v_2-) - R'(v_2-) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big[\rho'((v_2 - Y_i)-) - \mathbb{E}\,\rho'((v_2 - Y)-)\big].$$

By the weak law of large numbers, $\lim_{n\to\infty} \mathbb{P}(|Z_{nk}| \ge \epsilon) = 0$ for any fixed $\epsilon > 0$ and $k = 1, 2$. Consequently, with $\epsilon_1 := -R'(v_1+) > 0$ and $\epsilon_2 := R'(v_2-) > 0$,

$$\mathbb{P}\big(\widehat{v} \le v_1 \text{ or } \widehat{v} \ge v_2\big) \;\le\; \mathbb{P}\big(Z_{n1} \ge \epsilon_1 \text{ or } Z_{n2} \le -\epsilon_2\big) \le\; \mathbb{P}(Z_{n1} \ge \epsilon_1) + \mathbb{P}(Z_{n2} \le -\epsilon_2)$$

converges to 0 as $n \to \infty$. □

6.2 Loss Functions and Risks

The considerations in the previous section may be generalized as follows: Let Y be a random variable with values in a measurable space $(\mathcal{Y}, \mathcal{B})$. Now we want to make a "decision" v in a "decision space" $\mathbb{V}$ about the yet unobserved value Y. For a given *loss function* $L : \mathbb{V} \times \mathcal{Y} \to (-\infty, \infty]$ which is measurable in the second argument, the quality of a decision v is quantified by the *loss of* v,

$$L(v, Y),$$

or the *risk of* v,

$$R(v) \;:=\; \mathbb{E}\,L(v, Y) \;=\; \int L(v, y)\, P(dy),$$

i.e. the *mean loss of* v.

In the previous section we encountered already two and a half important examples: In all cases, $\mathcal{Y} = \mathbb{V} = \mathbb{R}$, the "decision" was a prediction v of Y, and $L(v, Y) = \rho(v - Y)$ for a given convex, coercive function $\rho : \mathbb{R} \to \mathbb{R}$.

Exercise 6.5. Let $Y \ge 0$ be the health costs which a randomly chosen customer of a health insurance company will cause next year. Let $e > 0$ be the annual premium he or she has to pay, and let $v \ge 0$ be his or her retention. Hence, the net revenues of the insurance company for this customer are given by

$$e - \max(Y - v, 0) \;=\; e + \min(v - Y, 0).$$

Now we would like to determine a retention v which is fair in the sense that

$$(*) \qquad \mathbb{E}\big(e - \max(Y - v, 0)\big) \;=\; 0.$$

(a) Show that there is a unique solution of $(*)$, provided that $e \le \mathbb{E}(Y) < \infty$.

(b) Determine a convex function $\rho : \mathbb{R} \to \mathbb{R}$ such that solving $(*)$ is equivalent to minimizing $\mathbb{E}\,\rho(v - Y)$ or $\mathbb{E}\big(\rho(v - Y) - \rho(-Y)\big)$.

Exercise 6.6. Let Y be a real-valued random variable. We identify a value $y \in \mathbb{R}$ with the indicator function $\mathbb{R} \ni t \mapsto 1_{[y \le t]}$ and would like to predict this function by some function $v : \mathbb{R} \to \mathbb{R}$. Suppose that our loss function is given by

$$L(v, y) \;:=\; \int \big(v(t) - 1_{[y \le t]}\big)^2 \, M(dt)$$

for some finite measure M on $\mathbb{R}$. Determine all functions v with minimal risk $R(v) = \mathbb{E}\, L(v, Y)$.

The special case of a finite set $\mathcal{Y}$. Here, Y is a categorical random variable whose distribution is determined by the probability weights

$$p(z) \;:=\; \mathbb{P}(Y = z), \quad z \in \mathcal{Y}.$$

Now our decision space $\mathbb{V}$ is the set $\mathbb{R}^{\mathcal{Y}}$ of all functions $v : \mathcal{Y} \to \mathbb{R}$, and we consider three different loss functions:

Example 6.7 (Least squares for categorical observations). We identify Y with the random indicator function $\mathcal{Y} \ni z \mapsto 1_{[z=Y]}$ which we would like to predict by a fixed function $v \in \mathbb{V}$. Let the loss of v be given by

$$L(v, Y) \;:=\; \sum_{z \in \mathcal{Y}} \big(v(z) - 1_{[z=Y]}\big)^2.$$

Here one can show that $R(v) = \mathbb{E}\, L(v, Y)$ is minimal if and only if

$$v(z) \;=\; p(z) \quad \text{for } z \in \mathcal{Y}.$$

Exercise 6.8. Prove the optimality result in Example 6.7.

Example 6.9 (Likelihood for categorical observations). We present two different loss functions whose definition is not so obvious at first glance. Both are related to likelihood methods as presented in the next section. We assume now that

$$p(z) \;>\; 0 \quad \text{for all } z \in \mathcal{Y}.$$

The two loss functions L_1, L_2 are given by

$$L_1(v, y) \;:=\; \sum_{z \in \mathcal{Y}} e^{v(z)} - v(y),$$
$$L_2(v, y) \;:=\; \log\Big(\sum_{z \in \mathcal{Y}} e^{v(z)}\Big) - v(y).$$

The corresponding risks are

$$R_1(v) := \mathbb{E}\, L_1(v, Y) \;=\; \sum_{z\in\mathcal{Y}} e^{v(z)} - \sum_{z\in\mathcal{Y}} p(z)v(z),$$
$$R_2(v) := \mathbb{E}\, L_2(v, Y) \;=\; \log\Big(\sum_{z\in\mathcal{Y}} e^{v(z)}\Big) - \sum_{z\in\mathcal{Y}} p(z)v(z).$$

In both cases, the sum $\sum_{z\in\mathcal{Y}} e^{v(z)} > 0$ appears, and this is strictly increasing in each value $v(z)$. But the second sum $\sum_{z\in\mathcal{Y}} p(z)v(z)$ prevents us from letting $v(z) \to -\infty$ for each $z \in \mathcal{Y}$.

For R_1, one can show that

$$R_1(v) \;\geq\; 1 - \sum_{z\in\mathcal{Y}} p(z)\log p(z)$$

with equality if and only if $v(z) = \log p(z)$ for all $z \in \mathcal{Y}$.

The risk function $R_2(v)$ is minimal at $v \in \mathbb{R}^{\mathcal{Y}}$ if and only if

$$\frac{e^{v(z)}}{\sum_{y\in\mathcal{Y}} e^{v(y)}} \;=\; p(z) \quad \text{for all } z \in \mathcal{Y}.$$

In other words, for some constant $c \in \mathbb{R}$,

$$v(z) \;=\; \log p(z) + c \quad \text{for all } z \in \mathcal{Y}.$$

A unique solution can be enforced by additional constraints, e.g.

(i) $\sum_{y\in\mathcal{Y}} e^{v(y)} = 1$ or
(ii) $v(y_o) = 0$ for a reference category $y_o \in \mathcal{Y}$ or
(iii) $\sum_{y\in\mathcal{Y}} v(y) \;=\; 0$.

Variant (iii) will be used in the context of logistic regression.

Exercise 6.10. Prove the optimality results in Example 6.9.

Exercise 6.11 (Mean squared loss in general). We have seen that the mean of a random variable $Y \in \mathbb{R}$ is an optimal predictor with respect to mean squared prediction error. This example as well as the settings in Exercises 6.6 and 6.8 can be viewed as special cases of the following abstract setting: We observe a random variable $Y \in \mathcal{Y}$ which can be mapped into a real Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ via a measurable mapping $T : \mathcal{Y} \to \mathbb{H}$ such that $\mathbb{E}(\|T(Y)\|^2) < \infty$. If $\mathbb{H}$ is our decision space, and if the loss of a decision $h \in \mathbb{H}$ is defined as

$$L(h, Y) \;:=\; \|h - T(Y)\|^2,$$

what is the unique minimizer h of $R(h) := \mathbb{E}\, L(h, Y)$?

Specify $\mathbb{H}$ and T for the three situations mentioned before.

How could you modify $L(h, Y)$ to obtain the same minimizer under the weaker assumption that $\mathbb{E}\,\|T(Y)\| < \infty$?

Empirical risk. In case of stochastically independent copies $Y_1, Y_2, \ldots, Y_n$ of Y with unknown distribution, one can estimate $R(v)$ by the empirical risk

$$\widehat{R}(v) \;:=\; \frac{1}{n}\sum_{i=1}^{n} L(v, Y_i).$$

Let P be the distribution of Y, and let $\widehat{P}$ be the empirical distribution of $Y_1, Y_2, \ldots, Y_n$. That is, $\widehat{P}(B) := \#\{i : Y_i \in B\}/n$ for $B \subset \mathcal{Y}$, and $\int h\, d\widehat{P} = n^{-1}\sum_{i=1}^{n} h(Y_i)$ for $h : \mathcal{Y} \to \mathbb{R}$. Then we may write

$$R(v) \;=\; \int L(v, y)\, P(dy) \quad \text{and} \quad \widehat{R}(v) \;=\; \int L(v, y)\, \widehat{P}(dy).$$

6.3 Maximum Likelihood Estimation

Suppose that the distribution P of Y is unknown, but let $(P_\theta)_{\theta\in\Theta}$ be a given family of probability distributions P_θ on $(\mathcal{Y}, \mathcal{B})$ which contains P or at least a "good approximation" of P. Now our goal is not a "decision" about Y but the definition and estimation of a "true parameter" $\theta_o \in \Theta$, having observed Y.

More precisely, suppose that P_θ admits a density function p_θ with respect to a measure M on $(\mathcal{Y}, \mathcal{B})$. For instance, let $\mathcal{Y} = \mathbb{R}^d$, and let p_θ be a (Lebesgue) probability density of P_θ in the usual sense. Or let $\mathcal{Y}$ be a countable set, and let p_θ be the weight function of P_θ, that is, $p_\theta(z) = P_\theta(\{z\})$ for $z \in \mathcal{Y}$. Now one may try to estimate the "true parameter" $\theta_o \in \Theta$ by means of the *negative log-likelihood*

$$L(\theta, Y) \;:=\; -\log p_\theta(Y).$$

The random functions $\theta \mapsto p_\theta(Y)$ and $\theta \mapsto \log p_\theta(Y)$ on the *parameter space* Θ are the so-called *likelihood function* and *log-likelihood function*, respectively.

Suppose that

(i) $P = P_{\theta_o}$ for some $\theta_o \in \Theta$,
(ii) $\int |\log p_\theta|\, dP_\theta < \infty$ for all $\theta \in \Theta$,
(iii) $P_\eta \neq P_\theta$ for all $\eta, \theta \in \Theta$ with $\eta \neq \theta$.

Then the parameter θ_o is the unique minimizer of

$$R(\theta) \;:=\; \mathbb{E}\, L(\theta, Y).$$

This follows immediately from the subsequent Lemma 6.12. If $L(\widehat{\theta}, Y) = \min_{\theta\in\Theta} L(\theta, Y)$ for some parameter $\widehat{\theta} = \widehat{\theta}(Y)$ in Θ, we call $\widehat{\theta}$ a *maximum likelihood estimator for* θ_o. If assumption (i) is not satisfied, one may view $\widehat{\theta}$ as an estimator of a minimizer of the risk $R(\cdot)$.

Lemma 6.12. *Let P and Q be probability distributions on $\mathcal{Y}$ with density functions p and q, respectively, with respect to some measure M. Further let $\int |\log p|\, dP < \infty$. Then,*

$$-\int \log(q)\, dP \;\geq\; -\int \log(p)\, dP$$

with equality if and only if $P = Q$.

Proof of Lemma 6.12. The asserted inequality is equivalent to the claim that the so-called *Kullback-Leibler divergence* $\int \log(p/q)\,dP$ is greater than or equal to zero, with equality if and only if $P = Q$. To verify this, we write

$$\begin{aligned}\int \log(p/q)\,dP &= -\int_{\{p>0\}} \log(q/p)\,p\,dM \\ &= -\int_{\{p>0\}} \log\Big(1+\frac{q-p}{p}\Big)\,p\,dM \\ &\geq -\int_{\{p>0\}} (q-p)\,dM \\ &= 1 - Q(\{p>0\}) \\ &\geq 0.\end{aligned}$$

Here we used the inequality $\log(1+t) \leq t$ for all $t \geq -1$, which is strict whenever $t \neq 0$. Hence the equality $\int \log(p/q)\,dP = 0$ implies that $M\big(\{p>0\} \cap \{p \neq q\}\big) = 0$. This implies already that $Q = P$ on the set $\{p > 0\}$. But since $1 = P(\{p>0\})$ and $Q(\mathcal{Y}) = 1$, this implies that $Q(\{p=0\}) = 0$, whence $Q = P$. □

Independent, identically distributed random variables. Suppose we observe stochastically independent copies $Y_1, Y_2, \ldots, Y_n$ of Y. Under the assumption that the distribution of Y equals P_{θ_o} for some unknown $\theta_o \in \Theta$, the distribution of the vector $\boldsymbol{Y} = (Y_i)_{i=1}^n$ is given by one of the density functions

$$\mathcal{Y}^n \ni \boldsymbol{y} \;\mapsto\; p_\theta(\boldsymbol{y}) := \prod_{i=1}^n p_\theta(y_i)$$

with respect to the product measure $M^{\otimes n}$ on $(\mathcal{Y}^n, \mathcal{B}^{\otimes n})$. The corresponding negative log-likelihood function for $\boldsymbol{Y}$ is given by

$$L(\theta, \boldsymbol{Y}) \;=\; -\sum_{i=1}^n \log p_\theta(Y_i).$$

In other words,

$$n^{-1} L(\theta, \boldsymbol{Y}) \;=\; \widehat{R}(\theta) = \int L(\theta, y)\,\widehat{P}(dy),$$

and this may be viewed as an estimator for $R(\theta) = \int L(\theta, y)\,P(dy)$. Hence, maximum likelihood estimation of θ_o, based on the observation vector $\boldsymbol{Y}$, is equivalent to minimization of the empirical risk $\widehat{R}(\cdot)$ for the single observations.

Example 6.13 (Bernoulli variables and binomial distributions). Let $\mathcal{Y} = \{0,1\}$, and set $p := \mathbb{P}(Y=1)$, i.e. $P = \mathrm{Bin}(1,p)$. Further let $\Theta = [0,1]$ and $P_\theta = \mathrm{Bin}(1,\theta)$. The weight function p_θ of P_θ is given by

$$p_\theta(0) \;=\; 1-\theta \quad \text{and} \quad p_\theta(1) \;=\; \theta.$$

Consequently,

$$L(\theta, y) \;=\; -(1-y)\log(1-\theta) - y\log\theta$$

and

$$R(\theta) \;=\; -(1-p)\log(1-\theta) - p\log\theta.$$

Since

$$R'(\theta) \;=\; \frac{\theta - p}{\theta(1-\theta)} \begin{cases} < 0 & \text{if } 0 < \theta < p, \\ > 0 & \text{if } p < \theta < 1, \end{cases}$$

the unique minimizer of $R(\cdot)$ is p.

If we observe independent copies $Y_1, Y_2, \ldots, Y_n$ of Y, then $\widehat{p} := n^{-1} \sum_{i=1}^n Y_i$ is a natural estimator for p. This is also the maximum likelihood estimator based on $\boldsymbol{Y}$ and the minimizer of the empirical risk $\widehat{R}(\cdot)$, because

$$n^{-1} L(\theta, \boldsymbol{Y}) = \widehat{R}(\theta) \;=\; -(1-\widehat{p}) \log(1-\theta) - \widehat{p} \log \theta.$$

Exercise 6.14 (Hardy–Weinberg model). We consider a population of diploid organisms and a particular gene with two potential alleles `A` and `a`. With respect to this gene, the individuals have a genotype in $\mathcal{Y} := \{\mathtt{AA}, \mathtt{Aa}, \mathtt{aa}\}$.

For a randomly chosen individual from that population, let Y be its genotype and $p(z) := \mathbb{P}(Y = z)$ for $z \in \mathcal{Y}$. Theoretical considerations (Hardy–Weinberg law) suggest that $p(\cdot) = p_{\theta_o}(\cdot)$ for an unknown parameter $\theta_o \in [0,1]$, where

$$p_\theta(\mathtt{AA}) \;:=\; \theta^2, \quad p_\theta(\mathtt{Aa}) \;:=\; 2\theta(1-\theta) \quad \text{and} \quad p_\theta(\mathtt{aa}) \;:=\; (1-\theta)^2.$$

Determine for this example the risk function $R(\theta) \;:=\; -\,\mathbb{E} \log p_\theta(Y)$ and its minimizer for arbitrary weight function $p(\cdot)$. That is, do not assume a priori that $p(\cdot) = p_{\theta_o}(\cdot)$ for some $\theta_o \in [0,1]$. Then determine the maximum likelihood estimator for θ_o, based on independent copies $Y_1, Y_2, \ldots, Y_n$ of Y.

Sample location parameters as maximum likelihood estimators. In Section 6.1, we represented the sample mean and sample quantiles as minimizers of empirical risk functions. They may also be viewed as maximum likelihood estimators, if we choose appropriate statistical models $(P_\theta)_{\theta \in \mathbb{R}}$. In general, let p_0 be a strictly positive probability density on $\mathbb{R}$, and for $\theta \in \mathbb{R}$ let P_θ be the distribution with density function

$$p_\theta \;:=\; p_0(\cdot - \theta).$$

Then a maximum likelihood estimator $\widehat{\theta} = \widehat{\theta}(\boldsymbol{Y})$ is a minimizer of

$$L(\theta, \boldsymbol{Y}) \;=\; -\sum_{i=1}^n \log p_0(Y_i - \theta).$$

This is also a minimizer of $\widehat{R}(\theta) = n^{-1} \sum_{i=1}^n \rho(\theta - Y_i)$ for a given convex, coercive function $\rho : \mathbb{R} \to \mathbb{R}$, provided that

$$p_0(y) \;=\; c_1 \exp(-c_2 \rho(-y))$$

for certain constants $c_1, c_2 > 0$.

In case of $\rho(t) = t^2$, we obtain the density function p_0 of a centered Gaussian distribution. In case of $\rho(t) = |t|$, we are dealing with Laplace distributions, while in case of $\rho(t) = (1-2\gamma)t + |t|$, we obtain non-symmetric Laplace distributions.

Exercise 6.15. Show that the function

$$\rho(t) \;:=\; \log(1+\cosh(t))$$

is strictly convex with $\lim_{|t|\to\infty} \rho(t) = \infty$. Show that the minimization of $\sum_{i=1}^n \rho(\theta - Y_i)$ with respect to $\theta \in \mathbb{R}$ corresponds to maximimum likelihood estimation for certain families of logistic distributions. The logistic distribution with mean μ and scale parameter $\sigma > 0$ is given by the density function $p_{\mu,\sigma}(y) := p_{0,1}((y-\mu)/\sigma)/\sigma$, where

$$p_{0,1}(t) \;:=\; \frac{e^t}{(1+e^t)^2} \;=\; \frac{1}{e^t + e^{-t} + 2}.$$

6.4 Application to Regression Problems

In regression settings, we consider observation pairs $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ and want to model and estimate the *conditional distributions* $\mathcal{L}(Y \mid X = x)$, $x \in \mathcal{X}$, or some aspects thereof. For a given decision space $\mathbb{V}$ and a given loss function $L : \mathbb{V} \times \mathcal{Y} \to \mathbb{R}$, we are looking for a regression function $f_o : \mathcal{X} \to \mathbb{V}$ such that

$$\mathbb{E}\, L(f_o(X), Y)$$

is minimal.

Concerning the distribution of X, we do not want to restrict ourselves. It should be possible that X or at least some components of X may be chosen arbitrarily by the experimenter. Thus we focus on the conditional distributions $\mathcal{L}(Y \mid X = x)$, $x \in \mathcal{X}$, and aim for functions f_o satisfying

$$f_o(x) \;\in\; \operatorname*{arg\,min}_{v \in \mathbb{V}} \mathbb{E}\big(L(v,Y) \,\big|\, X = x\big) \quad \text{for all } x \in \mathcal{X}.$$

For instance, in case of $\mathcal{Y} = \mathbb{V} = \mathbb{R}$ and $L(v,y) = \rho(v-y)$,

$$f_o(x) \;=\; \begin{cases} \mathbb{E}(Y \mid X = x) & \text{if } \rho(t) = t^2, \\ \text{Median}(Y \mid X = x) & \text{if } \rho(t) = |t|. \end{cases}$$

Now we want to estimate this optimal regression function f_o from independent observations (X_1,Y_1), (X_2,Y_2), …, (X_n,Y_n), where the X_i are treated as fixed points in $\mathcal{X}$, and $\mathcal{L}(Y_i) = \mathcal{L}(Y \mid X = X_i)$ for $i = 1,\dots,n$. Obviously, f_o minimizes the risk

$$R(f) = R(f, \boldsymbol{X}) \;:=\; \sum_{i=1}^n \mathbb{E}\, L(f(X_i), Y_i)$$

among all functions $f : \mathcal{X} \to \mathbb{V}$. Typically, this optimality property determines f_o uniquely on the set $\{X_1, X_2, \dots, X_n\}$, but at points $x \notin \{X_1,\dots,X_n\}$, the value $f_o(x)$ is only specified under additional model assumptions.

A naive estimator for f_o is given by a minimizer of the empirical risk, i.e. the observed loss

$$\left.\begin{matrix} \widehat{R}(f) = \widehat{R}(f, \boldsymbol{X}, \boldsymbol{Y}) \\ L(f) = L(f, \boldsymbol{X}, \boldsymbol{Y}) \end{matrix}\right\} \;:=\; \sum_{i=1}^n L(f(X_i), Y_i)$$

among all functions $f : \mathcal{X} \to \mathbb{V}$. But the resulting estimator is often useless. For instance, suppose that all X_i are pairwise different, that $\mathcal{Y} = \mathbb{V} = \mathbb{R}$ and $L(v, y) = \rho(v-y)$ with $\arg\min_{t\in\mathbb{R}} \rho(t) = \{0\}$. Then f minimizes the empirical risk $\widehat{R}(f)$ if and only if $f(X_i) = Y_i$ for $1 \le i \le n$. Two strategies with higher potential are:

(i) Restricting f to a particular family $\mathcal{F}$ of functions.

(ii) Minimizing

$$\widehat{R}(f) + \mathrm{Pen}(f)$$

in place of $\widehat{R}(f)$, where $\mathrm{Pen}(f)$ is a penalty term quantifying the "irregularity" of f.

Both strategies (i) and (ii) appeared already in the context of least squares estimation. Approach (ii) is called "regularization" or "penalization"; see Section 5.3.

Example 6.16 (Regression quantiles). We illustrate strategy (i) above with regression quantiles, introduced by Koenker and Basset (1982), without going into computational details. We consider observation pairs $(X, Y) \in [A, B] \times \mathbb{R}$ and want to estimate the γ-quantile $f_\gamma(x)$ of $\mathcal{L}(Y \mid X = x)$ for various values of $\gamma \in (0, 1)$ and $x \in [A, B]$. To this end, we assume that f_γ is an element of a given finite-dimensional vector space $\mathcal{F}$ of functions on $[A, B]$. Then we estimate f_γ by

$$\widehat{f_\gamma} \;\in\; \operatorname*{arg\,min}_{f\in\mathcal{F}} \sum_{i=1}^{n} \rho_\gamma(f(X_i) - Y_i),$$

where $\rho_\gamma(t) := (1 - 2\gamma)t + |t|$.

We applied this method to the baseball data from Example 4.8. Again, we consider the decimal logarithms of the yearly salaries (Y) and the number of seasons (X), where for one observation the value $X = 24$ has been replaced with $X = 20$. Figure 6.1 shows regression quantiles $\widehat{f_\gamma}$ for $\gamma = 0.1, 0.25, 0.5, 0.75, 0.9$, where $\mathcal{F}$ consists of all functions $f(x) = \sum_{i=0}^{3} a_i \log_{10}(x)^i$ with real parameters a_0, a_1, a_2, a_3. Figure 6.2 shows the resulting regression quantiles $\widehat{f_\gamma}$ in case of $\mathcal{F} = \mathcal{S}_3(0.5, 4.0, 11.0, 20.5)$. Close to the boundaries 0.5 and 20.5 one sees a weakness of this method: There is no guarantee that $\widehat{f_\gamma} \le \widehat{f_\eta}$ for $0 < \gamma < \eta < 1$.

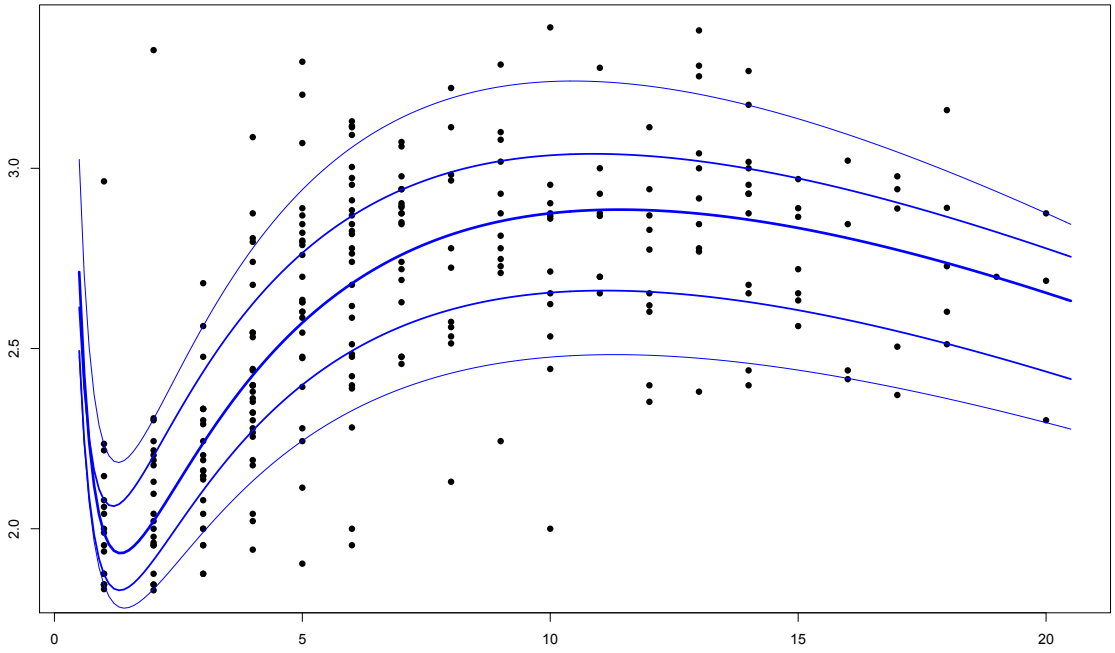

Figure 6.1: Regression quantiles for Baseball data in cubic model.

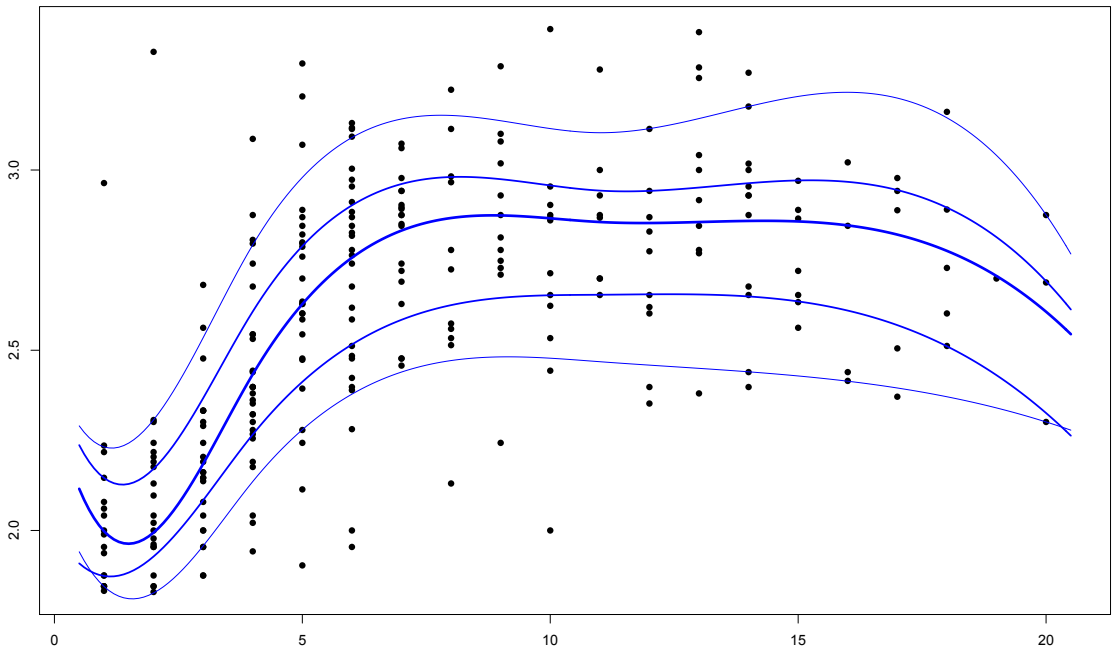

Figure 6.2: Regression quantiles for Baseball data in spline model.

Chapter 7

Logistic Regression and Related Models

In this chapter, we focus on categorical response variables Y with values in a finite set $\mathcal{Y}$. The goal is to model and estimate the conditional distributions $\mathcal{L}(Y \mid X = x)$, $x \in \mathcal{X}$. Our starting point will be a dichotomous variable Y.

7.1 Logistic Regression

Let $\mathcal{Y} = \{0, 1\}$. Here are two explicit examples for a dichotomous response Y:

• Credit defaults: Let $X \in \mathcal{X}$ describe various properties of a potential customer of a bank, i.e. a private person or some company, who is applying for a loan. This feature vector X may also include details of the loan itself, e.g. its amount. Then let Y indicate whether this customer will fail to pay the loan.

• Successes or failures of medical treatments: Let X be a vector of covariates containing features of a patient as well as specific information about an upcoming medical treatment for him or her. The response Y indicates whether the treatment will be successful or not.

The conditional distribution $\mathcal{L}(Y \mid X = x)$ is completely determined by the conditional probability

$$p(x) \;:=\; \mathbb{P}(Y = 1 \mid X = x) \;=\; \mathbb{E}(Y \mid X = x).$$

Since this number is always within $[0, 1]$, a linear model for $p(\cdot)$ doesn't make sense. Note also that the observations are heteroscedastic in general, because

$$\mathrm{Var}(Y \mid X = x) \;=\; p(x)(1 - p(x)).$$

A possible way out is to choose a monotone increasing, bijective mapping $\ell : \mathbb{R} \to (0, 1)$ and to assume that

$$p(x) \;=\; \ell(f(x))$$

with a regression function $f : \mathcal{X} \to \mathbb{R}$ as before. That means, f is an element of a finite-dimensional linear space $\mathcal{F}$ of real-valued functions on $\mathcal{X}$.

Concerning the so-called *inverse link function* ℓ, two choices are particularly popular:

- **Probit regression** with the standard Gaussian distribution function,

$$\ell(v) \;=\; \Phi(v).$$

- **Logistic regression** with the logistic function,

$$\ell(v) \;=\; \frac{\exp(v)}{1+\exp(v)} \;=\; \frac{1}{\exp(-v)+1}.$$

Here

$$f(x) \;=\; \mathrm{logit}(p(x)),$$

where

$$\mathrm{logit}(u) \;:=\; \log\Big(\frac{u}{1-u}\Big), \quad 0<u<1.$$

In the present section, we consider only logistic regression. A theoretical justification for this particular model will be given later. The regression function f can be interpreted in terms of odds and odds ratios: The odds of $Y=1$, given that $X=x$, are equal to

$$\frac{\mathbb{P}(Y=1\,|\,X=x)}{\mathbb{P}(Y=0\,|\,X=x)} \;=\; \frac{p(x)}{1-p(x)} \;=\; \exp(f(x)).$$

For two different points $x_1, x_2 \in \mathcal{X}$ we obtain the odds ratio

$$\frac{p(x_1)}{1-p(x_1)}\Big/\frac{p(x_2)}{1-p(x_2)} \;=\; \frac{p(x_1)\,(1-p(x_2))}{(1-p(x_1))\,p(x_2)} \;=\; \exp\big(f(x_1)-f(x_2)\big).$$

7.1.1 Maximum Likelihood Estimation

As in previous chapters, we consider stochastically independent observations (X_1,Y_1), (X_2,Y_2), $\ldots$, (X_n,Y_n) in $\mathcal{X}\times\{0,1\}$, where the X_i are viewed as fixed points (via conditioning, if necessary) while $\mathbb{P}(Y_i=1)=p(X_i)$ for some unknown function $p:\mathcal{X}\to[0,1]$.

Log-likelihood function and maximum likelihood estimator. Under the assumption that $p=\ell\circ f_*$ for some unknown function $f_*\in\mathcal{F}$, we obtain the following *negative log-likelihood function* $L=L(\cdot\,|\,\boldsymbol{X},\boldsymbol{Y}):\mathcal{F}\to\mathbb{R}$:

$$\begin{aligned} L(f) \;&:=\; -\sum_{i=1}^n \Big(Y_i\log\ell(f(X_i)) + (1-Y_i)\log\big(1-\ell(f(X_i))\big)\Big) \\ &=\; \sum_{i=1}^n \Big(\log\big(1+\exp(f(X_i))\big) - Y_i f(X_i)\Big). \end{aligned}$$

A *maximum likelihood estimator (MLE)* for f_* is any minimizer $\widehat{f}$ of the negative log-likelihood function,

$$\widehat{f}=\widehat{f}(\cdot\,|\,\boldsymbol{X},\boldsymbol{Y}) \;\in\; \operatorname*{arg\,min}_{f\in\mathcal{F}} L(f).$$

As we shall see later, $\widehat{f}$ is typically uniquely determined. The plausibility of this approach becomes visible by considering the expected negative log-likelihood function $R = R(\cdot \,|\, \boldsymbol{X}) : \mathcal{F} \to \mathbb{R}$:

$$\begin{aligned} R(f) \;&:=\; \mathbb{E}\, L(f) \\ &=\; -\sum_{i=1}^{n}\Big(p(X_i)\log \ell(f(X_i)) + (1-p(X_i))\log\big(1-\ell(f(X_i))\big)\Big) \\ &=\; \sum_{i=1}^{n}\Big(\log\big(1+\exp(f(X_i))\big) - p(X_i)f(X_i)\Big). \end{aligned}$$

Indeed, for fixed $p \in [0,1]$, the unique minimizer of $[0,1] \ni \theta \mapsto -p\log\theta - (1-p)\log(1-\theta)$ is equal to p; see Chapter 6. Consequently, if indeed $p = \ell \circ f_*$ for some $f_* \in \mathcal{F}$, then $R(f_*) \le R(f)$ with equality if and only if $f(\boldsymbol{X}) = f_*(\boldsymbol{X})$.

Exercise 7.1. Verify that the negative log-likelihood function can be written as

$$L(f) \;=\; \sum_{i=1}^{n} h\big((2Y_i - 1)f(X_i)\big)$$

with $h(r) := -\log \ell(r)$, $r \in \mathbb{R}$. Show also that h is strictly decreasing and strictly convex with $\lim_{r\to\infty} h(r) = 0$.

(Note that this representation is true for *any* link funktion $\ell : \mathbb{R} \to (0,1)$ such that $\ell(-v) = 1 - \ell(v)$ for $v \in \mathbb{R}$.)

Exercise 7.2. Suppose that the linear space $\mathcal{F}$ contains all constant functions. Show that an MLE $\widehat{f} \in \mathcal{F}$ satisfies necessarily the equation

$$\sum_{i=1}^{n} \ell(\widehat{f}(X_i)) \;=\; \sum_{i=1}^{n} Y_i.$$

Generalize this conclusion to other functions $g \in \mathcal{F}$.

Existence and uniqueness of the MLE. Suppose we have chosen basis functions $f_1, \ldots, f_p$ of $\mathcal{F}$. Then any function $f \in \mathcal{F}$ may be written as $f = \sum_{j=1}^{p} \theta_j f_j$ for some vector $\boldsymbol{\theta} = (\theta_j)_{j=1}^{p} \in \mathbb{R}^p$, and $f(X_i) = \boldsymbol{d}_i^\top \boldsymbol{\theta}$, where

$$\boldsymbol{d}_i \;:=\; (f_j(X_i))_{j=1}^{p}.$$

These vectors form the design matrix $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]^\top \in \mathbb{R}^{n\times p}$. Now, both L and R may be re-interpreted as functions on $\mathbb{R}^p$ and have the form

$$\widetilde{L}(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^{n}\big(a_i \boldsymbol{d}_i^\top \boldsymbol{\theta} - \log(1+\exp(\boldsymbol{d}_i^\top \boldsymbol{\theta}))\big)$$

with certain numbers $a_i \in [0,1]$, namely, $a_i = Y_i$ if $\widetilde{L} = L$ and $a_i = p(X_i)$ if $\widetilde{L} = R$. One can easily verify that $\ell(\cdot)$ is the derivative of the function $t \mapsto \log(1+\exp(t))$. This implies that gradient and Hessian matrix of $\tilde{L}$ are given by

$$\nabla \widetilde{L}(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^{n}(a_i - \ell(\boldsymbol{d}_i^\top \boldsymbol{\theta}))\, \boldsymbol{d}_i$$

and

$$D^2\widetilde{L}(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta})\, \boldsymbol{d}_i \boldsymbol{d}_i^\top,$$

respectively. Here the derivative of ℓ equals

$$\ell' \;=\; \ell(1-\ell) \;\in\; (0, 1/4].$$

Consequently, for arbitrary vectors $\boldsymbol{v} \in \mathbb{R}^p$,

$$\boldsymbol{v}^\top D^2\widetilde{L}(\boldsymbol{\theta})\boldsymbol{v} \;=\; \sum_{i=1}^{n} \ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta})(\boldsymbol{v}^\top \boldsymbol{d}_i)^2 \;\geq\; 0$$

with equality if and only if $\boldsymbol{v}$ is perpendicular to the space $\operatorname{span}(\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n)$. Consequently, $D^2\widetilde{L}$ is always positive semidefinite, whence $\widetilde{L}$ is a convex function.

Concerning strict convexity, suppose that $\operatorname{span}(\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n) = \mathbb{R}^p$, that means, the design matrix $\boldsymbol{D}$ has full column rank,

$$\operatorname{rank}(\boldsymbol{D}) \;=\; p. \tag{7.1}$$

Then the Hessian matrix $D^2\widetilde{L}(\boldsymbol{\theta})$ is positive definite for any $\boldsymbol{\theta} \in \mathbb{R}^p$, a sufficient condition for strict convexity of $\widetilde{L}$. If (7.1) is violated, there exists a nonzero vector $\boldsymbol{v} \in \mathbb{R}^p$ such that $\boldsymbol{v}^\top D^2\widetilde{L}(\boldsymbol{\theta})\boldsymbol{v} = 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$. But this implies that for any fixed $\boldsymbol{\eta} \in \mathbb{R}^p$, $\widetilde{L}(\boldsymbol{\eta} + t\boldsymbol{v})$ is linear in $t \in \mathbb{R}$, so $\widetilde{L}$ fails to be strictly convex. Hence condition (7.1) is necessary and sufficient for strict convexity of $\widetilde{L}$.

Under the assumption (7.1), there exists a unique or no minimizer of $\widetilde{L}$. As we shall see later in Lemma 7.7, a minimizer exists if and only if

$$\limsup_{r\to\infty} \nabla\widetilde{L}(r\boldsymbol{u})^\top \boldsymbol{u} \;>\; 0 \quad \text{for any } \boldsymbol{u} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}. \tag{7.2}$$

But

$$\nabla\widetilde{L}(r\boldsymbol{u})^\top \boldsymbol{u} \;=\; \sum_{i=1}^{n} (\ell(r\boldsymbol{d}_i^\top \boldsymbol{u}) - a_i)\boldsymbol{d}_i^\top \boldsymbol{u} \;\to\; \sum_{i=1}^{n} \big(1_{[\boldsymbol{d}_i^\top \boldsymbol{u} \geq 0]} - a_i\big)\boldsymbol{d}_i^\top \boldsymbol{u} \quad (r \to \infty),$$

because $\lim_{t\to-\infty} \ell(t) = 0$ and $\lim_{t\to\infty} \ell(t) = 1$. Consequently, (7.2) is equivalent to the condition

$$\sum_{i=1}^{n} \big(1_{[\boldsymbol{d}_i^\top \boldsymbol{u} \geq 0]} - a_i\big)\boldsymbol{d}_i^\top \boldsymbol{u} \;>\; 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}. \tag{7.3}$$

Since all summands $\big(1_{[\boldsymbol{d}_i^\top \boldsymbol{u} \geq 0]} - a_i\big)\boldsymbol{d}_i^\top \boldsymbol{u}$ in (7.3) are non-negative, one may reformulate this condition as follows: There exists *no* vector $\boldsymbol{u} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}$ such that

$$\begin{cases} a_i = 1 & \text{if } \boldsymbol{d}_i^\top \boldsymbol{u} > 0, \\ a_i = 0 & \text{if } \boldsymbol{d}_i^\top \boldsymbol{u} < 0. \end{cases} \tag{7.4}$$

On the other hand, if (7.1) is satisfied and if there exists a vector $\boldsymbol{u} \neq \boldsymbol{0}$ with property (7.4), then there exists no minimizer of $\widetilde{L}$.

Special case: Non-degenerate $p(\cdot)$. Suppose that condition (7.1) is satisfied, and let $0 < p(X_i) < 1$ for all i. Then the function $\widetilde{L} = R$ has a unique minimizer. Indeed, condition (7.4) would imply that for some vector $\boldsymbol{u} \neq \mathbf{0}$, $\boldsymbol{d}_i^\top \boldsymbol{u} = 0$ for all i, a contradiction to condition (7.1).

Special case: Multiple logistic regression. Suppose we observe $(\boldsymbol{X}_i, Y_i) \in \mathbb{R}^d \times \{0,1\}$, and let $\mathcal{F}$ consist of all affine functions f. That means, $f(\boldsymbol{x}) = a + \boldsymbol{b}^\top \boldsymbol{x}$ for certain parameters $a \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^d$. With $\boldsymbol{\theta} := [a, \boldsymbol{b}^\top]^\top$ we obtain the design vectors $\boldsymbol{d}_i = [1, \boldsymbol{X}_i^\top]^\top$. Now one can show that condition (7.1) is equivalent to the requirement that the vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ should not lie on a hyperplane in $\mathbb{R}^d$, see Exercise 7.4.

Under assumption (7.1), there exists a unique minimizer $\widehat{\boldsymbol{\theta}}$ of L if and only if there is no hyperplane in $\mathbb{R}^d$ separating the two sets $\{\boldsymbol{X}_i : Y_i = 0\}$ and $\{\boldsymbol{X}_i : Y_i = 1\}$ (weakly). That means, there may not exist a pair $(\boldsymbol{v}, r) \in (\mathbb{R}^d \setminus \{\mathbf{0}\}) \times \mathbb{R}$ such that

$$\{\boldsymbol{X}_i : Y_i = 0\} \subset \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{v}^\top \boldsymbol{x} \leq r\} \quad \text{and} \quad \{\boldsymbol{X}_i : Y_i = 1\} \subset \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{v}^\top \boldsymbol{x} \geq r\}.$$

In particular, the values $Y_1, Y_2, \ldots, Y_n$ may not be identical.

Exercise 7.3. Let $\boldsymbol{X} = (X_i)_{i=1}^n$ be a fixed vector with pairwise different components $X_i \in \mathbb{R}$, and let $\boldsymbol{Y} = (Y_i)_{i=1}^n$ be a vector of independent Bernoulli random variables with parameter $p \in (0,1)$. Compute the probability that the MLE for the standard model (all affine functions on $\mathbb{R}$) does not exist. What is its value in case of $p = 1/2$?

Exercise 7.4. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be vectors in $\mathbb{R}^d$, and let $\boldsymbol{D} := [\boldsymbol{d}_1\, \boldsymbol{d}_2\, \ldots\, \boldsymbol{d}_n]^\top$ with $\boldsymbol{d}_i := [1, \boldsymbol{x}_i^\top]^\top$. Show that the following four conditions are equivalent:

(i) $\operatorname{rank}(\boldsymbol{D}) \leq d$.

(ii) $\operatorname{span}(\boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{x}_3 - \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n - \boldsymbol{x}_1) \neq \mathbb{R}^d$.

(iii) $\operatorname{span}(\boldsymbol{x}_1 - \bar{\boldsymbol{x}}, \boldsymbol{x}_2 - \bar{\boldsymbol{x}}, \boldsymbol{x}_3 - \bar{\boldsymbol{x}}, \ldots, \boldsymbol{x}_n - \bar{\boldsymbol{x}}) \neq \mathbb{R}^d$, where $\bar{\boldsymbol{x}} := n^{-1} \sum_{i=1}^n \boldsymbol{x}_i$.

(iv) The vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ lie on a hyperplane in $\mathbb{R}^d$.

Exercise 7.5 (Newton procedure and iteratively reweighted least squares). Recall that the negative log-likelihood $L(\cdot)$ for logistic regression has the following first and second derivatives:

$$\nabla L(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^n (Y_i - \ell(\boldsymbol{d}_i^\top \boldsymbol{\theta}))\boldsymbol{d}_i, \quad D^2 L(\boldsymbol{\theta}) \;=\; \sum_{i=1}^n \ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta}) \boldsymbol{d}_i \boldsymbol{d}_i^\top.$$

(a) Determine the minimizer $\boldsymbol{h}_*$ of the second order Taylor approximation

$$\boldsymbol{h} \;\mapsto\; L(\boldsymbol{\theta}) + \nabla L(\boldsymbol{\theta})^\top \boldsymbol{h} + \frac{1}{2} \boldsymbol{h}^\top D^2 L(\boldsymbol{\theta}) \boldsymbol{h}$$

of $L(\boldsymbol{\theta} + \boldsymbol{h})$. This leads to the following algorithm: If $\boldsymbol{\theta}$ is our current candidate for $\widehat{\boldsymbol{\theta}}$, then $\boldsymbol{\psi}(\boldsymbol{\theta}) := \boldsymbol{\theta} + \boldsymbol{h}_*$ is the next and hopefully better candidate.

(b) Alternatively, one could consider the following target function: If $\boldsymbol{\theta}$ is our current candidate for $\widehat{\boldsymbol{\theta}}$, then we want to determine $\boldsymbol{h}$ such that

$$\sum_{i=1}^n \frac{\big(Y_i - \ell(\boldsymbol{d}_i^\top (\boldsymbol{\theta} + \boldsymbol{h}))\big)^2}{\ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta})}$$

is minimal. (Note that $\ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta}) = \mathrm{Var}(Y_i)$ if $\mathbb{P}(Y_i = 1) = \ell(\boldsymbol{d}_i^\top \boldsymbol{\theta})$.)

Suppose we replace $\ell(\boldsymbol{d}_i^\top(\boldsymbol{\theta}+\boldsymbol{h}))$ with $\ell(\boldsymbol{d}_i^\top \boldsymbol{\theta}) + \ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta})\boldsymbol{d}_i^\top \boldsymbol{h}$. Show that the resulting minimizer $\boldsymbol{h}_*$ is given by

$$\operatorname*{arg\,min}_{\boldsymbol{h}\in\mathbb{R}^p} \sum_{i=1}^n w_i \big(\tilde{Y}_i - \boldsymbol{d}_i^\top \boldsymbol{h}\big)^2$$

where $w_i := \ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta}) > 0$ and $\tilde{Y}_i := (Y_i - \ell(\boldsymbol{d}_i^\top \boldsymbol{\theta}))/w_i$.

(c) Show that the optimal vectors in (a) and (b) coincide.

(d) Verify that $\boldsymbol{h}_*$ may be determined with the following R commands:

$$\begin{aligned}
\boldsymbol{a} &\leftarrow \boldsymbol{D} \,\%*\%\, \boldsymbol{\theta},\\
\boldsymbol{p} &\leftarrow 1/(1+\mathsf{exp}(-\boldsymbol{a})),\\
\boldsymbol{v} &\leftarrow 1/\mathsf{sqrt}(2+\mathsf{exp}(\boldsymbol{a})+\mathsf{exp}(-\boldsymbol{a})),\\
\boldsymbol{h}_* &\leftarrow \mathsf{qr.solve}\big(\boldsymbol{v} * \boldsymbol{D}, (\boldsymbol{Y}-\boldsymbol{p})/\boldsymbol{v}\big),
\end{aligned}$$

where $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]^\top$ and $\boldsymbol{Y} = (Y_i)_{i=1}^n$.

(e) Parts (b) and (d) yield an iterative scheme, in which each step involves the minimization of a weighted sum of squares. (In the appendix, this approach is explained within a broader context.) Write your own program to compute the MLE $\widehat{\boldsymbol{\theta}}$.

Exercise 7.6 (Logistic regression with mis-specified model). Simulate a sample $\boldsymbol{X} = (X_i)_{i=1}^{100}$ from the standard Gaussian distribution. Then set $Y_i := F(X_i)$, where $F : \mathbb{R} \to [0,1]$ is given by

(a) $F(x) = \ell(x) = \exp(x)/(1+\exp(x))$;

(b) $F(x) = \Phi(x)$ (standard Gaussian distribution function);

(c) $F(x) = \max(0, 1-\exp(-x))$;

(d) $F(x) = \max\big(0, \min\big(1, (x+1)/2\big)\big)$.

Fit a standard logistic model, i.e. $F(x) = \ell(a+bx)$ for some real parameters a and b, to your "data" with R. Note that for the R function glm, the response variable Y may take values in $[0,1]$; just ignore the corresponding warning message. Then plot the true function F together with the fitted logistic function $\ell(\widehat{a} + \widehat{b}\,\cdot)$.

Lemma 7.7 (Coercivity of convex functions). *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex function. Then the following three properties of f are equivalent:*

(i) *The function f is coercive, that is,*

$$f(\boldsymbol{\theta}) \;\to\; \infty \quad \text{as } \|\boldsymbol{\theta}\| \to \infty.$$

(ii) *For any fixed $\boldsymbol{u} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}$,*

$$f(r\boldsymbol{u}) \;>\; f(\boldsymbol{0}) \quad \text{for sufficiently large } r > 0.$$

(iii) *The set of minimizers of f is a compact (and thus nonempty) set.*

If f is even differentiable, property (ii) is equivalent to the following one:

(ii') *For any fixed $\boldsymbol{u} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$,*

$$\lim_{r\to\infty} \nabla f(r\boldsymbol{u})^\top \boldsymbol{u} \;\in\; (0,\infty].$$

Proof of Lemma 7.7. Note first that f is necessarily continuous, which is a well-known fact from convex analysis.

We first show that conditions (i) and (ii) are equivalent. Suppose that condition (i) is satisfied. Then, for sufficiently large $R > 0$,

$$\gamma := \inf_{\boldsymbol{\theta}\in\mathbb{R}^p:\ \|\boldsymbol{\theta}\|\ge R} \big(f(\boldsymbol{\theta}) - f(\mathbf{0})\big) \;>\; 0.$$

In particular, for $\boldsymbol{u} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$ the difference $f(r\boldsymbol{u}) - f(\mathbf{0})$ is at least γ whenever $r \ge R/\|\boldsymbol{u}\|$. Thus, condition (ii) is satisfied too.

Suppose that condition (i) is violated. That means, there exist a real threshold γ and a sequence $(\boldsymbol{\theta}_n)_{n\ge1}$ in $\mathbb{R}^p \setminus \{\mathbf{0}\}$ such that $\lim_{n\to\infty} \|\boldsymbol{\theta}_n\| = \infty$ but $f(\boldsymbol{\theta}_n) \le \gamma$ for all $n \ge 1$. Since the unit sphere in $\mathbb{R}^p$ is compact, we may even assume that $\boldsymbol{u}_n := \|\boldsymbol{\theta}_n\|^{-1}\boldsymbol{\theta}_n$ converges to a unit vector $\boldsymbol{u}$ as $n \to \infty$. But for arbitrary numbers $r > 0$, the numbers $\lambda_n := r/\|\boldsymbol{\theta}_n\| > 0$ satisfy $\lim_{n\to\infty} \lambda_n = 0$, so continuity and convexity of f imply that

$$\begin{aligned}
f(r\boldsymbol{u}) - f(\mathbf{0}) \;&=\; \lim_{n\to\infty} \big(f(r\boldsymbol{u}_n) - f(\mathbf{0})\big) \\
&=\; \lim_{n\to\infty} \big(f((1-\lambda_n)\mathbf{0} + \lambda_n\boldsymbol{\theta}_n) - f(\mathbf{0})\big) \\
&\le\; \lim_{n\to\infty} \big((1-\lambda_n)f(\mathbf{0}) + \lambda_n f(\boldsymbol{\theta}_n) - f(\mathbf{0})\big) \\
&\le\; \lim_{n\to\infty} \lambda_n(\gamma - f(\mathbf{0})) \\
&=\; 0.
\end{aligned}$$

Hence, condition (ii) is violated too.

Now suppose for the moment that f is even differentiable. To prove equivalence of conditions (ii) and (ii'), note that for any fixed $\boldsymbol{u} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$, the function $r \mapsto h(r) := f(r\boldsymbol{u})$ is convex on $\mathbb{R}$ with derivative $h'(r) = \nabla f(r\boldsymbol{u})^\top \boldsymbol{u}$. Since h' is nondecreasing, the limit $\lim_{r\to\infty} h'(r)$ exists in $(-\infty,\infty]$. If condition (ii) is satisfied, then $h(r) > h(0)$ for some $r > 0$, and convexity of h implies that

$$\nabla f(r\boldsymbol{u})^\top \boldsymbol{u} = h'(r) \;\ge\; \frac{h(r) - h(0)}{r} \;>\; 0,$$

which proves condition (ii'). If condition (ii') is satisfied, then for sufficiently large $r > 0$, the derivative $h'(r)$ is strictly positive. But then convexity of h implies that for $s > r$,

$$f(s\boldsymbol{u}) = h(s) \;\ge\; h(r) + h'(r)(s-r) \;\to\; \infty \quad \text{as } s \to \infty,$$

so condition (ii) is satisfied too.

It remains to show that conditions (i) and (iii) are equivalent. Suppose that condition (i) is satisfied. Let $(\boldsymbol{\theta}_n)_{n\ge1}$ be a sequence in $\mathbb{R}^p$ such that $\lim_{n\to\infty} f(\boldsymbol{\theta}_n) = \gamma := \inf\{f(\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \mathbb{R}^p\}$. By

coercivity of f, this sequence is bounded, so we may replace it with a subsequence (if necessary) such that $\boldsymbol{\theta}_o := \lim_{n\to\infty} \boldsymbol{\theta}_n$ exists. Now, continuity of f implies that $f(\boldsymbol{\theta}_o) = \gamma$. In particular, the set of minimizers of f may be written as $\{f = f(\boldsymbol{\theta}_o)\}$. This set is nonempty, closed by continuity of f and bounded by coercivity of f. Hence condition (iii) is satisfied.

Suppose that condition (iii) is satisfied. Let $\boldsymbol{\theta}_o$ be a minimizer of f. If R is strictly larger than the maximum of $\|\boldsymbol{\theta} - \boldsymbol{\theta}_o\|$ over all minimizers of f, it follows from continuity of f that

$$\gamma := \min_{\boldsymbol{u}\in\mathbb{R}^p\colon \|\boldsymbol{u}\|=R} \big(f(\boldsymbol{\theta}_o + \boldsymbol{u}) - f(\boldsymbol{\theta}_o)\big) \;>\; 0.$$

But then it follows from convexity of f that for any $\boldsymbol{v} \in \mathbb{R}^p$ with $\|\boldsymbol{v}\| \geq R$ and $\boldsymbol{u} := R\|\boldsymbol{v}\|^{-1}\boldsymbol{v}$,

$$f(\boldsymbol{\theta}_o + \boldsymbol{v}) - f(\boldsymbol{\theta}_o) \;=\; f\Big(\boldsymbol{\theta}_o + \frac{\|\boldsymbol{v}\|}{R}\boldsymbol{u}\Big) - f(\boldsymbol{\theta}_o) \;\geq\; \frac{\|\boldsymbol{v}\|}{R}\big(f(\boldsymbol{\theta}_o + \boldsymbol{u}) - f(\boldsymbol{\theta}_o)\big) \;\geq\; \frac{\|\boldsymbol{v}\|\gamma}{R}.$$

This shows that $f(\boldsymbol{\theta}) \to \infty$ as $\|\boldsymbol{\theta} - \boldsymbol{\theta}_o\| \to \infty$. But since $\|\boldsymbol{\theta} - \boldsymbol{\theta}_o\|$ and $\|\boldsymbol{\theta}\|$ differ at most by $\|\boldsymbol{\theta}_o\|$, this is equivalent to condition (i). □

7.1.2 The Asymptotic Behavior of the Log-Likelihood Function

We consider a similar setting as at the end of the previous section, but embedded into a triangular scheme: After conditioning on covariates, if necessary, for each $n \in \mathbb{N}$ we observe $(\boldsymbol{d}_{n1}, Y_{n1})$, $(\boldsymbol{d}_{n2}, Y_{n2})$, …, $(\boldsymbol{d}_{nn}, Y_{nn})$ with fixed vectors $\boldsymbol{d}_{ni} \in \mathbb{R}^p$ and stochastically independent random variables $Y_{ni} \in \{0, 1\}$, and we write

$$p_{ni} \;:=\; \mathbb{P}(Y_{ni} = 1).$$

Now we investigate the asymptotic behaviour of the negative log-likelihood function $L_n : \mathbb{R}^p \to \mathbb{R}$ and its pointwise expectation R_n, that means,

$$\begin{aligned} L_n(\boldsymbol{\theta}) \;&=\; -\sum_{i=1}^{n} \big(Y_{ni}\boldsymbol{d}_{ni}^\top\boldsymbol{\theta} - \log(1 + \exp(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}))\big), \\ R_n(\boldsymbol{\theta}) \;&=\; -\sum_{i=1}^{n} \big(p_{ni}\boldsymbol{d}_{ni}^\top\boldsymbol{\theta} - \log(1 + \exp(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}))\big). \end{aligned}$$

Here and throughout the sequel, asymptotic statements refer to $n \to \infty$. Two first assumptions are:

(A.1) For sufficiently large n, the design matrix $\boldsymbol{D}_n := [\boldsymbol{d}_{n1}\, \boldsymbol{d}_{n2}\, \cdots\, \boldsymbol{d}_{nn}]^\top$ has rank p.

(A.2) For each $n \in \mathbb{N}$, the function R_n has a minimizer $\boldsymbol{\theta}_n \in \mathbb{R}^p$.

Assumption (A.1) implies that for sufficiently large n, both functions L_n and R_n are strictly convex. Together with (A.2), this implies that $\boldsymbol{\theta}_n$ is the *unique* minimizer of R_n. Here we emphasize that we work with the model of logistic regression without assuming that it is correct. If yes, assumption (A.2) is automatically fulfilled:

(A.2') For each $n \in \mathbb{N}$, there exists a vector $\boldsymbol{\theta}_n \in \mathbb{R}^p$ such that $p_{ni} = \ell(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}_n)$ for $1 \leq i \leq n$.

The next assumptions concern the gradient of L_n,

$$\nabla L_n(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^{n}(Y_{ni} - \ell(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}))\,\boldsymbol{d}_{ni},$$

more precisely, the matrix

$$\boldsymbol{\Gamma}_n^* \;:=\; \mathrm{Cov}\big(n^{-1/2}\nabla L_n(\boldsymbol{\theta}_n)\big) \;=\; \frac{1}{n}\sum_{i=1}^{n} p_{ni}(1-p_{ni})\,\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top.$$

In addition, we consider the Hessian matrix of the functions L_n and R_n,

$$D^2L_n(\boldsymbol{\theta}) = D^2R_n(\boldsymbol{\theta}) \;=\; n\boldsymbol{\Gamma}_n(\boldsymbol{\theta}) \quad \text{with} \quad \boldsymbol{\Gamma}_n(\boldsymbol{\theta}) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\ell'(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta})\,\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top.$$

(A.3) There exist symmetric matrices $\boldsymbol{\Gamma}^*, \boldsymbol{\Gamma} \in \mathbb{R}^{p\times p}$, where the latter is positive definite, such that

$$\boldsymbol{\Gamma}_n^* \;\to\; \boldsymbol{\Gamma}^* \quad \text{and} \quad \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n) \;\to\; \boldsymbol{\Gamma}.$$

(A.4) The design points $\boldsymbol{d}_{ni}$ fulfil the following Lindeberg type condition:

$$\Lambda_n := \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{d}_{ni}\|^2\min\Big(\frac{\|\boldsymbol{d}_{ni}\|}{\sqrt{n}},1\Big) \;\to\; 0.$$

If we replace assumption (A.2) with the stronger assumption (A.2'), then the matrices $\boldsymbol{\Gamma}_n^*$ and $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)$ in (A.3) coincide, whence $\boldsymbol{\Gamma}^* = \boldsymbol{\Gamma}$.

Special setting: Independent, identically distributed observations. Consider the simple setting with observations $(\boldsymbol{d}_{ni}, Y_{ni}) = (\boldsymbol{d}_i, Y_i)$ for all $n \geq 1$ and $1 \leq i \leq n$, where $(\boldsymbol{d}_1, Y_1)$, $(\boldsymbol{d}_2, Y_2)$, $(\boldsymbol{d}_3, Y_3)$, ... are independent copies of a random variable $(\boldsymbol{d}, Y) \in \mathbb{R}^p \times \{0,1\}$. Suppose that for some $\boldsymbol{\theta}_o \in \mathbb{R}^p$, $\mathbb{P}(Y = 1 \,|\, \boldsymbol{d} = \boldsymbol{w}) = \ell(\boldsymbol{w}^\top\boldsymbol{\theta}_o)$ for all $\boldsymbol{w} \in \mathbb{R}^p$, that $\mathbb{E}(\|\boldsymbol{d}\|^2)$ is finite and $\mathbb{E}(\boldsymbol{d}\boldsymbol{d}^\top)$ is positive definite. Then, conditional on $(\boldsymbol{d}_i)_{i\geq 1}$, Assumptions (A.1), (A.2') and (A.3-4) are satisfied almost surely with $\boldsymbol{\theta}_n = \boldsymbol{\theta}_o$ and

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^* \;=\; \mathbb{E}\big(\ell'(\boldsymbol{d}^\top\boldsymbol{\theta}_o)\,\boldsymbol{d}\boldsymbol{d}^\top\big) \;=\; \mathbb{E}\big(\mathrm{Var}(Y|\boldsymbol{d})\boldsymbol{d}\boldsymbol{d}^\top\big).$$

Indeed, Assumption (A.2) is obvious. By the law of large numbers, with probability one,

$$n^{-1}\boldsymbol{D}_n^\top\boldsymbol{D}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{d}_i\boldsymbol{d}_i^\top \;\to\; \mathbb{E}(\boldsymbol{d}\boldsymbol{d}^\top).$$

Since the limit is positive definite, the matrix $\boldsymbol{D}_n$ has rank p for sufficiently large n, which yields (A.1). Similary, with probability one,

$$\boldsymbol{\Gamma}_n^* = \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_o) = \frac{1}{n}\sum_{i=1}^{n}\ell'(\boldsymbol{d}_i^\top\boldsymbol{\theta}_o)\boldsymbol{d}_i\boldsymbol{d}_i^\top \;\to\; \boldsymbol{\Gamma} = \mathbb{E}\big(\ell'(\boldsymbol{d}^\top\boldsymbol{\theta}_o)\boldsymbol{d}\boldsymbol{d}^\top\big),$$

because $\ell' \leq 1/4$. This matrix is positive definite, because for any vector $\boldsymbol{v} \in \mathbb{R}^p$,

$$\boldsymbol{v}^\top \boldsymbol{\Gamma} \boldsymbol{v} \;=\; \mathbb{E}\big(\ell'(\boldsymbol{d}^\top \boldsymbol{\theta}_o)(\boldsymbol{v}^\top \boldsymbol{d})^2\big).$$

Since $\ell' > 0$, this is zero if and only if $\boldsymbol{v}^\top \boldsymbol{d} = 0$ almost surely, and by our assmption that $\mathbb{E}(\boldsymbol{d}\boldsymbol{d}^\top)$ is positive definite, this implies that $\boldsymbol{v} = 0$.

Finally, Assumption (A.4) is satisfied because

$$\Lambda_n \;\leq\; \Big(\max_{1\leq i\leq n} \frac{\|\boldsymbol{d}_i\|^2}{n}\Big)^{1/2} \cdot \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{d}_i\|^2 \;\to\; 0 \cdot \mathbb{E}(\|\boldsymbol{d}\|^2),$$

see Exercise 7.8.

Exercise 7.8. Let $Z_1, Z_2, Z_3, \ldots \geq 0$ be independent random variables such that $\mathbb{E}(Z_1) < \infty$. Then, with probability one,

$$n^{-1} \max\{Z_1, \ldots, Z_n\} \;\to\; 0.$$

(Hint: One can deduce from the strong law of large numbers or by a direct calculation that $Z_n/n \to 0$ almost surely. From this one can deduce the assertion about $\max\{Z_1, \ldots, Z_n\}$.)

Assumptions (A.1-4), combined with the Central Limit Theorem, imply an essential asymptotic property of the negative log-likelihood function L_n.

Theorem 7.9. *For $\Delta \in \mathbb{R}^p$ let us write*

$$L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - L_n(\boldsymbol{\theta}_n) \;=\; -\boldsymbol{Z}_n^\top \Delta + \Delta^\top \boldsymbol{\Gamma} \Delta/2 + r_n(\Delta)$$

with the random vector $\boldsymbol{Z}_n := n^{-1/2}\sum_{i=1}^n (Y_{ni} - p_{ni})\boldsymbol{d}_{ni}$ and a remainder $r_n(\Delta)$. Then, under Assumptions (A.1-4),

$$\boldsymbol{Z}_n \;\to_{\mathcal{L}}\; \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Gamma}^*)$$

and

$$\sup_{\Delta\,:\,\|\Delta\|\leq C} |r_n(\Delta)| \;\to\; 0 \quad \text{for any fixed } C > 0.$$

This theorem has various important consequences. The first one concerns the MLE of $\boldsymbol{\theta}_n$ and its asymptotic covariance matrix.

Theorem 7.10. *Under Assumptions (A.1-4), the negative log-likelihood function L_n has a unique minimizer $\widehat{\boldsymbol{\theta}}_n$ with asymptotic probability one, and*

$$\begin{aligned} n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \;&=\; \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1) \;\to_{\mathcal{L}}\; \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}^*\boldsymbol{\Gamma}^{-1}), \\ 2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) \;&=\; \boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1) \end{aligned}$$

with the random vector $\boldsymbol{Z}_n$ from Theorem 7.9. Furthermore, $\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n)$ is a consistent estimator of $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)$, that means,

$$\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n) \;\to_p\; 0.$$

Proof of Theorem 7.9. We start with the matrix-valued function $\mathbf{\Gamma}_n(\cdot)$. For arbitrary vectors $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \mathbb{R}^p$,

$$\mathbf{\Gamma}_n(\boldsymbol{\theta}) - \mathbf{\Gamma}_n(\widetilde{\boldsymbol{\theta}}) \;=\; \frac{1}{n}\sum_{i=1}^n \big(\ell'(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}) - \ell'(\boldsymbol{d}_{ni}^\top\widetilde{\boldsymbol{\theta}})\big)\, \boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top.$$

But $0 < \ell' = \ell(1-\ell) \le 1/4$ and $\ell'' = (1-2\ell)\ell' = u(1-u^2)/4$ with $u := 1-2\ell \in (-1,1)$ satisfies the inequality $|\ell''| \le (6\sqrt{3})^{-1} < 10^{-1}$. Hence, with the norm

$$\|\boldsymbol{A}\| \;:=\; \max\big\{\|\boldsymbol{A}\boldsymbol{v}\| : \boldsymbol{v} \in \mathbb{R}^p, \|\boldsymbol{v}\| = 1\big\}$$

of a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, we may conclude that

$$\begin{aligned}
\big\|\mathbf{\Gamma}_n(\boldsymbol{\theta}) - \mathbf{\Gamma}_n(\widetilde{\boldsymbol{\theta}})\big\| \;&\le\; \frac{1}{n}\sum_{i=1}^n \min\Big(\frac{1}{4}, \frac{\big|\boldsymbol{d}_{ni}^\top(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\big|}{10}\Big)\|\boldsymbol{d}_{ni}\|^2 \\
&\le\; \frac{1}{n}\sum_{i=1}^n \min\Big(\frac{1}{4}, \frac{\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\|\|\boldsymbol{d}_{ni}\|}{10}\Big)\|\boldsymbol{d}_{ni}\|^2 \\
&\le\; \max\Big(\frac{1}{4}, \frac{n^{1/2}\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\|}{10}\Big)\,\Lambda_n. \qquad (7.5)
\end{aligned}$$

Now to the main assertions: The difference

$$(L_n - R_n)(\boldsymbol{\theta}) \;=\; -\sum_{i=1}^n (Y_{ni} - p_{ni})\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}$$

is linear in $\boldsymbol{\theta} \in \mathbb{R}^p$. Together with Taylor's formula and the fact that $\nabla R_n(\boldsymbol{\theta}_n) = 0$ according to (A.2), this yields the representation

$$\begin{aligned}
L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - L_n(\boldsymbol{\theta}_n) \;&=\; (L_n - R_n)(n^{-1/2}\Delta) + R_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - R_n(\boldsymbol{\theta}_n) \\
&=\; (L_n - R_n)(n^{-1/2}\Delta) + n^{-1}\Delta^\top D^2 R_n(\boldsymbol{\theta}_n + \xi_{n,\Delta}\Delta)\Delta/2 \\
&=\; -\boldsymbol{Z}_n^\top\Delta + \Delta^\top\mathbf{\Gamma}_n(\boldsymbol{\theta}_n + \xi_{n,\Delta}\Delta)\Delta/2 \\
&=\; -\boldsymbol{Z}_n^\top\Delta + \Delta^\top\mathbf{\Gamma}\Delta/2 + r_n(\Delta)
\end{aligned}$$

with the random vector $\boldsymbol{Z}_n = n^{-1/2}\sum_{i=1}^n (Y_{ni} - p_{ni})\,\boldsymbol{d}_{ni}$, where

$$r_n(\Delta) \;:=\; \Delta^\top\big(\mathbf{\Gamma}_n(\boldsymbol{\theta}_n + \xi_{n,\Delta}\Delta) - \mathbf{\Gamma}\big)\Delta/2$$

and $0 \le \xi_{n,\Delta} \le n^{-1/2}$. Now it follows from (A.3-4) and (7.5) that for any fixed number $C > 0$,

$$\sup_{\Delta\in\mathbb{R}^p\,:\,\|\Delta\|\le C} |r_n(\Delta)| \;\le\; \frac{C^2}{2}\Big(\max\Big(\frac{1}{4}, \frac{C}{10}\Big)\Lambda_n + \|\mathbf{\Gamma}_n(\boldsymbol{\theta}_n) - \mathbf{\Gamma}\|\Big) \;\to\; 0.$$

It remains to verify that $\boldsymbol{Z}_n \to_{\mathcal{L}} \mathrm{N}_p(\mathbf{0}, \mathbf{\Gamma}^*)$. To this end we apply Lindeberg's Central Limit Theorem (Theorem A.16) for vector-valued random variables: One can write $\boldsymbol{Z}_n = \sum_{i=1}^n \boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}$ with the stochastically independent summands $\boldsymbol{Y}_{ni} := n^{-1/2}(Y_{ni} - p_{ni})\boldsymbol{d}_{ni}$, where $\mathbb{E}(\boldsymbol{Y}_{ni}) = \mathbf{0}$, and

$$\sum_{i=1}^n \mathbb{E}\big(\boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top\big) \;=\; \mathbf{\Gamma}_n^* \;\to\; \mathbf{\Gamma}^*$$

by (A.3). Furthermore, the Lindeberg condition of Theorem A.16 is fulfilled: Since $\|\boldsymbol{Y}_{ni}\| \le n^{-1/2}\|\boldsymbol{d}_{ni}\|$,

$$\mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^2 \min(\|\boldsymbol{Y}_{ni}\|, 1)\big) \;\le\; n^{-1}\|\boldsymbol{d}_{ni}\|^2 \min\big(n^{-1/2}\|\boldsymbol{d}_{ni}\|, 1\big),$$

so (A.4) implies that

$$\sum_{i=1}^{n} \mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^2 \min(1, \|\boldsymbol{Y}_{ni}\|)\big) \;\le\; \Lambda_n \;\to\; 0. \qquad \square$$

Proof of Theorem 7.10. The conclusions that the MLE $\widehat{\boldsymbol{\theta}}_n$ exists with asymptotic probability one, that $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) = \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1)$ and that $2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1)$ follow from Theorem 7.9 and general considerations in Section 7.2. Here is a heuristic argument: Suppose for the moment that

$$L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - L_n(\boldsymbol{\theta}_n) \;=\; -\boldsymbol{Z}_n^\top \Delta + 2^{-1}\Delta^\top \boldsymbol{\Gamma}\Delta$$

for all $\Delta \in \mathbb{R}^p$. The right-hand side can be rewritten as

$$-\boldsymbol{Z}_n^\top \Delta + 2^{-1}\Delta^\top \boldsymbol{\Gamma}\Delta \;=\; 2^{-1}(\Delta - \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n)^\top \boldsymbol{\Gamma}(\Delta - \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n) - 2^{-1}\boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n.$$

This would imply that $\Delta \mapsto L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta)$ has the unique minimizer $\widehat{\Delta}_n := \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n$, which is equivalent to L_n having the unique minimizer $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_n + n^{-1/2}\widehat{\Delta}_n$, and $n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) = \widehat{\Delta}_n$. Moreover, $2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) = \widehat{\Delta}_n^\top \boldsymbol{\Gamma}\widehat{\Delta}_n = \boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n$.

The claim about $\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n)$ is a consequence of inequality (7.5) in the proof of Theorem 7.9 and the fact that $n^{1/2}\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\| = O_p(1)$:

$$\big\|\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)\big\| \;\le\; \max\Big(\frac{1}{4}, \frac{n^{1/2}\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\|}{10}\Big)\Lambda_n \;=\; O_p(1)\Lambda_n \;=\; o_p(1).$$

$\square$

7.1.3 Likelihood-Based Statistical Procedures

The asymptotic properties of the negative log-likelihood function L_n or the MLE $\widehat{\boldsymbol{\theta}}_n$ imply various statistical procedures. In particular, one can construct tests and confidence regions for linear functions of the true parameter $\boldsymbol{\theta}_n$. These procedures are similar to those for linear models by means of student or F distributions. An important difference, however, is that in general we have only *asymptotic* validity, and for that reason we stick to the triangular scheme and assume throughout conditions (A.1), **(A.2')** and (A.3-4).

Unfortunately, many users of statistical software don't realize that albeit the output for logistic regression and other generalized models is rather similar to the one of linear models, the reported p-values and confidence intervals are based on asymptotics and thus should be taken with care. Sometimes, but not always, the corresponding software issues a warning if the input data are such that the asymptotic approximations are questionable.

In the following subsections, $\boldsymbol{\psi}$ is a given nonzero vector in $\mathbb{R}^p$, and $\boldsymbol{\Psi}$ is a given matrix in $\mathbb{R}^{p\times k}$ with rank $k \le p$. (In case of $k = p$ we tacitly assume that $\boldsymbol{\Psi} = \boldsymbol{I}_p$.)

Wald's Approach for Tests and Confidence Regions

We start with relatively simple procedures many of which are implemented in standard statistical software. But later on, we shall consider alternatives which tend to be more reliable.

It follows from Theorem 7.10 that

$$\widehat{\boldsymbol{\theta}}_n \;\sim_{\text{appr.}}\; \mathrm{N}_p(\boldsymbol{\theta}_n, \boldsymbol{\Sigma}_n)$$

with

$$\boldsymbol{\Sigma}_n \;:=\; \Big(\sum_{i=1}^{n} \ell'(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta}_n)\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top\Big)^{-1} \;=\; n^{-1}\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)^{-1}.$$

The estimator

$$\widehat{\boldsymbol{\Sigma}}_n \;:=\; \Big(\sum_{i=1}^{n} \ell'(\boldsymbol{d}_{ni}^\top \widehat{\boldsymbol{\theta}}_n)\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top\Big)^{-1} \;=\; n^{-1}\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n)^{-1}$$

is consistent in the sense that

$$\boldsymbol{\Sigma}_n^{-1}\widehat{\boldsymbol{\Sigma}}_n \;\approx\; \boldsymbol{I}_p.$$

This may be utilized as follows:

Simple linear functions of $\boldsymbol{\theta}_n$. Suppose we are interested in the real number $\boldsymbol{\psi}^\top\boldsymbol{\theta}_n$. On the one hand,

$$\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}}_n \;\sim_{\text{appr.}}\; \mathrm{N}(\boldsymbol{\psi}^\top\boldsymbol{\theta}_n, \sigma_{n,\psi}^2)$$

with

$$\sigma_{n,\psi} \;:=\; \sqrt{\boldsymbol{\psi}^\top\boldsymbol{\Sigma}_n\boldsymbol{\psi}}.$$

On the other hand, the corresponding standard error

$$\widehat{\sigma}_{n,\psi} \;:=\; \sqrt{\boldsymbol{\psi}^\top\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\psi}}$$

satisfies

$$\frac{\widehat{\sigma}_{n,\psi}}{\sigma_{n,\psi}} \;\approx\; 1.$$

In particular,

$$\frac{\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\psi}^\top\boldsymbol{\theta}_n}{\widehat{\sigma}_{n,\psi}} \;\sim_{\text{appr.}}\; \mathrm{N}(0,1).$$

For $\alpha \in (0,1)$, this yields the approximate $(1-\alpha)$-confidence interval

$$\big[\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}}_n \pm \widehat{\sigma}_{n,\psi}\Phi^{-1}(1-\alpha/2)\big]$$

for $\boldsymbol{\psi}^\top\boldsymbol{\theta}_n$. Moreover, for any given $\eta_o \in \mathbb{R}$,

$$2\Phi\Big(-\frac{|\boldsymbol{\psi}^\top\widehat{\boldsymbol{\theta}}_n - \eta_o|}{\widehat{\sigma}_{n,\psi}}\Big)$$

is an approximate p-value for the null hypothesis that $\boldsymbol{\psi}^\top\boldsymbol{\theta}_n = \eta_o$.

General linear functions of $\boldsymbol{\theta}_n$. Suppose we are interested in the vector $\boldsymbol{\Psi}^\top \boldsymbol{\theta}_n \in \mathbb{R}^k$. Here one can say that

$$\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}}_n \ \sim_{\text{appr.}}\ \mathrm{N}_k\big(\boldsymbol{\Psi}^\top \boldsymbol{\theta}_n, \boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_n \boldsymbol{\Psi}\big),$$

and

$$(\boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_n \boldsymbol{\Psi})^{-1} (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\Psi}) \ \approx\ \boldsymbol{I}_k.$$

Consequently, for any $\boldsymbol{\eta} \in \mathbb{R}^k$,

$$T_n(\boldsymbol{\eta}) \ :=\ (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\eta})^\top (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\Psi})^{-1} (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\eta})$$

is a test statistic such that

$$T_n(\boldsymbol{\Psi}^\top \boldsymbol{\theta}_n) \ \sim_{\text{appr.}}\ \chi^2_k.$$

This implies that the confidence ellipsoid

$$\boldsymbol{C}^{(W)} = \boldsymbol{C}^{(W)}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{\Psi}, \alpha) \ :=\ \big\{\boldsymbol{\eta} \in \mathbb{R}^k : T_n(\boldsymbol{\eta}) \le \chi^2_{k;1-\alpha}\big\}$$

for $\boldsymbol{\Psi}^\top \boldsymbol{\theta}_n$ has asymptotic confidence level $1-\alpha$. Moreover, with the distribution function $F_k(\cdot)$ of χ^2_k, for any given $\boldsymbol{\eta}_o \in \mathbb{R}^k$,

$$1 - F_k(T_n(\boldsymbol{\eta}_o))$$

as an asymptotic p-value of the null hypothesis that $\boldsymbol{\Psi}^\top \boldsymbol{\theta}_n = \boldsymbol{\eta}_o$. That means, for arbitrary $\alpha \in (0,1)$,

$$\mathbb{P}\big(1 - F_k(T_n(\boldsymbol{\Psi}^\top \boldsymbol{\theta}_n)) \le \alpha\big) \ \approx\ \alpha.$$

Exercise 7.11 (Comparing two symmetric, positive definite matrices). Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p\times p}$ be symmetric and positive definite. For any matrix $\boldsymbol{M} \in \mathbb{R}^{p\times p}$ let $\|\boldsymbol{M}\|_F$ and $\|\boldsymbol{M}\|$ be its Frobenius and operator norm, respectively,

$$\|\boldsymbol{M}\|_F \ :=\ \mathrm{trace}(\boldsymbol{M}^\top \boldsymbol{M})^{1/2}, \quad \|\boldsymbol{M}\| \ :=\ \max\{\|\boldsymbol{M}\boldsymbol{v}\| \colon \boldsymbol{v} \in \mathbb{R}^p, \|\boldsymbol{v}\| \le 1\},$$

where $\|\boldsymbol{w}\|$ is the standard Euclidean norm of a vector $\boldsymbol{w} \in \mathbb{R}^p$.

(a) Show that

$$\|\boldsymbol{A}^{-1}\boldsymbol{B} - \boldsymbol{I}_p\|_F \ \ge\ \|\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2} - \boldsymbol{I}_p\|_F.$$

Hint: Consider first a diagonal matrix $\boldsymbol{A}$.

(b) Show that

$$\|\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2} - \boldsymbol{I}_p\| \ =\ \max\left\{\max_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}} - 1,\ 1 - \min_{\boldsymbol{v}\neq\boldsymbol{0}} \frac{\boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}}\right\}.$$

Profile Likelihood

One may interpret the log-likelihood $-L_n(\boldsymbol{\theta})$ as a measure of plausibility of the null hypothesis that $\boldsymbol{\theta}_n = \boldsymbol{\theta}$. More precisely, it follows from Theorem 7.10 that

$$2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) \ =\ \boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}_n + o_p(1) \ \to_{\mathcal{L}}\ \chi^2_p,$$

because $\boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}_n = \|\boldsymbol{\Gamma}^{-1/2} \boldsymbol{Z}_n\|^2$ and $\boldsymbol{\Gamma}^{-1/2} \boldsymbol{Z}_n \to_{\mathcal{L}} \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I}_p)$ under (A.1), (A.2') and (A.3-4). Consequently,

$$C^{(L)} = C^{(L)}(\boldsymbol{Y}_n, \alpha) \;:=\; \big\{\boldsymbol{\theta} \in \mathbb{R}^p : 2L_n(\boldsymbol{\theta}) \le 2L_n(\widehat{\boldsymbol{\theta}}_n) + \chi^2_{p;1-\alpha}\big\}$$

is a confidence region for $\boldsymbol{\theta}_n$ with *asymptotic* confidence level (coverage probability) $1 - \alpha$. In fact, one can show that this confidence region and Wald's confidence ellipsoid (with $\boldsymbol{\Psi} = \boldsymbol{I}_p$) are asymptotically identical. Numerical experiments, however, show that for small and moderately large sample sizes, the present method is often more precise.

An obvious question is whether similar confidence regions may be constructed for $\boldsymbol{\Psi}^\top \boldsymbol{\theta}_n$. To this end, one needs sort of a negative log-likelihood function on $\mathbb{R}^k$: the negative *profile log-likelihood* at the point $\boldsymbol{\eta} \in \mathbb{R}^k$ is defined as

$$PL_n(\boldsymbol{\eta}) \;:=\; \inf_{\boldsymbol{\theta} \in \mathbb{R}^p \,:\, \boldsymbol{\Psi}^\top \boldsymbol{\theta} = \boldsymbol{\eta}} L_n(\boldsymbol{\theta}).$$

It follows from Exercises 7.12 and 7.13 that the negative profile log-likelihood function $PL_n : \mathbb{R}^k \to \mathbb{R}$ is well-defined and convex. If L_n is coercive, PL_n is coercive too, and the infimum in the definition of $PL_n(\boldsymbol{\eta})$ is a minimum.

Exercise 7.12 (Profile functions, I). Let $L : \mathbb{R}^p \to [-\infty, \infty]$ be an arbitrary function, and let $\boldsymbol{\Psi}$ be a matrix in $\mathbb{R}^{p \times k}$ with rank $k \le p$. Now let

$$PL : \mathbb{R}^k \to [-\infty, \infty], \quad PL(\boldsymbol{\eta}) \;:=\; \inf\big\{L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^p, \boldsymbol{\Psi}^\top \boldsymbol{\theta} = \boldsymbol{\eta}\big\}.$$

Verify the following properties of PL:

(a) PL is bounded from below if L is bounded from below. Precisely,

$$\inf_{\boldsymbol{\eta} \in \mathbb{R}^k} PL(\boldsymbol{\eta}) \;=\; \inf_{\boldsymbol{\theta} \in \mathbb{R}^p} L(\boldsymbol{\theta}).$$

(b) PL is coercive if L is coercive.

(c) There exists a matrix $\boldsymbol{A} \in \mathbb{R}^{p \times k}$ such that $\boldsymbol{\Psi}^\top \boldsymbol{A} = \boldsymbol{I}_k$. For any such matrix $\boldsymbol{A}$, the profile function PL may be written as

$$PL(\boldsymbol{\eta}) \;=\; \inf_{\boldsymbol{v} \in \mathbb{V}} L(\boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{v}),$$

where $\mathbb{V} := \{\boldsymbol{v} \in \mathbb{R}^p \colon \boldsymbol{\Psi}^\top \boldsymbol{v} = \boldsymbol{0}\}$.

(d) If L is continuous and coercive, then PL is continuous too, and the infimum in the definition of $PL(\boldsymbol{\eta})$ is a minimum.

Exercise 7.13 (Profile functions, II). Let $L : \mathbb{R}^p \to (-\infty, \infty]$ be a convex function which is bounded from below. Let $\boldsymbol{\Psi}$ and PL be as in Exercise 7.12. Verify the following properties of PL:

(a) PL is convex.

(b) Suppose that L is convex, continuous, coercive and even strictly convex on the set $\{L < \infty\}$. Then PL is strictly convex on the set $\{PL < \infty\}$.

Remark: In part (b) one can not dispense with coercivity of L. A counter-example is given by the function $L(\boldsymbol{\theta}) := \exp\bigl(\sqrt{1+\theta_1^2}+\theta_2\bigr)$ on $\mathbb{R}^2$ and $\boldsymbol{\Psi} := (1,0)^\top$.

The next result shows that the negative profile log-likelihood function PL_n has analogous properties as the negative log-likelihood function L_n.

Theorem 7.14. *For $\boldsymbol{w} \in \mathbb{R}^k$ let us write*

$$PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n + n^{-1/2}\boldsymbol{w}) - PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n) \;=\; -\boldsymbol{Z}_{n,\Psi}^\top\boldsymbol{w} + \boldsymbol{w}^\top\boldsymbol{\Gamma}_\Psi\boldsymbol{w}/2 + r_{n,\Psi}(\boldsymbol{w}),$$

where

$$\boldsymbol{\Gamma}_\Psi \;:=\; (\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi})^{-1} \quad \textit{and} \quad \boldsymbol{Z}_{n,\Psi} \;:=\; \boldsymbol{\Gamma}_\Psi\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n$$

with the random vector $\boldsymbol{Z}_n$ defined in Theorem 7.9. Then under conditions (A.1), (A.2') and (A.3-4),

$$\sup_{\boldsymbol{w}\,:\,\|\boldsymbol{w}\|\le C} |r_{n,\Psi}(\boldsymbol{w})| \;\to_p\; 0 \quad \textit{for any fixed } C > 0.$$

Furthermore, $\boldsymbol{Z}_{n,\Psi} \to_{\mathcal{L}} \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{\Gamma}_\Psi)$, and

$$2PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) \;=\; \boldsymbol{Z}_{n,\Psi}^\top\boldsymbol{\Gamma}_\Psi^{-1}\boldsymbol{Z}_{n,\Psi} + o_p(1) \;\to_{\mathcal{L}}\; \chi^2_k.$$

This theorem follows directly from Theorem 7.9 and the general results in Section 7.2. The last part implies that

$$C^{(L)} = C^{(L)}(\boldsymbol{Y}_n, \alpha) \;:=\; \bigl\{\boldsymbol{\eta} \in \mathbb{R}^k : 2PL_n(\boldsymbol{\eta}) \le 2L_n(\widehat{\boldsymbol{\theta}}_n) + \chi^2_{k;1-\alpha}\bigr\}$$

defines a confidence region for $\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n$ with asymptotic coverage probability $1-\alpha$. Furthermore, for any $\boldsymbol{\eta} \in \mathbb{R}^k$,

$$1 - F_k\bigl(2PL_n(\boldsymbol{\eta}) - 2L_n(\widehat{\boldsymbol{\theta}}_n)\bigr)$$

is an asymptotic p-value for the null hypothesis that $\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n = \boldsymbol{\eta}$.

Special case: Tests of simplified models. We return temporarily to the original description of logistic regression. Suppose we want to test the null hypothesis that the underlying regression function $f = \operatorname{logit} p$ lies in a given linear subspace $\mathcal{F}_o$ of $\mathcal{F}$, where $\dim(\mathcal{F}) - \dim(\mathcal{F}_o) = k$. Let $\widehat{f}$ be the MLE of f in the full model, and let $\widehat{f}_o$ be the MLE under the null hypothesis. Then

$$1 - F_k\bigl(2L(\widehat{f}_o) - 2L(\widehat{f})\bigr)$$

is an approximate p-value of the null hypothesis that $f \in \mathcal{F}_o$.

To verify this claim, suppose we have chosen a basis $f_1, \ldots, f_p$ of $\mathcal{F}$ such that $f_1, \ldots, f_{p_o}$ span the smaller space $\mathcal{F}_o$. With the resulting design matrix $\boldsymbol{D} = [f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})]$ and $\boldsymbol{\Psi} = [\boldsymbol{0}_{p_o\times p-p_o}, \boldsymbol{I}_{p-p_o}]^\top \in \mathbb{R}^{p\times(p-p_o)}$, one can rewrite the pair $\bigl(L(\widehat{f}), L(\widehat{f}_o)\bigr)$ as $\bigl(L(\widehat{\boldsymbol{\theta}}), PL(\boldsymbol{0})\bigr)$.

7.1.4 Returning from Asymptopia

The term "asymptopia" was presumably coined by the statistician David Freedman for statistical inference justified by asymptotics, referring to the novels "Utopia" by Thomas Morus and "Ecotopia" by Ernest Callenbach.

In specific applications there is no triangular array of observations, just one data set. An obvious question is how to interpret the conditions (A.1), (A.2') and (A.3-4) for a single data set with observations $(\boldsymbol{d}_1, Y_1), (\boldsymbol{d}_2, Y_2), \ldots, (\boldsymbol{d}_n, Y_n)$.

Condition (A.1) is no problem; we just assume that the design matrix $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]^\top$ has rank p.

Condition (A.2') is a model assumption the plausibility of which may be checked graphically at least; see the subsequent data example. Hence we assume that for given design vectors $\boldsymbol{d}_i$, the observations Y_i are independent with

$$\mathbb{P}(Y_i = 1) \;=\; \ell(\boldsymbol{d}_i^\top \boldsymbol{\theta}_*)$$

for $1 \le i \le n$ with an unknown parameter $\boldsymbol{\theta}_* \in \mathbb{R}^p$.

Condition (A.3) is superfluous in principle. By means of a linear transformation of $\mathbb{R}^p$ we may achieve that

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}_*) = \frac{1}{n} \sum_{i=1}^{n} \ell'(\boldsymbol{d}_i^\top \boldsymbol{\theta}_*) \boldsymbol{d}_i \boldsymbol{d}_i^\top \;=\; \boldsymbol{I}_p.$$

To this end, one could simply replace $\boldsymbol{d}_i$ with $\boldsymbol{\Gamma}(\boldsymbol{\theta}_*)^{-1/2} \boldsymbol{d}_i$ and any potential parameter $\boldsymbol{\theta}$ with $\boldsymbol{\Gamma}(\boldsymbol{\theta}_*)^{1/2} \boldsymbol{\theta}$.

But this trick for condition (A.3) necessitates a modification of condition (A.4). Now we require that

$$\Lambda \;:=\; \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\Gamma}(\boldsymbol{\theta}_*)^{-1/2} \boldsymbol{d}_i\|^2 \min\Big(\frac{\|\boldsymbol{\Gamma}(\boldsymbol{\theta}_*)^{-1/2} \boldsymbol{d}_i\|}{\sqrt{n}}, 1 \Big)$$

is "rather small". Since we do not know $\boldsymbol{\theta}_*$, one could consider

$$\widehat{\Lambda} \;:=\; \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}})^{-1/2} \boldsymbol{d}_i\|^2 \min\Big(\frac{\|\boldsymbol{\Gamma}(\widehat{\boldsymbol{\theta}})^{-1/2} \boldsymbol{d}_i\|}{\sqrt{n}}, 1 \Big)$$

as a diagnostic quantity, similar to the maximal leverage in Chapter 4.

7.1.5 A Data Example

A data set provided by PD Dr. Bürk (Lübeck), contains data of all heart surgeries that have been performed at the University Hospital Lübeck in a certain time period. In particular, the data set contains the variate $Y = \mathrm{mortality}$ which specifies whether the patient died within a certain time window as a consequence of this intervention. Furthermore, the values of 21 covariates have been reported. These describe properties of the patients or the circumstances of the surgery. Table 7.1 contains a list of all covariates involved. Most of them are dichotomous with yes/no coded as 1/0. Numerical covariates are $X(1)$ and $X(17)$. Covariate $X(3)$ has been treated as numerical as well, although it is an ordinal feature with values in $\{1, 2, 3, 4, 5\}$.

Variable	Meaning
$X(1)$	age in years
$X(2)$	gender (1 = female, 0 = male)
$X(3)$	ASA score (American Society of Anesthesiologists), classifies the physical well-being (1 = completely healthy, 2 = slightly sick, 3 = seriously sick, 4 = in life-threatening condition, 5 = about to die, 6 = brain dead)
$X(4)$	risk factor: cerebral (yes/no)
$X(5)$	risk factor: cardiovascular (yes/no)
$X(6)$	risk factor: pulmonal (yes/no)
$X(7)$	risk factor: renal (yes/no)
$X(8)$	risk factor: hepatic (yes/no)
$X(9)$	risk factor: immunological (yes/no)
$X(10)$	risk factor: metabolic (yes/no)
$X(11)$	risk factor: non-cooperative, unreliable (yes/no)
$X(12)$	etiology: maligne (yes/no)
$X(13)$	etiology: vascular (yes/no)
$X(14)$	antibiotics (yes/no)
$X(15)$	surgery indicated (yes/no)
$X(16)$	emergency surgery (yes/no)
$X(17)$	duration of operation (in minutes)
$X(18)$	septic operation (yes/no)
$X(19)$	experienced surgeon (yes/no)
$X(20)$	blood transfusion (yes/no)
$X(21)$	intensive care (yes/no)
Y	mortality (1 = died, survived = 0)

Table 7.1: Variables for data example.

First analysis. The data set contains $21'556$ observations, with $Y = 1$ in 662 cases. With these observations, we estimated the parameters a and $b(j)$ for the model

$$\operatorname{logit} \mathbb{P}(Y = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) \;=\; a + \boldsymbol{b}^\top \boldsymbol{x} = a + \sum_{j=1}^{21} b(j) x(j).$$

Table 7.2 contains the point estimates $\widehat{b}(j)$ together with the standard errors and p-values via profile log-likelihood. In addition, adjusted p-values via the Bonferroni-Holm method are reported.

A graphical presentation of the results. Figure 7.1 shows a rug plot of the pairs $(\widehat{Z}_i, Y_i)$ with $\widehat{Z}_i := \widehat{a} + \widehat{\boldsymbol{b}}^\top \boldsymbol{X}_i$. In addition, one sees the graph of a monotone function $\widehat{\ell} : \mathbb{R} \to \mathbb{R}$ minimizing the sum $\sum_{i=1}^n (Y_i - \widehat{\ell}(\widehat{Z}_i))^2$ (black step function) as well as the graph of the logistic function ℓ (smooth blue curve). The explicit computation of $\widehat{\ell}$ is discussed in Chapter 10. The fact that the step function $\widehat{\ell}$ coincides with the logistic function quite well indicates that the logistic model fits the data reasonably well.

ROC curves. Logistic regression is often viewed as a means to determine a promising *discriminant function* $x \mapsto \widehat{f}(x)$ on $\mathcal{X}$. In case of $\mathcal{X} = \mathbb{R}^d$ and $\widehat{f}(\boldsymbol{x}) = \widehat{a} + \widehat{\boldsymbol{b}}^\top \boldsymbol{x}$, one also calls $\boldsymbol{x} \mapsto \widehat{\boldsymbol{b}}^\top \boldsymbol{x}$

j	$\widehat{b}(j)$	(st.error)	p-value	adj. p-value
1	0.0382	(0.0041)	< 0.0001	< 0.0001
2	0.1066	(0.0996)	0.2841	1.0000
3	1.3152	(0.0738)	< 0.0001	< 0.0001
4	-0.1925	(0.1166)	0.0971	0.5823
5	0.0475	(0.1325)	0.7193	1.0000
6	0.2527	(0.1021)	0.0135	0.1485
7	0.4596	(0.1112)	< 0.0001	0.0006
8	-0.1640	(0.1053)	0.1175	0.5877
9	-0.3265	(0.3021)	0.2652	1.0000
10	0.2701	(0.1226)	0.0297	0.2457
11	-0.2168	(0.1256)	0.0818	0.5727
12	0.3442	(0.1417)	0.0159	0.1590
13	0.3525	(0.1322)	0.0084	0.1003
14	0.7018	(0.1185)	< 0.0001	< 0.0001
15	-1.6171	(0.2102)	< 0.0001	< 0.0001
16	1.1675	(0.1368)	< 0.0000	< 0.0000
17	0.0014	(0.0006)	0.0273	0.2457
18	1.2064	(0.1629)	< 0.0000	< 0.0001
19	-0.0840	(0.1220)	0.4927	1.0000
20	0.7382	(0.1131)	< 0.0001	< 0.0001
21	2.0286	(0.1345)	< 0.0001	< 0.0001

Table 7.2: Logistic regression analysis, 1.

a discriminant function. Now this function $\widehat{f}$ is used like a test statistic: For a *future* case (X, Y) of which only X is observed initially, one *predicts* that

$$Y \;=\; \begin{cases} 1 & \text{if } \widehat{f}(X) > c, \\ 0 & \text{if } \widehat{f}(X) \le c. \end{cases}$$

Here c is a threshold yet to be chosen. This is like a *medical test* with unknown *sensitivity* $\mathrm{Sens}(c) := \mathbb{P}(\widehat{f}(X) > c \,|\, Y = 1)$ and unknown *specifity* $\mathrm{Spec}(c) := \mathbb{P}(\widehat{f}(X) \le c \,|\, Y = 0)$, where the data (X_i, Y_i), $1 \le i \le n$, and thus $\widehat{f}$ are viewed as fixed. Now, these quantities are estimated by

$$\widehat{\mathrm{Sens}}(c) \;:=\; \frac{\#\{i : Y_i = 1, \widehat{f}(X_i) > c\}}{\#\{i : Y_i = 1\}},$$
$$\widehat{\mathrm{Spec}}(c) \;:=\; \frac{\#\{i : Y_i = 0, \widehat{f}(X_i) \le c\}}{\#\{i : Y_i = 0\}}.$$

The empirical ROC curve (**r**eceiver **o**perating **c**haracteristic) for this family of tests is the curve

$$c \;\mapsto\; \big(1 - \widehat{\mathrm{Spec}}(c), \widehat{\mathrm{Sens}}(c)\big).$$

Figure 7.2 shows this curve for our specific data example. From this curve one can guess, for instance, that for a suitable threshold (which is not visible from the curve), the estimated sensitivity as well as the estimated specifity are between 0.879 and 0.880. Some people use the area under the ROC curve as a measure for the discriminatory power of these tests.

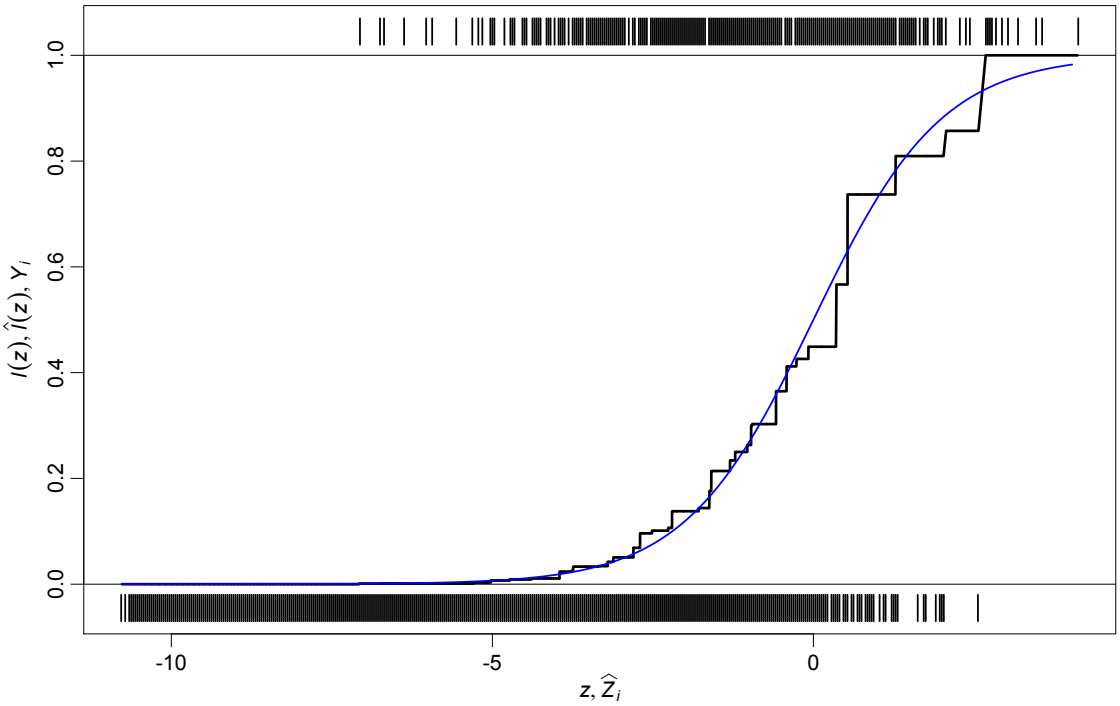

Figure 7.1: Logistic regression analysis, 2.

A likelihood ratio test. Sometimes there is a group of related covariates none of which has a significant influence on the response individually, but the whole group is relevant. In the present data example, this is not the case, but we illustrate this concept with the risk factors, i.e. the covariates $X(4)$, $X(5)$, …, $X(11)$: We compare the minimum of the negative log-likelihood function for the full model with the corresponding minimum for the reduced model in which $b(4) = b(5) = \cdots = b(11) = 0$. In other words, we compare $L(\widehat{\boldsymbol{\theta}})$ with $PL(\mathbf{0})$, where

$$\boldsymbol{\Psi} \;=\; \begin{bmatrix} \mathbf{0}_{4\times 8} \\ \boldsymbol{I}_8 \\ \mathbf{0}_{10\times 8} \end{bmatrix}.$$

An approximate p-value for the null hypothesis that $b(4) = b(5) = \cdots = b(11) = 0$ is then

$$1 - F_8\big(2PL(\mathbf{0}) - 2L(\widehat{\boldsymbol{\theta}})\big),$$

where F_8 stands for the distribution function of χ^2_8. Our data yield $2L(\widehat{\boldsymbol{\theta}}) \approx 3303.66$ and $2PL(\mathbf{0}) \approx 3342.75$. Hence the p-value equals

$$1 - F_8(39.09) \;\approx\; 4.729 \cdot 10^{-6}.$$

Exercise 7.15. The ROC curve just described could be too optimistic, because one and the same data set is used twice: To estimate the discriminant function $\widehat{f}$, and to estimate sensitivity and specifity of the predictor $X \mapsto 1_{[\widehat{f}(X)>c]}$ for various values of the threshold $c \in \mathbb{R}$.

Write a program to compute and display "cross-validated" ROC curves for multiple logistic regression. That is, for given data $\boldsymbol{Y} \in \{0,1\}^n$ and $[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]^\top \in \mathbb{R}^{n\times d}$, it should compute the

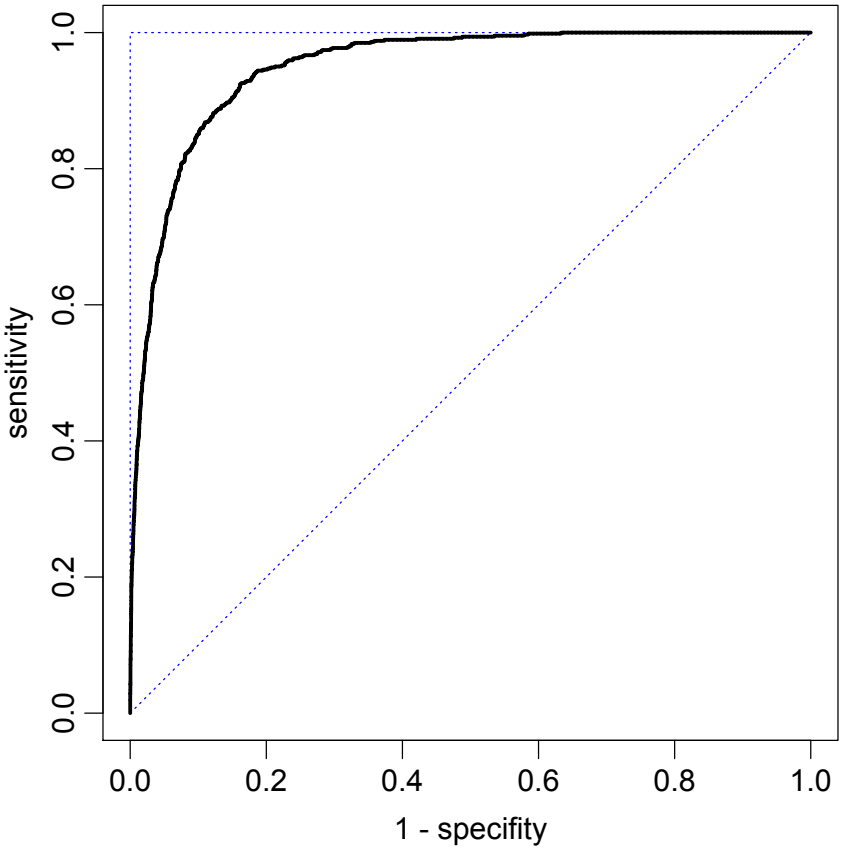

Figure 7.2: Empirical ROC curve for the data example.

estimators

$$\widehat{\mathrm{Sens}}(c) := \frac{\#\{i \le n \colon Y_i = 1, \widehat{f}_{-i}(\boldsymbol{X}_i) > c\}}{\#\{i \le i \colon Y_i = 1\}},$$
$$\widehat{\mathrm{Spec}}(c) := \frac{\#\{i \le n \colon Y_i = 0, \widehat{f}_{-i}(\boldsymbol{X}_i) \le c\}}{\#\{i \le i \colon Y_i = 0\}}$$

for c in a sufficiently rich set of thresholds, and it should display the resulting ROC curve $c \mapsto \big(1 - \widehat{\mathrm{Spec}}(c), \widehat{\mathrm{Sens}}(c)\big)$. Here $\widehat{f}_{-i}$ is computed as $\widehat{f}$ with the reduced data $(Y_k)_{k \ne i}$ and $(\boldsymbol{X}_k)_{k \ne i}$.

Illustrate your program with the data set Buerk.txt used before.

Exercise 7.16. The data set IrishEd.txt lists various features of 435 irish persons in the year 1967. The features are:

sex:	1: male, 2: female
DVRT:	achieved points in a personality test during primary school
edlevel:	later achieved education level
lvcert:	certificate when leaving secondary school: 1: passed, 2: failed
fathocc:	score calculated based on the profession of the father
schltype:	type of secondary school: 1: secondary school, 2: vocational school

(a) Use a statistical software package of your choice to fit a logistic model with lvcert as response and all other covariates, except edlevel. Interpret briefly the output.

(b) Repeat the analysis of part (a), this time with the additional covariate edlevel. Your software should issue a warning. Why? Which condition for the existence of a unique MLE is violated?

(c) Perform a "residual" analysis as follows: Plot the pairs $(\widehat{f}(X_i), Y_i)$. Then add some non-parametric least squares fit for these pairs, treating the numbers $\widehat{f}(X_i)$ as values of a real covariate. Compare this or these fits with the logistic function ℓ which interpolates the fitted probabilities $\ell(\widehat{f}(X_i))$.

Exercise 7.17. Write a program that performs a logistic regression analysis for data given by a design matrix $\boldsymbol{D} = [\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_p]$ and an observation vector $\boldsymbol{Y} \in \{0,1\}^n$. Your program should compute for each column $\boldsymbol{D}_j$ of $\boldsymbol{D}$ the profile likelihood ratio p-value of the null hypothesis that the parameter θ_j for $\boldsymbol{D}_j$ equals 0. In addition to these p single p-values, the program should compute adjusted p-values.

Exercise 7.18 (Wald confidence bands). Consider the model of simple logistic regression with regression functions $f(x) = a + bx$, $x \in \mathbb{R}$. Start from observation vectors $\boldsymbol{X} \in \mathbb{R}^n$, $\boldsymbol{Y} \in \{0,1\}^n$ and the corresponding MLE $\widehat{\boldsymbol{\theta}} = (\widehat{a}, \widehat{b})^\top$.

(a) Determine Wald's $(1-\alpha)$-confidence ellipsoid for $\boldsymbol{\theta}$.

(b) Deduce from that a $(1-\alpha)$-confidence band for the true regression function f, that means, simultaneous $(1-\alpha)$-confidence intervals for $f(x)$, $x \in \mathbb{R}$. (Hint: Lemma 3.27.)

(c) Implement this confidence band in a computer program. Your program should have the data vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and the confidence level $1-\alpha$ as input arguments, and it should (enable to) plot the fitted function $\ell \circ \widehat{f}$ as well as the $(1-\alpha)$-confidence band for $\ell \circ f$.

(d) Apply your program to a simulated or real data set.

Exercise 7.19 (Comparison of Wald's method and profile likelihood). As in Exercise 7.18, we consider simple logistic regression with $f(x) = a + bx$, $x \in \mathbb{R}$. Simulate the power function of tests of the null hypothesis "$b = 0$" versus "$b \neq 0$" based on Wald's method and based on profile likelihood as follows:

Choose a vector $\boldsymbol{X} \in [-2,2]^n$ with equispaced components, symmetric around 0. Then generate N times a vector $\boldsymbol{Y} \in \{0,1\}^n$ with independent components such that $\mathbb{P}(Y_i = 1) = \ell(bX_i)$ (logistic model with $a = 0$). Approximate the power function at b by the proportion of simulations in which the tests rejected the null hypothesis at level $\alpha = 0.05$. Plot these powers versus b. Discuss your results briefly.

(Specifically, you could choose $N = 1'000$, $b = 0, 0.5, 1.0, 1.5$ and $n = 25, 100$. If you are more patient, you could choose larger values of N, say, $N = 5'000$ or $N = 10'000$, and a finer grid of values for b.)

7.1.6 Case–Control Studies

In this section, we shall see that the model of logistic regression is applicable in a situation in which the data do *not* follow that model. The starting point are observations $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \{0,1\}$.

Considerations about the standard model. Suppose that $\boldsymbol{X}$ is a random vector with distribution Q, and for some parameters $a_o \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^d$,

$$\mathbb{P}(Y = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}) \;=\; \ell(a_o + \boldsymbol{b}^\top \boldsymbol{x})$$

for arbitrary $\boldsymbol{x} \in \mathbb{R}^d$. Then

$$\begin{aligned} \mathbb{P}(Y = 1) \;&=\; \int \ell(a_o + \boldsymbol{b}^\top \boldsymbol{x})\, Q(d\boldsymbol{x}), \\ \mathbb{P}(Y = 0) \;&=\; \int (1 - \ell(a_o + \boldsymbol{b}^\top \boldsymbol{x}))\, Q(d\boldsymbol{x}), \end{aligned}$$

and the conditional distributions $Q_y := \mathcal{L}(\boldsymbol{X} \,|\, Y = y)$ are given by

$$\begin{aligned} Q_1(B) \;&=\; \int_B \frac{\ell(a_o + \boldsymbol{b}^\top \boldsymbol{x})}{\mathbb{P}(Y = 1)}\, Q(d\boldsymbol{x}), \\ Q_0(B) \;&=\; \int_B \frac{1 - \ell(a_o + \boldsymbol{b}^\top \boldsymbol{x})}{\mathbb{P}(Y = 0)}\, Q(d\boldsymbol{x}). \end{aligned}$$

Hence,

$$\frac{dQ_1}{dQ_0}(\boldsymbol{x}) \;=\; \frac{\ell(a_o + \boldsymbol{b}^\top \boldsymbol{x})}{1 - \ell(a_o + \boldsymbol{b}^\top \boldsymbol{x})} \frac{\mathbb{P}(Y = 0)}{\mathbb{P}(Y = 1)} \;=\; \exp\bigl(a_o - \operatorname{logit} \mathbb{P}(Y = 1) + \boldsymbol{b}^\top \boldsymbol{x}\bigr).$$

Since $\int (dQ_1/dQ_0)(\boldsymbol{x})\, Q_0(d\boldsymbol{x}) = 1$, the parameter a_o is given by

$$a_o \;=\; \operatorname{logit} \mathbb{P}(Y = 1) - C(\boldsymbol{b}) \quad \text{with} \quad C(\boldsymbol{b}) \;:=\; \log \int \exp(\boldsymbol{b}^\top \boldsymbol{x})\, Q_0(d\boldsymbol{x}).$$

From cross-sectional to case–control studies. In many applications, $\boldsymbol{X}$ and Y describe an individual from a population. If one draws a random sample from that population, the resulting observations $(\boldsymbol{X}_1, Y_1)$, $(\boldsymbol{X}_2, Y_2)$, …, $(\boldsymbol{X}_n, Y_n)$ are independent copies of a generic random pair $(\boldsymbol{X}, Y)$. In biomedical application, people talk about *cross-sectional studies*.

Sometimes the probability $\mathbb{P}(Y = 1)$ is rather small, and it could happen that a simple random sample contains too few observations with $Y_i = 1$. A possible way out are *case–control studies*. That means, for $y = 0, 1$ one draws a random sample of fixed size N_y from the subpopulation of individuals with $Y = y$. This leads to observations $(\boldsymbol{X}_1, Y_1)$, $(\boldsymbol{X}_2, Y_2)$, …, $(\boldsymbol{X}_n, Y_n)$ with fixed values Y_i and independent random vectors

$$\boldsymbol{X}_i \;\sim\; \begin{cases} Q_0 & \text{if } Y_i = 0, \\ Q_1 & \text{if } Y_i = 1. \end{cases}$$

The observations with $Y_i = 1$ are often referred to as “cases”, while observations with $Y_i = 0$ are the “controls”.

Validity of logistic regression analysis. One may analyze the data from a case–control study by means of logistic regression to do inference about the parameter vector $\boldsymbol{b}$, although the model assumptions are not fulfilled. This has been shown by Prentice and Pyke (1979). We dispense with a complete formal proof, but present a simple heuristic argument:

Suppose we perform a case–control study, but the group size N_1 is random with distribution $\mathrm{Bin}(n, p_o)$ for a given $p_o \in (0,1)$, while $N_0 = n - N_1$. Then the resulting observations follow the logistic regression model with parameter $\big(\mathrm{logit}\, p_o - C(\boldsymbol{b}), \boldsymbol{b}\big)$ in place of $(a_o, \boldsymbol{b})$. By means of the negative log-likelihood function

$$L(a, \boldsymbol{b}) \;:=\; -\sum_{i=1}^{n} \Big(Y_i (a + \boldsymbol{b}^\top \boldsymbol{X}_i) - \log\big(1 + \exp(a + \boldsymbol{b}^\top \boldsymbol{X}_i)\big) \Big)$$

and the negative profile log-likelihood function

$$L(\boldsymbol{b}) \;:=\; \min_{a \in \mathbb{R}} L(a, \boldsymbol{b})$$

one can construct asymptotically valid tests and confidence regions for $\boldsymbol{b}$. By the way, $\widehat{a}(\boldsymbol{b}) := \arg\min_{a \in \mathbb{R}} L(a, \boldsymbol{b})$ is the unique solution a of the equation

$$\frac{1}{n} \sum_{i=1}^{n} \ell(a + \boldsymbol{b}^\top \boldsymbol{X}_i) \;=\; \bar{Y}_n;$$

see also Exercise 7.2.

7.2 General Asymptotic Considerations

In this section we consider the following scenario: For $n = 1, 2, 3, \ldots$ let $L_n : \mathbb{R}^p \to (-\infty, \infty]$ be a random convex function which is bounded from below. We assume that for each n there exists a fixed parameter $\boldsymbol{\theta}_n \in \mathbb{R}^p$ such that $L_n(\boldsymbol{\theta}_n) < \infty$, and that for arbitrary $\Delta \in \mathbb{R}^p$ we can write

$$L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) \;=\; L_n(\boldsymbol{\theta}_n) - \boldsymbol{Z}_n^\top \Delta + \Delta^\top \boldsymbol{\Gamma} \Delta / 2 + r_n(\Delta)$$

with

- a random vector $\boldsymbol{Z}_n \in \mathbb{R}^p$ such that $\boldsymbol{Z}_n = O_p(1)$,
- a fixed symmetric, positive definite matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times p}$ and
- a random remainder term $r_n(\Delta)$ such that

$$\sup_{\Delta \in \mathbb{R}^p : \|\Delta\| \le C} \big| r_n(\Delta) \big| \;\to_p\; 0 \quad \text{for any fixed } C > 0.$$

(Again, asymptotic statements refer to $n \to \infty$, unless specified differently.)

These properties of $L_n(\cdot)$ imply various important consequences:

Theorem 7.20 (Asymptotics of M estimators). *Under the conditions on $(L_n)_n$ just stated, the function L_n has a minimizer $\widehat{\boldsymbol{\theta}}_n$ with asymptotic probability one. Furthermore, this minimizer satisfies the equations*

$$\begin{aligned} n^{1/2}\big(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\big) \;&=\; \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}_n + o_p(1), \\ 2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) \;&=\; \boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}_n + o_p(1). \end{aligned}$$

Theorem 7.21 (Asymptotics of profile functions). *Let $\boldsymbol{\Psi} \in \mathbb{R}^{p\times k}$ be a fixed matrix with rank $k \le p$, and for $\boldsymbol{\eta} \in \mathbb{R}^k$ let*

$$PL_n(\boldsymbol{\eta}) \;:=\; \inf\bigl\{L_n(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^p, \boldsymbol{\Psi}^\top\boldsymbol{\theta} = \boldsymbol{\eta}\bigr\}.$$

Under the conditions on $(L_n)_n$ just stated, this defines a random convex function $PL_n : \mathbb{R}^k \to (-\infty,\infty]$ which is bounded from below, and $PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n) < \infty$. Moreover, for $\boldsymbol{w} \in \mathbb{R}^k$ we may write

$$PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n + n^{-1/2}\boldsymbol{w}) \;=\; PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n) - \boldsymbol{Z}_{n,\Psi}^\top\boldsymbol{w} + \boldsymbol{w}^\top\boldsymbol{\Gamma}_\Psi\boldsymbol{w}/2 + r_{n,\Psi}(\boldsymbol{w})$$

with the random vector $\boldsymbol{Z}_{n,\Psi} := \boldsymbol{\Gamma}_\Psi\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n = O_p(1)$, the symmetric and positive definite matrix $\boldsymbol{\Gamma}_\Psi := (\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi})^{-1}$ and a random remainder term $r_{n,\Psi}(\boldsymbol{w})$ such that

$$\sup_{\boldsymbol{w}\in\mathbb{R}^k\,:\,\|\boldsymbol{w}\|\le C} \bigl|r_{n,\Psi}(\boldsymbol{w})\bigr| \;\to_p\; 0 \quad \text{for any fixed } C > 0.$$

Remark 7.22. Theorem 7.21 shows that PL_n has similar properties as L_n. If we apply Theorem 7.20 to PL_n in place of L_n, then we obtain the respresentation

$$2PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n) - 2\inf_{\boldsymbol{\eta}\in\mathbb{R}^k} PL_n(\boldsymbol{\eta}) \;=\; \boldsymbol{Z}_{n,\Psi}^\top\boldsymbol{\Gamma}_\Psi^{-1}\boldsymbol{Z}_{n,\Psi} + o_p(1).$$

Moreover, the left hand side equals $2PL_n(\boldsymbol{\Psi}^\top\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n)$ whenever $\widehat{\boldsymbol{\theta}}_n$ exists.

Proof of Theorem 7.20. At first, we assume that $\boldsymbol{\Gamma} = \boldsymbol{I}_p$. To simplify our arguments, we look through a $\sqrt{n}$-magnifying glass and define

$$\begin{aligned} H_n(\Delta) \;&:=\; L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - L_n(\boldsymbol{\theta}_n), \\ \check{H}_n(\Delta) \;&:=\; -\boldsymbol{Z}_n^\top\Delta + \|\Delta\|^2/2 \end{aligned}$$

for $\Delta \in \mathbb{R}^p$. Then $r_n(\Delta) = H_n(\Delta) - \check{H}_n(\Delta)$, and we set

$$\rho_n(C) \;:=\; \max_{\|\Delta\|\le C} \bigl|r_n(\Delta)\bigr|.$$

A vector $\widehat{\Delta}_n \in \mathbb{R}^p$ minimizes H_n if and only if $\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_n + n^{-1/2}\widehat{\Delta}_n$ minimizes L_n. Hence, we have to show that H_n has a minimizer $\widehat{\Delta}_n$ with asymptotic probability one, and that

$$\widehat{\Delta}_n \;=\; \boldsymbol{Z}_n + o_p(1), \quad 2H_n(\widehat{\Delta}_n) \;=\; -\|\boldsymbol{Z}_n\|^2 + o_p(1).$$

The function $\check{H}_n$ has the unique minimizer

$$\operatorname*{arg\,min}_{\Delta\in\mathbb{R}^p} \check{H}_n(\Delta) \;=\; \boldsymbol{Z}_n,$$

because

$$\check{H}_n(\Delta) \;=\; \|\Delta - \boldsymbol{Z}_n\|^2/2 - \|\boldsymbol{Z}_n\|^2/2.$$

In addition,

$$\min_{\Delta\in\mathbb{R}^p} \check{H}_n(\Delta) \;=\; \check{H}_n(\boldsymbol{Z}_n) \;=\; -\|\boldsymbol{Z}_n\|^2/2,$$

and for $\boldsymbol{w} \in \mathbb{R}^p$ we get

$$\check{H}_n(\boldsymbol{Z}_n + \boldsymbol{w}) - \check{H}_n(\boldsymbol{Z}_n) \;=\; \|\boldsymbol{w}\|^2/2. \tag{7.6}$$

By assumption, $\boldsymbol{Z}_n = O_p(1)$. For any fixed $\varepsilon > 0$ (small) and $C > 0$ (large) let

$$A_{n,\varepsilon,C} \;:=\; \big\{\|\boldsymbol{Z}_n\| < C \text{ and } \rho_n(C+\varepsilon) < \varepsilon^2/4\big\}.$$

Now suppose that this event occurs. Then H_n is a real-valued, convex function on the open ball

$$U \;:=\; \{\Delta : \|\Delta\| < C + \varepsilon\} \;\supset\; \{\boldsymbol{Z}_n + \boldsymbol{w} : \boldsymbol{w} \in \mathbb{R}^p, \|\boldsymbol{w}\| \le \varepsilon\}.$$

In particular, H_n is continuous on U. Moreover, it follows from (7.6) that

$$\begin{aligned}
\min_{\|\boldsymbol{w}\|=\varepsilon} H_n(\boldsymbol{Z}_n + \boldsymbol{w}) \;&\ge\; \min_{\|\boldsymbol{w}\|=\varepsilon} \check{H}_n(\boldsymbol{Z}_n + \boldsymbol{w}) - \rho_n(C+\varepsilon) \\
&=\; \check{H}_n(\boldsymbol{Z}_n) + \varepsilon^2/2 - \rho_n(C+\varepsilon) \\
&\ge\; H_n(\boldsymbol{Z}_n) + \varepsilon^2/2 - 2\rho_n(C+\varepsilon) \\
&>\; H_n(\boldsymbol{Z}_n).
\end{aligned}$$

By convexity of H_n on $\mathbb{R}^p$, this implies that even

$$\inf_{\|\boldsymbol{w}\| \ge \varepsilon} H_n(\boldsymbol{Z}_n + \boldsymbol{w}) \;=\; \min_{\|\boldsymbol{w}\|=\varepsilon} H_n(\boldsymbol{Z}_n + \boldsymbol{w}) \;>\; H_n(\boldsymbol{Z}_n).$$

To verify the latter claim, note that for $\boldsymbol{w} = r\boldsymbol{u}$ with a unit vector $\boldsymbol{u} \in \mathbb{R}^p$ and a number $r \ge \varepsilon$,

$$\begin{aligned}
H_n(\boldsymbol{Z}_n + r\boldsymbol{u}) \;&=\; H_n(\boldsymbol{Z}_n) + \big(H_n(\boldsymbol{Z}_n + r\boldsymbol{u}) - H_n(\boldsymbol{Z}_n + 0\boldsymbol{u})\big) \\
&\ge\; H_n(\boldsymbol{Z}_n) + \frac{r}{\varepsilon}\big(H_n(\boldsymbol{Z}_n + \varepsilon\boldsymbol{u}) - H_n(\boldsymbol{Z}_n + 0\boldsymbol{u})\big) \\
&\ge\; H_n(\boldsymbol{Z}_n) + \big(H_n(\boldsymbol{Z}_n + \varepsilon\boldsymbol{u}) - H_n(\boldsymbol{Z}_n + 0\boldsymbol{u})\big) \\
&\ge\; \min_{\|\boldsymbol{w}\|=\varepsilon} H_n(\boldsymbol{Z}_n + \boldsymbol{w}).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\Big|\inf_{\Delta \in \mathbb{R}^p} H_n(\Delta) + \|\boldsymbol{Z}_n\|^2/2\Big| \;&=\; \Big|\min_{\|\boldsymbol{w}\| \le \varepsilon} H_n(\boldsymbol{Z}_n + \boldsymbol{w}) - \min_{\|\boldsymbol{w}\| \le \varepsilon} \check{H}_n(\boldsymbol{Z}_n + \boldsymbol{w})\Big| \\
&\le\; \rho_n(C+\varepsilon) \;<\; \varepsilon^2/4.
\end{aligned}$$

Consequently, in case of the event $A_{n,\varepsilon,C}$ occuring, the function H_n has a minimizer $\widehat{\Delta}_n$, and this minimizer satisfies necessarily

$$\|\widehat{\Delta}_n - \boldsymbol{Z}_n\| \;\le\; \varepsilon \quad \text{and} \quad \big|2H_n(\widehat{\Delta}_n) + \|\boldsymbol{Z}_n\|^2\big| \;<\; \varepsilon^2/2.$$

But note that

$$\begin{aligned}
\mathbb{P}(A_{n,\varepsilon,C}) \;&\ge\; 1 - \mathbb{P}(\|\boldsymbol{Z}_n\| \ge C) - \mathbb{P}\big(\rho_n(C+\varepsilon) \ge \varepsilon^2/4\big) \\
&\ge\; 1 - \sup_{m \ge n} \mathbb{P}(\|\boldsymbol{Z}_m\| \ge C) - o(1) \\
&\to\; 1 - \limsup_{m \to \infty} \mathbb{P}(\|\boldsymbol{Z}_m\| \ge C),
\end{aligned}$$

and the right hand side gets arbitrarily close to one as $C \to \infty$. This proves the theorem in the special case of $\mathbf{\Gamma} = \boldsymbol{I}_p$.

In the case of an arbitrary symmetric, positive definite matrix $\mathbf{\Gamma}$, note that

$$-\boldsymbol{Z}_n^\top \Delta + 2^{-1}\Delta^\top \mathbf{\Gamma}\Delta \;=\; -(\mathbf{\Gamma}^{-1/2}\boldsymbol{Z}_n)^\top(\mathbf{\Gamma}^{1/2}\Delta) + \|\mathbf{\Gamma}^{1/2}\Delta\|^2/2.$$

Hence, for arbitrary $\boldsymbol{\theta}, \Delta \in \mathbb{R}^p$, we introduce

$$\begin{aligned}(\tilde{\boldsymbol{\theta}}_n, \tilde{\boldsymbol{\theta}}, \tilde{\Delta}) \;&:=\; (\mathbf{\Gamma}^{1/2}\boldsymbol{\theta}_n, \mathbf{\Gamma}^{1/2}\boldsymbol{\theta}, \mathbf{\Gamma}^{1/2}\Delta),\\ \tilde{\boldsymbol{Z}}_n \;&:=\; \mathbf{\Gamma}^{-1/2}\boldsymbol{Z}_n,\end{aligned}$$

and

$$\begin{aligned}\tilde{L}_n(\tilde{\boldsymbol{\theta}}) \;&:=\; L_n(\mathbf{\Gamma}^{-1/2}\tilde{\boldsymbol{\theta}}) &&=\; L_n(\boldsymbol{\theta}),\\ \tilde{r}_n(\tilde{\Delta}) \;&:=\; \tilde{L}_n(\tilde{\boldsymbol{\theta}}_n + n^{-1/2}\tilde{\Delta}) - \tilde{L}_n(\tilde{\boldsymbol{\theta}}_n) + \tilde{\boldsymbol{Z}}_n^\top\tilde{\Delta} - \|\tilde{\boldsymbol{Z}}_n\|^2/2 &&=\; r_n(\Delta).\end{aligned}$$

Note that $\tilde{\boldsymbol{\theta}}$ is a minimizer of $\tilde{L}_n$ if and only if $\boldsymbol{\theta} = \mathbf{\Gamma}^{-1/2}\tilde{\boldsymbol{\theta}}$ is a minimizer of L_n, and the infima of $\tilde{L}_n$ and of L_n over $\mathbb{R}^p$ coincide. The assumption that $\|\boldsymbol{Z}_n\| = O_p(1)$ implies that $\|\tilde{\boldsymbol{Z}}_n\| = O_p(1)$ too. Moreover, for any fixed $C > 0$,

$$\sup_{\tilde{\Delta}:\|\tilde{\Delta}\|\le C} |\tilde{r}_n(\tilde{\Delta})| \;=\; \sup_{\Delta:\|\mathbf{\Gamma}^{1/2}\Delta\|\le C} |r_n(\Delta)| \;\le\; \sup_{\Delta:\|\Delta\|\le \lambda_{\min}(\mathbf{\Gamma})^{-1/2}C} |r_n(\Delta)| \;\to_p\; 0,$$

where $\lambda_{\min}(\mathbf{\Gamma})$ denotes the smallest eigenvalue of $\mathbf{\Gamma}$. Hence, the previous considerations for $\mathbf{\Gamma} = \boldsymbol{I}_p$ imply that with asymptotic probability one, there exists a minimizer $\widehat{\boldsymbol{\theta}}_n$ of L_n such that

$$\begin{aligned}n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \;&=\; \mathbf{\Gamma}^{-1/2}\tilde{\boldsymbol{Z}}_n + o_p(1) \;=\; \mathbf{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1),\\ 2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) \;&=\; \|\tilde{\boldsymbol{Z}}_n\|^2 + o_p(1) \quad\;\; =\; \boldsymbol{Z}_n^\top\mathbf{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1).\end{aligned}$$

□

Proof of Theorem 7.21. At first, we assume that $\mathbf{\Gamma} = \boldsymbol{I}_p$ and $\mathbf{\Psi}^\top\mathbf{\Psi} = \boldsymbol{I}_k$. As in the proof of Theorem 7.20, we work with the local negative log-likelihood function H_n, its quadratic approximation $\check{H}_n$ and the error bounds $\rho_n(C)$. If we define the profile functions

$$\begin{aligned}PH_n(\boldsymbol{w}) \;&:=\; \inf\big\{H_n(\Delta) : \Delta \in \mathbb{R}^p, \mathbf{\Psi}^\top\Delta = \boldsymbol{w}\big\},\\ P\check{H}_n(\boldsymbol{w}) \;&:=\; \inf\big\{\check{H}_n(\Delta) : \Delta \in \mathbb{R}^p, \mathbf{\Psi}^\top\Delta = \boldsymbol{w}\big\},\end{aligned}$$

then

$$PL_n(\mathbf{\Psi}^\top\boldsymbol{\theta}_n + n^{-1/2}\boldsymbol{w}) - PL_n(\mathbf{\Psi}^\top\boldsymbol{\theta}_n) \;=\; PH_n(\boldsymbol{w}) - PH_n(\boldsymbol{0}).$$

Hence it suffices to show that

$$\sup_{\|\boldsymbol{w}\|\le C} \big|PH_n(\boldsymbol{w}) - P\check{H}_n(\boldsymbol{w})\big| \;\to_p\; 0 \tag{7.7}$$

for any fixed $C > 0$, and

$$P\check{H}_n(\boldsymbol{w}) - P\check{H}_n(\boldsymbol{0}) \;=\; -\boldsymbol{Z}_{n,\Psi}^\top\boldsymbol{w} + \|\boldsymbol{w}\|^2/2 \tag{7.8}$$

for arbitrary $\boldsymbol{w} \in \mathbb{R}^k$, where $\boldsymbol{Z}_{n,\Psi} := \boldsymbol{\Psi}^\top \boldsymbol{Z}$.

Let us start with (7.8). Note that $\boldsymbol{\Psi}^\top \Delta = \boldsymbol{\Psi}^\top \Delta'$ if and only if $\boldsymbol{\Psi}^\top(\Delta - \Delta') = \boldsymbol{0}$. Thus we represent the space $\mathbb{R}^p$ as the direct sum of the two orthogonal subspaces

$$\mathbb{V}_0 \;:=\; \{\boldsymbol{v} \in \mathbb{R}^p \colon \boldsymbol{\Psi}^\top \boldsymbol{v} = \boldsymbol{0}\} \quad \text{and} \quad \mathbb{V}_1 \;:=\; \boldsymbol{\Psi}\mathbb{R}^k.$$

Recall that

$$\check{H}_n(\Delta) \;=\; \|\Delta - \boldsymbol{Z}_n\|^2/2 - \|\boldsymbol{Z}_n\|^2/2.$$

If we write $\Delta = \Delta_0 + \Delta_1$ and $\boldsymbol{Z}_n = \boldsymbol{Z}_{n0} + \boldsymbol{Z}_{n1}$ with $\Delta_0, \boldsymbol{Z}_{n0} \in \mathbb{V}_0$ and $\Delta_1, \boldsymbol{Z}_{n1} \in \mathbb{V}_0$, then $\boldsymbol{\Psi}^\top \Delta = \boldsymbol{\Psi}^\top \Delta_1 = \boldsymbol{w}$ if and only if

$$\Delta_1 \;=\; \boldsymbol{\Psi}\boldsymbol{w}.$$

Consequently, $\{\Delta \in \mathbb{R}^p \colon \boldsymbol{\Psi}^\top \Delta = \boldsymbol{w}\} = \{\boldsymbol{\Psi}\boldsymbol{w} + \Delta_0 \colon \Delta_0 \in \mathbb{V}_0\}$, and Pythagoras' equality leads to

$$\begin{aligned} P\check{H}_n(\boldsymbol{w}) \;&=\; \inf_{\Delta_0 \in \mathbb{V}_0} \big(\|\Delta_0 - \boldsymbol{Z}_{n0}\|^2/2 + \|\boldsymbol{\Psi}\boldsymbol{w} - \boldsymbol{Z}_{n1}\|^2/2 - \|\boldsymbol{Z}_n\|^2/2\big) \\ &=\; \begin{cases} \check{H}_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0}), \\ \|\boldsymbol{\Psi}\boldsymbol{w} - \boldsymbol{Z}_{n1}\|^2/2 - \|\boldsymbol{Z}_n\|^2/2. \end{cases} \end{aligned}$$

Moreover, $\boldsymbol{Z}_{n1}$ is the orthogonal projection of $\boldsymbol{Z}_n$ onto $\mathbb{V}_1$ which can be written as

$$\boldsymbol{Z}_{n1} \;=\; \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \boldsymbol{Z}_n \;=\; \boldsymbol{\Psi}\boldsymbol{Z}_{n,\Psi}.$$

Hence,

$$\begin{aligned} P\check{H}_n(\boldsymbol{w}) \;&=\; \|\boldsymbol{\Psi}\boldsymbol{w} - \boldsymbol{\Psi}\boldsymbol{Z}_{n,\Psi}\|^2/2 - \|\boldsymbol{Z}_n\|^2/2 \\ &=\; -\boldsymbol{Z}_{n,\Psi}^\top \boldsymbol{w} + \|\boldsymbol{w}\|^2/2 + \|\boldsymbol{Z}_{n1}\|^2/2 - \|\boldsymbol{Z}_n\|^2/2 \\ &=\; -\boldsymbol{Z}_{n,\Psi}^\top \boldsymbol{w} + \|\boldsymbol{w}\|^2/2 - \|\boldsymbol{Z}_{n0}\|^2/2. \end{aligned}$$

In particular, $P\check{H}_n(\boldsymbol{0}) = -\|\boldsymbol{Z}_{n0}\|^2/2$, and this leads to (7.8).

To prove (7.7), note that for any $\boldsymbol{v} \in \mathbb{V}_0$,

$$\check{H}_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0} + \boldsymbol{v}) \;=\; \check{H}_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0}) + \|\boldsymbol{v}\|^2/2.$$

Thus, we fix arbitrary constants $C > 0$ (large) and $\varepsilon > 0$ (small) and consider the event

$$B_{n,\varepsilon,C} \;:=\; \big\{\|\boldsymbol{Z}_{n0}\| < C \text{ and } \rho_n(2C + \varepsilon) < \varepsilon^2/4\big\}.$$

If this event $B_{n,\varepsilon,C}$ occurs, then for $\boldsymbol{w} \in \mathbb{R}^k$ with $\|\boldsymbol{w}\| \le C$ and $\boldsymbol{v} \in \mathbb{V}_0$ with $\|\boldsymbol{v}\| = \varepsilon$,

$$\begin{aligned} &H_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0} + \boldsymbol{v}) - H_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0}) \\ &\quad\ge\; \check{H}_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0} + \boldsymbol{v}) - \check{H}_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0}) - 2\rho_n(2C + \varepsilon) \\ &\quad=\; \varepsilon^2/2 - 2\rho_n(2C + \varepsilon) \;>\; 0. \end{aligned}$$

Together with convexity of H_n, this implies that

$$PH_n(\boldsymbol{w}) \;=\; \min_{\boldsymbol{v} \in \mathbb{V} : \|\boldsymbol{v}\| \le \varepsilon} H_n(\boldsymbol{\Psi}\boldsymbol{w} + \boldsymbol{Z}_{n0} + \boldsymbol{v}),$$

and thus,

$$\begin{aligned}&\big|PH_n(\boldsymbol{w}) - P\check{H}_n(\boldsymbol{w})\big| \\ &= \Big|\min_{\boldsymbol{v}\in\mathbb{V}:\|\boldsymbol{v}\|\le\varepsilon} H_n(\boldsymbol{\Psi}\boldsymbol{w}+\boldsymbol{Z}_{n0}+\boldsymbol{v}) - \min_{\boldsymbol{v}\in\mathbb{V}:\|\boldsymbol{v}\|\le\varepsilon} \check{H}_n(\boldsymbol{\Psi}\boldsymbol{w}+\boldsymbol{Z}_{n0}+\boldsymbol{v})\Big| \\ &\le\ \rho_n(2C+\varepsilon)\ <\ \varepsilon^2/4.\end{aligned}$$

Since $\liminf_{n\to\infty}\mathbb{P}(B_{n,\varepsilon,C}) = \liminf_{n\to\infty}\mathbb{P}(\|\boldsymbol{Z}_{n0}\| < C) \to 1$, as $C\to\infty$, these considerations prove (7.7).

In case of arbitrary matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$, we proceed similarly as in the proof of Theorem 7.20: At first, for arbitrary $\boldsymbol{\theta},\Delta\in\mathbb{R}^p$, we introduce $(\tilde{\boldsymbol{\theta}}_n,\tilde{\boldsymbol{\theta}},\tilde{\Delta}) := (\boldsymbol{\Gamma}^{1/2}\boldsymbol{\theta}_n,\boldsymbol{\Gamma}^{1/2}\boldsymbol{\theta},\boldsymbol{\Gamma}^{1/2}\Delta)$, $\tilde{L}_n(\tilde{\boldsymbol{\theta}}) := L_n(\boldsymbol{\Gamma}^{-1/2}\tilde{\boldsymbol{\theta}}) = L_n(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{Z}}_n := \boldsymbol{\Gamma}^{-1/2}\boldsymbol{Z}_n$. This amounts to a linear transformation of $\mathbb{R}^p$. Furthermore, we introduce

$$\tilde{\boldsymbol{\Psi}}\ :=\ \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Psi}\boldsymbol{\Gamma}_{\Psi}^{1/2}\ =\ \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Psi}(\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi})^{-1/2},$$

and for arbitrary $\boldsymbol{\eta},\boldsymbol{w}\in\mathbb{R}^k$ we define

$$\tilde{\boldsymbol{\eta}}\ :=\ \boldsymbol{\Gamma}_{\Psi}^{1/2}\boldsymbol{\eta}\quad\text{and}\quad\tilde{\boldsymbol{w}}\ :=\ \boldsymbol{\Gamma}_{\Psi}^{1/2}\boldsymbol{w},$$

respectively. The rationale for the latter linear transformation of $\mathbb{R}^k$ is that $\tilde{\boldsymbol{\Psi}}^\top\tilde{\boldsymbol{\Psi}} = \boldsymbol{I}_k$, and the equation $\boldsymbol{\Psi}^\top\boldsymbol{\theta} = \boldsymbol{\eta}$ is equivalent to $\tilde{\boldsymbol{\Psi}}^\top\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\eta}}$. Moreover, $\tilde{\boldsymbol{Z}}_{n,\tilde{\Psi}} := \tilde{\boldsymbol{\Psi}}^\top\tilde{\boldsymbol{Z}}_n$ satisfies the equations $\tilde{\boldsymbol{Z}}_{n,\tilde{\Psi}}^\top\tilde{\boldsymbol{w}} = \boldsymbol{Z}_{n,\Psi}^\top\boldsymbol{w}$ and $\|\tilde{\boldsymbol{Z}}_{n,\tilde{\Psi}}\|^2 = \boldsymbol{Z}_{n\Psi}^\top\boldsymbol{\Gamma}_{\Psi}^{-1}\boldsymbol{Z}_{n,\Psi}$, while $\|\tilde{\boldsymbol{w}}\|^2 = \boldsymbol{w}^\top\boldsymbol{\Gamma}_{\Psi}\boldsymbol{w}$ ... □

7.3 Methods for a Multicategorical Response

In this section, we treat two methods for the general case that $\mathcal{Y} = \{0,1,\ldots,K\}$ for some $K\ge 1$.

7.3.1 Multinomial Logit Models

A justification of logit models. A standard model of multivariate statistics is as follows: Suppose that the joint distribution of random variables $\boldsymbol{X}\in\mathbb{R}^d$ and $Y\in\mathcal{Y} := \{0,1,\ldots,K\}$ with $K\ge 1$ is given by the weights

$$w_y\ :=\ \mathbb{P}(Y=y)\ >\ 0$$

and

$$\mathcal{L}(\boldsymbol{X}\,|\,Y=y)\ =\ \mathrm{N}_d(\boldsymbol{\mu}_y,\boldsymbol{\Sigma})$$

with certain vectors $\boldsymbol{\mu}_0,\ldots,\boldsymbol{\mu}_K\in\mathbb{R}^d$ and a symmetric, positive definite matrix $\boldsymbol{\Sigma}\in\mathbb{R}^{d\times d}$. Then one can easily verify that

$$\begin{aligned}\mathbb{P}(Y=y\,|\,\boldsymbol{X}=\boldsymbol{x})\ &=\ \frac{w_y\exp\big(-(\boldsymbol{x}-\boldsymbol{\mu}_y)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y)/2\big)}{\sum_{z=0}^K w_z\exp\big(-(\boldsymbol{x}-\boldsymbol{\mu}_z)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_z)/2\big)} \\ &=\ \frac{\exp(a_y+\boldsymbol{b}_y^\top\boldsymbol{x})}{\sum_{z=0}^K\exp(a_z+\boldsymbol{b}_z^\top\boldsymbol{x})}\end{aligned}$$

with $a_y := \log(w_y) - \boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y/2$ and $\boldsymbol{b}_y := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y$. These conditional probabilities remain unchanged if we replace the pairs $(a_y, \boldsymbol{b}_y)$ with $(a_y - a_0, \boldsymbol{b}_y - \boldsymbol{b}_0)$. For $K = 1$ we arrive at a logistic regression model! For general $K \geq 1$, we arrive at the multinomial logit model described later.

The previous considerations render logit models plausible, but they also indicate potential starting values for the parameters $a_y, \boldsymbol{b}_y$. Namely, we replace

$$\begin{aligned} w_y \text{ with } \widehat{w}_y &:= N_y/n, \quad \text{where } N_y := \#\{i \leq n : Y_i = y\}, \\ \boldsymbol{\mu}_y \text{ with } \widehat{\boldsymbol{\mu}}_y &:= \frac{1}{N_y} \sum_{i:Y_i=y} \boldsymbol{X}_i, \\ \boldsymbol{\Sigma} \text{ with } \widehat{\boldsymbol{\Sigma}} &:= \frac{1}{n-K-1} \sum_{y=0}^{K} \sum_{i:\,Y_i=y} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_y)(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_y)^\top. \end{aligned}$$

The multinomial logit model. For a given finite-dimensional linear space $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$, we assume that there exist functions $f_1, \ldots, f_K \in \mathcal{F}$ such that

$$\mathbb{P}(Y = y \,|\, X = x) \;=\; \frac{\exp(f_y(x))}{\sum_{z=0}^{K} \exp(f_z(x))} \quad \text{for all } x \in \mathcal{X}, y \in \{0, 1, \ldots, K\},$$

where $f_0(x) := 0$. The functions $f_1, \ldots, f_K$ or corresponding parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ in $\mathbb{R}^{\dim(\mathcal{F})}$ can be estimated via maximum likelihood, and all previous considerations for the negative log-likelihood function may be adapted to the case $K \geq 1$.

7.3.2 The Ordinal Logit Model

In the previous section, the elements $0, 1, \ldots, K$ of $\mathcal{Y}$ did not have a specific meaning, except that 0 took the role of a reference category. But in some applications, Y is an *ordinal* covariate, that means, the values $0, 1, \ldots, K$ represent a canonical order. For instance, in clinical studies one could replace a dichotomous response with values 0 ('healthy') and 1 ('ill') with an ordinal response taking the values 0 ('healthy'), 1 ('slightly ill') and 2 ('seriously ill'). Of course, one could still work with the multinomial logit model, but there is an alternative approach.

Logistic regression via a 'latent response'. Suppose that underlying our data is a linear model with $X \in \mathcal{X}$ and $\widetilde{Y} \in \mathbb{R}$, where

$$\widetilde{Y} \;=\; f(X) + Z$$

with a regression function $f \in \mathcal{F}$ and a random error Z which is stochastically independent from X and follows the logistic distribution function $\ell(\cdot)$. Suppose that instead of $\tilde{Y}$ we only observe

$$Y \;:=\; 1_{[\widetilde{Y} \geq 0]}.$$

Then $(X, Y) \in \mathcal{X} \times \{0, 1\}$ adheres to a logistic regression model, because

$$\mathbb{P}(Y = 1 \,|\, X = x) \;=\; \mathbb{P}(f(x) + Z \geq 0) \;=\; \mathbb{P}(Z \geq -f(x)) \;=\; \ell(f(x)).$$

Note that $1 - \ell(-z) = \ell(z)$ for $z \in \mathbb{R}$.

We may rewrite $f \in \mathcal{F}$ as $f(x) = f_o(x) - a$, where $a := -f(x_o)$ and $f_o(x) := f(x) - f(x_o)$ for a fixed reference value $x_o \in \mathcal{X}$. Then

$$Y \;=\; 1_{[f_o(X)+Z \geq a]}$$

and

$$\begin{aligned}
\mathbb{P}(Y = 1 \,|\, X = x) \;&=\; \ell(f_o(x) - a) \;=\; 1 - \ell(a - f_o(x)), \\
\mathbb{P}(Y = 0 \,|\, X = x) \;&=\; 1 - \ell(f_o(x) - a) \;=\; \ell(a - f_o(x)).
\end{aligned}$$

The general case. The previous construction may be generalized to observations $(X, Y) \in \mathcal{X} \times \{0, 1, \ldots, K\}$: We assume that for certain thresholds

$$-\infty =: a_0 \;<\; a_1 \;<\; \cdots \;<\; a_K \;<\; a_{K+1} := \infty$$

and a function

$$f_o \;\in\; \mathcal{F}_o := \{f - f(x_o) : f \in \mathcal{F}\}$$

the conditional distribution of Y, given X, is given by

$$\mathbb{P}(Y = y \,|\, X = x) \;=\; \ell(a_{y+1} - f_o(x)) - \ell(a_y - f_o(x))$$

for $y = 0, 1, \ldots, K$. Again this corresponds to a latent response

$$f_o(X) + Z,$$

and

$$Y \;=\; y \quad \text{if and only if} \quad a_y \;\leq\; f_o(X) + Z \;<\; a_{y+1}.$$

Figure 7.3 illustrates this construction in case of $K = 2$ and $a_1 = -1.5$, $a_2 = 1.5$. The horizontal axis represents the potential values of $f_o(x)$. Vertically one can see for each value $f_o(x)$ the intervals $[0, \ell(a_1 - f_o(x))]$ (light gray), $[\ell(a_1 - f_o(x)), \ell(a_2 - f_o(x))]$ (gray) and $[\ell(a_2 - f_o(x)), 1]$ (dark gray). The lengths of these intervals are the probabilities $\mathbb{P}(Y = 0 \,|\, X = x)$, $\mathbb{P}(Y = 1 \,|\, X = x)$ and $\mathbb{P}(Y = 2 \,|\, X = x)$, respectively.

Again, one may estimate the parameters $\boldsymbol{a} = (a_y)_{y=1}^K$ and f_o via maximum likelihood. The corresponding negative log-likelihood function is given by

$$L(\boldsymbol{a}, f) \;=\; -\sum_{i=1}^{n} \log\big(\ell(a_{Y_i+1} - f(X_i)) - \ell(a_{Y_i} - f(X_i))\big).$$

Exercise 7.23 (Convexity of the negative log-likelihood function). **(a)** Show that

$$h(\boldsymbol{x}) \;:=\; -\log(\ell(x_2) - \ell(x_1))$$

with $\log(z) := -\infty$ for $z \leq 0$ defines a continuous, convex function $h : \mathbb{R}^2 \to (0, \infty]$ which is strictly convex on $\{h < \infty\} = \{\boldsymbol{x} \in \mathbb{R}^2 : x_1 < x_2\}$. Precisely, for $\boldsymbol{x} \in \mathbb{R}^2$ with $x_1 < x_2$,

$$\nabla h(\boldsymbol{x}) \;=\; \frac{1}{\ell(x_2) - \ell(x_1)} \begin{pmatrix} \ell'(x_1) \\ -\ell'(x_2) \end{pmatrix}$$

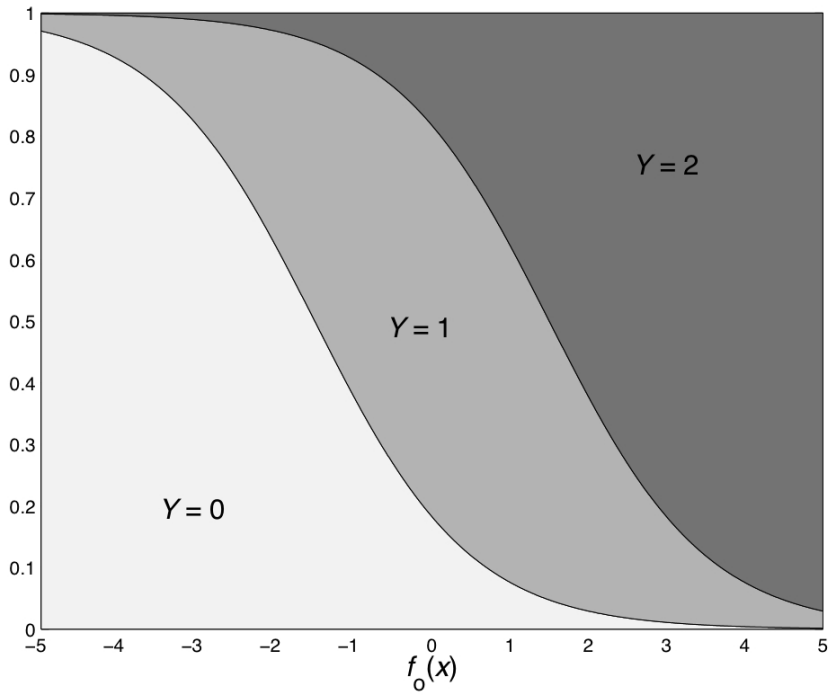

Figure 7.3: Illustration of the ordinal logit model for $K = 2$ and $a_1 = -1.5$, $a_2 = 1.5$.

and

$$D^2 h(\boldsymbol{x}) \;=\; \begin{pmatrix} \ell'(x_1) & 0 \\ 0 & \ell'(x_2) \end{pmatrix} + \frac{\ell'(x_1)\ell'(x_2)}{\big(\ell(x_2) - \ell(x_1)\big)^2} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix}^\top .$$

(b) Show that

$$L(\boldsymbol{a}, f) \;:=\; -\sum_{i=1}^{n} \log\big(\ell(a_{Y_i+1} - f(X_i)) - \ell(a_{Y_i} - f(X_i))\big)$$

defines a convex function $L : \mathbb{R}^K \times \mathcal{F} \to (0, \infty]$, where $L(\boldsymbol{a}, f) < \infty$ if and only if $a_1 < a_2 < \cdots < a_K$.

Exercise 7.24 (Ordinal logit model for a ternary response). Implement the MLE for a simple ordinal logit model with covariate $X \in \mathbb{R}$ and $Y \in \{0, 1, 2\}$. That means, we assume that

$$\mathbb{P}(Y = y \,|\, X = x) = \ell(a_{y+1} - bx) - \ell(a_y - bx)$$

with the parameter $\boldsymbol{\theta} = (a_1, a_2, b)^\top$, where $-\infty =: a_0 < a_1 < a_2 < a_3 := \infty$. Write a program which computes the MLE $\widehat{\boldsymbol{\theta}}$ for given observation vectors $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \{0, 1, 2\}^n$. For the minimization of the negative log-likelihood function, use your own implementation of a Newton–Raphson method or the built-in function optim of R.

Apply your program to the data set USPresidential, where the response variable is party with the order 'Democrat' < 'Independent' < 'Republican'.

7.4 Poisson Regression

Logistic regression is a special instance of the general class of *generalized linear models*, see also Section 7.5. In the present section we present another member of this family. We consider observations $(X, Y) \in \mathcal{X} \times \mathbb{N}_0$ and assume that

$$\mathcal{L}(Y \,|\, X = x) \;=\; \mathrm{Poiss}\big(\exp(f(x))\big) \quad \text{for all } x \in \mathcal{X}$$

with an unknown regression function $f \in \mathcal{F}$. Here $\mathrm{Poiss}(\lambda)$ is the Poisson distribution with parameter (= mean = variance) $\lambda \geq 0$, that means,

$$\begin{aligned} \mathbb{P}(Y = k \,|\, X = x) \;&=\; \exp\big(-\exp(f(x))\big) \frac{\exp(f(x))^k}{k!} \\ &=\; \exp\big(k f(x) - \exp(f(x)) - \log(k!)\big) \quad \text{for } k \in \mathbb{N}_0. \end{aligned}$$

Hence, the resulting negative log-likelihood function is given by

$$L(f) \;:=\; \sum_{i=1}^{n} \big(\exp(f(X_i)) - Y_i f(X_i)\big)$$

plus the additional term $\sum_{i=1}^{n} \log(Y_i!)$ which does not depend on f. Since we are mainly interested in *differences* of the negative log-likelihood function, we ignore the latter term.

Specific examples for this model are:

• Insurance cases: An insurance company is dividing its customers into several groups by means of certain features summarized as X, and for each group, Y is the number of cases in a future time period.

• In a medical experiment with cancer cells, several cell cultures are treated with different kinds or doses of therapeutic agents. Here X describes the treatment for one particular cell culture, and Y could be the number (or concentration) of cells having survived this treatment.

Having specified a basis $f_1, \ldots, f_p$ for $\mathcal{F}$, we may reinterpret L as a function on $\mathbb{R}^p$,

$$L(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^{n} (\exp(\boldsymbol{d}_i^\top \boldsymbol{\theta}) - Y_i \boldsymbol{d}_i^\top \boldsymbol{\theta}),$$

where $\boldsymbol{D} = [f_1(\boldsymbol{X}), \ldots, f_p(\boldsymbol{X})] = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n]^\top$, and the corresponding risk function $R := \mathbb{E}\, L$ is given by

$$R(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} (\exp(\boldsymbol{d}_i^\top \boldsymbol{\theta}) - \mathbb{E}(Y_i) \boldsymbol{d}_i^\top \boldsymbol{\theta}).$$

Exercise 7.25. For nonnegative numbers $a_1, a_2, \ldots, a_n$ and vectors $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n \in \mathbb{R}^p$, we define the function $\widetilde{L} : \mathbb{R}^p \to \mathbb{R}$ via

$$\widetilde{L}(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^{n} (\exp(\boldsymbol{d}_i^\top \boldsymbol{\theta}) - a_i \boldsymbol{d}_i^\top \boldsymbol{\theta}).$$

(a) Derive the gradient and Hessian matrix of this function.

(b) Show that $\widetilde{L}$ is convex, and that it is strictly convex if and only if the vectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n$ span $\mathbb{R}^p$.

(c) Derive necessary and sufficient conditions for coercivity of $\widetilde{L}$.

For the subsequent asymptotic considerations, we consider a triangular scheme of independent observations $(\boldsymbol{d}_{ni}, Y_{ni}) \in \mathbb{R}^p \times \mathbb{N}_0$, $1 \leq i \leq n$, for each $n \in \mathbb{N}$, where the design vectors $\boldsymbol{d}_{ni}$ are viewed as fixed. The negative log-likelihood function $L_n : \mathbb{R}^p \to \mathbb{R}$ is given by

$$L_n(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^{n} \big(\exp(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta}) - Y_{ni} \boldsymbol{d}_{ni}^\top \boldsymbol{\theta} \big),$$

and its first and second derivatives are given by

$$\begin{aligned} \nabla L_n(\boldsymbol{\theta}) \;&=\; -\sum_{i=1}^{n} (Y_{ni} - \exp(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta})) \boldsymbol{d}_{ni}, \\ D^2 L_n(\boldsymbol{\theta}) \;&=\; n \boldsymbol{\Gamma}_n(\boldsymbol{\theta}) \quad \text{with } \boldsymbol{\Gamma}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \exp(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta}) \boldsymbol{d}_{ni} \boldsymbol{d}_{ni}^\top, \end{aligned}$$

see Exercise 7.25. As shown in the latter exercise, L_n and $R_n := \mathbb{E}\, L_n$ are strictly convex if and only if the design matrix $\boldsymbol{D}_n = [\boldsymbol{d}_{n1}\, \boldsymbol{d}_{n2}\, \ldots\, \boldsymbol{d}_{nn}]^\top$ has rank p. Similarly as in the setting of logistic regression, we formulate four regularity assumptions under which the asymptotic behaviour of L_n and related objects can be derived:

(B.1) For sufficiently large n, the design matrix $\boldsymbol{D}_n$ has rank p.

(B.2) For each $n \in \mathbb{N}$, there exists a vector $\boldsymbol{\theta}_n \in \mathbb{R}^p$ such that $Y_{ni} \sim \mathrm{Poiss}(\lambda_{ni})$ with $\lambda_{ni} = \exp(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta}_n)$ for $1 \leq i \leq n$.

(B.3) $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)$ converges to a symmetric, positive definite matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times p}$.

(B.4) There exists a constant $\varepsilon > 0$ such that

$$\Lambda_n(\varepsilon) := \frac{1}{n} \sum_{i=1}^{n} \lambda_{ni} \exp(\varepsilon \|\boldsymbol{d}_{ni}\|) \;=\; O(1).$$

Assumptions (B.1-2) imply that for sufficiently large n, the risk function $R_n = \mathbb{E}\, L_n$,

$$R_n(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} \big(\exp(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta}) - \lambda_{ni} \boldsymbol{d}_{ni}^\top \boldsymbol{\theta} \big)$$

is strictly convex with unique minimizer $\boldsymbol{\theta}_n$.

Theorem 7.26. *For $\Delta \in \mathbb{R}^p$ write*

$$L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - L_n(\boldsymbol{\theta}_n) \;=\; -\boldsymbol{Z}_n^\top \Delta + \Delta^\top \boldsymbol{\Gamma} \Delta / 2 + r_n(\Delta)$$

with the random vector $\boldsymbol{Z}_n := n^{-1/2} \sum_{i=1}^{n} (Y_{ni} - \lambda_{ni}) \boldsymbol{d}_{ni}$ and a remainder $r_n(\Delta)$. Then under the assumptions (B.1-4),

$$\boldsymbol{Z}_n \;\to_{\mathcal{L}}\; \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Gamma}),$$

and

$$\sup_{\Delta\,:\,\|\Delta\|\le C} |r_n(\Delta)| \;\to\; 0 \quad \textit{for any fixed } C > 0.$$

In particular, the log-likelihood function L_n has a unique minimizer $\widehat{\boldsymbol{\theta}}_n$ with asymptotic probability one, and

$$n^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \;=\; \boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1) \;\to_{\mathcal{L}}\; \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Gamma}^{-1}).$$

Moreover, $\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n)$ is a consistent estimator of $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)$, i.e.

$$\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n) \;\to_p\; 0.$$

By means of this theorem and the general results in Section 7.2, one can adapt all tests and confidence regions which we introduced for logistic regression to Poisson regression.

Proof of Theorem 7.26. Again, we start with the matrix-valued function $\boldsymbol{\Gamma}_n(\cdot)$. Here,

$$\begin{aligned}
\big\|\boldsymbol{\Gamma}_n(\boldsymbol{\theta}) - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)\big\| \;&\le\; \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{d}_{ni}\|^2 \big|\exp(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}) - \exp(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}_n)\big| \\
&\le\; \frac{1}{n}\sum_{i=1}^n \lambda_{ni}\|\boldsymbol{d}_{ni}\|^2 \big|\exp(\boldsymbol{d}_{ni}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_n)) - 1\big| \\
&\le\; \frac{1}{n}\sum_{i=1}^n \lambda_{ni}\|\boldsymbol{d}_{ni}\|^2 \big(\exp\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\|\boldsymbol{d}_{ni}\|\big) - 1\big).
\end{aligned}$$

Now we use the inequality

$$\big|z^2(\exp(\delta z) - 1)\big| \;\le\; \frac{(3/e)^3\delta}{(\varepsilon-\delta)^3}\,\exp(\varepsilon|z|) \quad \text{for } z \in \mathbb{C} \text{ and } 0 < \delta < \varepsilon;$$

see Exercise 7.27. If we apply this inequality to $z = \|\boldsymbol{d}_{ni}\|$ and the constant ε in (B.4), then we obtain for any $\delta \in (0, \varepsilon)$ the inequality

$$\sup_{\boldsymbol{\theta}\,:\,\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|\le\delta} \big\|\boldsymbol{\Gamma}_n(\boldsymbol{\theta}) - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)\big\| \;\le\; \frac{(3/e)^3\delta}{(\varepsilon-\delta)^3}\,\Lambda_n(\varepsilon). \tag{7.9}$$

For any fixed $\boldsymbol{\theta} \in \mathbb{R}^p$, the expectation of $L_n(\boldsymbol{\theta})$ is equal to

$$R_n(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^n \big(\exp(\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}) - \lambda_{ni}\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}\big),$$

and $\boldsymbol{\theta}_n$ is the minimizer of R_n by assumption (B.2), so $\nabla R_n(\boldsymbol{\theta}_n) = 0$. Moreover, the difference $(L_n - R_n)(\boldsymbol{\theta}) = -\sum_{i=1}^n (Y_{ni} - \lambda_{ni})\boldsymbol{d}_{ni}^\top\boldsymbol{\theta}$ is linear in $\boldsymbol{\theta}$, whence

$$\begin{aligned}
L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - L_n(\boldsymbol{\theta}_n) \;&=\; -\boldsymbol{Z}_n^\top\Delta + R_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - R_n(\boldsymbol{\theta}_n) \\
&=\; -\boldsymbol{Z}_n^\top\Delta + \Delta^\top\boldsymbol{\Gamma}\Delta/2 + r_n(\Delta)
\end{aligned}$$

with

$$r_n(\Delta) \;:=\; \Delta^\top\big(\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n + \xi_{n,\Delta}\Delta) - \boldsymbol{\Gamma}\big)\Delta/2$$

and $0 \le \xi_{n,\Delta} \le n^{-1/2}$. Consequently, conditions (B.3-4) entail that

$$\sup_{\Delta\,:\,\|\Delta\|\le C} |r_n(\Delta)| \;\le\; \frac{C^2}{2}\Big(\frac{2n^{-1/2}C}{(\varepsilon - n^{-1/2}C)_+^3}\,\Lambda_n(\varepsilon) + \|\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n) - \boldsymbol{\Gamma}\|\Big) \;\to\; 0.$$

It remains to show that the asymptotic distribution of $\boldsymbol{Z}_n$ equals $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Gamma})$. For this purpose we use characteristic functions and show that

$$\mathbb{E}\exp(\boldsymbol{i}\boldsymbol{t}^\top \boldsymbol{Z}_n) \;\to\; \exp(-\boldsymbol{t}^\top\boldsymbol{\Gamma}\boldsymbol{t}/2) \quad \text{for each } \boldsymbol{t} \in \mathbb{R}^p.$$

This is sufficient, because the limit on the right hand side equals $\int_{\mathbb{R}^p} \exp(\boldsymbol{i}\boldsymbol{t}^\top\boldsymbol{z})\,\mathrm{N}_p(\mathbf{0},\boldsymbol{\Gamma})(d\boldsymbol{z})$. Recall that

$$\mathbb{E}\exp(z(Y-\lambda)) \;=\; \exp\big(\lambda(\exp(z) - 1 - z)\big) \quad \text{for } z \in \mathbb{C} \text{ and } Y \sim \mathrm{Poiss}(\lambda);$$

see Exercise 7.28. Applying this formula to $Y = Y_{ni}$, $\lambda = \lambda_{ni}$ and $z = z_{ni} := n^{-1/2}\boldsymbol{i}\boldsymbol{t}^\top\boldsymbol{d}_{ni}$ yields the equation

$$\begin{aligned}
\mathbb{E}\exp(\boldsymbol{i}\boldsymbol{t}^\top \boldsymbol{Z}_n) \;&=\; \exp\Big(\sum_{i=1}^n \lambda_{ni}(\exp(z_{ni}) - 1 - z_{ni})\Big) \\
&=\; \exp\Big(\sum_{i=1}^n \lambda_{ni} z_{ni}^2/2 + \rho_n\Big) \\
&=\; \exp\Big(-\boldsymbol{t}^\top\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n)\boldsymbol{t}/2 + \rho_n\Big) \\
&=\; \exp\Big(-\boldsymbol{t}^\top\boldsymbol{\Gamma}\boldsymbol{t}/2 + o(1) + \rho_n\Big),
\end{aligned}$$

where

$$\rho_n \;:=\; \sum_{i=1}^n \lambda_{ni}\big(\exp(z_{ni}) - 1 - z_{ni} - z_{ni}^2/2\big).$$

Hence, it suffices to show that $\rho_n \to 0$. Since $\mathrm{Re}\, z_{ni} = 0$ and $|z_{ni}| \le \|\boldsymbol{t}\|\|\boldsymbol{d}_{ni}\|$, it follows from part (b) and part (a) of Exercise 7.27 that

$$\big|\exp(z_{ni}) - 1 - z_{ni} - z_{ni}^2/2\big| \;\le\; \frac{|z_{ni}|^3}{6} \;\le\; \frac{\|\boldsymbol{t}\|^3\|\boldsymbol{d}_{ni}\|^3}{6n^{3/2}} \le\; \frac{\tilde{c}(\varepsilon)}{n^{3/2}}\|\boldsymbol{t}\|^3\exp(\varepsilon\|\boldsymbol{d}_{ni}\|)$$

with $\tilde{c}(\varepsilon) := \varepsilon^{-3}(3/e)^3/6$, so

$$|\rho_n| \;\le\; \tilde{c}(\varepsilon)\|\boldsymbol{t}\|^3 n^{-1/2}\Lambda_n(\varepsilon) \;\to\; 0.$$

□

Exercise 7.27. **(a)** Show that for arbitrary $z \in \mathbb{C}$, $k \in \mathbb{N}_0$ and $0 \le \delta < \varepsilon$, the following inequalities hold true:

$$\big|z^k\exp(z\delta)\big| \;\le\; \frac{c_k}{(\varepsilon-\delta)^k}\,\exp(\varepsilon|z|),$$

$$\big|z^k(\exp(z\delta) - 1)\big| \;\le\; \frac{c_{k+1}\delta}{(\varepsilon-\delta)^{k+1}}\,\exp(\varepsilon|z|),$$

where $c_j := (j/e)^j$.

(b) Show that

$$\Big|\exp(z) - \sum_{j=0}^{k} \frac{z^j}{j!}\Big| \;\le\; \frac{|z|^{k+1}}{(k+1)!}\,\exp\big(\max\{\mathrm{Re}\,z, 0\}\big)$$

for arbitrary $k \in \mathbb{N}_0$. Proposal: Show at first that

$$\exp(z) - \sum_{j=0}^{k} \frac{z^j}{j!} \;=\; \frac{z^{k+1}}{k!}\int_0^1 \exp(tz)(1-t)^k\,dt.$$

Exercise 7.28. Let Y have a Poisson distribution with parameter $\lambda \ge 0$. Show that

$$\mathbb{E}\exp\big(z(Y-\lambda)\big) \;=\; \exp\big(\lambda(\exp(z) - 1 - z)\big) \quad \text{for arbitrary } z \in \mathbb{C}.$$

Exercise 7.29 (Poisson regression). Fit a Poisson regression model to the data set Blau.txt with the function glm in R. The response Y should be the number of days a pupil missed school. Interpret your results. Would you consider the Poisson model as plausible? In particular, what do you think about the implicit assumption that $\mathrm{Var}(Y \mid X) = \mathbb{E}(Y \mid X)$?

7.5 Complements

In the previous sections our main focus was on logistic regression and Poisson regression. Both models are special cases of *generalized linear models* which will now be explained briefly.

For statistical inference, i.e. point estimators, tests and confidence regions, we considered mainly likelihood methods. But for these tests and confidence regions we can only guarantee approximate validity. There are at least two procedures with guaranteed finite sample validity which will be described in the context of generalized linear models.

7.5.1 Generalized Linear Models

We consider a generic observation (X, Y) and given observations (X_1, Y_1), …, (X_n, Y_n) with values in $\mathcal{X} \times \mathcal{Y}$. We also write $\boldsymbol{X} = (X_i)_{i=1}^n$ and $\boldsymbol{Y} = (Y_i)_{i=1}^n$. Here we consider $\boldsymbol{X}$ (after conditioning, if necessary) as a fixed vector and $\boldsymbol{Y}$ as a random vector with stochastically independent components.

A generalized linear model has essentially two ingredients:

- A family $(Q_\phi)_{\phi\in\mathbb{R}^L}$ of probability distributions Q_ϕ on $\mathcal{Y}$; here we assume that

$$\mathcal{L}(Y \mid X = x) \;=\; Q_{f_*(x)} \quad \text{and} \quad \mathcal{L}(Y_i) \;=\; Q_{f_*(X_i)}$$

for an unknown regression function $f_* : \mathcal{X} \to \mathbb{R}^L$.

- A finite-dimensional linear space $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}^L$; here we assume that f_* belongs to this space or may be approximated sufficiently well by some function in $\mathcal{F}$.

Example 7.30 (Logistic regression). Here $\mathcal{Y} = \{0, 1\}$, $L = 1$ and

$$Q_\phi \;=\; (1 - \ell(\phi))\delta_0 + \ell(\phi)\delta_1 = \mathrm{Bin}(1, \ell(\phi)).$$

Example 7.31 (Multinomial logit model). Here $\mathcal{Y} = \{0, 1, \ldots, L\}$, $L \geq 2$, and for $\phi \in \mathbb{R}^L$,

$$Q_\phi \;=\; \sum_{y=0}^{L} \exp(\phi_y)\delta_y \Big/ \sum_{z=0}^{L} \exp(\phi_z), \quad \phi_0 := 0.$$

The unknown function f_* is a tuple $(f_{*,y})_{y=1}^L$ of functions $f_{*,y} : \mathcal{X} \to \mathbb{R}$.

Example 7.32 (Poisson regression). Here $\mathcal{Y} = \mathbb{N}_0$, $L = 1$ and

$$Q_\phi \;=\; \mathrm{Poiss}(\exp(\phi)).$$

Example 7.33 (Linear models). Suppose that $\mathcal{Y} = \mathbb{R}$ and

$$Y \;=\; f_*(X) + \varepsilon$$

with an unknown regression function $f_* \in \mathcal{F}$ and a random error ε such that $\mathcal{L}(\varepsilon \,|\, X) \equiv \mathrm{N}(0, \sigma^2)$, $\sigma > 0$ unknown.

With $L = 2$ and

$$Q_\phi \;=\; \mathrm{N}(\phi_1, \exp(\phi_2)^2)$$

one could write

$$\mathcal{L}(Y \,|\, X = x) \;=\; Q_{\tilde{f}_*(x)}$$

where $\tilde{f}_*(x) := (f_*(x), \log \sigma)$.

Exercise 7.34. We consider fixed design vectors $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n \in \mathbb{R}^p$ and stochastically independent observations $Y_1, Y_2, \ldots, Y_n$ with $Y_i \sim \mathrm{N}(\boldsymbol{d}_i^\top \boldsymbol{\theta}, \sigma^2)$. Here, $\boldsymbol{\theta} \in \mathbb{R}^p$ and $\sigma > 0$ are unknown parameters. Determine the maximum likelihood estimator $(\widehat{\boldsymbol{\theta}}, \widehat{\sigma})$ of $(\boldsymbol{\theta}, \sigma)$.

7.5.2 Exact Confidence Regions for f_*

For any $f \in \mathcal{F}$, let $T(f, \boldsymbol{X}, \boldsymbol{Y})$ be a test statistic for the null hypothesis that $f_* = f$, larger values indicating a violation. For instance let

$$T(f, \boldsymbol{X}, \boldsymbol{Y}) \;:=\; L(f) - \inf_{g \in \mathcal{F}} L(g)$$

with the negative log-likelihood function $L = L(\cdot, \boldsymbol{X}, \boldsymbol{Y})$,

$$L(f) = L(f, \boldsymbol{X}, \boldsymbol{Y}) \;:=\; -\sum_{i=1}^{n} \log p_{f(X_i)}(Y_i),$$

where p_ϕ is the density function of Q_ϕ with respect to some measure M on $\mathcal{Y}$.

Now let $\kappa_\alpha(f)$ be the $(1-\alpha)$-quantile of $\mathcal{L}_f(T(f, \boldsymbol{X}, \boldsymbol{Y}))$, where the subscript f indicates that we consider the distribution of $\boldsymbol{Y}$ in case of $f_* = f$. Then

$$C_\alpha(\boldsymbol{X}, \boldsymbol{Y}) \;:=\; \big\{ f \in \mathcal{F} : T(f, \boldsymbol{X}, \boldsymbol{Y}) \leq \kappa_\alpha(f) \big\}$$

is a $(1-\alpha)$-confidence region for f_*, because

$$\mathbb{P}(f_* \in C_\alpha(\boldsymbol{X}, \boldsymbol{Y})) \;=\; \mathbb{P}\big(T(f_*, \boldsymbol{X}, \boldsymbol{Y}) \leq \kappa_\alpha(f_*)\big) \;\geq\; 1 - \alpha$$

by definition of $\kappa_\alpha(f)$.

The explicit computation of this confidence region is a nontrivial task in general. Already the determination of single quantiles $\kappa_\alpha(f)$, $f \in \mathcal{F}$, could be difficult and require Monte Carlo simulations or approximations. Nevertheless, these confidence regions are worth being mentioned and will be revisited in the chapter about bootstrap methods.

Exercise 7.35 (Getting rid of the constant term). We consider multiple logistic or Poisson regression, that means, we consider $\mathcal{X} = \mathbb{R}^d$ and regression functions of the form $f(\boldsymbol{x}) = a + \boldsymbol{b}^\top \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^d$. Suppose we observe a data matrix $\underline{\boldsymbol{X}} = (\boldsymbol{X}_i)_{i=1}^n \in (\mathbb{R}^d)^n$ (viewed as fixed) and a random vector $\boldsymbol{Y} = (Y_i)_{i=1}^n$ in $\{0,1\}^n$ or $\mathbb{N}_0^n$ with stochastically independent components

$$Y_i \sim \begin{cases} \mathrm{Bin}(1, \ell(a + \boldsymbol{b}^\top \boldsymbol{X}_i)) & \text{(logistic model)}, \\ \mathrm{Poiss}(\exp(a + \boldsymbol{b}^\top \boldsymbol{X}_i)) & \text{(Poisson model)}. \end{cases}$$

(a) Let $Y_+ := \sum_{i=1}^n Y_i$. Show that for arbitrary $a \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^d$ and $m \in \mathbb{N}_0$,

$$\mathcal{L}_{a,\boldsymbol{b}}(\boldsymbol{Y} \,|\, Y_+ = m) \;=\; \mathcal{L}_{0,\boldsymbol{b}}(\boldsymbol{Y} \,|\, Y_+ = m).$$

(b) Explain how to construct a confidence region for $\boldsymbol{b}$ with guaranteed confidence level $1 - \alpha$.

7.5.3 Permutation Tests of Association

The null hypothesis that there is no true association between the X- and Y-values may be stated precisely as follows:

Null hypothesis H_o: The distribution of $(\boldsymbol{X}, \boldsymbol{Y})$ does not change if we replace $\boldsymbol{Y}$ with $\sigma\boldsymbol{Y} := (Y_{\sigma(i)})_{i=1}^n$ with an arbitrary fixed permutation $\sigma \in \mathcal{S}_n$.[1]

An equivalent formulation of H_o is: For a random permutation $S \sim \mathrm{Unif}(\mathcal{S}_n)$ which is stochastically independent from $(\boldsymbol{X}, \boldsymbol{Y})$,

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}) \;=\; \mathcal{L}(\boldsymbol{X}, S\boldsymbol{Y}).$$

Here are two special cases of the general null hypothesis H_o:

Null hypothesis H_o': The pairs (X_i, Y_i) are independent copies of a random variable (X, Y), where X and Y are stochastically independent.

Null hypothesis H_o'': The points $X_1, \ldots, X_n$ are fixed, and $Y_1, \ldots, Y_n$ are independent, identically distributed random variables.

Exact p-values for the null hypothesis H_o may be achieved via permutation tests: One chooses a test statistic $T : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ which quantifies the apparent association between X- and

[1] $\mathcal{S}_n$ is the set of all bijections $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Y-values. Then, depending on the choice of T and the working hypothesis, one computes the left-sided p-value

$$\pi_\ell(\boldsymbol{X},\boldsymbol{Y}) \;:=\; \#\big\{\sigma\in\mathcal{S}_n : T(\boldsymbol{X},\sigma\boldsymbol{Y}) \le T(\boldsymbol{X},\boldsymbol{Y})\big\}/n!$$

or the right-sided p-value

$$\pi_r(\boldsymbol{X},\boldsymbol{Y}) \;:=\; \#\big\{\sigma\in\mathcal{S}_n : T(\boldsymbol{X},\sigma\boldsymbol{Y}) \ge T(\boldsymbol{X},\boldsymbol{Y})\big\}/n!$$

or the two-sided p-value

$$\pi_t(\boldsymbol{X},\boldsymbol{Y}) \;:=\; 2\cdot\min\big\{\pi_\ell(\boldsymbol{X},\boldsymbol{Y}),\pi_r(\boldsymbol{X},\boldsymbol{Y})\big\}.$$

Indeed, for arbitrary $\alpha\in(0,1)$,

$$\mathbb{P}(\pi(\boldsymbol{X},\boldsymbol{Y})\le\alpha) \;\le\; \alpha \quad \text{under } H_o.$$

In special cases, such as Fisher's exact test or Wilcoxon's rank sum test, the p-values above may be computed explicitly. Otherwise, one could replace them with Monte-Carlo versions: One simulates stochastically independent permutations $S_1,S_2,\ldots,S_M$ with distribution $\mathrm{Unif}(\mathcal{S}_n)$ which are independent from $(\boldsymbol{X},\boldsymbol{Y})$ and computes

$$\begin{aligned}
\widehat{\pi}_\ell(\boldsymbol{X},\boldsymbol{Y}) \;&:=\; \Big(\#\big\{j\in\{1,2,\ldots,M\} : T(\boldsymbol{X},S_j\boldsymbol{Y})\le T(\boldsymbol{X},\boldsymbol{Y})\big\}+1\Big)\Big/(M+1),\\
\widehat{\pi}_r(\boldsymbol{X},\boldsymbol{Y}) \;&:=\; \Big(\#\big\{j\in\{1,2,\ldots,M\} : T(\boldsymbol{X},S_j\boldsymbol{Y})\ge T(\boldsymbol{X},\boldsymbol{Y})\big\}+1\Big)\Big/(M+1),
\end{aligned}$$

or $\widehat{\pi}_t(\boldsymbol{X},\boldsymbol{Y}) := 2\cdot\min\big\{\widehat{\pi}_\ell(\boldsymbol{X},\boldsymbol{Y}),\widehat{\pi}_r(\boldsymbol{X},\boldsymbol{Y})\big\}$. For arbitrary $\alpha\in(0,1)$,

$$\left.\begin{matrix}\mathbb{P}\big(\widehat{\pi}_\ell(\boldsymbol{X},\boldsymbol{Y})\le\alpha\big)\\ \mathbb{P}\big(\widehat{\pi}_r(\boldsymbol{X},\boldsymbol{Y})\le\alpha\big)\end{matrix}\right\} \;\le\; \frac{\lfloor(M+1)\alpha\rfloor}{M+1} \;\le\; \alpha \quad \text{under } H_o,$$

and this implies that

$$\mathbb{P}\big(\widehat{\pi}_t(\boldsymbol{X},\boldsymbol{Y})\le\alpha\big) \;\le\; \alpha \quad \text{under } H_o.$$

To explain the inequalities for $\widehat{\pi}_\ell$ and $\widehat{\pi}_r$ we introduce the permutation $S_0 := \mathrm{id}$ and write

$$\widehat{\pi}_{\ell\,(r)}(\boldsymbol{X},\boldsymbol{Y}) \;=\; \#\big\{j\in\{0,1,\ldots,M\} : T(\boldsymbol{X},S_j\boldsymbol{Y}) \;\le(\ge)\, T(\boldsymbol{X},S_0\boldsymbol{Y})\big\}/(M+1).$$

Then the asserted inequalities follow from the fact that under H_o, the tuple $\big(T(\boldsymbol{X},S_j\boldsymbol{Y})\big)_{j=0}^M$ is *exchangeable*. That means, its distribution remains invariant if its components are permuted randomly.

Exercise 7.36 (Exact p-values for standard logistic or Poisson regression). We consider multiple logistic or Poisson regression as in Exercise 7.35. Implement a permutation test of the null hypothesis that $\boldsymbol{b}=\boldsymbol{0}$, based on the log-likelihood ratio statistic

$$T(\underline{\boldsymbol{X}},\boldsymbol{Y}) := 2L(\widehat{\boldsymbol{\theta}}_o) - 2L(\widehat{\boldsymbol{\theta}}).$$

Here $\widehat{\boldsymbol{\theta}}$ denotes the MLE of $\boldsymbol{\theta}=(a,\boldsymbol{b})$, while $\widehat{\boldsymbol{\theta}}_o$ denotes the MLE under the null hypothesis. In case of logistic regression, $\widehat{\boldsymbol{\theta}}_o=(\mathrm{logit}(\bar{Y}),\boldsymbol{0})$, in case of Poisson regression, $\widehat{\boldsymbol{\theta}}_o=(\log(\bar{Y}),\boldsymbol{0})$.

Chapter 8

Bootstrap Methods

The name "*bootstrap method*" refers to the idiom that someone "drags himself up by his own bootstraps" (similar to the Duke of Münchhausen, who got out of a swamp by pulling his own hair). These procedures are applicable in many situations and yield tests and confidence regions with given approximate confidence level. After the pioneering paper by Bradley Efron (1979), numerous authors developed and analyzed bootstrap procedures. A good overview is provided by the paper of Bickel and Freedman (1981) and the monograph of Beran and Ducharme (1991).

8.1 Bootstrap Methods for I.I.D. Observations

For $n \in \mathbb{N}$ let $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$ be stochastically independent random variables with unknown distribution P_n on $\mathcal{Y}$. Suppose we are interested in a certain parameter $\theta(P_n) \in \Theta$. To construct a confidence region for $\theta(P_n)$, we choose a mapping $T_n : \mathcal{Y}^n \times \Theta \to (-\infty, \infty]$, where $T_n(\boldsymbol{Y}_n, \theta)$ serves as a test statistic for the null hypothesis that $\theta(P_n) = \theta$, and $\boldsymbol{Y}_n := (Y_{ni})_{i=1}^n$. Let $\kappa_{n,\alpha}(P_n)$ be the $(1-\alpha)$-quantile of $\mathcal{L}\big(T_n(\boldsymbol{Y}_n, \theta(P_n))\big)$. Suppose that an oracle would not tell us $\theta(P_n)$ but the quantile $\kappa_{n,\alpha}(P_n)$. Then we could compute the confidence region

$$C_{n,\alpha}^{\text{oracle}}(\boldsymbol{Y}_n) \;:=\; \big\{\theta \in \Theta : T_n(\boldsymbol{Y}_n, \theta) \le \kappa_{n,\alpha}(P_n)\big\}.$$

Indeed,

$$\mathbb{P}\big(\theta(P_n) \in C_{n,\alpha}^{(\text{oracle})}(\boldsymbol{Y}_n)\big) \;=\; \mathbb{P}\big(T_n(\boldsymbol{Y}_n, \theta(P_n)) \le \kappa_{n,\alpha}(P_n)\big) \;\ge\; 1-\alpha.$$

Classical statistical procedures such as, for instance, student confidence intervals for the mean of a univariate distribution rely on the fact that a proper choice of the test statistic T_n and, if needed, additional assumptions on P_n lead to a quantile $\kappa_{n,\alpha}(P_n)$ which does not depend on P_n. In this case we do not need an oracle. In other situations, we rely on the fact that the unknown quantile $\kappa_{n,\alpha}(P_n)$ converges to a known quantile κ_α. If we replace $\kappa_{n,\alpha}(P_n)$ in $C_{n,\alpha}^{(\text{oracle})}(\boldsymbol{Y}_n)$ with κ_α, we obtain a confidence region with *asymptotic* confidence level $1-\alpha$.

At this point we should explain why we consider triangular arrays, i.e. let P_n depend on the sample size. In the "older days", many authors proved limit theorems for a fixed distribution P and sample

size n tending to infinity. For instance, suppose that one can show that in case of $P_n = P$ for all n,

$$\mathcal{L}\big(T_n(\boldsymbol{Y}_n, \theta(P))\big) \;\to_w\; R$$

for some continuous distribution R on the real line with unique $(1-\alpha)$-quantile κ_α. If the latter quantile is known, then

$$C_{n,\alpha}^{\text{classical}}(\boldsymbol{Y}_n) \;:=\; \big\{\theta \in \Theta : T_n(\boldsymbol{Y}_n, \theta) \le \kappa_\alpha\big\}$$

defines a confidence region for $\theta(P)$ with asymptotic confidence level $1-\alpha$ in the sense that

$$\mathbb{P}\big(\theta(P) \in C_{n,\alpha}^{\text{classical}}(\boldsymbol{Y}_n)\big) \;=\; \mathbb{P}\big(T_n(\boldsymbol{Y}_n, \theta(P)) \le \kappa_\alpha\big) \;\to\; 1-\alpha.$$

At first glance this is wonderful. The problem is, however, that the minimal sample size n such that the true coverage probability is sufficiently close to the nominal confidence level $1-\alpha$ may depend very sensitively on P. A statistician is typically dealing with different distributions but just one sample for each of them. The purpose of asymptotics is to show that certain procedures work for sufficiently large sample size and a larger collection of distributions. In other words, it is important to show that the weak convergence above is to some extent uniform in P.

The idea of bootstrap procedures is to replace the unknown quantile $\kappa_{n,\alpha}(P_n)$ with $\kappa_{n,\alpha}(\widehat{P}_n)$, where $\widehat{P}_n(\cdot) = \widehat{P}_n(\cdot\,; \boldsymbol{Y}_n)$ is an estimator for the distribution P_n. This approach is based on the hope that $\kappa_{n,\alpha}(P_n)$ and $\kappa_{n,\alpha}(\widehat{P}_n)$ are quite similar, so the coverage probability of

$$C_{n,\alpha}(\boldsymbol{Y}_n) \;:=\; \big\{\theta \in \Theta : T_n(\boldsymbol{Y}_n, \theta) \le \kappa_{n,\alpha}(\widehat{P}_n)\big\}$$

is close to $1-\alpha$ or even larger. In other words, we replace $\kappa_{n,\alpha}(P_n)$ by the $(1-\alpha)$-quantile of the random distribution $\mathcal{L}\big(T_n(\boldsymbol{Y}_n^*, \theta(\widehat{P}_n)) \,\big|\, \boldsymbol{Y}_n\big)$. Here $\boldsymbol{Y}_n^* = (Y_{ni}^*)_{i=1}^n$ is a random sample such that conditional on $\boldsymbol{Y}_n$, its elements Y_{ni}^* are independent with distribution $\widehat{P}_n$.

Examples for $\theta(P_n)$ and $T(\boldsymbol{Y}_n, \theta)$. Let $\mathcal{Y} = \mathbb{R}^d$. We assume that $\int \|y\|^2 \, P_n(dy) < \infty$. Now we consider the mean vector and the covariance matrix of P_n,

$$\begin{aligned}
\mu_n = \mu(P_n) \;&:=\; \int y \, P_n(dy) \quad \text{and} \\
\Sigma_n = \Sigma(P_n) \;&:=\; \int (y-\mu_n)(y-\mu_n)^\top \, P_n(dy).
\end{aligned}$$

Natural estimators for μ_n and Σ_n are given by

$$\begin{aligned}
\widehat{\mu}_n \;&:=\; \bar{Y}_n = n^{-1} \sum_{i=1}^n Y_{ni} \quad \text{and} \\
\widehat{\Sigma}_n \;&:=\; (n-1)^{-1} \sum_{i=1}^n (Y_{ni} - \bar{Y}_n)(Y_{ni} - \bar{Y}_n)^\top,
\end{aligned}$$

respectively.

Suppose we are interested primarily in $\theta(P_n) := \mu(P_n) \in \Theta := \mathbb{R}^d$. Possible test statistics would be, for instance,

$$\begin{aligned} T_{n1}(\boldsymbol{Y}_n, \mu) &:= \|\widehat{\mu}_n - \mu\|, \\ T_{n2}(\boldsymbol{Y}_n, \mu) &:= (\widehat{\mu}_n - \mu)^\top \widehat{\Sigma}_n^{-1} (\widehat{\mu}_n - \mu), \\ T_{n3}(\boldsymbol{Y}_n, \mu) &:= \max_{j=1,\dots,d} \widehat{\Sigma}_n(j,j)^{-1/2} \big|\widehat{\mu}_n(j) - \mu(j)\big|. \end{aligned}$$

When working with T_{n2}, we assume that Σ_n is nonsingular and $\widehat{\Sigma}_n$ is nonsingular with high probability. In connection with T_{n3} we assume that Σ_n has strictly positive diagonal elements and that $\widehat{\Sigma}_n$ shares this property with high probability. Confidence regions based on T_{n1} are Euclidean confidence balls, T_{n2} leads to confidence ellipsoids, and T_{n3} yields confidence rectangles.

Suppose we are interested primarily in $\theta(P_n) := \Sigma(P_n)$ and assume that the latter matrix is positive definite. Then a potential test statistic would be, for instance,

$$T_n(\boldsymbol{Y}_n, \Sigma) := \|\widehat{\Sigma}_n^{-1/2} \Sigma \widehat{\Sigma}_n^{-1/2} - I_d\|.$$

Suppose we are interested in the correlation

$$\theta(P_n) := \rho_{12}(P_n) := \frac{\Sigma_n(1,2)}{\sqrt{\Sigma_n(1,1)\Sigma_n(2,2)}}.$$

If Σ_n is positive definite, $\rho_{12}(P_n) \in (-1,1)$. A natural estimator for $\rho_{12}(P_n)$ is given by the sample correlation coefficient $\widehat{\rho}_n := \widehat{\Sigma}_n(1,2)/\sqrt{\widehat{\Sigma}_n(1,1)\widehat{\Sigma}_n(2,2)}$, and potential test statistics would be

$$\begin{aligned} T_{n1}(\boldsymbol{Y}_n, \rho) &:= |\widehat{\rho}_n - \rho|, \\ T_{n2}(\boldsymbol{Y}_n, \rho) &:= \big|\operatorname{artanh}(\widehat{\rho}_n) - \operatorname{artanh}(\rho)\big|. \end{aligned}$$

The latter proposal is suggested by certain considerations in multivariate statistics.

Examples for $\widehat{P}_n$. The distribution P_n is often estimated by the empirical distribution $\widehat{P}_n^{\mathrm{emp}}$ of the observations Y_{ni}, i.e.

$$\widehat{P}_n^{\mathrm{emp}}(B) := \#\{i : Y_{ni} \in B\}/n.$$

In this case, the principle of the bootstrap method may be described as follows: Imagine that the distribution P_n describes a certain large "population" from which a random sample of size n has been drawn. Now we want to estimate the relation between sample and population. To this end, we treat the sample as an artificial population, and we draw artificial samples from this population, with replacement.

To generate such a bootstrap sample $\boldsymbol{Y}_n^* = (Y_{ni}^*)_{i=1}^n$ explicitly, we just simulate independent indices $I(1), I(2), \dots, I(n) \sim \mathrm{Unif}\{1, 2, \dots, n\}$, independent from $\boldsymbol{Y}_n$, and then we set

$$Y_{ni}^* := Y_{n,I(i)}, \quad 1 \le i \le n.$$

In case of $\mathcal{Y} = \mathbb{R}^d$, there is a potential modification: The empirical distribution $\widehat{P}_n^{\mathrm{emp}}$ has the property that $\mu(\widehat{P}_n^{\mathrm{emp}}) = \widehat{\mu}_n = \bar{Y}_n$, but

$$\Sigma(\widehat{P}_n^{\mathrm{emp}}) \;=\; \frac{n-1}{n}\,\widehat{\Sigma}_n.$$

However, since $\widehat{\Sigma}_n$ is an unbiased estimator of $\Sigma(P_n)$ (as known from multivariate statistics), we consider

$$\widehat{P}_n \;:=\; \widehat{P}_n^{\mathrm{emp}} \star \mathrm{N}_d(0, n^{-1}\widehat{\Sigma}_n).$$

That means, for given data $\boldsymbol{Y}_n$, $\widehat{P}_n$ is the distribution of

$$Y^* \;:=\; Y_{n,I} + n^{-1/2}\widehat{\Sigma}_n^{1/2} Z$$

with stochastically independent random variables $I \sim \mathrm{Unif}\{1, 2, \ldots, n\}$ and $Z \sim \mathrm{N}_d(0, I_d)$.

Monte Carlo variants of the bootstrap confidence regions. Typically, no explicit formulae for the bootstrap quantile $\kappa_{n,\alpha}(\widehat{P}_n)$ are available. Then one resorts to the following Monte Carlo variant of $C_{n,\alpha}(\boldsymbol{Y}_n)$: For given data $\boldsymbol{Y}_n$, we simulate independent copies $\boldsymbol{Y}_n^{(1)}, \boldsymbol{Y}_n^{(2)}, \ldots, \boldsymbol{Y}_n^{(M)}$ of $\boldsymbol{Y}_n^*$. With these bootstrap samples we compute the values

$$T_n(\boldsymbol{Y}_n^{(s)}, \theta(\widehat{P}_n)), \quad 1 \le s \le M,$$

and sort them. This leads to the random values $\tau_{n,1} \le \tau_{n,2} \le \cdots \le \tau_{n,M}$. Then we define

$$\widehat{\kappa}_{n,\alpha} \;:=\; \tau_{n,\lceil (M+1)(1-\alpha)\rceil}$$

and

$$\widehat{C}_{n,\alpha}(\boldsymbol{Y}_n) \;:=\; \big\{\theta \in \Theta : T_n(\boldsymbol{Y}_n, \theta) \le \widehat{\kappa}_{n,\alpha}\big\}.$$

The rationale behind the factor $M + 1$ of $1 - \alpha$ is as follows: Suppose for the moment that the $M + 1$ random variables

$$T_n(\boldsymbol{Y}_n, \theta(P_n)),\; T_n(\boldsymbol{Y}_n^{(1)}, \theta(\widehat{P}_n)), \ldots, T_n(\boldsymbol{Y}_n^{(M)}, \theta(\widehat{P}_n))$$

are independent and identically distributed. (At least asymptotically this is often true.) Then the probability that $T_n(\boldsymbol{Y}_n, \theta(P_n)) \le \tau_{n,k}$ is at least $k/(M + 1)$; see Exercise 8.1.

Exercise 8.1 (Monte Carlo critical values). The following inequality is due to Jöckel (1986). Let $(T_0, T_1, T_2, \ldots, T_M) \in \mathbb{R}^{M+1}$ be a random tuple which is exchangeable, that is, its distribution is invariant under arbitrary permutations of its components. Further let $T_{(1)} \le T_{(2)} \le \cdots \le T_{(M)}$ be the order statistics of $T_1, T_2, \ldots, T_M$. (Note that T_0 is excluded!) Show that for $1 \le k \le M$,

$$\mathbb{P}(T_0 \le T_{(k)}) \;\ge\; \frac{k}{M+1}$$

with equality if $T_0, T_1, \ldots, T_M$ are pairwise different almost surely.

Validity of bootstrap methods. A standard strategy to verify validity of a bootstrap procedure consists of two steps:

Step 1: One identifies conditions on the sequence $(P_n)_n$ under which $\mathcal{L}\big(T_n(\boldsymbol{Y}_n, \theta(P_n))\big)$ converges weakly to a continuous distribution R with unique $(1-\alpha)$-quantile κ_α.

Step 2: One shows that the sequence $(\widehat{P}_n)_n$ satisfies the conditions in Step 1 in probability. That means, for each n there exists an event A_n in terms of $\boldsymbol{Y}_n$ such that $\mathbb{P}(A_n) \to 1$, and along this sequence $(A_n)_n$, the sequence $(\widehat{P}_n)_n$ satisfies the same conditions as $(P_n)_n$ in Step 1.

Step 1 implies that $\kappa_{n,\alpha}(P_n) \to \kappa_\alpha$, whereas Step 2 leads to $\kappa_{n,\alpha}(\widehat{P}_n) \to_p \kappa_\alpha$. Both steps together imply that

$$\begin{aligned}\mathbb{P}\big(\theta(P_n) \in C_{n,\alpha}(\boldsymbol{Y}_n)\big) &= \mathbb{P}\big(T_n(\boldsymbol{Y}_n, \theta(P_n)) \le \kappa_{n,\alpha}(\widehat{P}_n)\big) \\ &= \mathbb{P}\big(T_n(\boldsymbol{Y}_n, \theta(P_n)) + o_p(1) \le \kappa_\alpha\big) \\ &\to 1-\alpha.\end{aligned}$$

Note that we do not show directly that $C_{n,\alpha}(\boldsymbol{Y}_n)$ and $C_{n,\alpha}^{\mathrm{oracle}}(\boldsymbol{Y}_n)$ behave similarly in some sense. Instead we take a "detour via asymptopia", showing that the true distibution $\mathcal{L}\big(T_n(\boldsymbol{Y}_n, \theta(P_n))\big)$ converges weakly to a reasonable limit R and that its bootstrap estimator $\mathcal{L}\big(T_n(\boldsymbol{Y}_n^*, \theta(\widehat{P}_n)) \,\big|\, \boldsymbol{Y}_n\big)$ converges weakly to the same limit in probability. In some special settings, people have been able to show that the difference between $\mathcal{L}\big(T_n(\boldsymbol{Y}_n, \theta(P_n))\big)$ and $\mathcal{L}\big(T_n(\boldsymbol{Y}_n^*, \theta(\widehat{P}_n)) \,\big|\, \boldsymbol{Y}_n\big)$ with respect to a suitable distance measure is of smaller order than the difference between both distributions and the limit R. This implies that the bootstrap quantile $\kappa_{n,\alpha}(\widehat{P}_n)$ is a better surrogate for $\kappa_{n,\alpha}(P_n)$ than κ_α. But such considerations are beyond the scope of this course.

Validity of bootstrap methods for the mean. Let us illustrate the standard strategy outlined above in a special case. We consider distributions P_n on $\mathbb{R}^d$ and are interested in the mean vectors $\mu_n = \mu(P_n)$. The following lemma and corollary comprise Step 1 above:

Lemma 8.2. *Suppose that the covariance matrix of P_n converges to a fixed nonzero matrix Σ,*

$$\Sigma(P_n) \to \Sigma, \tag{8.1}$$

and suppose that the following Lindeberg condition is satisfied:

$$\Lambda_n(P_n) := \mathbb{E}\Big(\|Y_{n1} - \mu_n\|^2 \min\big(n^{-1/2}\|Y_{n1} - \mu_n\|, 1\big)\Big) \to 0. \tag{8.2}$$

Then

$$Q_n := \mathcal{L}\big(\sqrt{n}(\widehat{\mu}_n - \mu_n)\big) \to_w \mathrm{N}_d(0, \Sigma)$$

and

$$\mathbb{E}\,\|\widehat{\Sigma}_n - \Sigma\|_F \to 0.$$

If Σ is nonsingular, then $\widehat{\Sigma}_n$ is nonsingular with asymptotic probability one, and

$$\tilde{Q}_n := \mathcal{L}\big(\sqrt{n}\,\widehat{\Sigma}_n^{-1/2}(\widehat{\mu}_n - \mu_n)\big) \to_w \mathrm{N}_d(0, I_d).$$

The distribution $\tilde{Q}_n$ can be viewed as a distribution on the compactified space $\mathbb{R}^d \cup \{'\infty'\}$, where $'\infty'$ is reserved for the (rare) event that $\widehat{\Sigma}_n$ is singular. Lemma 8.2 is essentially a consequence of Lindeberg's CLT (Theorem A.16). It has the following implication:

Corollary 8.3. *Under the conditions of Lemma 8.2,*

$$R_{n1} := \mathcal{L}\big(\sqrt{n}\|\widehat{\mu}_n - \mu_n\|\big) \;\to_w\; R_1 := \mathcal{L}(\|Z\|), \quad \text{with } Z \sim \mathrm{N}_d(0,\Sigma).$$

where $Z \sim \mathrm{N}_d(0,\Sigma)$. If Σ is nonsingular, then

$$R_{n2} := \mathcal{L}\big(n(\widehat{\mu}_n - \mu_n)^\top \widehat{\Sigma}_n^{-1}(\widehat{\mu}_n - \mu_n)\big) \;\to_w\; R_2 := \chi^2_d.$$

If Σ has strictly positive diagonal terms, then

$$\begin{aligned} R_{n3} := \mathcal{L}\Big(\sqrt{n}\max_{j=1,\ldots,d} \widehat{\Sigma}_n(j,j)^{-1/2}\big|\widehat{\mu}_n(j) - \mu_n(j)\big|\Big) \\ \to_w\; R_3 := \mathcal{L}\Big(\max_{j=1,\ldots,d} \Sigma(j,j)^{-1/2}|Z(j)|\Big). \end{aligned}$$

Note that the limiting distributions R_1, R_2, R_3 in this corollary are continuous with unique quantiles. It remains to perform Step 2. This is essentially accomplished with the next lemma:

Lemma 8.4. *Let $\widehat{P}_n$ be the empirical distribution $\widehat{P}_n^{\mathrm{emp}}$ or the "smoothed" empirical distribution $\widehat{P}_n^{\mathrm{emp}} \star \mathrm{N}_d(0, n^{-1}\widehat{\Sigma}_n)$. Under the conditions of Lemma 8.2,*

$$\mathbb{E}\big\|\Sigma(\widehat{P}_n) - \Sigma\big\|_F \;\to\; 0 \quad \text{and} \quad \mathbb{E}\,\Lambda_n(\widehat{P}_n) \;\to\; 0.$$

This lemma implies that there exist numbers $\varepsilon_n > 0$ such that $\varepsilon_n \to 0$ and

$$\mathbb{P}(A_n) \;\to\; 1 \quad \text{with} \quad A_n \;:=\; \big\{\big\|\Sigma(\widehat{P}_n) - \Sigma\big\|_F \le \varepsilon_n \text{ and } \Lambda_n(\widehat{P}_n) \le \varepsilon_n\big\}.$$

Proof of Lemma 8.2. We apply Theorem A.16 with $\boldsymbol{Y}_{ni} := n^{-1/2}(Y_{ni} - \mu_n)$. The assumptions of Theorem A.16 are satisfied with $\boldsymbol{\Sigma}_n = \Sigma_n$ and $\Lambda_n = \Lambda_n(P_n)$. In particular,

$$Z_n := \sqrt{n}(\widehat{\mu}_n - \mu_n) \;=\; \sum_{i=1}^n \boldsymbol{Y}_{ni} \;\to_{\mathcal{L}}\; \mathrm{N}_d(0,\Sigma),$$

$$\mathbb{E}(\|Z_n\|^2) \;=\; \mathrm{trace}(\Sigma_n) \;\to\; \mathrm{trace}(\Sigma)$$

and

$$\mathbb{E}\,\|\check{\Sigma}_n - \Sigma_n\|_F \;\to\; 0$$

with $\check{\Sigma}_n := \sum_{i=1}^n \boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top$. But

$$\begin{aligned} \widehat{\Sigma}_n \;&=\; \frac{1}{n-1}\sum_{i=1}^n (Y_{ni} - \mu_n)(Y_{ni} - \mu_n)^\top - \frac{n}{n-1}(\widehat{\mu}_n - \mu_n)(\widehat{\mu}_n - \mu_n)^\top \\ &=\; \frac{n}{n-1}\check{\Sigma}_n - \frac{1}{n-1}Z_n Z_n^\top, \end{aligned}$$

whence

$$\begin{aligned}\mathbb{E}\,\|\widehat{\Sigma}_n - \Sigma_n\|_F \;&\le\; \frac{n}{n-1}\,\mathbb{E}\,\|\check{\Sigma}_n - \Sigma_n\|_F + \frac{1}{n-1}\|\Sigma_n\|_F + \frac{1}{n-1}\mathrm{trace}(\Sigma_n)\\ &\to\; 0,\end{aligned}$$

by the triangle inequality for $\|\cdot\|_F$ and the fact that $\|Z_n Z_n^\top\|_F = \|Z_n\|^2$. Consequently, the distribution $Q_n = \mathcal{L}(Z_n)$ converges weakly to $\mathrm{N}_d(0,\Sigma)$, and Slutsky's lemma (Exercise A.13) implies that $\tilde{Q}_n = \mathcal{L}(\widehat{\Sigma}_n^{-1/2} Z_n)$ converges weakly to $\mathrm{N}_d(0, I_d)$. □

Proof of Lemma 8.4. Recall that

$$\Sigma\big(\widehat{P}_n^{\mathrm{emp}} \star \mathrm{N}_d(0, n^{-1}\widehat{\Sigma}_n)\big) - \Sigma_n \;=\; \widehat{\Sigma}_n - \Sigma_n$$

and

$$\Sigma(\widehat{P}_n^{\mathrm{emp}}) - \Sigma_n \;=\; \frac{n-1}{n}\widehat{\Sigma}_n - \Sigma_n \;=\; \frac{n-1}{n}(\widehat{\Sigma}_n - \Sigma_n) - \frac{1}{n}\Sigma_n.$$

Thus, Lemma 8.2 implies that

$$\mathbb{E}\,\|\Sigma(\widehat{P}_n) - \Sigma_n\|_F \;\le\; \mathbb{E}\,\|\widehat{\Sigma}_n - \Sigma_n\|_F + n^{-1}\|\Sigma_n\|_F \;\to\; 0.$$

For the claim about $\Lambda_n(\widehat{P}_n)$ we use a simple inequality which is provided in Exercise 8.5 below. This implies that

$$\begin{aligned}\Lambda_n(\widehat{P}_n^{\mathrm{emp}}) \;&=\; \frac{1}{n}\sum_{i=1}^n \|Y_{ni} - \widehat{\mu}_n\|^2 \min\big(n^{-1/2}\|Y_{ni} - \widehat{\mu}_n\|, 1\big)\\ &\le\; \frac{8}{n}\sum_{i=1}^n \|Y_{ni} - \mu_n\|^2 \min\big(n^{-1/2}\|Y_{ni} - \mu_n\|, 1\big) + 8\|\widehat{\mu}_n - \mu_n\|^2,\end{aligned}$$

so

$$\mathbb{E}\,\Lambda_n(\widehat{P}_n^{\mathrm{emp}}) \;\le\; 8\Lambda_n(P_n) + 8n^{-1}\mathrm{trace}(\Sigma_n) \;\to\; 0.$$

For the smoothed empirical distribution $\widehat{P}_n = \widehat{P}_n^{\mathrm{emp}} \star \mathrm{N}_d(0, n^{-1}\widehat{\Sigma}_n)$, the inequality in Exercise 8.5 yields

$$\Lambda_n(\widehat{P}_n) \;\le\; 8\Lambda_n(\widehat{P}_n^{\mathrm{emp}}) + 8n^{-1}\mathrm{trace}(\widehat{\Sigma}_n) \;\to_p\; 0.$$ □

Exercise 8.5. Show that for arbitrary vectors $a, b \in \mathbb{R}^d$ and numbers $c > 0$,

$$\|a+b\|^2 \min(c\|a+b\|, 1) \;\le\; 8\|a\|^2 \min(c\|a\|, 1) + 8\|b\|^2 \min(c\|b\|, 1).$$

Exercise 8.6 (Bootstrap confidence rectangles for the mean). Let $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_n]^\top$ be a data matrix consisting of independent, identically distributed observations $Y_1, Y_2, \ldots, Y_n \in \mathbb{R}^d$ with unknown mean μ and covariance matrix Σ. We assume that the diagonal elements of Σ are strictly positive and would like to determine a confidence rectangle for μ, based on the test statistic

$$T(\boldsymbol{Y}, \mu) \;:=\; \max_{j=1,\ldots,d} \widehat{\Sigma}(j,j)^{-1/2}\big|\overline{Y}(j) - \mu(j)\big|.$$

Write a program performing this task. The input should be the data matrix $\boldsymbol{Y}$, the confidence level $1-\alpha$ (with default 95%) and the number M of Monte Carlo simulations (with default 1999).

8.2 Bootstrap Methods for Regression Models

In the context of regression models, we consider observations $(\boldsymbol{d}_1, Y_1)$, $(\boldsymbol{d}_2, Y_2)$, …, $(\boldsymbol{d}_n, Y_n)$ with fixed vectors $\boldsymbol{d}_i \in \mathbb{R}^p$ and stochastically independent random variables $Y_i \in \mathcal{Y}$. We always assume that the design matrix $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]^\top \in \mathbb{R}^{n \times p}$ has rank p. Moreover, we assume that the distribution of the data vector $\boldsymbol{Y} = (Y_i)_{i=1}^n$ depends on a certain parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ and possibly further nuisance parameters. Now we are interested in

$$\boldsymbol{\Psi}^\top \boldsymbol{\theta} \ \in \ \mathbb{R}^k$$

for a given matrix $\boldsymbol{\Psi} \in \mathbb{R}^{p \times k}$ of rank k.

8.2.1 Logistic and Poisson Regression

In logistic and Poisson regression, the distribution of $\boldsymbol{Y}$ is completely characterized by $\boldsymbol{\theta}$, namely,

$$\mathcal{L}(Y_i) \ = \ \begin{cases} \mathrm{Bin}(1, \ell(\boldsymbol{d}_i^\top \boldsymbol{\theta})) & \text{if } \mathcal{Y} = \{0,1\} \quad \text{(logistic regression)}, \\ \mathrm{Poiss}(\exp(\boldsymbol{d}_i^\top \boldsymbol{\theta})) & \text{if } \mathcal{Y} = \mathbb{N}_0 \quad \text{(Poisson regression)}. \end{cases}$$

An obvious step is to replace the unknown parameter $\boldsymbol{\theta}$ with its MLE $\widehat{\boldsymbol{\theta}}$. This leads to *parametric bootstrap procedures*.

Bootstrap confidence regions. For arbitrary $\boldsymbol{\eta} \in \mathbb{R}^k$ let $T(\cdot, \boldsymbol{\eta}) : \mathcal{Y}^n \to \mathbb{R}$ be a test statistic for the null hypothesis that $\boldsymbol{\Psi}^\top \boldsymbol{\theta} = \boldsymbol{\eta}$. For $\boldsymbol{\theta}_o \in \mathbb{R}^p$, the $(1-\alpha)$-quantile of $\mathcal{L}_{\boldsymbol{\theta}_o}(T(\boldsymbol{Y}, \boldsymbol{\Psi}^\top \boldsymbol{\theta}_o))$ is denoted with $\kappa_\alpha(\boldsymbol{\theta}_o)$. Here the subscript $\boldsymbol{\theta}_o$ refers to the distribution in case of $\boldsymbol{\theta} = \boldsymbol{\theta}_o$. Now we estimate the unknown quantile $\kappa_\alpha(\boldsymbol{\theta})$ by $\kappa_\alpha(\widehat{\boldsymbol{\theta}})$ and define the bootstrap confidence region

$$C_\alpha(\boldsymbol{Y}) \ := \ \big\{ \boldsymbol{\eta} \in \mathbb{R}^k : T(\boldsymbol{Y}, \boldsymbol{\eta}) \le \kappa_\alpha(\widehat{\boldsymbol{\theta}}) \big\}.$$

This is our proxy for the 'confidence region'

$$C_\alpha^{\mathrm{oracle}}(\boldsymbol{Y}) \ := \ \big\{ \boldsymbol{\eta} \in \mathbb{R}^k : T(\boldsymbol{Y}, \boldsymbol{\eta}) \le \kappa_\alpha(\boldsymbol{\theta}) \big\}$$

which would be available if we had access to an oracle revealing the value $\kappa_\alpha(\boldsymbol{\theta})$.

Bootstrap tests. For a fixed $\boldsymbol{\eta} \in \mathbb{R}^k$, we would like to test whether $\boldsymbol{\Psi}^\top \boldsymbol{\theta} = \boldsymbol{\eta}$. To this end let

$$G(r \,|\, \boldsymbol{\theta}_o) \ := \ \mathbb{P}_{\boldsymbol{\theta}_o}\big(T(\boldsymbol{Y}, \boldsymbol{\Psi}^\top \boldsymbol{\theta}_o) < r\big).$$

If an oracle would provide us with the distribution function $G(\cdot \,|\, \boldsymbol{\theta})$ (but not with $\boldsymbol{\theta}$!), a valid p-value would be given by

$$1 - G\big(T(\boldsymbol{Y}, \boldsymbol{\eta}) \,\big|\, \boldsymbol{\theta}\big).$$

As a surrogate we compute the p-value

$$1 - G\big(T(\boldsymbol{Y}, \boldsymbol{\eta}) \,\big|\, \widehat{\boldsymbol{\theta}}\big)$$

or a Monte Carlo version of it. One could also replace the MLE $\widehat{\boldsymbol{\theta}}$ with the MLE $\widehat{\boldsymbol{\theta}}(\boldsymbol{\eta})$ minimizing the negative log-likelihood under the constraint $\boldsymbol{\Psi}^\top \boldsymbol{\theta} = \boldsymbol{\eta}$.

Comparison with classical methods and asymptotic validity. To prove asymptotic validity of the bootstrap methods, we resort to a triangular scheme of observations as in Chapter 7. Thus we consider observation vectors $\boldsymbol{Y} = \boldsymbol{Y}_n = (Y_{ni})_{i=1}^n$, parameter vectors $\boldsymbol{\theta} = \boldsymbol{\theta}_n \in \mathbb{R}^p$ with fixed dimension p, and we use the test statistic

$$T_n(\boldsymbol{Y}_n, \boldsymbol{\eta}) \;:=\; 2PL_n(\boldsymbol{\eta}) - 2L_n(\widehat{\boldsymbol{\theta}}_n)$$

or

$$T_n(\boldsymbol{Y}_n, \boldsymbol{\eta}) \;:=\; n(\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\eta})^\top (\boldsymbol{\Psi}^\top \boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n)^{-1} \boldsymbol{\Psi})^{-1} (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\eta}).$$

Then, under conditions (A.1), (A.2'), (A.3-4) or (B.1-4),

$$T_n(\boldsymbol{Y}_n, \boldsymbol{\Psi}^\top \boldsymbol{\theta}_n) \;\to_{\mathcal{L}}\; \chi^2_k.$$

Hence, the classical confidence regions via Wald's method or profile likelihood are similar to the bootstrap confidence regions. The former replace the unknown quantile $\kappa_{n,\alpha}(\boldsymbol{\theta}_n)$ with $\chi^2_{k;1-\alpha}$ and the unknown distribution function $G_n(\cdot \,|\, \boldsymbol{\theta}_n)$ with the distribution function $F_k(\cdot)$ of χ^2_k.

Conditions (A.1), (A.2'), (A.3-4) or (B.1-4) imply that the bootstrap methods work, too. The reason is that these conditions remain valid if the parameters $\boldsymbol{\theta}_n$ are replaced with $\widehat{\boldsymbol{\theta}}_n$, where usual convergence is replaced with convergence in probability. Precisely:

In case of logistic regression, suppose that conditions (A.1), (A.2') and (A.3-4) are satisfied. Conditions (A.1) and (A.4) concern only the design vectors, so they are valid in the bootstrap world, too. By construction, condition (A.2') is valid in the bootstrap world as well. Condition (A.3) corresponds to the requirement that $\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n) \to_p \boldsymbol{\Gamma}$. The latter is fulfilled, because $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n) \to \boldsymbol{\Gamma}$ by (A.3) and $\boldsymbol{\Gamma}_n(\widehat{\boldsymbol{\theta}}_n) - \boldsymbol{\Gamma}_n(\boldsymbol{\theta}_n) \to_p 0$ by Theorem 7.10.

In case of Poisson regression, suppose that conditions (B.1-4) are satisfied. Then, conditions (B.1-2) carry over to the bootstrap world, whereas condition (B.3) in the bootstrap world is a consequence of the original condition (B.3) and Theorem 7.26. Concerning (B.4), we fix an arbitrary number $\varepsilon' \in (0, \varepsilon)$. Then

$$\begin{aligned}
\widehat{\Lambda}_n \;&:=\; \frac{1}{n} \sum_{i=1}^n \exp(\boldsymbol{d}_{ni}^\top \widehat{\boldsymbol{\theta}}_n) \exp(\varepsilon' \|\boldsymbol{d}_{ni}\|) \\
&=\; \frac{1}{n} \sum_{i=1}^n \exp(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta}_n) \exp\bigl(\boldsymbol{d}_{ni}^\top (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) + \varepsilon' \|\boldsymbol{d}_{ni}\|\bigr) \\
&\le\; \frac{1}{n} \sum_{i=1}^n \exp(\boldsymbol{d}_{ni}^\top \boldsymbol{\theta}_n) \exp\bigl(\bigl(\varepsilon' + \|\boldsymbol{\theta}_n - \widehat{\boldsymbol{\theta}}_n\|\bigr) \|\boldsymbol{d}_{ni}\|\bigr) \\
&\le\; \Lambda_n \quad \text{if } \|\boldsymbol{\theta}_n - \widehat{\boldsymbol{\theta}}_n\| \le \varepsilon - \varepsilon'.
\end{aligned}$$

The latter inequality is satisfied with asymptotic probability one, whence condition (B.4) is satisfied in the bootstrap world, too.

Exercise 8.7 (Bootstrap confidence band for logistic regression). In Exercise 7.18, you implemented a confidence band for the regression function $f(x) = a + bx$ in simple logistic regression, using Wald's method. Now implement a bootstrap version of this procedure, replacing the asymptotic quantile $\chi^2_{2;1-\alpha}$ with the bootstrap quantile $\kappa_\alpha(\widehat{a}, \widehat{b}, \boldsymbol{X})$. Here $\kappa_\alpha(a, b, \boldsymbol{X})$ is the $(1-\alpha)$-quantile of the distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \widehat{\Sigma}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ with $\boldsymbol{\theta} = (a, b)^\top$.

Exercise 8.8 (Bootstrap test for logistic regression). We consider multiple logistic regression, that means, we consider $\mathcal{X} = \mathbb{R}^d$ and assume that the unknown regression function f is given by $f(x) = a + b^\top x$ with an unknown parameter $(a, b) \in \mathbb{R} \times \mathbb{R}^d$. Suppose we observe $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ with $X_i \in \mathbb{R}^d$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ with $Y_i \in \{0, 1\}$. Implement the (Monte Carlo) bootstrap p-value for testing $H_o : b = 0$ versus $H_1 : b \neq 0$ with the likelihood ratio statistic

$$T(\boldsymbol{X}, \boldsymbol{Y}) \ := \ 2L(\mathrm{logit}(\bar{Y}), \boldsymbol{0}) - 2L(\widehat{a}, \widehat{b}).$$

Here, $(\widehat{a}, \widehat{b})$ is the maximum likelihood estimator for (a, b), and $(\mathrm{logit}(\bar{Y}), 0)$ is the MLE under the null hypothesis H_o.

8.2.2 Bootstrap Methods for Linear Models

Now we discuss the classical linear model with observation vector

$$\boldsymbol{Y} \ = \ \boldsymbol{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

Here $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$ is a random vector with stochastically independent components such that $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$. This implies that $\mathbb{E}(\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}}) = \boldsymbol{\Psi}^\top \boldsymbol{\theta}$. In case of $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$, one can write $\mathrm{Var}(\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}}) = \sigma^2 \boldsymbol{\Gamma}_\Psi$ with

$$\boldsymbol{\Gamma}_\Psi \ := \ \boldsymbol{\Psi}^\top (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\Psi} \ = \ \big(\boldsymbol{\psi}_i^\top (\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}_j\big)_{i,j=1}^k,$$

where $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_k$ denote the columns of $\boldsymbol{\Psi}$. Hence, for a hypothetical value $\boldsymbol{\eta} \in \mathbb{R}^k$ of $\boldsymbol{\Psi}^\top \boldsymbol{\theta}$, we consider the following test statistics:

$$\begin{aligned} T_1(\boldsymbol{Y}, \boldsymbol{\eta}) \ &:= \ \sqrt{(\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\eta})^\top \boldsymbol{\Gamma}_\Psi^{-1} (\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\eta})}, \\ T_2(\boldsymbol{Y}, \boldsymbol{\eta}) \ &:= \ \max_{j=1,\ldots,k} \frac{\big|\boldsymbol{\psi}_j^\top \widehat{\boldsymbol{\theta}} - \eta_j\big|}{\sqrt{\boldsymbol{\Gamma}_{\Psi,jj}}}. \end{aligned}$$

Eventually, $T_1(\boldsymbol{Y}, \cdot)$ yields confidence ellipsoids while $T_2(\boldsymbol{Y}, \cdot)$ leads to confidence rectangles for $\boldsymbol{\Psi}^\top \boldsymbol{\theta}$. In the sequel, $T(\boldsymbol{Y}, \cdot)$ stands for one of these two test statistics.

For given matrices $\boldsymbol{D}$ and $\boldsymbol{\Psi}$, the distribution of $T(\boldsymbol{Y}, \boldsymbol{\Psi}^\top \boldsymbol{\theta})$ depends only on the distribution of $\boldsymbol{\varepsilon}$, because

$$\boldsymbol{\Psi}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{\Psi}^\top \boldsymbol{\theta} \ = \ \big(\boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\Psi}\big)^\top \boldsymbol{\varepsilon},$$

and

$$T_1(\boldsymbol{Y}, \boldsymbol{\Psi}^\top \boldsymbol{\theta}) \ = \ \|\boldsymbol{A}^\top \boldsymbol{\varepsilon}\|, \quad T_2(\boldsymbol{Y}, \boldsymbol{\Psi}^\top \boldsymbol{\theta}) \ = \ \|\boldsymbol{B}^\top \boldsymbol{\varepsilon}\|_\infty$$

with the Euclidean norm $\|\cdot\|$, the maximum norm $\|\cdot\|_\infty$, and the matrices

$$\begin{aligned} \boldsymbol{A} \ &:= \ \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\Psi} \boldsymbol{\Gamma}_\Psi^{-1/2} \ \in \ \mathbb{R}^{n \times k}, \\ \boldsymbol{B} \ &:= \ [\boldsymbol{b}_1 \, \boldsymbol{b}_2 \, \ldots \, \boldsymbol{b}_k] \ \in \ \mathbb{R}^{n \times k} \quad \text{with} \quad \boldsymbol{b}_j \ := \ \boldsymbol{\Gamma}_{\Psi,jj}^{-1/2} \boldsymbol{D}(\boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{\psi}_j. \end{aligned}$$

The precise definition of these matrices is irrelevant for our theoretical considerations. The only properties which will be used later are that

$$\boldsymbol{A}\mathbb{R}^k, \boldsymbol{B}\mathbb{R}^k \ \subset \ \boldsymbol{D}\mathbb{R}^p$$

and

$$\operatorname{trace}(\boldsymbol{A}^\top \boldsymbol{A}) \;=\; \operatorname{trace}(\boldsymbol{B}^\top \boldsymbol{B}) \;=\; k.$$

In connection with $T_2(\boldsymbol{Y}, \boldsymbol{\eta})$ we do not need the assumption that $\boldsymbol{\Psi}$ has rank k. Here, the condition that all columns $\boldsymbol{\psi}_j$ are different from $\mathbf{0}$ is sufficient.

8.2.3 The Residual Bootstrap

Suppose that the components $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ of $\boldsymbol{\varepsilon}$ are not only independent but even identically distributed with an unknown distribution function F. We also assume that

$$\mu(F) \;=\; 0 \quad \text{and} \quad 0 \;<\; \sigma(F) \;<\; \infty.$$

Here and throughout this section, we identify a distribution on the real line with its distribution function. For simplicity, we also assume that the column space $\boldsymbol{D}\mathbb{R}^p$ contains the constant vector $\mathbf{1}_n$.

For $T \;=\; T_1, T_2$, the distribution of $T(\boldsymbol{Y}, \boldsymbol{\Psi}^\top \boldsymbol{\theta})$ depends only on the distribution function F. Hence, let $\kappa_\alpha(F) = \kappa_\alpha(F, \boldsymbol{D}, \boldsymbol{\Psi})$ be the $(1-\alpha)$-quantile of

$$\mathcal{L}\big(\|\boldsymbol{A}^\top \boldsymbol{\varepsilon}\|\big) \quad \text{or} \quad \mathcal{L}\big(\|\boldsymbol{B}^\top \boldsymbol{\varepsilon}\|_\infty\big).$$

Now this unknown quantile is estimated by $\kappa_\alpha(\widehat{F})$, where $\widehat{F} = \widehat{F}(\cdot, \boldsymbol{Y}, \boldsymbol{D})$ is a suitable estimator for F. The following lemma shows that $\kappa_\alpha(\widehat{F})$ is a good surrogate for $\kappa_\alpha(F)$, as soon as $\widehat{F}$ is close to F. Here we quantify the distance between distributions or distribution functions with the Mallows distances. The latter are treated in detail in Section A.8.

Lemma 8.9 (Bickel and Freedman, 1983). *Let F and G be distribution functions on $\mathbb{R}$ such that $\mu(F) = \mu(G) = 0$ and $\sigma(F), \sigma(G) < \infty$. Further, let $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ be random vectors in $\mathbb{R}^n$ with distribution $F^{\otimes n}$ and $G^{\otimes n}$, respectively. Then, for any fixed matrix $\boldsymbol{A} \in \mathbb{R}^{n \times k}$,*

$$\left.\begin{array}{c} d_{M,2}\Big(\mathcal{L}\big(\|\boldsymbol{A}^\top \boldsymbol{\delta}\|\big), \mathcal{L}\big(\|\boldsymbol{A}^\top \boldsymbol{\varepsilon}\|\big)\Big) \\ d_{M,2}\Big(\mathcal{L}\big(\|\boldsymbol{A}^\top \boldsymbol{\delta}\|_\infty\big), \mathcal{L}\big(\|\boldsymbol{A}^\top \boldsymbol{\varepsilon}\|_\infty\big)\Big) \end{array}\right\} \;\le\; \sqrt{\operatorname{trace}(\boldsymbol{A}^\top \boldsymbol{A})}\, d_{M,2}(F, G).$$

Proof of Lemma 8.9. We consider a special coupling of $F^{\otimes n}$ and $G^{\otimes n}$: Let $U_1, U_2, \ldots, U_n$ be stochastically independent and uniformly distributed on $(0, 1)$. Then

$$\boldsymbol{\delta} \;:=\; \big(F^{-1}(U_i)\big)_{i=1}^n \quad \text{and} \quad \boldsymbol{\varepsilon} \;:=\; \big(G^{-1}(U_i)\big)_{i=1}^n$$

follow the intended distributions, and

$$\mathbb{E}\big((\boldsymbol{\delta} - \boldsymbol{\varepsilon})(\boldsymbol{\delta} - \boldsymbol{\varepsilon})^\top\big) \;=\; d_{M,2}(F, G)^2\, \boldsymbol{I}_n.$$

By definition of $d_{M,2}$ and the triangle inequality for norms,

$$\begin{aligned} &d_{M,2}\Big(\mathcal{L}\big(\|\boldsymbol{A}^\top \boldsymbol{\delta}\|_{(\infty)}\big), \mathcal{L}\big(\|\boldsymbol{A}^\top \boldsymbol{\varepsilon}\|_{(\infty)}\big)\Big)^2 \\ &\qquad \le\; \mathbb{E}\Big[\big(\|\boldsymbol{A}^\top \boldsymbol{\delta}\|_{(\infty)} - \|\boldsymbol{A}^\top \boldsymbol{\varepsilon}\|_{(\infty)}\big)^2\Big] \;\le\; \mathbb{E}\big[\|\boldsymbol{A}^\top (\boldsymbol{\delta} - \boldsymbol{\varepsilon})\|_{(\infty)}^2\big]. \end{aligned}$$

But

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{A}^\top(\boldsymbol{\delta}-\boldsymbol{\varepsilon})\|_\infty^2\big] &\le \mathbb{E}\big[\|\boldsymbol{A}^\top(\boldsymbol{\delta}-\boldsymbol{\varepsilon})\|^2\big] \\ &= \mathbb{E}\ \mathrm{trace}\big(\boldsymbol{A}^\top(\boldsymbol{\delta}-\boldsymbol{\varepsilon})(\boldsymbol{\delta}-\boldsymbol{\varepsilon})^\top\boldsymbol{A}\big) \\ &= \mathrm{trace}(\boldsymbol{A}^\top\boldsymbol{A})\, d_{M,2}(F,G)^2.\end{aligned}$$

□

Estimation of the error distribution function F. We discuss this topic in a triangular array setting. That means, we consider the random vector

$$\boldsymbol{Y}_n = \boldsymbol{D}_n\boldsymbol{\theta}_n + \boldsymbol{\varepsilon}_n$$

with a design matrix $\boldsymbol{D}_n \in \mathbb{R}^{n\times p(n)}$ of rank $p(n)$ such that $\boldsymbol{D}_n\mathbb{R}^{p(n)}$ contains $\mathbf{1}_n$, an unknown parameter vector $\boldsymbol{\theta}_n \in \mathbb{R}^{p(n)}$, and an error vector $\boldsymbol{\varepsilon}_n = (\varepsilon_{ni})_{i=1}^n$ with distribution $F_n^{\otimes n}$ and an unknown distribution function F_n. Again, we assume that $\mu(F_n) = 0$ and $\sigma(F_n) < \infty$. Suppose that our primary interest is in $\boldsymbol{\Psi}_n^\top\boldsymbol{\theta}_n$ with a given matrix $\boldsymbol{\Psi}_n \in \mathbb{R}^{p(n)\times k(n)}$ of rank $k(n)$. Under the assumption that

$$k(n) = O(1),$$

it is only important that our estimator $\widehat{F}_n = \widehat{F}_n(\cdot, \boldsymbol{Y}_n, \boldsymbol{D}_n)$ satisfies the following condition:

$$d_{M,2}(\widehat{F}_n, F_n) \to_p 0. \tag{8.3}$$

With the LSE $\widehat{\boldsymbol{\theta}}_n$ and the corresponding residual vector $\widehat{\boldsymbol{\varepsilon}}_n = \boldsymbol{Y}_n - \boldsymbol{D}_n\widehat{\boldsymbol{\theta}}_n$, we estimate the distribution function F_n by the empirical distribution function $\widehat{F}_n$ of the residuals, given by

$$\widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^n 1_{[\widehat{\varepsilon}_{ni}\le x]}.$$

This estimator $\widehat{F}_n$ satisfies (8.3) as soon as $p(n)$ is small in comparison to n:

Lemma 8.10 (Bickel and Freedman, 1983). *Suppose that*

$$p(n)/n \to 0 \quad \textit{and} \quad d_{M,2}(F_n, F) \to 0$$

for a fixed distribution function F such that $\mu(F) = 0$ and $\sigma(F) < \infty$. Then the empirical distribution function $\widehat{F}_n$ of the residuals satisfies condition (8.3).

If we combine lemmas 8.9 and 8.10, then we may conclude that the residual bootstrap works well whenever the ratio p/n is small (and k stays bounded). Comparing this with the results in Section 4.2, it is remarkable that we do not need any condition on the maximal leverage,

$$\max_{i=1,2,\ldots,n} H_{ii} \quad \text{with} \quad \boldsymbol{H} = \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top.$$

Proof of Lemma 8.10. By the triangle inequality for $d_{M,2}$,

$$d_{M,2}(\widehat{F}_n, F_n) \le d_{M,2}(\widehat{F}_n, \check{F}_n)) + d_{M,2}(\check{F}_n, F) + d_{M,2}(F_n, F),$$

where $\check{F}_n$ is the empirical distribution (function) of the actual errors ε_{ni}, $1 \le i \le n$.

To get an upper bound for $d_{M,2}(\widehat{F}_n, \check{F}_n)$, we consider a very simple coupling of these two distributions, namely,

$$R_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{(\widehat{\varepsilon}_{ni},\varepsilon_{ni})}.$$

This yields the bound

$$\begin{aligned} d_{M,2}(\widehat{F}_n, \check{F}_n)^2 \;&\le\; \frac{1}{n}\sum_{i=1}^{n}(\widehat{\varepsilon}_{ni}-\varepsilon_{ni})^2 \;=\; \frac{1}{n}\,\|\widehat{\boldsymbol{\varepsilon}}_n - \boldsymbol{\varepsilon}_n\|^2 \\ &=\; \frac{1}{n}\,\big\|(\boldsymbol{I}_n - \boldsymbol{H}_n)\boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_n\big\|^2 \;=\; \frac{1}{n}\,\|\boldsymbol{H}_n\boldsymbol{\varepsilon}_n\|^2 \;=\; \frac{1}{n}\,\mathrm{trace}(\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top \boldsymbol{H}_n). \end{aligned}$$

Here $\boldsymbol{H}_n$ stands for the hat matrix $\boldsymbol{D}_n(\boldsymbol{D}_n^\top\boldsymbol{D}_n)^{-1}\boldsymbol{D}_n^\top$, and in the last step we used the fact that $\boldsymbol{H}_n^\top = \boldsymbol{H}_n = \boldsymbol{H}_n^2$. Together with the equality $\mathrm{trace}(\boldsymbol{H}_n) = p(n)$, these considerations show that

$$\mathbb{E}\big[d_{M,2}(\widehat{F}_n, \check{F}_n)^2\big] \;\le\; \frac{1}{n}\,\mathbb{E}\,\mathrm{trace}(\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top\boldsymbol{H}_n) \;=\; \frac{\sigma(F_n)^2 p(n)}{n} \;\to\; 0,$$

because $\sigma(F_n)^2 \to \sigma(F)^2$.

Since $d_{M,2}(F_n, F) \to 0$, it remains to show that $d_{M,2}(\check{F}_n, F) \to_p 0$. The dimension $p(n)$ is irrelevant for $\check{F}_n$. On the one hand, it is well-known that

$$\mathbb{P}\big(\|\check{F}_n - F_n\|_\infty \ge \eta\big) \;\le\; C_1\exp(-C_2 n\eta^2) \quad \text{for all } \eta \ge 0$$

with universal constants $C_1 \ge 2$ and $C_2 > 0$. On the other hand, we may apply Theorem A.16 to $\boldsymbol{Y}_{ni} := n^{-1/2}\varepsilon_{ni}$. Indeed, $\mathbb{E}(\boldsymbol{Y}_{ni}) = 0$ for $1 \le i \le n$, and

$$\sum_{i=1}^{n}\mathbb{E}(\boldsymbol{Y}_{ni}^2) \;=\; \sigma(F_n)^2 \to \sigma(F)^2.$$

Furthermore,

$$\begin{aligned} \limsup_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\big(\boldsymbol{Y}_{ni}^2\min(|\boldsymbol{Y}_{ni}|,1)\big) \;&=\; \limsup_{n\to\infty}\int x^2\min(n^{-1/2}|x|,1)\,F_n(dx) \\ &\le\; \inf_{\delta>0}\limsup_{n\to\infty}\int x^2\min(\delta|x|,1)\,F_n(dx) \\ &=\; \inf_{\delta>0}\int x^2\min(\delta|x|,1)\,F(dx) \\ &=\; 0. \end{aligned}$$

Consequently, the second moment $\tau^2(\check{F}_n) := \int x^2\,\check{F}_n(dx)$ satisfies

$$\mathbb{E}\big|\tau^2(\check{F}_n) - \sigma(F_n)^2\big| \;=\; \mathbb{E}\Big|\sum_{i=1}^{n}\boldsymbol{Y}_{ni}^2 - \sigma(F_n)^2\Big| \;\to\; 0.$$

(We work with $\tau^2(\cdot)$ instead of $\sigma(\cdot)^2$, because the mean of $\check{F}_n$ may differ from 0.) Hence, there exists a sequence $(\delta_n)_n$ of positive numbers tending to 0 such that the event

$$A_n \;:=\; \Big\{\|\check{F}_n - F_n\|_\infty + \big|\tau^2(\check{F}_n) - \sigma(F_n)^2\big| \le \delta_n\Big\}$$

has asymptotic probability 1. But along $(A_n)_n$,

$$\big|\check{F}_n(x) - F(x)\big| \;\le\; \big|F_n(x) - F(x)\big| + \delta_n \;\to\; 0$$

for any continuity point x of F, and

$$\big|\tau^2(\check{F}_n) - \sigma(F)^2\big| \;\le\; \big|\sigma(F_n)^2 - \sigma(F)^2\big| + \delta_n \;\to\; 0.$$

Hence, we can deduce from Theorem A.31 that $d_{M,2}(\check{F}_n, F) \to 0$ along $(A_n)_n$. □

Refinements. The deviation of the true coverage probabilities of the bootstrap confidence sets just defined from the nominal level $1-\alpha$ becomes typically smaller, if the test statistics $T(\boldsymbol{Y}, \boldsymbol{\eta})$ are divided by $\widehat{\sigma} = \|\widehat{\boldsymbol{\varepsilon}}\|/\sqrt{n-p}$, a so-called *studentization*.

In addition, it can be beneficial to replace the empirical distribution $\widehat{F}$ of the residuals with

$$\widehat{F}^{\mathrm{s}} \;:=\; \widehat{F} \star \mathrm{N}\Big(0, \frac{p}{n}\,\widehat{\sigma}^2\Big).$$

For the variance of the latter distribution $\widehat{F}^{\mathrm{s}}$ is precisely the unbiased estimate $\widehat{\sigma}^2$ of $\sigma(F)^2$. That means, for given data $\boldsymbol{Y}$ we simulate an error vector $\boldsymbol{\varepsilon}^* \sim (\widehat{F}^{\mathrm{s}})^{\otimes n}$ as follows:

$$\boldsymbol{\varepsilon}^* \;:=\; \big(\widehat{\varepsilon}_{J_i} + (p/n)^{1/2}\widehat{\sigma} Z_i\big)_{i=1}^n$$

with independent random variables $J_1, \ldots, J_n \sim \mathrm{Unif}\{1, \ldots, n\}$ and $Z_1, \ldots, Z_n \sim \mathrm{N}(0,1)$.

Exercise 8.11 (Estimation of the error distribution). Choose fixed numbers $X_1, \ldots, X_n$ forming a regular grid in $[-3, 3]$, and then simulate random variables

$$Y_i \;=\; 2X_i + 1 + \varepsilon_i, \quad 1 \le i \le n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random errors with distribution (function) F such that $\mu(F) = 0$ and $\sigma(F) = 1$.

(a) Estimate F by the empirical distribution (function) $\widehat{F}$ of the residuals resulting from the model $Y = a + bX + \varepsilon$. Visualize F and $\widehat{F}$. Do this for $n \in \{100, 1000\}$ and the following distributions F:

(a.1) $\mathrm{Unif}[-c_1, c_1]$,

(a.2) $\mathcal{L}(c_2\xi)$ with $\xi \sim t_3$,

(a.3) centered Gamma distribution with shape parameter 2 and scale parameter c_3.

The parameters c_1, c_2, c_3 should be chosen such that $\sigma(F) = 1$.

(b) Repeat part (a), but this time with the smoothed distribution (function) $\widehat{F}^{\mathrm{s}}$ in place of $\widehat{F}$. Do you see an improvement?

8.2.4 Wild Bootstrap

It is remarkable that the residual bootstrap allows us to avoid the assumption of Gaussian errors or of small leverages. But in numerous applications the assumption of homoscedastic errors is questionable, not to mention identically distributed errors. For such situations, Jeff Wu (1986), Rudolf Beran (1986) and Regina Liu (1988) proposed a procedure which had been analyzed later by Enno Mammen (1993) and is nowadays called "wild bootstrap". This name refers to the ambitious idea to estimate n different error distributions from n observations.

Precisely, one estimates the unknown distribution $\mathcal{L}(\boldsymbol{\varepsilon})$ by $\mathcal{L}(\boldsymbol{\varepsilon}^* \,|\, \boldsymbol{Y})$, where $\boldsymbol{\varepsilon}^*$ is constructed as follows: Let $\xi, \xi_1, \xi_2, \xi_3, \ldots$ be independent and identically distributed random variables with

$$\mathbb{E}(\xi) \;=\; 0 \quad \text{and} \quad \mathrm{Var}(\xi) \;=\; 1,$$

also independent from $\boldsymbol{Y}$. Then we define

$$\boldsymbol{\varepsilon}^* \;:=\; \bigl(\xi_i \widehat{\varepsilon}_i\bigr)_{i=1}^n.$$

In other words, for each index i, the distribution of ε_i is estimated by $\mathcal{L}(\xi\,\widehat{\varepsilon}_i \,|\, \boldsymbol{Y})$. Of course, these estimators are not very accurate for a single index i. But for our purposes it is only important that $\mathcal{L}(\boldsymbol{A}^\top \boldsymbol{\varepsilon}^* \,|\, \boldsymbol{Y})$ is a good approximation for $\mathcal{L}(\boldsymbol{A}^\top \boldsymbol{\varepsilon})$. Here $\boldsymbol{A} \in \mathbb{R}^{n\times k}$ is a given matrix such that

$$\boldsymbol{A}\mathbb{R}^k \;\subset\; \boldsymbol{D}\mathbb{R}^p \quad \text{and} \quad \mathrm{trace}(\boldsymbol{A}^\top \boldsymbol{A}) \;=\; k.$$

Validity of wild bootstrap. Again, we consider a triangular scheme of observation vectors $\boldsymbol{Y}_n = \boldsymbol{D}_n \boldsymbol{\theta}_n + \boldsymbol{\varepsilon}_n$. Here the design matrix $\boldsymbol{D}_n = [\boldsymbol{d}_{n1}, \boldsymbol{d}_{n2}, \ldots, \boldsymbol{d}_{nn}]^\top \in \mathbb{R}^{n\times p(n)}$ has rank $p(n)$, and $\boldsymbol{\varepsilon}_n$ has independent components such that

$$\mathbb{E}(\varepsilon_{ni}) \;=\; 0 \quad \text{and} \quad \sigma_{ni} \;:=\; \mathrm{Std}(\varepsilon_{ni}) \;<\; \infty.$$

Now we are interested in the distributions

$$\mathcal{L}(\boldsymbol{A}_n^\top \boldsymbol{\varepsilon}_n) \quad \text{and} \quad \mathcal{L}(\boldsymbol{A}_n^\top \boldsymbol{\varepsilon}_n^* \,|\, \boldsymbol{Y}_n)$$

with given matrices $\boldsymbol{A}_n = [\boldsymbol{a}_{n1}, \boldsymbol{a}_{n2}, \ldots, \boldsymbol{a}_{nn}]^\top \in \mathbb{R}^{n\times k}$ with fixed dimension k, where $\boldsymbol{A}_n\mathbb{R}^k \subset \boldsymbol{D}_n\mathbb{R}^{p(n)}$ and

$$\mathrm{trace}(\boldsymbol{A}_n^\top \boldsymbol{A}_n) \;=\; \sum_{i=1}^n \|\boldsymbol{a}_{ni}\|^2 \;=\; k.$$

Furthermore, $\boldsymbol{\varepsilon}_n^* = (\xi_i \widehat{\varepsilon}_{ni})_{i=1}^n$.

We use three conditions on ξ, $\boldsymbol{\varepsilon}_n$, $\boldsymbol{A}_n$ and the hat matrix

$$\boldsymbol{H}_n = (H_{n,ij})_{i,j=1}^n \;=\; \boldsymbol{D}_n(\boldsymbol{D}_n^\top \boldsymbol{D}_n)^{-1}\boldsymbol{D}_n^\top$$

resulting from the subsequent proof:

(W.0) $\mathbb{E}(\xi) = 0,\ \mathbb{E}(\xi^2) = 1$ and $\mathbb{E}(|\xi|^3) < \infty$;

(W.1) $$\boldsymbol{\Sigma}_n := \sum_{i=1}^n \sigma_{ni}^2\, \boldsymbol{a}_{ni}\boldsymbol{a}_{ni}^\top \;=\; O(1) \quad \text{and} \quad \sum_{i=1}^n \mathbb{E}(|\varepsilon_{ni}|^3)\|\boldsymbol{a}_{ni}\|^3 \;\to\; 0;$$

(W.2) $$\sum_{i=1}^n H_{n,ii}\|\boldsymbol{a}_{ni}\|^2 \sum_{j=1}^n H_{n,jj}\sigma_{nj}^2 \;\to\; 0.$$

Theorem 8.12. *Under conditions (W.0-2),*

$$d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n), \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma}_n)\big) \;\to\; 0$$

and

$$d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n), \mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^* \,|\, \boldsymbol{Y}_n)\big) \;\to_p\; 0.$$

Remarks on (W.1-2). Condition (W.2) involves the leverages $H_{n,ii}$. The standardized design matrix

$$\tilde{\boldsymbol{D}}_n \;:=\; \boldsymbol{D}_n(\boldsymbol{D}_n^\top\boldsymbol{D}_n)^{-1/2} \;=\; [\tilde{\boldsymbol{d}}_{n1}, \tilde{\boldsymbol{d}}_{n2}, \ldots, \tilde{\boldsymbol{d}}_{nn}]^\top$$

has orthonormal columns, that means, $\tilde{\boldsymbol{D}}_n^\top\tilde{\boldsymbol{D}}_n = \boldsymbol{I}_{p(n)}$, and one may write

$$\boldsymbol{H}_n \;=\; \tilde{\boldsymbol{D}}_n\tilde{\boldsymbol{D}}_n^\top, \quad H_{n,ij} \;=\; \tilde{\boldsymbol{d}}_{ni}^\top\tilde{\boldsymbol{d}}_{nj}.$$

By assumption, $\boldsymbol{A}_n = \tilde{\boldsymbol{D}}_n\boldsymbol{C}_n$ with a certain matrix $\boldsymbol{C}_n \in \mathbb{R}^{p(n)\times k}$, and $k = \mathrm{trace}(\boldsymbol{A}_n^\top\boldsymbol{A}_n) = \mathrm{trace}(\boldsymbol{C}_n^\top\boldsymbol{C}_n)$. In particular, $\boldsymbol{a}_{ni} = \boldsymbol{C}_n^\top\tilde{\boldsymbol{d}}_{ni}$, whence

$$\|\boldsymbol{a}_{ni}\| \;\le\; \sqrt{kH_{n,ii}}.$$

Thus, a sufficient condition for (W.1-2) is:

(W.3) $$\kappa_n := \max_{1\le i\le n} \mathbb{E}(|\varepsilon_{ni}|^3) \;=\; O(1) \quad \text{and} \quad \lambda_n := \max_{1\le i\le n} H_{n,ii} \;=\; o(p(n)^{-1}).$$

To see this, note that by Jensen's inequality, $\sigma_{ni} \le \mathbb{E}(|\varepsilon_{ni}|^3)^{1/3}$, whence

$$\begin{aligned}
\mathrm{trace}(\boldsymbol{\Sigma}_n) \;&\le\; \kappa_n^{2/3}\sum_{i=1}^n \|\boldsymbol{a}_{ni}\|^2 \;=\; k\kappa_n^{2/3} \;=\; O(1),\\
\sum_{i=1}^n \mathbb{E}(|\varepsilon_{ni}|^3)\|\boldsymbol{a}_{ni}\|^3 \;&\le\; \kappa_n\sum_{i=1}^n \|\boldsymbol{a}_{ni}\|^2 \max_{1\le j\le n}\|\boldsymbol{a}_{nj}\| \;\le\; k^{3/2}\kappa_n\lambda_n^{1/2} \;=\; o(1),\\
\sum_{i=1}^n H_{n,ii}\|\boldsymbol{a}_{ni}\|^2 \;&\le\; \lambda_n\sum_{i=1}^n \|\boldsymbol{a}_{ni}\|^2 \;=\; k\lambda_n \;=\; o(p(n)^{-1}),\\
\sum_{j=1}^n H_{n,jj}\sigma_{nj}^2 \;&\le\; \kappa_n^{2/3}\sum_{j=1}^n H_{n,jj}^2 \;=\; \kappa_n^{2/3}p(n) \;=\; O(p(n)).
\end{aligned}$$

Proof of Theorem 8.12. At first, we show that

$$d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^* \,|\, \boldsymbol{Y}_n), \mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**} \,|\, \boldsymbol{Y}_n)\big) \;\to_p\; 0,$$

where

$$\boldsymbol{\varepsilon}_n^{**} \;:=\; (\xi_i\varepsilon_{ni})_{i=1}^n.$$

On the one hand,

$$\begin{aligned}\mathbb{E}\Big(d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^* \,|\, \boldsymbol{Y}_n), \mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**} \,|\, \boldsymbol{Y}_n)\big)^2\Big) \;&\le\; \mathbb{E}\,\mathbb{E}\big(\|\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^* - \boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**}\|^2 \,\big|\, \boldsymbol{Y}_n\big)\\ &=\; \mathbb{E}\big(\|\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^* - \boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**}\|^2\big),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\big(\|\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^* - \boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**}\|^2\big) \;&=\; \mathbb{E}\Big(\Big\|\sum_{i=1}^n \xi_i(\widehat{\varepsilon}_{ni} - \varepsilon_{ni})\boldsymbol{a}_{ni}\Big\|^2\Big)\\ &=\; \sum_{i,j=1}^n \mathbb{E}\big(\xi_i\xi_j(\widehat{\varepsilon}_{ni} - \varepsilon_{ni})(\widehat{\varepsilon}_{nj} - \varepsilon_{nj})\big)\boldsymbol{a}_{ni}^\top\boldsymbol{a}_{nj}\\ &=\; \sum_{i=1}^n \mathbb{E}\big((\widehat{\varepsilon}_{ni} - \varepsilon_{ni})^2\big)\|\boldsymbol{a}_{ni}\|^2.\end{aligned}$$

On the other hand,

$$\widehat{\varepsilon}_{ni} - \varepsilon_{ni} \;=\; -(\boldsymbol{H}_n\boldsymbol{\varepsilon}_n)_i \;=\; -\sum_{j=1}^n \tilde{\boldsymbol{d}}_{ni}^\top\tilde{\boldsymbol{d}}_{nj}\varepsilon_{nj},$$

so

$$\mathbb{E}\big((\widehat{\varepsilon}_{ni} - \varepsilon_{ni})^2\big) \;=\; \sum_{j=1}^n(\tilde{\boldsymbol{d}}_{ni}^\top\tilde{\boldsymbol{d}}_{nj})^2\sigma_{nj}^2 \;\le\; \|\tilde{\boldsymbol{d}}_{ni}\|^2\sum_{j=1}^n\|\tilde{\boldsymbol{d}}_{nj}\|^2\sigma_{nj}^2 \;=\; H_{n,ii}\sum_{j=1}^n H_{n,jj}\sigma_{nj}^2.$$

Consequently,

$$\mathbb{E}\Big(d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^* \,|\, \boldsymbol{Y}_n), \mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**} \,|\, \boldsymbol{Y}_n)\big)^2\Big) \;\le\; \sum_{i=1}^n H_{n,ii}\|\boldsymbol{a}_{ni}\|^2\sum_{j=1}^n H_{n,jj}\sigma_{nj}^2 \;=\; o(1)$$

by assumption (W.2).

It remains to show that

$$d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**} \,|\, \boldsymbol{\varepsilon}_n), \mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n)\big) \;\to_p\; 0,$$

and by the triangular inequality for $d_{M,2}(\cdot,\cdot)$, this is a consequence of

$$d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n), \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma}_n)\big) \;\to\; 0, \tag{8.4}$$

$$d_{M,2}\big(\mathcal{L}(\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n^{**} \,|\, \boldsymbol{\varepsilon}_n), \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma}_n)\big) \;\to_p\; 0. \tag{8.5}$$

But both claims are direct consequences of Corollary A.35. Note first that

$$\boldsymbol{A}_n^\top\boldsymbol{\varepsilon}_n \;=\; \sum_{i=1}^n \boldsymbol{Y}_{ni} \quad\text{with}\quad \boldsymbol{Y}_{ni} \;:=\; \varepsilon_{ni}\boldsymbol{a}_{ni} \;\in\; \mathbb{R}^k,$$

so $\mathbb{E}(\boldsymbol{Y}_{ni}) = \boldsymbol{0}$ and

$$\begin{aligned}\sum_{i=1}^n \mathrm{Var}(\boldsymbol{Y}_{ni}) \;&=\; \sum_{i=1}^n \sigma_{ni}^2\,\boldsymbol{a}_{ni}\boldsymbol{a}_{ni}^\top \;=\; \boldsymbol{\Sigma}_n \;=\; O(1),\\ \sum_{i=1}^n \mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^3\big) \;&=\; \sum_{i=1}^n \mathbb{E}(|\varepsilon_{ni}|^3)\|\boldsymbol{a}_{ni}\|^3 \;\to\; 0.\end{aligned}$$

This yields claim (8.4), and we also know that

$$\mathbb{E}\Big\|\sum_{i=1}^{n} \varepsilon_{ni}^2\, \boldsymbol{a}_{ni}\boldsymbol{a}_{ni}^\top - \boldsymbol{\Sigma}_n\Big\|_F \;\to\; 0.$$

The latter conclusion is important for

$$\boldsymbol{A}_n^\top \boldsymbol{\varepsilon}_n^{**} \;=\; \sum_{i=1}^{n} \boldsymbol{Y}_{ni}^* \quad \text{with} \quad \boldsymbol{Y}_{ni}^* \;:=\; \xi_i \varepsilon_{ni} \boldsymbol{a}_{ni}.$$

Indeed, $\mathbb{E}(\boldsymbol{Y}_{ni}^* \,|\, \boldsymbol{\varepsilon}_n) = \boldsymbol{0}$ and

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{Var}(\boldsymbol{Y}_{ni}^* \,|\, \boldsymbol{\varepsilon}_n) \;&=\; \sum_{i=1}^{n} \varepsilon_{ni}^2\, \boldsymbol{a}_{ni}\boldsymbol{a}_{ni}^\top \;=\; \boldsymbol{\Sigma}_n + o_p(1), \\
\sum_{i=1}^{n} \mathbb{E}\big(\|\boldsymbol{Y}_{ni}^*\|^3 \,\big|\, \boldsymbol{\varepsilon}_n\big)\Big) \;&=\; \mathbb{E}(|\xi|^3) \sum_{i=1}^{n} |\varepsilon_{ni}|^3 \|\boldsymbol{a}_{ni}\|^3 \;\to_p\; 0.
\end{aligned}$$

Hence, Corollary A.35 yields claim (8.5). □

Choosing $\mathcal{L}(\xi)$. To apply Theorem 8.12, we have to make sure that $\mathbb{E}(\xi) = 0$, $\mathbb{E}(\xi^2) = 1$ and $\mathbb{E}(|\xi|^3) < \infty$. Obvious choices of $\mathcal{L}(\xi)$ would be

$$\mathrm{N}(0,1) \quad \text{or} \quad \mathrm{Unif}\{-1,1\}.$$

But refined considerations indicate that one should also aim for the additional constraint

$$\mathbb{E}(\xi^3) \;=\; 1. \tag{8.6}$$

Examples for random variables ξ with these properties are constructed in Exercise 8.14.

To motivate (8.6), let us consider the special case that $p(n) = 1$ and $\boldsymbol{d}_{ni} = 1 = \boldsymbol{A}_n$ for all n and $1 \le i \le n$. That means, we want to estimate the common mean θ_n of the random variables $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$:

Lemma 8.13. *Let $Z_n := \sqrt{n}(\bar{Y}_n - \theta_n)$ and $Z_n^* := n^{-1/2} \sum_{i=1}^n \xi_i (Y_{ni} - \bar{Y}_n)$, where $Y_{ni} = \theta_n + \varepsilon_{ni}$ and $\mathbb{E}\big(|\varepsilon_{ni}|^{3+\delta}\big) \le \kappa$ for $1 \le i \le n$ and certain constants κ, $\delta > 0$. Then, $\mathbb{E}(Z_n) = 0 = \mathbb{E}(Z_n^* \,|\, \boldsymbol{Y}_n)$ and*

$$\begin{aligned}
\mathbb{E}(Z_n^2) \;&=\; \frac{1}{n}\sum_{i=1}^{n} \sigma_{ni}^2, \\
\mathbb{E}((Z_n^*)^2 \,|\, \boldsymbol{Y}_n) \;&=\; \frac{1}{n}\sum_{i=1}^{n} (\varepsilon_{ni} - \bar{\varepsilon}_n)^2 \;=\; \mathbb{E}(Z_n^2) + o_p(1), \\
\sqrt{n}\;\mathbb{E}(Z_n^3) \;&=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\varepsilon_{ni}^3), \\
\sqrt{n}\;\mathbb{E}((Z_n^*)^3 \,|\, \boldsymbol{Y}_n) \;&=\; \mathbb{E}(\xi^3)\frac{1}{n}\sum_{i=1}^{n} (\varepsilon_{ni} - \bar{\varepsilon}_n)^3 \;=\; \mathbb{E}(\xi^3)\sqrt{n}\;\mathbb{E}(Z_n^3) + o_p(1).
\end{aligned}$$

Under the additional assumption that $n^{-1}\sum_{i=1}^n \sigma_{ni}^2$ and $n^{-1}\sum_{i=1}^n \mathbb{E}(\varepsilon_{ni}^3)$ stay bounded away from 0, this lemma implies that

$$\frac{\mathrm{skewness}(Z_n^* \,|\, \boldsymbol{Y}_n)}{\mathrm{skewness}(Z_n)} \;\to_p\; \mathbb{E}(\xi^3).$$

Here $\mathrm{skewness}(Z) := \mathbb{E}(Z^3)/\,\mathrm{Std}(Z)^3$. Lemma 8.13 follows from elementary calculations and Theorem A.36 in the appendix.

Exercise 8.14. Construct a random variable ξ satisfying

$$\mathbb{E}(\xi) \;=\; 0, \quad \mathbb{E}(\xi^2) \;=\; 1 \quad \text{and} \quad \mathbb{E}(\xi^3) \;=\; 1$$

in three different ways:

(a) $\mathcal{L}(\xi) = (1-p)\delta_a + p\delta_b$ with suitable constants $p \in (0,1)$ and $a < 0 < b$.

(b) $\xi \;=\; aZ_1 + b(Z_2^2 - 1)$ with independent random variables $Z_1, Z_2 \sim \mathrm{N}(0,1)$ and suitable constants a, b.

(c) $\xi = G - c$ with a random variable $G \sim \mathrm{Gamma}(a,b)$ and suitable constants $a, b, c > 0$.

8.3 Exact Tests and Confidence Regions in the Linear Model

Similarly as in Section 7.5.2, we describe a procedure which is related to the wild bootstrap and yields exact confidence regions for $\boldsymbol{\theta}$. It is a variation of sign tests as treated in the textbook Dümbgen (2015). We assume that

$$Y_i \;=\; \boldsymbol{d}_i^\top \boldsymbol{\theta} + \varepsilon_i$$

with independent random variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ such that

$$\mathcal{L}(-\varepsilon_i) \;=\; \mathcal{L}(\varepsilon_i). \tag{8.7}$$

For a hypothetical value $\boldsymbol{\eta} \in \mathbb{R}^p$ of $\boldsymbol{\theta}$ let $T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta})$ be a test statistic of type

$$T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta}) \;=\; \widetilde{T}\big((Y_i - \boldsymbol{d}_i^\top \boldsymbol{\eta})_{i=1}^n, \boldsymbol{D}\big)$$

with some function $\widetilde{T}(\cdot, \boldsymbol{D}) : \mathbb{R}^n \to \mathbb{R}$. For instance,

$$T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta}) \;:=\; \Big\| n^{-1/2} \sum_{i=1}^n (Y_i - \boldsymbol{d}_i^\top \boldsymbol{\eta})\, \boldsymbol{d}_i \Big\|.$$

can be written as $\widetilde{T}\big((Y_i - \boldsymbol{d}_i^\top \boldsymbol{\eta})_{i=1}^n, \boldsymbol{D}\big)$ with

$$\widetilde{T}(\boldsymbol{z}, \boldsymbol{D}) \;:=\; \Big\| n^{-1/2} \sum_{i=1}^n z_i\, \boldsymbol{d}_i \Big\|.$$

In general, large values of $T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta})$ are considered as evidence against the null hypothesis that $\boldsymbol{\theta} = \boldsymbol{\eta}$. Further, let $\xi_1, \xi_2, \ldots, \xi_n$ and $\boldsymbol{Y}$ be stochastically independent with $\xi_i \sim \mathrm{Unif}\{-1, 1\}$.

The random vector $(Y_i - \boldsymbol{d}_i^\top \boldsymbol{\theta})_{i=1}^n = (\varepsilon_i)_{i=1}^n$ has the same distribution as $(\xi_i(Y_i - \boldsymbol{d}_i^\top \boldsymbol{\theta}))_{i=1}^n = (\xi_i \varepsilon_i)_{i=1}^n$. With the left-continuous distribution function

$$G(r \,|\, \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta}) \;:=\; \mathbb{P}\Big(\widetilde{T}\big((\xi_i(Y_i - \boldsymbol{d}_i^\top \boldsymbol{\eta}))_{i=1}^n, \boldsymbol{D}\big) < r \;\Big|\; \boldsymbol{Y}\Big),$$

a p-value of the null hypothesis that $\boldsymbol{\theta} = \boldsymbol{\eta}$ is given by

$$1 - G\big(T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta}) \,|\, \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta}\big).$$

Furthermore,

$$\Big\{\boldsymbol{\eta} \in \mathbb{R}^p : G(T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta}) \,|\, \boldsymbol{\eta}) < 1 - \alpha\Big\}$$

defines a $(1-\alpha)$-confidence region for $\boldsymbol{\theta}$. The only problem with this approach is that the explicit computation of the latter region is highly nontrivial.

Assumption (8.7) could be replaced with the weaker assumption that

$$\mathbb{P}(\varepsilon_i < 0) \;=\; \mathbb{P}(\varepsilon_i > 0) \;=\; 1/2. \tag{8.8}$$

In this case, one could restrict one's attention to test statistics of type

$$T(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta}) \;=\; \widetilde{T}\Big(\big(\operatorname{sign}(Y_i - \boldsymbol{d}_i^\top \boldsymbol{\eta})\big)_{i=1}^n, \boldsymbol{D}\Big),$$

for instance,

$$T(\boldsymbol{Y}, \boldsymbol{\eta}) \;:=\; \Big\| n^{-1/2} \sum_{i=1}^n \operatorname{sign}(Y_i - \boldsymbol{d}_i^\top \boldsymbol{\eta})\, \boldsymbol{d}_i \Big\|.$$

Then one could replace $G(r \,|\, \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{\eta})$ with

$$G(r \,|\, \boldsymbol{D}) \;:=\; \mathbb{P}\big[\tilde{T}\big((\xi_i)_{i=1}^n, \boldsymbol{D}\big) < r\big],$$

and this would lead to p-values and confidence regions for $\boldsymbol{\theta}$, assuming only (8.8).

8.4 Bootstrap Failures and Subsampling

Soon after the bootstrap had been introduced, statisticians discovered some situations in which it does *not* work. Some authors claimed that a variation of the original bootstrap, called *subsampling* or *m-out-of-n bootstrap*, may solve the problem. In the present section we illustrate potential failures of the bootstrap method in i.i.d. settings and show that subsampling is *not* a reliable solution. The material in this section is mainly from Dümbgen (1993).

The general framework. As in Section 8.1, we consider a sequence $(P_n)_n$ of distributions on $\mathbb{R}^d$, and for each n we observe a sample $\boldsymbol{Y}_n = (Y_{ni})_{i=1}^n$ of independent random variables $Y_{ni} \sim P_n$. We are mainly interested in functions of the mean vector $\mu_n = \mu(P_n)$, and we assume that $\int \|x\|^2 \, P_n(dx)$ is finite, so the covariance matrix $\Sigma_n = \Sigma(P_n)$ is well-defined. As usual, the parameters μ_n and Σ_n are estimated by the sample mean $\widehat{\mu}_n = n^{-1} \sum_{i=1}^n Y_{ni}$ and the sample covariance $\widehat{\Sigma}_n := (n-1)^{-1} \sum_{i=1}^n (Y_{ni} - \bar{Y}_n)(Y_{ni} - \bar{Y}_n)^\top$. Now we consider the distributions

$$Q_n \;:=\; \mathcal{L}\big(\sqrt{n}(\widehat{\mu}_n - \mu_n)\big),$$

and

$$R_n \;:=\; \mathcal{L}\big(\sqrt{n}\big(g(\widehat{\mu}_n) - g(\mu_n)\big)\big)$$

for a given function

$$g : \mathbb{R}^d \to \mathbb{R}.$$

If an oracle would reveal the γ-quantile $q_{n,\gamma}$ of R_n for any $\gamma \in (0,1)$ we wish, we could compute the lower $(1-\alpha)$-confidence bound

$$g(\widehat{\mu}_n) - n^{-1/2} q_{n,1-\alpha}$$

for $g(\mu_n)$, the upper $(1-\alpha)$-confidence bound

$$g(\widehat{\mu}_n) - n^{-1/2} q_{n,\alpha}$$

for $g(\mu_n)$, or we could combine the lower and upper $(1-\alpha/2)$-confidence bounds to get a $(1-\alpha)$-confidence interval for $g(\mu_n)$.

Two versions of the bootstrap. Let $\widehat{P}_n$ be the empirical distribution $\widehat{P}_n^{\mathrm{emp}}$ of $\boldsymbol{Y}_n$ or its smoothed version $\widehat{P}_n^{\mathrm{emp}} \star \mathrm{N}_d(0, n^{-1}\widehat{\Sigma}_n)$. Now consider a random vector $\boldsymbol{Y}_n^* = (Y_{ni}^*)_{i=1}^{m_n}$ such that, conditional on $\boldsymbol{Y}_n$, the m_n components Y_{ni}^* are independent with distribution $\widehat{P}_n$. Then we estimate Q_n by

$$\widehat{Q}_n \;:=\; \mathcal{L}\big(\sqrt{m_n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \,\big|\, \boldsymbol{Y}_n\big)$$

and R_n by

$$\widehat{R}_n \;:=\; \mathcal{L}\big(\sqrt{m_n}\big(g(\widehat{\mu}_n^*) - g(\widehat{\mu}_n)\big) \,\big|\, \boldsymbol{Y}_n\big),$$

where $\widehat{\mu}_n^* := m_n^{-1} \sum_{i=1}^{m_n} Y_{ni}^*$.

As to the bootstrap sample size m_n, the classical bootstrap (resampling) uses $m_n = n$. On the other hand, *subsampling* or *m-out-of-n bootstrap* means that $m_n \le n$ with

$$m_n \;\to\; \infty \quad \text{but} \quad m_n/n \;\to\; 0.$$

If we think about $\boldsymbol{Y}_n$ as a sample from a a large population, i.e. a population size much larger than the sample size n, it seems to be natural to work with bootstrap samples from the 'population' $\boldsymbol{Y}_n$ of size $m_n \ll n$. At the end of the day, however, the main question is whether $\widehat{R}_n$ is a consistent estimator of R_n.

Asymptotics for Q_n and $\widehat{Q}n$. The following two facts follow from Lemmas 8.2 and 8.4 with straightforward modifications: Suppose that for a fixed nonzero matrix $\Sigma \in \mathbb{R}^{d\times d}$,

(A.0) $$\Sigma_n \;\to\; \Sigma \quad \text{and} \quad \mathbb{E}\big(\|Y_{n1} - \mu_n\|^2 \min\big\{m_n^{-1/2}\|Y_{n1} - \mu_n\|, 1\big\}\big) \;\to\; 0.$$

Then,

$$Q_n \;\to_w\; \mathrm{N}_d(0, \Sigma),$$
$$\widehat{Q}_n \;\to_{w,p}\; \mathrm{N}_d(0, \Sigma).$$

Note that the second part of assumption (A.0) is more restrictive in case of subsampling. So one should have good reasons to use subsampling instead of resampling. As soon as $d_{M,2}(P_n, P) \to 0$ for a fixed distribution P with finite second moments, (A.0) is satisfied with $\Sigma = \Sigma(P)$.

In the subsequent three scenarios we always assume (A.0) and consider different assumptions on $(\mu)_n$ and g.

Asymptotics I for R_n and $\widehat{R}_n$. In addition to (A.0), suppose that for a fixed vector $\mu \in \mathbb{R}^d$,

(A.1) $\mu_n \to \mu$, and g is continuously differentiable on a neighborhood of μ.

Then

$$\begin{aligned} R_n &\to_w \mathrm{N}\big(0, \nabla g(\mu)^\top \Sigma \nabla g(\mu)\big), \\ \widehat{R}_n &\to_{w,p} \mathrm{N}\big(0, \nabla g(\mu)^\top \Sigma \nabla g(\mu)\big). \end{aligned}$$

This follows essentially from the following observation: If g is continuously differentiable in a convex neighborhood U of μ, then for vectors $\mu_0, \mu_1 \in U$,

$$\begin{aligned} \big|g(\mu_1) - g(\mu_0) - \nabla g(\mu)^\top(\mu_1 - \mu_0)\big| &= \Big|\int_0^1 \big(\nabla g((1-t)\mu_0 + t\mu_1) - \nabla g(\mu)\big)^\top (\mu_1 - \mu_0)\, dt\Big| \\ &\le \sup_{t\in[0,1]} \big\|\nabla g((1-t)\mu_0 + t\mu_1) - \nabla g(\mu)\big\|\, \|\mu_1 - \mu_0\| \\ &= o\big(\|\mu_1 - \mu_0\|\big) \quad \text{as } \mu_0, \mu_1 \to \mu. \end{aligned}$$

Hence, with

$$Z_n := \sqrt{n}(\widehat{\mu}_n - \mu_n) \quad \text{and} \quad Z_n^* := \sqrt{m_n}(\widehat{\mu}_n^* - \widehat{\mu}_n)$$

we may write

$$\begin{aligned} \sqrt{n}\big(g(\widehat{\mu}_n) - g(\mu_n)\big) &= \sqrt{n}\big(g(\mu_n + n^{-1/2} Z_n) - g(\mu_n)\big) \\ &= \nabla g(\mu)^\top Z_n + o_p(1), \\ \sqrt{m_n}\big(g(\widehat{\mu}_n^*) - g(\widehat{\mu}_n)\big) &= \sqrt{m_n}\big(g(\widehat{\mu}_n + m_n^{-1/2} Z_n^*) - g(\widehat{\mu}_n)\big) \\ &= \nabla g(\mu)^\top Z_n^* + o_p(1), \end{aligned}$$

because $\mu_n \to \mu$ and $Z_n, Z_n^* = O_p(1)$.

Consequently, under conditions (A.0) and (A.1), both bootstrap versions yield asymptotically valid confidence sets, provided that $\Sigma \nabla g(\mu) \neq 0$.

Asymptotics II for R_n and $\widehat{R}_n$. In addition to (A.0), suppose that for any fixed $\mu \in \mathbb{R}^d$, the mapping g satisfies the following condition:

(A.2) There exists a continuous mapping $Dg(\mu, \cdot) : \mathbb{R}^d \to \mathbb{R}$ such that for $v \in \mathbb{R}^d$,

$$g(\mu + v) = g(\mu) + Dg(\mu, v) + o(\|v\|) \quad \text{as } v \to 0,$$

and

$$Dg(\mu, tv) = tDg(\mu, v) \quad \text{for all } t \ge 0.$$

Assumption (A.2) is weaker than differentiability of g at μ. Indeed, g is differentiable at μ if and only if $Dg(\mu,\cdot)$ is linear, see the two examples later.

Now suppose that for some fixed $\mu \in \mathbb{R}^d$,

$$\mu_n \;=\; \mu + n^{-1/2}\Delta_n, \quad \text{where } \Delta_n \to \Delta \in \mathbb{R}^d.$$

Then,

$$\begin{aligned}\sqrt{n}\big(g(\widehat{\mu}_n) - g(\mu_n)\big) \;&=\; \sqrt{n}\Big(g\big(\mu + n^{-1/2}(\Delta_n + Z_n)\big) - g\big(\mu + n^{-1/2}\Delta_n\big)\Big)\\ &=\; \sqrt{n}\, Dg\big(\mu, n^{-1/2}(\Delta_n + Z_n)\big) - \sqrt{n}\, Dg\big(\mu, n^{-1/2}\Delta_n\big) + o_p(1)\\ &=\; Dg(\mu, \Delta_n + Z_n) - Dg(\mu, \Delta_n) + o_p(1).\end{aligned}$$

The second last step follows from (A.3) together with the facts that $n^{-1/2}(\Delta_n + Z_n) = O_p(n^{-1/2})$ and $n^{-1/2}\Delta_n = O(n^{-1/2})$. In the last step we used the homogeneity property of $Dg(\mu,\cdot)$. Hence,

$$R_n \;\to_w\; \mathcal{L}\big(Dg(\mu, \Delta + Z) - Dg(\mu, \Delta)\big).$$

As to the bootstrap,

$$\begin{aligned}&\sqrt{m_n}\big(g(\widehat{\mu}_n^*) - g(\widehat{\mu}_n)\big)\\ &=\; \sqrt{m_n}\Big(g\big(\mu + n^{-1/2}(Z_n + \Delta_n) + m_n^{-1/2}Z_n^*\big) - g\big(\mu + n^{-1/2}(Z_n + \Delta_n)\big)\Big)\\ &=\; \sqrt{m_n}\, Dg\big(\mu, n^{-1/2}(Z_n + \Delta_n) + m_n^{-1/2}Z_n^*\big) - \sqrt{m_n}\, Dg\big(\mu, n^{-1/2}(Z_n + \Delta_n)\big) + o_p(1)\\ &=\; Dg\big(\mu, \sqrt{m_n/n}(Z_n + \Delta_n) + Z_n^*\big) - Dg\big(\mu, \sqrt{m_n/n}(Z_n + \Delta_n)\big) + o_p(1).\end{aligned}$$

Consequently, the usual bootstrap yields an estimated distribution $\widehat{R}_n$ which behaves asymptotically like the random measure

$$\mathcal{L}\big(Dg(\mu, \Delta + Z + Z^*) - Dg(\mu, \Delta + Z) \,\big|\, Z\big),$$

where Z, Z^* are independent with distribution $\mathrm{N}_d(0,\Sigma)$. The subsampling version yields an estimated distribution $\widehat{R}_n$ converging weakly to the fixed distribution

$$\mathcal{L}(Dg(\mu, Z^*)).$$

If g is differentiable at μ, that is, $Dg(\mu,\cdot)$ is linear, then

$$\begin{aligned}Dg(\mu, \Delta + Z) - Dg(\mu, \Delta) \;&=\; Dg(\mu, Z),\\ Dg(\mu, \Delta + Z + Z^*) - Dg(\mu, \Delta + Z) \;&=\; Dg(\mu, Z^*),\end{aligned}$$

so the usual bootstrap and subsampling yield distributions $\widehat{R}_n$ converging weakly in probability to the same limit as R_n. If $Dg(\mu,\cdot)$ is nonlinear, however, the usual bootstrap fails in general. Subsampling fails too in general, unless $\Delta = 0$, that is, $(\mu)_n$ is essentially constant.

Example 8.15. Let $g(\mu) := \|\mu - \mu_o\|$ for a given $\mu_o \in \mathbb{R}^d$. Then assumption (A.3) holds true with

$$Dg(\mu, v) \;=\; \begin{cases}\|\mu - \mu_o\|^{-1}(\mu - \mu_o)^\top v & \text{if } \mu \neq \mu_o,\\ \|v\| & \text{if } \mu = \mu_o.\end{cases}$$

Thus g is continuously differentiable on $\mathbb{R}^d \setminus \{\mu_o\}$. Indeed, if $\mu \neq \mu_o$, then as $\mathbb{R}^d \ni v \to 0$,

$$\begin{aligned} g(\mu+v) - g(\mu) &= \frac{\|\mu-\mu_o+v\|^2 - \|\mu-\mu_o\|^2}{\|\mu-\mu_o+v\| + \|\mu-\mu_o\|} \\ &= \frac{2(\mu-\mu_o)^\top v + \|v\|^2}{\|\mu-\mu_o+v\| + \|\mu-\mu_o\|} \\ &= \frac{2(\mu-\mu_o)^\top v + O(\|v\|^2)}{2\|\mu-\mu_o\| + O(\|v\|)} \\ &= \|\mu-\mu_o\|^{-1}(\mu-\mu_o)^\top v + O(\|v\|^2). \end{aligned}$$

If $\mu = \mu_o$, then $g(\mu+v) - g(\mu) = \|v\|$ for all $v \in \mathbb{R}^d$.

Example 8.16. Let $g(\mu) := \max(\mu_1, \ldots, \mu_d)$. Here assumption (A.3) is satisfied with

$$Dg(\mu, v) = \max_{j \in J(\mu)} v_j, \quad J(\mu) := \big\{j \in \{1,\ldots,d\} : \mu_j = g(\mu)\big\}.$$

This mapping $Dg(\mu, \cdot)$ is linear if and only if $J(\mu)$ consists of one index. Indeed, let $\delta := g(\mu) - \max_{k \notin J(\mu)} \mu_k > 0$. If $v \in \mathbb{R}^d$ with $\|v\| \leq 2\delta$, then for arbitrary indices $j \in J(\mu)$ and $k \notin J(\mu)$,

$$(\mu+v)_j - (\mu+v)_k \geq \delta - 2\|v\| \geq 0,$$

whence

$$g(\mu+v) = \max_{j \in J(\mu)} (\mu+v)_j = g(\mu) + \max_{j \in J(\mu)} v_j.$$

Asymptotics III for R_n and $\widehat{R}_n$. Looking at Asymptotics I and II, it looks as if subsampling works whenever the usual bootstrap works, and at least in some special scenarios, subsampling works while the usual bootstrap does not. However, it may happen that the usual boostrap works while subsampling does not. Suppose that (A.0) and (A.2) are satisfied. In addition:

(A.3) Suppose that $\mu_n = \mu_o + r_n\Delta_n$ with a point μ_o at which g is *not* differentiable, with a sequence of numbers $r_n > 0$ such that $r_n \to 0$ but $r_n\sqrt{n} \to \infty$, and with a sequence of unit vectors Δ_n converging to some unit vector Δ. Moreover, suppose that there exists a nonzero vector $\gamma(\Delta) \in \mathbb{R}^d$ such that

$$g(\mu_n + v_n) - g(\mu_n) = \gamma(\Delta)^\top v_n + o(\|v_n\|)$$

for arbitrary sequences $(v_n)_n$ in $\mathbb{R}^d$ such that $v_n = o(r_n)$.

On the one hand,

$$\begin{aligned} \sqrt{n}\big(g(\widehat{\mu}_n) - g(\mu_n)\big) &= \sqrt{n}\big(g(\mu_n + n^{-1/2}Z_n) - g(\mu_n)\big) \\ &= \gamma(\Delta)^\top Z_n + o_p(1), \end{aligned}$$

because $n^{-1/2}Z_n = O_p(n^{-1/2}) = o_p(r_n)$. On the other hand, if $r_n\sqrt{m_n} \to \infty$, then

$$\begin{aligned} \sqrt{m_n}\big(g(\widehat{\mu}_n^*) - g(\widehat{\mu}_n)\big) &= \sqrt{m_n}\big(g(\mu_n + n^{-1/2}Z_n + m_n^{-1/2}Z_n^*) - g(\mu_n + n^{-1/2}Z_n)\big) \\ &= \gamma(\Delta)^\top Z_n^* + o_p(1), \end{aligned}$$

because $n^{-1/2}Z_n + m_n^{-1/2}Z_n^* = O_p(m_n^{-1/2}) = o_p(r_n)$, whereas $r_n\sqrt{m_n} \to \lambda \in [0,\infty)$ implies that

$$\begin{aligned}
&\sqrt{m_n}\big(g(\widehat{\mu}_n^*) - g(\widehat{\mu}_n)\big) \\
&\quad = \sqrt{m_n}\big(g(\mu_o + r_n\Delta_n + n^{-1/2}Z_n + m_n^{-1/2}Z_n^*) - g(\mu_o + r_n\Delta_n + n^{-1/2}Z_n)\big) \\
&\quad = Dg(\mu_o, \lambda\Delta + Z_n^*) - Dg(\mu_o, \lambda\Delta) + o_p(1),
\end{aligned}$$

because $r_n\Delta_n = m_n^{-1/2}(\lambda\Delta + o(1))$, $n^{-1/2}Z_n = o_p(m_n^{-1/2})$ and $m_n^{-1/2}Z_n^* = O_p(m_n^{-1/2})$. All in all, we see that

$$\begin{aligned}
R_n &\to_w \mathrm{N}\big(0, \gamma(\Delta)^\top\Sigma\gamma(\Delta)\big), \\
\widehat{R}_n &\to_{w,p} \begin{cases} \mathrm{N}\big(0, \gamma(\Delta)^\top\Sigma\gamma(\Delta)\big) & \text{if } r_n\sqrt{m_n} \to \infty, \\ \mathcal{L}\big(Dg(\mu_o, \lambda\Delta + Z) - Dg(\mu_o, \lambda\Delta)\big) & \text{if } r_n\sqrt{m_n} \to \lambda \in [0,\infty). \end{cases}
\end{aligned}$$

Consequently, subsampling fails in the present setting if $m_n \to \infty$ too slowly, whereas usual resampling would work.

Concerning the assumptions about g in the present setting, they are satisfied in the first example, when $g(\mu) = \|\mu - \mu_o\|$, with $\gamma(\Delta) = \Delta$. Indeed,

$$\begin{aligned}
g(\mu_n + v_n) - g(\mu_n) &= \frac{\|\mu_n + v_n - \mu_o\|^2 - \|\mu_n - \mu_o\|^2}{\|\mu_n + v_n - \mu_o\| + \|\mu_n - \mu_o\|} \\
&= \frac{2r_n\Delta_n^\top v_n + \|v_n\|^2}{r_n\|\Delta_n + r_n^{-1}v_n\| + r_n} \\
&= \frac{2\Delta^\top v_n + o(v_n)}{2 + o(1)} = \Delta^\top v_n + o(\|v_n\|),
\end{aligned}$$

provided that $v_n = o(r_n)$.

In the second example, when $g(\mu) = \max(\mu_1, \ldots, \mu_d)$, let $\mu_o \in \mathbb{R}^d$ such that the set $J(\mu_o)$ consists of at least 2 indices. Let $\Delta \in \mathbb{R}^d$ such that for some $j_o \in J(\mu_o)$,

$$\Delta_{j_o} > \Delta_j \quad \text{for all } j \in J(\mu_o) \setminus \{j_o\}.$$

Then (A.4) is satisfied with $\gamma(\Delta) = (1_{[j=j_o]})_{j=1}^d$. To see this, let $(v_n)_n$ be any sequence in $\mathbb{R}^d$ such that $v_n = o(r_n)$. For any index $j \in \{1, \ldots, d\} \setminus J(\mu_o)$,

$$(\mu_n + v_n)_j - (\mu_n + v_n)_{j_o} = \mu_{oj} - \mu_{oj_o} + O(r_n) < 0$$

for sufficiently large n, while for any index $j \in J(\mu_o) \setminus \{j_o\}$,

$$(\mu_n + v_n)_j - (\mu_n + v_n)_{j_o} = r_n(\Delta_{nj} - \Delta_{nj_o}) + o(r_n) = r_n(\Delta_j - \Delta_{j_o} + o(1)) < 0$$

for sufficiently large n. Applying the same reasoning to $(0)_n$ instead of $(v_n)_n$ shows that for sufficiently large n,

$$g(\mu_n + v_n) - g(\mu_n) = \gamma(\Delta)^\top(\mu_n + v_n) - \gamma(\Delta)^\top\mu_n = \gamma(\Delta)^\top v_n.$$

A possible, though conservative solution. The previous considerations show that neither the usual resampling nor subsampling are reliable in connection with non-smooth functions g. Suppose that g is at least Lipschitz-continuous on $\mathbb{R}^d$. Now consider random samples $\boldsymbol{Y}_n, \boldsymbol{Y}_n^*$ and $\boldsymbol{Y}_n^{(1)}, \ldots, \boldsymbol{Y}_n^{(b)}$ with $\boldsymbol{Y}_n$ and $\boldsymbol{Y}_n^*$ as before in the classical bootstrap, while conditional on $\boldsymbol{Y}_n$, the samples $\boldsymbol{Y}_n^{(1)}, \ldots, \boldsymbol{Y}_n^{(b)}$ are just independent copies of $\boldsymbol{Y}_n^*$. Let $\widehat{\mu}_n$, $\widehat{\mu}_n^*$ and $\widehat{\mu}_n^{(1)}, \ldots, \widehat{\mu}_n^{(b)}$ be the corresponding sample means, and set

$$Z_n \;=\; \sqrt{n}(\widehat{\mu}_n - \mu_n), \quad Z_n^* \;=\; \sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \quad \text{and} \quad Z_n^{(s)} \;:=\; \sqrt{n}(\widehat{\mu}_n^{(s)} - \widehat{\mu}_n)$$

for $1 \le s \le b$. For $\gamma \in (1/2, 1)$, we estimate the γ-quantile $q_{n,\gamma}$ of R_n by

$$\widehat{q}_{n,\gamma,b} \;:=\; \max_{1 \le s \le b} \widehat{q}_{n,\gamma}(2\widehat{\mu}_n - \widehat{\mu}_n^{(s)}),$$

where $\widehat{q}_{n,\gamma}(x)$ denotes the γ-quantile of

$$\mathcal{L}\Big(\sqrt{n}\big(g(x + n^{-1/2}Z_n^*) - g(x)\big) \,\Big|\, \boldsymbol{Y}_n\Big).$$

Similarly, if $\gamma \in (0, 1/2)$, we estimate the γ-quantile $q_{n,\alpha}$ of R_n by

$$\widehat{q}_{n,\gamma,b} \;:=\; \min_{1 \le s \le b} \widehat{q}_{n,\gamma}(2\widehat{\mu}_n - \widehat{\mu}_n^{(s)}).$$

The rationale behind this is that

$$\sqrt{n} \min_{1 \le s \le b} \|2\widehat{\mu}_n - \widehat{\mu}_n^{(s)} - \mu_n\| \;=\; \min_{1 \le s \le b} \|Z_n - Z_n^{(s)}\| \;\to_p\; 0 \quad \text{as } n, b \to \infty.$$

This result is proved in Exercise 8.17. It implies that among the b random points $2\widehat{\mu}_n - \widehat{\mu}_n^{(s)}$, $1 \le s \le b$, there is one with distance $o_p(n^{-1/2})$ to the unknown true parameter μ_n, provided that $n, b \to \infty$. This fact, combined with the Lipschitz-continuity of g, implies that

$$\widehat{q}_{n,\gamma,b} \begin{cases} \ge\; q_{n,\gamma} + o_p(1) & \text{if } \gamma > 1/2 \\ \le\; q_{n,\gamma} + o_p(1) & \text{if } \gamma < 1/2 \end{cases}$$

as $n, b \to \infty$. In that sense, the confidence bounds with $\widehat{q}_{n,\gamma,b}$ in place of $q_{n,\gamma}$ are asymptotically conservative.

Exercise 8.17. For each integer $n \ge 1$ let $Z_n, Z_n^{(1)}, Z_n^{(2)}, Z_n^{(3)}, \ldots$ be random vectors in $\mathbb{R}^d$ such that for any fixed integer $b \ge 1$,

$$\mathcal{L}(Z_n, Z_n^{(1)}, \ldots, Z_n^{(b)}) \;\to_w\; Q^{\otimes(b+1)} \quad \text{as } n \to \infty,$$

where Q is an arbitrary probability distribution on $\mathbb{R}^d$. Show that

$$\min_{1 \le s \le b} \|Z_n - Z_n^{(s)}\| \;\to_p\; 0 \quad \text{as } n, b \to \infty.$$

Chapter 9

Empirical Likelihood

Empirical likelihood methods, introduced by Art Owen (1990, 1991, 2001), are a nonparametric methods based on data-driven models.

9.1 Empirical Likelihood for I.I.D. Observations

Let $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$ be stochastically independent random variables with unknown distribution P_n on $\mathcal{Y}$. Suppose we are interested in a certain parameter $\theta(P_n) \in \Theta$. To construct a confidence region for $\theta(P_n)$, we consider all discrete distributions

$$\widehat{P}_{n,\boldsymbol{p}} \;=\; \sum_{i=1}^{n} p_i \delta_{Y_{ni}}$$

with a weight vector $\boldsymbol{p}$ in the n-dimensional unit simplex

$$\Pi_n \;:=\; \big\{\boldsymbol{p} \in [0,\infty)^n : p_+ = 1\big\}.$$

Here we use the notation $b_+ := \sum_{i=1}^n b_i$ for $\boldsymbol{b} \in \mathbb{R}^n$. The plausibility of such a distribution $\widehat{P}_{n,\boldsymbol{p}}$ is measured by the negative log-likelihood $S(\boldsymbol{p})$, where for arbitrary vectors $\boldsymbol{p} \in \mathbb{R}^n$,

$$S(\boldsymbol{p}) \;:=\; -\sum_{i=1}^{n} \log(p_i) \;\in\; (-\infty, \infty]$$

with the convention that $\log(a) = -\infty$ for $a \le 0$. One can also write

$$S(\boldsymbol{p}) \;=\; -\log\big(\prod_{i=1}^{n} \max(p_i, 0)\big).$$

The smaller $S(\boldsymbol{p})$, the more plausible is $\widehat{P}_{n,\boldsymbol{p}}$. The uniquely most plausible distribution is the empirical distribution

$$\widehat{P}_n \;=\; \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_{ni}},$$

which is $\widehat{P}_{n,\boldsymbol{p}}$ with $\boldsymbol{p} = (1/n)_{i=1}^n$. This could be derived from Lemma 6.12, applied to the distributions P and Q on $\{1, \ldots, n\}$ with $P(\{i\}) = 1/n$ and $Q(\{i\}) = p_i$. Alternatively, for

arbitrary scalars $\lambda > 0$,

$$\operatorname*{arg\,min}_{\boldsymbol{p}\in[0,\infty)^n}\big(S(\boldsymbol{p}) + \lambda p_+\big) \;=\; \operatorname*{arg\,min}_{\boldsymbol{p}\in[0,\infty)^n} \sum_{i=1}^n (-\log(p_i) + \lambda p_i) \;=\; (1/\lambda)_{i=1}^n,$$

and for $\lambda = n$ we obtain a vector in Π_n (Optimization via Lagrange's method).

Assuming that $\theta(\widehat{P}_{n,\boldsymbol{p}})$ is well-defined for all $\boldsymbol{p} \in \Pi_n$, we define the *negative empirical log-likelihood function* $L_n : \Theta \to [n\log n, \infty]$ by

$$L_n(\eta) \;:=\; \inf\big\{S(\boldsymbol{p}) : \boldsymbol{p} \in \Pi_n, \theta(\widehat{P}_{n,\boldsymbol{p}}) = \eta\big\}$$

with the convention that $\inf(\emptyset) = \infty$. (Strictly speaking, L_n is a negative profile log-likelihood function.) Under certain assumptions on $\theta(\cdot)$, one can show that

$$2L_n(\theta(P_n)) - 2n\log(n) \;\to_{\mathcal{L}}\; \chi^2_d \quad (n \to \infty)$$

for a certain integer $d > 0$. In this case,

$$C_{n,\alpha}(\boldsymbol{Y}_n) \;:=\; \big\{\eta \in \Theta : L_n(\eta) \le n\log(n) + \chi^2_{d;1-\alpha}/2\big\}$$

defines a confidence region for $\theta(P_n)$ with asymptotic confidence level $1 - \alpha$.

9.1.1 Analytical Properties of an Empirical Likelihood Function

In this section we consider a set $\mathcal{M} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ with $n \ge 2$ points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^d$ and investigate the function $L : \mathbb{R}^d \to [n\log n, \infty]$ with

$$L(\boldsymbol{\eta}) \;:=\; \inf\Big\{S(\boldsymbol{p}) : \boldsymbol{p} \in \Pi_n, \sum_{i=1}^n p_i\boldsymbol{y}_i = \boldsymbol{\eta}\Big\}.$$

With the matrix

$$\underline{\boldsymbol{Y}} \;:=\; [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]^\top \;\in\; \mathbb{R}^{n\times d}$$

one can write

$$L(\boldsymbol{\eta}) \;=\; \inf\big\{S(\boldsymbol{p}) : \boldsymbol{p} \in \Pi_n, \underline{\boldsymbol{Y}}^\top\boldsymbol{p} = \boldsymbol{\eta}\big\}.$$

Exercise 9.1. Show that L is a convex function on $\mathbb{R}^d$.

Before presenting further properties of L, we collect a few geometric facts in the next lemma.

Lemma 9.2. ***(a)*** *The convex hull of $\mathcal{M}$, i.e. the smallest convex set containing $\mathcal{M}$, is equal to*

$$\operatorname{conv}(\mathcal{M}) \;=\; \big\{\underline{\boldsymbol{Y}}^\top\boldsymbol{p} : \boldsymbol{p} \in \Pi_n\big\}.$$

(b) *For any point $\boldsymbol{y}_o \in \operatorname{conv}(\mathcal{M})$, the spaces*

$$\operatorname{span}(\mathcal{M} - \boldsymbol{y}_o),\; \operatorname{span}(\mathcal{M} - \mathcal{M}),\; \operatorname{span}\big(\operatorname{conv}(\mathcal{M}) - \operatorname{conv}(\mathcal{M})\big)$$

and

$$\mathbb{V}(\mathcal{M}) := \big\{\underline{\boldsymbol{Y}}^\top\boldsymbol{b} : \boldsymbol{b} \in \mathbb{R}^n, b_+ = 0\big\}$$

are identical.

(c) *The linear space $\mathbb{V}(\mathcal{M})$ is equal to $\mathbb{R}^d$ if and only if*

$$\operatorname{span}\left\{\begin{pmatrix}\boldsymbol{y}\\1\end{pmatrix} : \boldsymbol{y}\in\mathcal{M}\right\} \;=\; \mathbb{R}^{d+1}.$$

(d) *For an arbitrary point $\boldsymbol{\eta}\in\mathbb{R}^d$, the following two statements are equivalent:*

(d.1) For each $\boldsymbol{v}\in\mathbb{V}(\mathcal{M})$ there exists a number $t>0$ such that $\boldsymbol{\eta}+t\boldsymbol{v}\in\operatorname{conv}(\mathcal{M})$.

(d.2) There exists a vector $\boldsymbol{p}\in\Pi_n\cap(0,1)^n$ such that $\boldsymbol{\eta}=\underline{\boldsymbol{Y}}^\top\boldsymbol{p}$.

Remark. A point $\boldsymbol{\eta}$ with the properties (d.1-2) in Lemma 9.2 is called an *internal point of* $\operatorname{conv}(\mathcal{M})$. In particular, the arithmetic mean $\bar{\boldsymbol{y}} = n^{-1}\sum_{i=1}^n \boldsymbol{y}_i$ is always an internal point of $\operatorname{conv}(\mathcal{M})$. In case of $\mathbb{V}(\mathcal{M})=\mathbb{R}^d$, a point $\boldsymbol{\eta}$ is an internal point of $\operatorname{conv}(\mathcal{M})$ if and only if it is an *interior point* of $\operatorname{conv}(\mathcal{M})$.

Exercise 9.3. Prove parts (a) and (b) of Lemma 9.2.

Proof of Lemma 9.2 (c-d). According to part (b), $\mathbb{V}(\mathcal{M})=\mathbb{R}^d$ if and only if there exist vectors $\boldsymbol{z}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_d$ in $\mathcal{M}$ such that

$$\operatorname{span}(\boldsymbol{z}_1-\boldsymbol{z}_0,\boldsymbol{z}_2-\boldsymbol{z}_0,\ldots,\boldsymbol{z}_d-\boldsymbol{z}_0) \;=\; \mathbb{R}^d.$$

This es equivalent to the inequality

$$\det\big[\boldsymbol{z}_1-\boldsymbol{z}_0,\boldsymbol{z}_2-\boldsymbol{z}_0,\ldots,\boldsymbol{z}_d-\boldsymbol{z}_0\big] \;\neq\; 0.$$

On the other hand,

$$\operatorname{span}\Big\{\begin{pmatrix}\boldsymbol{y}\\1\end{pmatrix} : \boldsymbol{y}\in\mathcal{M}\Big\} \;=\; \mathbb{R}^{d+1}$$

if and only if there exist vectors $\boldsymbol{z}_0,\boldsymbol{z}_1,\ldots,\boldsymbol{z}_d\in\mathcal{M}$ such that

$$\det\begin{bmatrix}\boldsymbol{z}_1, & \boldsymbol{z}_2, & \ldots, & \boldsymbol{z}_d, & \boldsymbol{z}_0\\ 1, & 1, & \ldots, & 1, & 1\end{bmatrix} \;\neq\; 0.$$

But

$$\begin{aligned}\det\big[\boldsymbol{z}_1-\boldsymbol{z}_0,\boldsymbol{z}_2-\boldsymbol{z}_0,\ldots,\boldsymbol{z}_d-\boldsymbol{z}_0\big] \;&=\; \det\begin{bmatrix}\boldsymbol{z}_1-\boldsymbol{z}_0, & \boldsymbol{z}_2-\boldsymbol{z}_0, & \ldots, & \boldsymbol{z}_d-\boldsymbol{z}_0, & \boldsymbol{z}_0\\ 0, & 0, & \ldots, & 0, & 1\end{bmatrix}\\ &=\; \det\begin{bmatrix}\boldsymbol{z}_1, & \boldsymbol{z}_2, & \ldots, & \boldsymbol{z}_d, & \boldsymbol{z}_0\\ 1, & 1, & \ldots, & 1, & 1\end{bmatrix}.\end{aligned}$$

This proves part (c).

It remains to verify part (d). Suppose that $\boldsymbol{\eta}\in\mathbb{R}^d$ satisfies condition (d.1). In particular, $\boldsymbol{\eta}\in\operatorname{conv}(\mathcal{M})$, so $\boldsymbol{\eta}-\bar{\boldsymbol{y}}\in\mathbb{V}(\mathcal{M})$ by part (b). Consequently, there exists a number $t>0$ such that $\boldsymbol{\eta}+t(\boldsymbol{\eta}-\bar{\boldsymbol{y}})\in\operatorname{conv}(\mathcal{M})$. That means, $\boldsymbol{\eta}+t(\boldsymbol{\eta}-\bar{\boldsymbol{y}})=\underline{\boldsymbol{Y}}^\top\boldsymbol{q}$ for a suitable vector $\boldsymbol{q}\in\Pi_n$. But then, $\boldsymbol{\eta}=\underline{\boldsymbol{Y}}^\top\boldsymbol{p}$ with

$$\boldsymbol{p} \;:=\; \Big(\frac{q_i+t/n}{1+t}\Big)_{i=1}^n \;\in\; \Pi_n\cap(0,1)^n.$$

Suppose that condition (d.2) is fulfilled, i.e. $\boldsymbol{\eta} = \underline{\boldsymbol{Y}}^\top \boldsymbol{p}$ for some $\boldsymbol{p} \in \Pi_n \cap (0,1)^n$. For arbitrary vectors $\boldsymbol{v} = \underline{\boldsymbol{Y}}^\top \boldsymbol{b}$ with $b_+ = 0$ and numbers t, $\boldsymbol{\eta} + t\boldsymbol{v} = \underline{\boldsymbol{Y}}^\top(\boldsymbol{p} + t\boldsymbol{b})$, and $\boldsymbol{p} + t\boldsymbol{b} \in \Pi_n$, provided that $|t|$ is sufficiently small. □

After these preparations, we are ready for an explicit representation of a negative empirical log-likelihood function:

Theorem 9.4. ***(a)*** *The function L is convex, and on $\mathrm{conv}(\mathcal{M})$ it is continuous. Its unique minimizer is $\bar{\boldsymbol{y}}$ with $L(\bar{\boldsymbol{y}}) = n \log n$.*

(b) *For arbitrary $\boldsymbol{\eta} \in \mathbb{R}^d$,*

$$L(\boldsymbol{\eta}) \;=\; -\inf\Big\{ S\big(\big(n + \boldsymbol{\lambda}^\top(\boldsymbol{y}_i - \boldsymbol{\eta})\big)_{i=1}^n\big) : \boldsymbol{\lambda} \in \mathbb{R}^d \Big\}.$$

The latter infimum is finite if and only if $\boldsymbol{\eta}$ is an internal point of $\mathrm{conv}(\mathcal{M})$. In this case, $L(\boldsymbol{\eta}) = S(\boldsymbol{p})$ with

$$\boldsymbol{p} \;=\; \Big(\frac{1}{n + \boldsymbol{\lambda}^\top(\boldsymbol{y}_i - \boldsymbol{\eta})}\Big)_{i=1}^n \;\in\; \Pi_n \quad \textit{and} \quad \boldsymbol{\lambda} \;\in\; \operatorname*{arg\,min}_{\boldsymbol{\lambda} \in \mathbb{R}^d} S\big(\big(n + \boldsymbol{\lambda}^\top(\boldsymbol{y}_i - \boldsymbol{\eta})\big)_{i=1}^n\big).$$

This representation is quite useful for the numerical computation of $L(\boldsymbol{\eta})$, particularly in situations when $n \gg d$. For instead of minimizing $S(\cdot)$ on a subset of Π_n, one has to minimize the convex function $\boldsymbol{\lambda} \mapsto S\big(\big(n + \boldsymbol{\lambda}^\top(\boldsymbol{y}_i - \boldsymbol{\eta})\big)_{i=1}^n\big)$ on $\mathbb{R}^d$.

Proof of Theorem 9.4. Convexity of L has been shown in Exercise 9.1. Concerning continuity of L on $\mathrm{conv}(\mathcal{M})$, note first that $L(\boldsymbol{\eta})$ is the *minimum* of $S(\cdot)$ on the set $\big\{\boldsymbol{p} \in \Pi_n : \underline{\boldsymbol{Y}}^\top \boldsymbol{p} = \boldsymbol{\eta}\big\}$, because the latter set is compact, and $S(\cdot)$ is continuous. For $\boldsymbol{\eta} \in \mathrm{conv}(\mathcal{M})$, let $(\boldsymbol{\eta}^{(m)})_m$ be a sequence in $\mathrm{conv}(\mathcal{M})$ such that

$$\lim_{m\to\infty} \boldsymbol{\eta}^{(m)} \;=\; \boldsymbol{\eta} \quad \text{and} \quad \lim_{m\to\infty} L(\boldsymbol{\eta}^{(m)}) \;=\; \liminf_{\mathrm{conv}(\boldsymbol{M}) \ni \boldsymbol{\gamma} \to \boldsymbol{\eta}} L(\boldsymbol{\gamma}).$$

Now we write $L(\boldsymbol{\eta}^{(m)}) = S(\boldsymbol{p}^{(m)})$ for some $\boldsymbol{p}^{(m)} \in \Pi_n$ with $\underline{\boldsymbol{Y}}^\top \boldsymbol{p}^{(m)} = \boldsymbol{\eta}^{(m)}$. Since Π_n is comapct, we may assume without loss of generality that $(\boldsymbol{p}^{(m)})_m$ converges to a vector $\boldsymbol{p} \in \Pi_n$. But the latter vector fulfils automatically the equation $\underline{\boldsymbol{Y}}^\top \boldsymbol{p} = \boldsymbol{\eta}$. Hence,

$$L(\boldsymbol{\eta}) \;\le\; S(\boldsymbol{p}) \;=\; \lim_{m\to\infty} S(\boldsymbol{p}^{(m)}) \;=\; \liminf_{\mathrm{conv}(\mathcal{M}) \ni \boldsymbol{\gamma} \to \boldsymbol{\eta}} L(\boldsymbol{\gamma}).$$

On the other hand, let $L(\boldsymbol{\eta}) = S(\boldsymbol{p}) < \infty$ for some $\boldsymbol{p} \in \Pi_n$ with $\underline{\boldsymbol{Y}}^\top \boldsymbol{p} = \boldsymbol{\eta}$. In particular, $\boldsymbol{p} \in (0,1)^n$. Now let $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(e)}$ be basis vectors of $\mathbb{V}(\mathcal{M})$, and let us write $\boldsymbol{b}^{(j)} = \underline{\boldsymbol{Y}}^\top \boldsymbol{\beta}^{(j)}$ with $\beta_+^{(j)} = 0$. Then

$$\begin{aligned}
\limsup_{\mathrm{conv}(\mathcal{M}) \ni \boldsymbol{\gamma} \to \boldsymbol{\eta}} L(\boldsymbol{\gamma}) \;&\le\; \limsup_{c_1,\ldots,c_e \to 0} L\Big(\boldsymbol{\eta} + \sum_{j=1}^e c_j \boldsymbol{b}^{(j)}\Big) \\
&\le\; \limsup_{c_1,\ldots,c_e \to 0} S\Big(\boldsymbol{p} + \sum_{j=1}^e c_j \boldsymbol{\beta}^{(j)}\Big) \;=\; S(\boldsymbol{p}) \;=\; L(\boldsymbol{\eta}).
\end{aligned}$$

This proves continuity of $L(\cdot)$ on $\mathrm{conv}(\mathcal{M})$. Since $S(\boldsymbol{p}) \geq n \log n$ for all $\boldsymbol{p} \in \Pi_n$ with equality if and only if $\boldsymbol{p} = (1/n)_{i=1}^n$, $L(\boldsymbol{\eta}) \geq n \log n$ with equality if and only if $\boldsymbol{\eta} = \bar{\boldsymbol{y}}$.

Now we turn to the special representation of $L(\boldsymbol{\eta})$. Without loss of generality let $\boldsymbol{\eta} = \mathbf{0}$, otherwise we could replace each vector $\boldsymbol{y}_i$ with $\boldsymbol{y}_i - \boldsymbol{\eta}$.

Suppose that $L(\mathbf{0}) < \infty$, that means, $\mathbf{0}$ is an internal point of $\mathrm{conv}(\mathcal{M})$. The definition of $L(\mathbf{0})$ and convexity of $S(\cdot)$ imply that the following properties of $\boldsymbol{p} \in \mathbb{R}^n$ are equivalent:

(ELLF.1) $\boldsymbol{p} \in \Pi_n \cap (0,1)^n$ with $\underline{\boldsymbol{Y}}^\top \boldsymbol{p} = \mathbf{0}$ and $L(\mathbf{0}) = S(\boldsymbol{p})$.

(ELLF.2) $\boldsymbol{p} \in \Pi \cap (0,1)^n$ with $\underline{\boldsymbol{Y}}^\top \boldsymbol{p} = \mathbf{0}$ and

$$\frac{d}{dt}\Big|_{t=0} S(\boldsymbol{p} + t\boldsymbol{\delta}) \;=\; 0 \quad \text{for all } \boldsymbol{\delta} \in \mathbb{R}^n \text{ with } \delta_+ = 0, \underline{\boldsymbol{Y}}^\top \boldsymbol{\delta} = \mathbf{0}.$$

This follows from the fact that the linear space $\mathrm{span}\big\{\boldsymbol{q} - \boldsymbol{p} : \boldsymbol{q} \in \Pi_n, \underline{\boldsymbol{Y}}^\top \boldsymbol{q} = \mathbf{0}\big\}$ coincides with $\big\{\boldsymbol{\delta} \in \mathbb{R}^n : \delta_+ = 0, \underline{\boldsymbol{Y}}^\top \boldsymbol{\delta} = \mathbf{0}\big\}$. But

$$\frac{d}{dt}\Big|_{t=0} S(\boldsymbol{p} + t\boldsymbol{\delta}) \;=\; \sum_{i=1}^n \frac{\delta_i}{p_i}.$$

Hence, (ELLF.2) is equivalent to

(ELLF.2') $\boldsymbol{p} \in \Pi \cap (0,1)^n$ with $\underline{\boldsymbol{Y}}^\top \boldsymbol{p} = \mathbf{0}$ and

$$(1/p_i)_{i=1}^n \;\perp\; \{\mathbf{1}_n, \boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(d)\}^\perp.$$

Here $\boldsymbol{Y}(1), \boldsymbol{Y}(2), \ldots, \boldsymbol{Y}(d)$ denote the columns of $\underline{\boldsymbol{Y}}$. The condition on $(1/p_i)_{i=1}^n$ in (ELLF.2') is equivalent to $(1/p_i)_{i=1}^n$ being a linear combination of the vectors $\mathbf{1}_n$, $\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(d)$, so

$$\boldsymbol{p} \;=\; \big((\kappa + \boldsymbol{\lambda}^\top \boldsymbol{y}_i)^{-1}\big)_{i=1}^n$$

for certain $\kappa \in \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^d$. Together with $\sum_{i=1}^n p_i \boldsymbol{y}_i = \mathbf{0}$ and $p_+ = 1$, we find out that

$$0 \;=\; \sum_{i=1}^n \frac{\boldsymbol{\lambda}^\top \boldsymbol{y}_i}{\kappa + \boldsymbol{\lambda}^\top \boldsymbol{y}_i} \;=\; \sum_{i=1}^n \Big(1 - \frac{\kappa}{\kappa + \boldsymbol{\lambda}^\top \boldsymbol{y}_i}\Big) \;=\; n - \kappa,$$

whence $\kappa = n$. Consequently,

$$\boldsymbol{p} \;=\; \big((n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i)^{-1}\big)_{i=1}^n$$

for some $\boldsymbol{\lambda} \in \mathbb{R}^d$ such that

$$\min_{i=1,\ldots,n} (n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i) \;>\; 0 \quad \text{and} \quad \sum_{i=1}^n \frac{\boldsymbol{y_i}}{n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i} \;=\; \mathbf{0}. \tag{9.1}$$

On the other hand, let $\boldsymbol{\lambda}$ be a vector in $\mathbb{R}^d$ satisfying (9.1), and let $\boldsymbol{p} := \big((n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i)^{-1}\big)_{i=1}^n$. Then $\boldsymbol{p}$ belongs automatically to Π_n, because

$$\sum_{i=1}^n p_i \;=\; \frac{1}{n} \sum_{i=1}^n \frac{n}{n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i} \;=\; \frac{1}{n} \sum_{i=1}^n \frac{n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i}{n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i} \;=\; 1.$$

These considerations show that condition (ELLF.1) on $\boldsymbol{p} \in \mathbb{R}^n$ is equivalent to the following condition:

(ELLF.3) $\boldsymbol{p} = \big((n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i)^{-1}\big)_{i=1}^n$ for some $\boldsymbol{\lambda} \in \mathbb{R}^d$ satisfying (9.1).

Now we consider the function $h : \mathbb{R}^d \to (-\infty, \infty]$ with

$$h(\boldsymbol{\lambda}) \;=\; S\big((n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i)_{i=1}^n\big).$$

This is a convex function on $\mathbb{R}^d$, where $h(\boldsymbol{\lambda}) < \infty$ if and only if $n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i > 0$ for all indices i. For such vectors $\boldsymbol{\lambda}$,

$$\nabla h(\boldsymbol{\lambda}) \;=\; -\sum_{i=1}^n \frac{\boldsymbol{y}_i}{n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i}.$$

Consequently, $\boldsymbol{\lambda} \in \mathbb{R}^d$ is a minimizer of h if and only if it satisfies condition (9.1). The corresponding minimum equals

$$h(\boldsymbol{\lambda}) \;=\; -S(\boldsymbol{p}) \quad \text{with} \quad \boldsymbol{p} \;:=\; \big((n + \boldsymbol{\lambda}^\top \boldsymbol{y}_i)^{-1}\big)_{i=1}^n.$$

It remains to show that $\inf\big\{h(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathbb{R}^d\big\}$ is equal to $-\infty$ whenever $\mathbf{0}$ is not an internal point of $\mathrm{conv}(\mathcal{M})$. At first, let $\mathbf{0} \notin \mathrm{conv}(\mathcal{M})$. Then $\boldsymbol{y}_o := \arg\min_{\boldsymbol{y} \in \mathrm{conv}(\mathcal{M})} \|\boldsymbol{y}\|$ satisfies the inequality

$$\min_{\boldsymbol{y} \in \mathrm{conv}(\mathcal{M})} \boldsymbol{y}_o^\top \boldsymbol{y} \;=\; \|\boldsymbol{y}_o\|^2 \;>\; 0.$$

Then, for $t \geq 0$ we have the inequality

$$h(t\boldsymbol{y}_o) = -\sum_{i=1}^n \log(n + t\boldsymbol{y}_o^\top \boldsymbol{y}_i) \;\leq\; -n \log(n + t\|y_o\|^2),$$

and the right-hand side tends to $-\infty$ as $t \to \infty$.

Now let $\mathbf{0}$ be a point in $\mathrm{conv}(\mathcal{M})$, but not an internal point of $\mathrm{conv}(\mathcal{M})$. If we replace $\mathbb{R}^d$ with $\mathbb{V}(\mathcal{M}) = \mathrm{span}(\mathcal{M} - \mathbf{0}) = \mathrm{span}(\mathcal{M})$, then $\mathrm{conv}(\mathcal{M})$ has nonempty interior, and $\mathbf{0}$ is a boundary point of $\mathrm{conv}(\mathcal{M})$. It is known from convex analysis that there exists a vector $\boldsymbol{y}_o \in \mathbb{V}(\mathcal{M})$ such that $\boldsymbol{y}_o^\top \boldsymbol{y} \geq 0$ for all $\boldsymbol{y} \in \mathrm{conv}(\mathcal{M})$, and $\boldsymbol{y}_o^\top \boldsymbol{y}_I > 0$ for at least one index $I \in \{1, 2, \ldots, n\}$. Consequently, for $t \geq 0$ we get the inequality

$$h(t\boldsymbol{y}_o) = -\sum_{i=1}^n \log(n + t\boldsymbol{y}_o^\top \boldsymbol{y}_i) \;\leq\; -(n-1)\log(n) - \log(n + t\boldsymbol{y}_o^\top \boldsymbol{y}_I),$$

and the right-hand side converges to $-\infty$ as $t \to \infty$. □

The compact representation of $L(\boldsymbol{\eta})$ in Theorem 9.4 leads to a rather short proof of the following result:

Theorem 9.5. *For $\boldsymbol{\eta} \in \mathbb{R}^d$ let*

$$\begin{aligned}
\boldsymbol{Z} \;&:=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\eta}), \\
M \;&:=\; \frac{1}{\sqrt{n}} \max_{i=1,\ldots,n} \|\boldsymbol{y}_i - \boldsymbol{\eta}\|, \\
\widehat{\boldsymbol{\Gamma}} \;&:=\; \frac{1}{n} \sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\eta})(\boldsymbol{y}_i - \boldsymbol{\eta})^\top.
\end{aligned}$$

Then, for any constant $\kappa > 0$ and fixed symmetric, positive definite matrix $\mathbf{\Gamma} \in \mathbb{R}^{d\times d}$,

$$2L(\boldsymbol{\eta}) - 2n\log(n) \;=\; \boldsymbol{Z}^\top \mathbf{\Gamma}^{-1}\boldsymbol{Z} + \rho(\mathcal{M}, \boldsymbol{\eta}, \mathbf{\Gamma}),$$

where $\rho(\mathcal{M}, \boldsymbol{\eta}, \mathbf{\Gamma}) \to 0$ as

$$M \;\to\; 0, \quad \|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\| \;\to\; 0 \quad \textit{and} \quad \|\boldsymbol{Z}\| \;\le\; \kappa.$$

Proof of Theorem 9.5. Without loss of generality let $\boldsymbol{\eta} = \mathbf{0}$. Otherwise we could just replace each $\boldsymbol{y}_i$ with $\boldsymbol{y}_i - \boldsymbol{\eta}$. By means of Theorem 9.4 we obtain the representation

$$2L(\mathbf{0}) - 2L(\bar{\boldsymbol{y}}) \;=\; -\inf_{\Delta\in\mathbb{R}^d} H(\Delta) \quad \text{with} \quad H(\Delta) \;:=\; -2\sum_{i=1}^n \log\Big(1 + \frac{\boldsymbol{y}_i^\top \Delta}{\sqrt{n}}\Big).$$

Now we use the elementary inequality

$$\big|\log(1+x) - x + x^2/2\big| \;\le\; \frac{|x|^3}{3(1-|x|)^+} \quad \text{for } x \in \mathbb{R},$$

see Exercise 9.6. This implies that

$$\begin{aligned} H(\Delta) \;&=\; -\frac{2}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{y}_i^\top\Delta + \frac{1}{n}\sum_{i=1}^n (\boldsymbol{y}_i^\top\Delta)^2 + r_1(\Delta) \\ &=\; -2\boldsymbol{Z}^\top\Delta + \Delta^\top\widehat{\mathbf{\Gamma}}\Delta + r_1(\Delta) \\ &=\; \check{H}(\Delta) + r_1(\Delta) + r_2(\Delta), \end{aligned}$$

where

$$\check{H}(\Delta) \;:=\; -2\boldsymbol{Z}^\top\Delta + \Delta^\top\mathbf{\Gamma}\Delta$$

and

$$\begin{aligned} |r_1(\Delta)| \;&\le\; \frac{2M\|\Delta\|}{3(1-M\|\Delta\|)^+}\frac{1}{n}\sum_{i=1}^n (\boldsymbol{y}_i^\top\Delta)^2 \;=\; \frac{2M\|\Delta\|\Delta^\top\widehat{\mathbf{\Gamma}}\Delta}{3(1-M\|\Delta\|)^+} \;\le\; \frac{2M\|\Delta\|^3\lambda_{\max}(\widehat{\mathbf{\Gamma}})}{3(1-M\|\Delta\|)^+}, \\ |r_2(\Delta)| \;&\le\; \|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|\|\Delta\|^2. \end{aligned}$$

Thus, under the stated conditions, $\sup_{\|\Delta\|\le C}\big|r_1(\Delta) + r_2(\Delta)\big|$ converges to 0 for any fixed $C > 0$. This implies that

$$-\inf_{\Delta\in\mathbb{R}^d} H(\Delta) \;=\; -\min_{\Delta\in\mathbb{R}^d} \check{H}(\Delta) + o(1) \;=\; -\check{H}(\mathbf{\Gamma}^{-1}\boldsymbol{Z}) + o(1) \;=\; \boldsymbol{Z}^\top\mathbf{\Gamma}^{-1}\boldsymbol{Z} + o(1),$$

by similar arguments as in the proof of Theorem 7.20. □

Exercise 9.6. Show that

$$x - \frac{x^2}{2} \;\le\; \log(1+x) \;\le\; x - \frac{x^2}{2} + \frac{x^3}{3} \quad \text{for } x \ge 0$$

and

$$x - \frac{x^2}{2} + \frac{x^3}{3(1+x)} \;\le\; \log(1+x) \;\le\; x - \frac{x^2}{2} \quad \text{for } -1 < x \le 0.$$

9.1.2 Inference about the Mean

Now we discuss the special case that $\mathcal{Y} = \mathbb{R}^d$ and $\theta(P_n) := \int \boldsymbol{y}\, P_n(d\boldsymbol{y}) =: \boldsymbol{\mu}_n$. In this case, the previous results imply the following one:

Corollary 9.7. *Suppose that the distributions $P_n = \mathcal{L}(\boldsymbol{Y}_{n1})$ fulfil the following conditions:*

$$\boldsymbol{\Sigma}_n := \mathrm{Var}(\boldsymbol{Y}_{n1}) \;\to\; \boldsymbol{\Sigma} \quad \textit{with } \lambda_{\min}(\boldsymbol{\Sigma}) > 0,$$

$$\Lambda_n := \mathbb{E}\Big(\|\boldsymbol{Y}_{n1} - \boldsymbol{\mu}_n\|^2 \min\big\{ n^{-1/2} \|\boldsymbol{Y}_{n1} - \boldsymbol{\mu}_n\|, 1 \big\} \Big) \;\to\; 0.$$

Then,

$$2L_n(\boldsymbol{\mu}_n) - 2n \log n \;\to_{\mathcal{L}}\; \chi^2_d.$$

This result implies that the confidence set

$$\big\{ \boldsymbol{\eta} \in \mathbb{R}^d : L_n(\boldsymbol{\eta}) \le n \log n + \chi^2_{d;1-\alpha}/2 \big\}$$

contains the unknown mean $\boldsymbol{\mu}_n$ with asymptotic probability $1 - \alpha$.

Proof of Corollary 9.7. This result follows directly from Theorem 9.5 and the Central Limit Theorem (Theorem A.16), because

$$\begin{aligned}
\boldsymbol{Z}_n \;&:= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\boldsymbol{Y}_{ni} - \boldsymbol{\mu}_n) \;\to_{\mathcal{L}}\; \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}), \\
M_n \;&:= \frac{1}{\sqrt{n}} \max_{1 \le i \le n} \|\boldsymbol{Y}_{ni} - \boldsymbol{\mu}_n\| \;\to_p\; 0, \\
\widehat{\boldsymbol{\Sigma}}_n \;&:= \frac{1}{n} \sum_{i=1}^n (\boldsymbol{Y}_{ni} - \boldsymbol{\mu}_n)(\boldsymbol{Y}_{ni} - \boldsymbol{\mu}_n)^\top \;\to_p\; \boldsymbol{\Sigma}.
\end{aligned}$$

□

Numerical example. The explicit computation and visualization of the confidence region

$$C_{n,\alpha} \;:=\; \big\{ \boldsymbol{\eta} \in \mathbb{R}^d : L_n(\boldsymbol{\eta}) \le n \log n + \chi^2_{d;1-\alpha}/2 \big\}$$

is a non-trivial task. But in dimensions $d \le 2$, one can approximate this set sufficiently well by calculating $L_n(\boldsymbol{\eta})$ for all vectors $\boldsymbol{\eta}$ in a fine grid. Figure 9.1 depicts a sample of size $n = 50$ from a bivariate Gaussian distribution with mean $0 \in \mathbb{R}^2$ and covariances $\Sigma(1,1) = \Sigma(2,2) = 1$, $\Sigma(1,2) = 0.7$. In addition one sees the true mean (green star), the sample mean (solid black point) and the boundary of $C_{n,\alpha}$ for $\alpha = 0.5, 0.1, 0.05, 0.01$ (blue lines).

9.2 Empirical Likelihood for Linear Regression

Now we consider a linear regression setting with observations $(\boldsymbol{d}_{n1}, Y_{n1}), \ldots, (\boldsymbol{d}_{nn}, Y_{nn})$, consisting of fixed vectors $\boldsymbol{d}_{ni} \in \mathbb{R}^p$ and random variables

$$Y_{ni} \;=\; \boldsymbol{d}_{ni}^\top \boldsymbol{\theta}_n + \varepsilon_{ni}$$

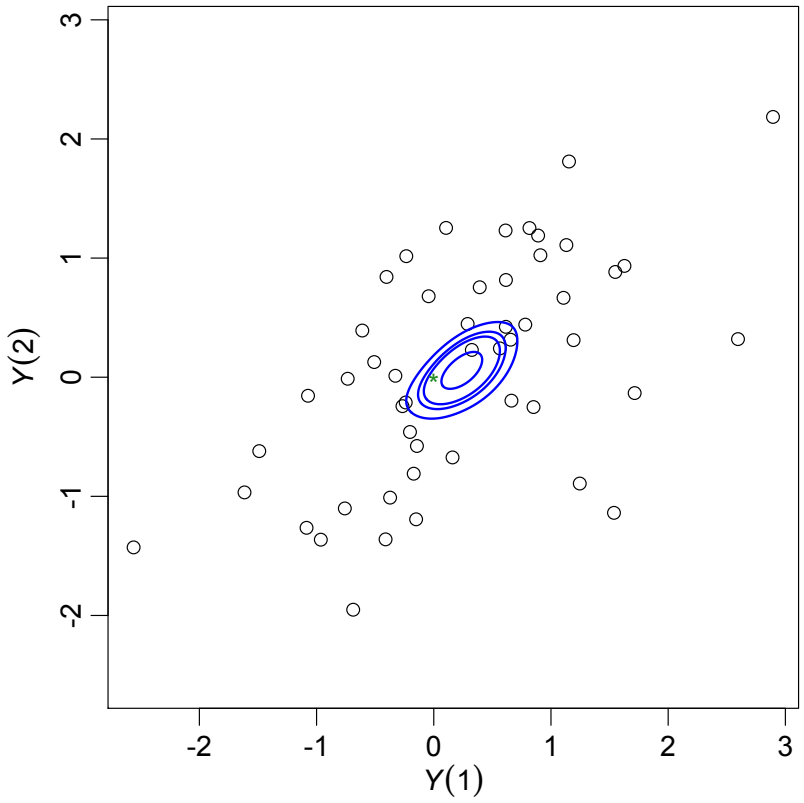

Figure 9.1: Confidence regions for a mean via empirical likelihood.

with a fixed unknown parameter $\boldsymbol{\theta}_n \in \mathbb{R}^p$ and stochastically independent random errors ε_{n1}, ε_{n2}, $\ldots$, ε_{nn}, where

$$\mathbb{E}(\varepsilon_{ni}) \;=\; 0 \quad \text{and} \quad \sigma_{ni} := \mathrm{Std}(\varepsilon_{ni}) \;<\; \infty$$

for $1 \le i \le n$. We also assume that the design matrix $\boldsymbol{D}_n = [\boldsymbol{d}_{n1}\, \boldsymbol{d}_{n2}\, \ldots\, \boldsymbol{d}_{nn}]^\top$ has rank p for sufficiently large n.

To motivate the definition of the negative empirical log-likelihood function, we first discuss a particular estimation problem: For a probability weight vector $\boldsymbol{p} \in \Pi_n \cap (0,1)^n$ and any vector $\boldsymbol{\theta} \in \mathbb{R}^p$, we consider the sum of squares

$$Q_{\boldsymbol{p}}(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^{n} p_i (Y_{ni} - \boldsymbol{d}_{ni}^\top \boldsymbol{\theta})^2.$$

This is a strictly convex function on $\boldsymbol{\theta}$ with

$$\nabla Q_{\boldsymbol{p}}(\boldsymbol{\theta}) \;=\; -2 \sum_{i=1}^{n} p_i (Y_{ni} - \boldsymbol{d}_{ni}^\top \boldsymbol{\theta}) \boldsymbol{d}_{ni} \quad \text{and} \quad D^2 Q_{\boldsymbol{p}}(\boldsymbol{\theta}) \;=\; 2 \sum_{i=1}^{n} p_i \boldsymbol{d}_{ni} \boldsymbol{d}_{ni}^\top.$$

In particular, the unique minimizer of $Q_{\boldsymbol{p}}(\cdot)$ is given by the equation

$$\sum_{i=1}^{n} p_i \boldsymbol{Y}_{ni}(\boldsymbol{\theta}) \;=\; \boldsymbol{0}$$

with the *residual vectors*

$$\boldsymbol{Y}_{ni}(\boldsymbol{\theta}) \;:=\; (Y_{ni} - \boldsymbol{d}_{ni}^\top \boldsymbol{\theta}) \boldsymbol{d}_{ni} \;\in\; \mathbb{R}^p.$$

An explicit formula for this minimizer is given by

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{p}} \;=\; (\boldsymbol{D}_n^\top \operatorname{diag}(\boldsymbol{p}) \boldsymbol{D}_n)^{-1} \boldsymbol{D}_n^\top \operatorname{diag}(\boldsymbol{p}) \boldsymbol{Y}_n,$$

and for the special weight vector $\boldsymbol{p} = (1/n)_{i=1}^n$, we obtain the usual LSE $\widehat{\boldsymbol{\theta}}_n$.

The residual vectors $\boldsymbol{Y}_{ni}(\boldsymbol{\theta})$ play a particular role, because $\mathbb{E}\, \boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n) = \mathbf{0}$, and

$$\mathbb{E}\Big(\sum_{i=1}^n p_i \boldsymbol{Y}_{ni}(\boldsymbol{\theta}) \Big) \;=\; (\boldsymbol{D}_n^\top \operatorname{diag}(\boldsymbol{p}) \boldsymbol{D}_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta})$$

is equal to $\mathbf{0}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_n$. Thus we define

$$L_n(\boldsymbol{\theta}) \;:=\; \inf\Big\{ S(\boldsymbol{p}) : \boldsymbol{p} \in \Pi_n, \sum_{i=1}^n p_i \boldsymbol{Y}_{ni}(\boldsymbol{\theta}) = \mathbf{0} \Big\}.$$

Theorem 9.4 provides us with the alternative representation

$$L_n(\boldsymbol{\theta}) \;=\; - \inf_{\boldsymbol{\lambda} \in \mathbb{R}^p} S\big(\big(n + \boldsymbol{\lambda}^\top \boldsymbol{Y}_{ni}(\boldsymbol{\theta}) \big)_{i=1}^n \big),$$

and one can verify easily that $L_n(\cdot)$ is a convex function on $\mathbb{R}^p$.

The subsequent results show that under certain conditions, likelihood-based procedures as introduced in the context of logistic and Poisson regression are applicable with the empirical log-likelihood function in place of a usual log-likelihood function. In particular, the wild bootstrap as well as empirical likelihood are able to deal with heteroscedastic errors. Our asymptotic considerations refer to the following conditions:

(E.1) For a fixed symmetric, positive definite matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times p}$,

$$\frac{1}{n} \sum_{i=1}^n \sigma_{ni}^2 \, \boldsymbol{d}_{ni} \boldsymbol{d}_{ni}^\top \;\to\; \boldsymbol{\Gamma}.$$

(E.2)

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}(|\varepsilon_{ni}|^3) \|\boldsymbol{d}_{ni}\|^3 \;=\; O(1).$$

(E.3) For a fixed symmetric, positive definite matrix $\boldsymbol{\Gamma}_o \in \mathbb{R}^{p \times p}$,

$$\frac{1}{n} \sum_{i=1}^n \boldsymbol{d}_{ni} \boldsymbol{d}_{ni}^\top \;\to\; \boldsymbol{\Gamma}_o.$$

(E.4)

$$\frac{1}{n} \sum_{i=1}^n \|\boldsymbol{d}_{ni}\|^4 \;=\; O(1).$$

Theorem 9.8. *Under conditions (E.1–2),*

$$2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) \;=\; \boldsymbol{Z}_n^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}_n + o_p(1) \;\to_{\mathcal{L}}\; \chi_p^2,$$

where $\boldsymbol{Z}_n := n^{-1/2} \sum_{i=1}^n \varepsilon_{ni} \boldsymbol{d}_{ni} \to_{\mathcal{L}} \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Gamma})$.

Under conditions (E.1–4),

$$2L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - 2L_n(\boldsymbol{\theta}_n) \;=\; -2\boldsymbol{Z}_{n*}^\top\Delta + \Delta^\top\boldsymbol{\Gamma}_*\Delta + r_n(\Delta)$$

for arbitrary $\Delta \in \mathbb{R}^p$, *where* $\boldsymbol{\Gamma}_* := \boldsymbol{\Gamma}_o\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_o$, $\boldsymbol{Z}_{n*} := \boldsymbol{\Gamma}_o\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n \to_{\mathcal{L}} \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Gamma}_*)$, *and*

$$\sup_{\Delta\,:\,\|\Delta\|\le C} |r_n(\Delta)| \;\to_p\; 0 \quad \textit{for any fixed } C > 0.$$

Proof of Theorem 9.8. For both parts, we apply Theorem 9.5 with p in place of d and $\boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n) = \varepsilon_{ni}\boldsymbol{d}_{ni}$ or $\boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n + n^{-1/2}\Delta) = \varepsilon_{ni}\boldsymbol{d}_{ni} + n^{-1/2}\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top\Delta$ in place of $\boldsymbol{y}_i - \boldsymbol{\eta}$.

Conditions (E.1–2) and Lindeberg's CLT imply that

$$\begin{aligned}
\boldsymbol{Z}_n \;&=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_{ni}\boldsymbol{d}_{ni} \;\to_{\mathcal{L}}\; \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Gamma}),\\
M_n \;&:=\; \frac{1}{\sqrt{n}}\max_{1\le i\le n}\|\varepsilon_{ni}\boldsymbol{d}_{ni}\| \;\to_p\; 0,\\
\widehat{\boldsymbol{\Gamma}}_n \;&:=\; \frac{1}{n}\sum_{i=1}^n \varepsilon_{ni}^2\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top \;\to_p\; \boldsymbol{\Gamma}.
\end{aligned}$$

Hence, Theorem 9.5 implies that

$$2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n) \;=\; \boldsymbol{Z}_n^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n + o_p(1),$$

and $\boldsymbol{Z}_n^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n = \|\boldsymbol{\Gamma}^{-1/2}\boldsymbol{Z}_n\|^2 \to_{\mathcal{L}} \chi_p^2$.

To verify the second part, we consider the auxiliary objects

$$\begin{aligned}
\boldsymbol{Z}_n(\Delta) \;&:=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n + n^{-1/2}\Delta),\\
M_n(\Delta) &:=\; \frac{1}{\sqrt{n}}\max_{1\le i\le n}\|\boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n + n^{-1/2}\Delta)\|,\\
\widehat{\boldsymbol{\Gamma}}_n(\Delta) &:=\; \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n + n^{-1/2}\Delta)\boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n + n^{-1/2}\Delta)^\top.
\end{aligned}$$

Note that $\boldsymbol{Y}_{ni}(\boldsymbol{\theta}_n + n^{-1/2}\Delta) = \varepsilon_{ni}\boldsymbol{d}_{ni} - n^{-1/2}\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top\Delta$, whence

$$\boldsymbol{Z}_n(\Delta) \;=\; \boldsymbol{Z}_n - \boldsymbol{\Gamma}_{no}\Delta \quad \text{with} \quad \boldsymbol{\Gamma}_{no} \;:=\; \frac{1}{n}\sum_{i=1}^n \boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top.$$

Thus, for any constant $C > 0$ and arbitrary vectors Δ with $\|\Delta\| \le C$,

$$\begin{aligned}
\|\boldsymbol{Z}_n(\Delta) - \boldsymbol{Z}_n\| \;&\le\; \lambda_{\max}(\boldsymbol{\Gamma}_{no})C \;=\; O(1),\\
\|\boldsymbol{Z}_n(\Delta) - (\boldsymbol{Z}_n - \boldsymbol{\Gamma}_o\Delta)\| &\le\; \|\boldsymbol{\Gamma}_{no} - \boldsymbol{\Gamma}_o\|C \;=\; o(1).
\end{aligned}$$

Moreover,

$$\begin{aligned}
M_n(\Delta) \;&\le\; M_n + Cn^{-1}\max_{1\le i\le n}\|\boldsymbol{d}_{ni}\|^2\\
&\le\; M_n + Cn^{-1/2}\Big(\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{d}_{ni}\|^4\Big)^{1/2}\\
&=\; M_n + O(n^{-1/2}) \;\to_p\; 0,
\end{aligned}$$

and

$$\begin{aligned}
\big\|\widehat{\boldsymbol{\Gamma}}_n(\Delta) - \widehat{\boldsymbol{\Gamma}}_n\big\| \;&=\; \Big\|\frac{1}{n}\sum_{i=1}^n \big(n^{-1}(\boldsymbol{d}_{ni}^\top\Delta)^2 - 2n^{-1/2}\varepsilon_{ni}\boldsymbol{d}_{ni}^\top\Delta\big)\boldsymbol{d}_{ni}\boldsymbol{d}_{ni}^\top\Big\| \\
&\le\; \frac{C^2}{n^2}\sum_{i=1}^n \|\boldsymbol{d}_{ni}\|^4 + \frac{2C}{n^{3/2}}\sum_{i=1}^n |\varepsilon_{ni}|\|\boldsymbol{d}_{ni}\|^3 \\
&\le\; O(n^{-1}) + \frac{2C}{n^{3/2}}\Big(\sum_{i=1}^n |\varepsilon_{ni}|^3\|\boldsymbol{d}_{ni}\|^3\Big)^{1/3}\Big(\sum_{j=1}^n \|\boldsymbol{d}_{nj}\|^3\Big)^{2/3} \\
&=\; O(n^{-1}) + \frac{2C}{\sqrt{n}}\Big(\frac{1}{n}\sum_{i=1}^n |\varepsilon_{ni}|^3\|\boldsymbol{d}_{ni}\|^3\Big)^{1/3}\Big(\frac{1}{n}\sum_{j=1}^n \|\boldsymbol{d}_{nj}\|^3\Big)^{2/3} \\
&\le\; O(n^{-1}) + \frac{2C}{\sqrt{n}}\Big(\frac{1}{n}\sum_{i=1}^n |\varepsilon_{ni}|^3\|\boldsymbol{d}_{ni}\|^3\Big)^{1/3}\Big(\frac{1}{n}\sum_{j=1}^n \|\boldsymbol{d}_{nj}\|^4\Big)^{1/2} \\
&=\; O_p(n^{-1/2}).
\end{aligned}$$

Consequently, we can apply Theorem 9.5 simultaneously for all $\Delta \in \mathbb{R}^p$ with $\|\Delta\| \le C$, and this leads to

$$\begin{aligned}
&2L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - 2L_n(\boldsymbol{\theta}_n) \\
&\quad=\; \big(2L_n(\boldsymbol{\theta}_n + n^{-1/2}\Delta) - 2L_n(\widehat{\boldsymbol{\theta}}_n)\big) - \big(2L_n(\boldsymbol{\theta}_n) - 2L_n(\widehat{\boldsymbol{\theta}}_n)\big) \\
&\quad=\; (\boldsymbol{Z}_n - \boldsymbol{\Gamma}_o\Delta)^\top\boldsymbol{\Gamma}^{-1}(\boldsymbol{Z}_n - \boldsymbol{\Gamma}_o\Delta) - \boldsymbol{Z}_n^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n + r_n(\Delta) \\
&\quad=\; -2(\boldsymbol{\Gamma}_o\boldsymbol{\Gamma}^{-1}\boldsymbol{Z}_n)^\top\Delta + \Delta^\top\boldsymbol{\Gamma}_o\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_o\Delta + r_n(\Delta) \\
&\quad=\; -2\boldsymbol{Z}_{n*}^\top\Delta + \Delta^\top\boldsymbol{\Gamma}_*\Delta + r_n(\Delta),
\end{aligned}$$

where $\sup\big\{|r_n(\Delta)| : \|\Delta\| \le C\big\} \to_p 0$. □

Example 9.9 (Simple linear regression). In the left panel of Figure 9.2, one sees a scatter plot of simulated data (X_i, Y_i), $1 \le i \le n = 200$ with $X_i = i/n$ and

$$Y_i \;=\; 1 + X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, X_i^2).$$

In addition, the true regression function $f(x) = 1 + x$ (green line), the estimated regression function $\widehat{f}(x) = \widehat{a} + \widehat{b}x$ (blue straight line) and simultaneous confidence intervals

$$\big[\min\{a + bx : (a,b)^\top \in C_{0.05}\}, \max\{a + bx : (a,b)^\top \in C_{0.05}\}\big]$$

for $x \in \mathbb{R}$ (blue curves) are depicted, where

$$C_\alpha \;:=\; \big\{\boldsymbol{\theta} \in \mathbb{R}^2 : L(\boldsymbol{\theta}) \le n\log n + \chi^2_{2;1-\alpha}/2\big\}.$$

The right panel of Figure 9.2 depicts the boundary of the confidence region C_α for the confidence levels $1 - \alpha = 0.5, 0.9, 0.95, 0.99$ (blue curves), the LSE $\widehat{\boldsymbol{\theta}}$ (black dot) and the true parameter $\boldsymbol{\theta} = (1,1)^\top$ (green star).

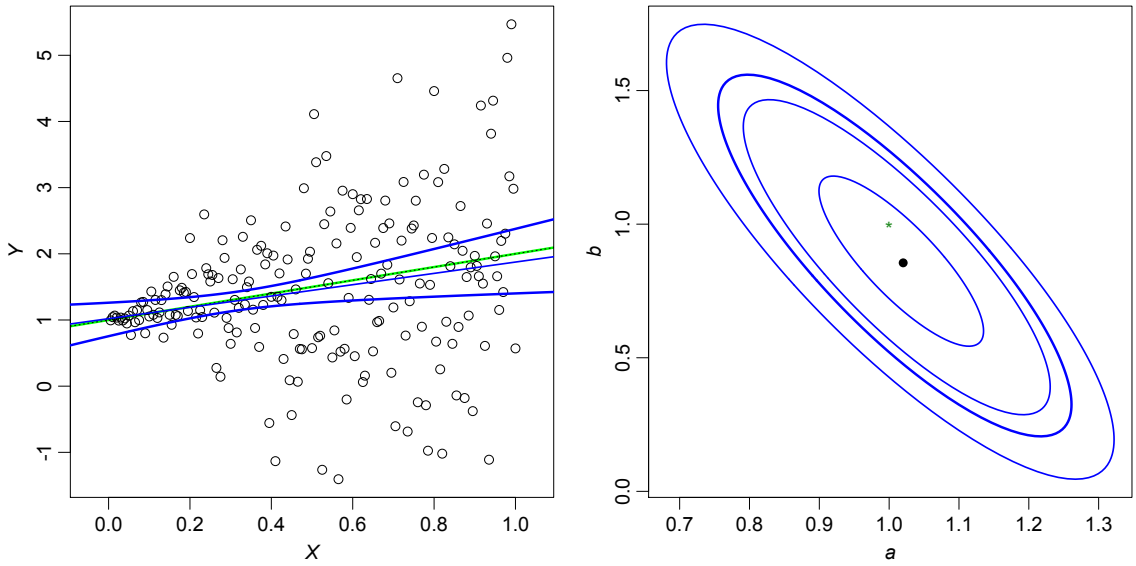

Figure 9.2: Confidence regions for simple linear regression via empirical likelihood.

Example 9.10 (Quadratic regression). In the upper panel of Figure 9.3, one sees a scatter plot of simulated data (X_i, Y_i), $1 \leq i \leq n = 200$ with $X_i = i/n$ and

$$Y_i \;=\; 1 + X_i - X_i^2 + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, (X_i/4)^2).$$

In addition, the true regression function $f(x) = 1 + x - x^2$ (green curve), the estimated regression function $\widehat{f}(x) = \widehat{a}_0 + \widehat{a}_1 x + \widehat{a}_2 x^2$ (central blue curve) and simultaneous confidence intervals

$$\big[\min\{a_0 + a_1 x + a_2 x^2 : (a_0, a_1, a_2)^\top \in C_{0.05}\}, \max\{a_0 + a_1 x + a_2 x^2 : (a_0, a_1, a_2)^\top \in C_{0.05}\}\big]$$

for $x \in \mathbb{R}$ (outer blue curves) are depicted, where

$$C_\alpha \;:=\; \big\{\boldsymbol{\theta} \in \mathbb{R}^3 : L(\boldsymbol{\theta}) \leq n \log n + \chi^2_{3;1-\alpha}/2\big\}.$$

The lower panel of Figure 9.3 depicts the true first derivative $f'(x) = 1 - 2x$ (green line), the estimated first derivative $\widehat{f}'(x) = \widehat{a}_1 - 2\widehat{a}_2 x$ (blue line) and simultaneous 95%-confidence intervals for $f'(x)$ (outer blue curves). In particular, the 95%-confidence interval for the unique point $x_* = 0.5$ with $f'(x_*) = 0$ is equal to $[0.440, 0.572]$.

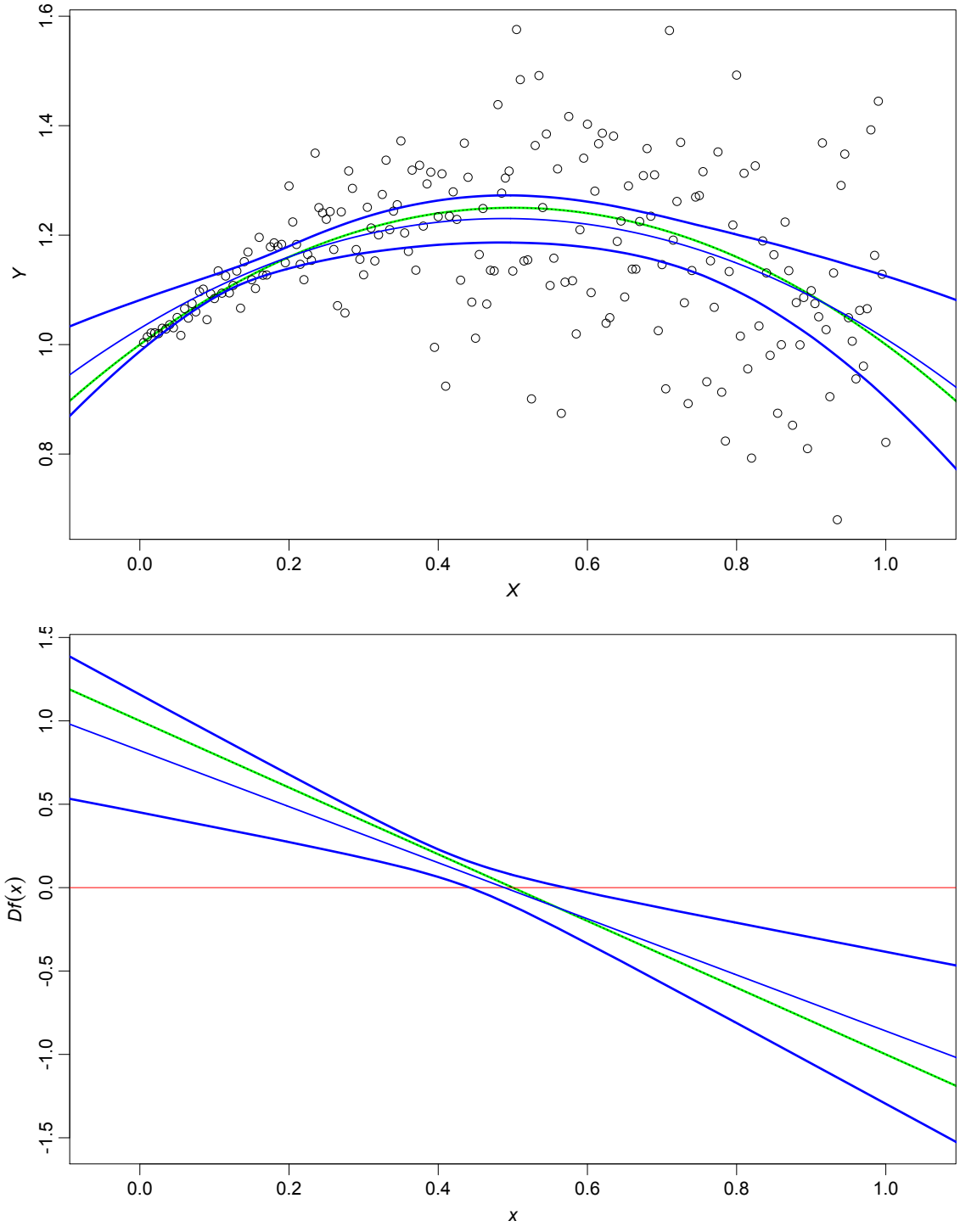

Figure 9.3: Confidence regions for quadratic regression via empirical likelihood.

Chapter 10

Isotonic Regression

In the context of checking the output of logistic regression graphically, we touched upon so-called isotonic regression. The material in this chapter provides a brief introduction to this topic. For a broader view on the field of statistical inference under qualitative constraints, we refer to the monographs of Robertson et al. (1988) and Groeneboom and Jongbloed (2014).

The general setting are observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ with values in $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R} \times \mathbb{R}$. After conditioning, if necessary, we assume that $X_1, \ldots, X_n$ are fixed values, and the Y_i are independent random variables with distributions $Q(\cdot \,|\, X_i)$. Here $\big(Q(\cdot \,|\, x)\big)_{x \in \mathcal{X}}$ is an unknown family of distributions on $\mathcal{Y}$. Now we want to estimate a real-valued certain feature $f(x)$ of $Q(\cdot \,|\, x)$. Instead of assuming that f belongs to a finite-dimensional space of functions, we only assume that f belongs to

$$\mathcal{F}_\uparrow \;:=\; \big\{ f : \mathcal{X} \to \mathbb{R} \text{ isotonic} \big\},$$

where 'isotonic' is a synonym for 'monotone increasing'.

Example 10.1 (Isotonic means). If our goal is to estimate the conditional mean function μ, given by $\mu(x) := \int y \, Q(dy \,|\, x)$, we could aim for a function $\widehat{\mu}$ in

$$\operatorname*{arg\,min}_{f \in \mathcal{F}_\uparrow} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

Example 10.2 (Isotonic quantiles). If our goal is to estimate a conditional γ-quantile function q_γ for a given $\gamma \in (0,1)$, that is,

$$Q\big((-\infty, q_\gamma(x)) \,\big|\, x\big) \;\le\; \gamma \;\le\; Q\big((-\infty, q_\gamma(x)] \,\big|\, x\big),$$

we could aim for a function $\widehat{q}_\gamma$ in

$$\operatorname*{arg\,min}_{f \in \mathcal{F}_\uparrow} \sum_{i=1}^{n} \rho_\gamma(f(X_i) - Y_i)$$

over all $f \in \mathcal{F}_\uparrow$, where $\rho_\gamma(t) := (1 - 2\gamma)t + |t|$ for $t \in \mathbb{R}$.

Example 10.3 (Isotonic binary regression). Suppose that $\mathcal{Y} = \{0, 1\}$. Then $Q(\cdot \,|\, x)$ is uniquely determined by the mean function $p(x) := Q(\{1\} \,|\, x)$, and it could be estimated by minimizer $\widehat{p}$ of

the negative log-likelihood, that is, a function in

$$\operatorname*{arg\,min}_{f\in\mathcal{F}_\uparrow\,:\,0\le f\le 1}\sum_{i=1}^{n}\big[-Y_i\log(f(X_i))-(1-Y_i)\log(1-f(X_i))\big],$$

where $\log(0) := -\infty$ and $0\cdot\log(0) := 0$.

Example 10.4 (Isotonic Poisson regression). Suppose that $\mathcal{Y} = \mathbb{N}_0$ and $Q(\cdot\,|\,x) = \mathrm{Poiss}(\lambda(x))$ for some unknown function $\lambda : \mathcal{X} \to [0,\infty)$. Again, this could be estimated by a minimizer of the negative log-likelihood, i.e. a function $\widehat{\lambda}$ in

$$\operatorname*{arg\,min}_{f\in\mathcal{F}_\uparrow\,:\,f\ge 0}\sum_{i=1}^{n}\big[f(X_i)-Y_i\log(f(X_i))\big].$$

10.1 The Pool-Adjacant-Violators Algorithm (PAVA)

In all examples we have seen so far, the goal is to find a function in the set

$$\operatorname*{arg\,min}_{f\in\mathcal{F}_\uparrow} L(f)$$

for some target function $L = L(\cdot, \boldsymbol{X}, \boldsymbol{Y})$. Typically, the set of minimizers of L over $\mathcal{F}_\uparrow$ is not a singleton. Indeed, in our examples, $L(f)$ depends only on $f(\boldsymbol{X})$. Thus we reparametrize the problem as follows: Let $x_1 < \cdots < x_m$ be the different elements of $\{X_1,\dots,X_n\}$. On the one hand, if $f \in \mathcal{F}_\uparrow$, then the vector $\boldsymbol{f} := (f(x_j))_{j=1}^m$ belongs to the convex cone

$$\mathbb{R}^m_\uparrow \;:=\; \big\{\boldsymbol{f}\in\mathbb{R}^m : f_1\le\cdots\le f_m\big\}.$$

On the other hand, for any vector $\boldsymbol{f} \in \mathbb{R}^m_\uparrow$ there exists a (non-unique) function $f \in \mathcal{F}_\uparrow$ such that $f(x_j) = f_j$ for $1 \le j \le m$.

Now we aim for a vector $\widehat{\boldsymbol{f}}$ in the set

$$\operatorname*{arg\,min}_{\boldsymbol{f}\in\mathbb{R}^m_\uparrow} L(\boldsymbol{f}),$$

where

$$L(\boldsymbol{f}) \;:=\; \sum_{j=1}^{m} L_j(f_j)$$

with certain functions $L_j = L_j(\cdot, \boldsymbol{X}, \boldsymbol{Y}) : \mathbb{R} \to (-\infty,\infty]$. Here is our general assumption on these functions L_j:

(L) For any nonvoid set $S \subset \{1,\dots,m\}$ let $L_S(t) := \sum_{j\in S} L_j(t)$. There exists a number $\xi_S \in \mathbb{R}$ such that L_S is decreasing on $(-\infty,\xi_S]$ and increasing on $[\xi_S,\infty)$.

Let us review the explicit examples we have seen before:

Example 10.1 (continued). For $f \in \mathcal{F}_\uparrow$ and $\boldsymbol{f} := (f(x_j))_{j=1}^m \in \mathbb{R}^m_\uparrow$,

$$\sum_{i=1}^n (Y_i - f(X_i))^2 \;=\; \sum_{j=1}^m \sum_{i:X_i=x_j} (Y_i - f_j)^2 \;=\; \sum_{j=1}^m w_j(f_j - \bar{y}_j)^2 + \|\boldsymbol{Y}\|^2 - \sum_{j=1}^m w_j \bar{y}_j^2,$$

where

$$w_j \;:=\; \#\{i : X_i = x_j\} \quad \text{and} \quad \bar{y}_j \;:=\; \frac{1}{w_j} \sum_{i:X_i=x_j} Y_j.$$

Thus estimating isotonic means amounts to minimizing $L(\boldsymbol{f}) = \sum_{j=1}^m L_j(f_j)$ over all $\boldsymbol{f} \in \mathbb{R}^m_\uparrow$, where $L_j(t) := w_j(t - \bar{y}_j)^2$. For any nonvoid set $S \subset \{1, \ldots, m\}$,

$$L_S(t) \;=\; w_S(t - \bar{y}_S)^2 + \sum_{j \in S} w_j \bar{y}_j^2 - w_S \bar{y}_S^2,$$

where $w_S = \sum_{j \in S} w_j$ and

$$\bar{y}_S \;=\; w_S^{-1} \sum_{j \in S} w_j \bar{y}_j \;=\; \text{sample mean of } \big(Y_i : X_i = x_j \text{ for some } j \in S\big).$$

Thus condition (L) is satisfied with $\xi_S = \bar{y}_S$.

Example 10.2 (continued). With $f \in \mathcal{F}_\uparrow$ and the corresponding $\boldsymbol{f} \in \mathbb{R}^m_\uparrow$,

$$\sum_{i=1}^n \rho_\gamma(f(X_i) - Y_i) \;=\; \sum_{j=1}^m L_j(f_j)$$

with

$$L_j(t) \;:=\; \sum_{i:X_i=x_j} \rho_\gamma(t - Y_i).$$

Condition (L) is satisfied with ξ_S being any sample γ-quantile of $\big(Y_i : X_i = x_j \text{ for some } j \in S\big)$, see Lemma 6.1.

Example 10.3 (continued). With $f \in \mathcal{F}_\uparrow$ with $0 \le f \le 1$, $\boldsymbol{f} \in \mathbb{R}^m_\uparrow$, w_j and $\bar{y}_j$ as before,

$$\sum_{i=1}^n \big[-Y_i \log(f(X_i)) - (1 - Y_i) \log(1 - f(X_i))\big] \;=\; \sum_{j=1}^m L_j(f_j)$$

with $L_j(t) := -w_j\big[\bar{y}_j \log(t) + (1 - \bar{y}_j) \log(1 - t)\big]$ for $t \in [0, 1]$ and $L_j(t) := \infty$ for $t \notin [0, 1]$. This leads to

$$L_S(t) \;=\; -w_S\big[\bar{y}_S \log(t) + (1 - \bar{y}_S) \log(1 - t)\big]$$

for $t \in [0, 1]$, and $L_S(t) = \infty$ for $t \notin [0, 1]$. Since

$$\frac{d}{dt} L_S(t) \;=\; w_S \frac{t - \bar{y}_S}{t(1 - t)}$$

for $t \in (0, 1)$, condition (L) holds true with $\xi_S = \bar{y}_S$.

Example 10.4 (continued). With $f \in \mathcal{F}_\uparrow$ with $f \geq 0$, $\boldsymbol{f} \in \mathbb{R}^m_\uparrow$, w_j and $\bar{y}_j$ as before,

$$\sum_{i=1}^{n} \big[f(X_i) - Y_i \log(f(X_i)) \big] \;=\; \sum_{j=1}^{m} L_j(f_j)$$

with $L_j(t) := w_j[t - \bar{y}_j \log(t)]$ for $t \geq 0$ and $L_j(t) := \infty$ for $t < 0$. This leads to

$$L_S(t) \;=\; w_S[t - \bar{y}_S \log(t)]$$

for $t \geq 0$ and $L_S(t) = \infty$ for $t < 0$. Here,

$$\frac{d}{dt} L_S(t) \;=\; w_S(1 - \bar{y}_S/t)$$

for $t > 0$. Thus, condition (L) holds true with $\xi_S = \bar{y}_S$.

Abstract description of the PAVA

The idea is to work with piecewise constant vectors in $\mathbb{R}^m$. Precisely, let $\mathcal{S}$ be a partition of $\{1, \ldots, m\}$ into index intervals $S = \{a, \ldots, b\}$ with $1 \leq a \leq b \leq m$. Then we consider the linear space

$$\mathbb{R}^m_{\mathcal{S}} \;:=\; \big\{ \boldsymbol{f} \in \mathbb{R}^m : f_i = f_j \text{ whenever } i, j \in S \text{ for some } S \in \mathcal{S} \big\}.$$

Note that for $\boldsymbol{f} \in \mathbb{R}^m_{\mathcal{S}}$, one can write

$$L(\boldsymbol{f}) \;=\; \sum_{S \in \mathcal{S}} L_S(f_{j(S)}),$$

where $j(S)$ denotes any index in S. In particular,

$$L(\boldsymbol{f}) \;\geq\; L(\widehat{\boldsymbol{f}}_{\mathcal{S}}),$$

where $\widehat{\boldsymbol{f}}_{\mathcal{S}} = (\widehat{f}_{\mathcal{S},j})_{j=1}^m$ is given by

$$\widehat{f}_{\mathcal{S},j} \;:=\; \xi_S \quad \text{for } j \in S, S \in \mathcal{S}.$$

Now the algorithm works as follows:

Start. We start with the finest partition $\mathcal{S} = \big\{\{1\}, \{2\}, \ldots, \{m\}\big\}$.

Induction step. Let $\mathcal{S}$ be our current partition. Suppose there exist two sets $S = \{a, \ldots, b\}$ and $T = \{b+1, \ldots, c\}$ in $\mathcal{S}$ such that $\xi_S \geq \xi_T$ ("adjacent violators"). Then we replace $\mathcal{S}$ with a coarser partition by replacing the two sets S and T with the one set $S \cup T$ (pool the adjacent violators).

Termination. If $\mathcal{S}$ satisfies $\xi_S < \xi_T$ whenever $S, T \in \mathcal{S}$ with $S < T$ element-wise, the algorithm stops and returns $\widehat{\boldsymbol{f}} := \widehat{\boldsymbol{f}}_{\mathcal{S}}$.

Note that $\#\mathcal{S}$ equals m in the beginning, and $\#\mathcal{S}$ decreases by one in each induction step. Hence, the algorithm terminates after at most $m - 1$ repetitions of the induction step. By construction, the resulting vector $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{f}}_{\mathcal{S}}$ belongs to $\mathbb{R}^m_\uparrow \cap \mathbb{R}^m_{\mathcal{S}}$. That it minimizes $L(\boldsymbol{f})$ over all $\boldsymbol{f} \in \mathbb{R}^m_\uparrow$ is not so obvious. This fact follows from the next lemma.

Lemma 10.5. *Suppose that $\mathcal{S}$ is a partition of $\{1,\dots,m\}$ such that $\xi_S \geq \xi_T$ for two sets $S = \{a,\dots,b\}$ and $T = \{b+1,\dots,c\}$ in $\mathcal{S}$. Let $\tilde{\mathcal{S}}$ be the partition resulting from replacing S and T with $S \cup T$. Then, for any vector $\boldsymbol{f} \in \mathbb{R}^m_\uparrow \cap \mathbb{R}^m_{\mathcal{S}}$, there exists a vector $\tilde{\boldsymbol{f}} \in \mathbb{R}^m_\uparrow \cap \mathbb{R}^m_{\tilde{\mathcal{S}}}$ such that $L(\tilde{\boldsymbol{f}}) \leq L(\boldsymbol{f})$.*

This lemma shows that the minimum of $L(\boldsymbol{f})$ over all $\boldsymbol{f} \in \mathbb{R}^m_\uparrow \cap \mathbb{R}^m_{\mathcal{S}}$ remains unchanged whenever we pool two adjacent violators in $\mathcal{S}$. We start with the finest partition $\mathcal{S} = \big\{\{1\},\dots,\{m\}\big\}$, so $\mathbb{R}^m_\uparrow \cap \mathbb{R}^m_{\mathcal{S}} = \mathbb{R}^m_\uparrow$, and we continue this pooling until eventually the minimizer $\widehat{\boldsymbol{f}}_{\mathcal{S}}$ over all $\boldsymbol{f} \in \mathbb{R}^m_{\mathcal{S}}$ belongs also to $\mathbb{R}^m_\uparrow$. This implies that the final vector $\widehat{\boldsymbol{f}}_{\mathcal{S}}$ minimizes $L(\boldsymbol{f})$ among all $\boldsymbol{f} \in \mathbb{R}^m_\uparrow$.

Proof of Lemma 10.5. Let $\boldsymbol{f} \in \mathbb{R}^m_\uparrow \cap \mathbb{R}^m_{\mathcal{S}}$. Now let $\tilde{\boldsymbol{f}}$ be given by

$$\tilde{f}_j \;:=\; \begin{cases} f_j & \text{for } j \in \{1,\dots,m\} \setminus \{a,\dots,c\}, \\ \theta & \text{for } j \in \{a,\dots,c\}, \end{cases}$$

where $f_a \leq \theta \leq f_b$. Then, $\tilde{\boldsymbol{f}} \in \mathbb{R}^m_\uparrow \cap \mathbb{R}^m_{\tilde{\mathcal{S}}}$, and

$$L(\boldsymbol{f}) - L(\tilde{\boldsymbol{f}}) \;=\; L_S(f_a) - L_S(\theta) + L_T(f_c) - L_T(\theta).$$

If $f_a \leq \xi_T$, then $\xi_T \leq \xi_S$ implies that L_S is decreasing on $[f_a, \min(\xi_T, f_c)]$, and L_T is increasing on $[\min(f_c, \xi_T), f_c]$. Thus, choosing $\theta = \min(f_c, \xi_T)$ leads to a vector $\tilde{\boldsymbol{f}}$ such that $L(\tilde{\boldsymbol{f}}) \leq L(\boldsymbol{f})$.

If $f_a \geq \xi_T$, then L_T is increasing on $[f_a, f_c]$. Hence, the choice $\theta = f_a$ leads to a vector $\tilde{\boldsymbol{f}}$ such that $L(\tilde{\boldsymbol{f}}) \leq L(\boldsymbol{f})$. □

Refinement. Suppose that the minimizers ξ_S, $\emptyset \neq S \subset \{1,\dots,m\}$, fulfill the *Cauchy mean value condition*. That is, for arbitrary disjoint nonvoid sets $S, T \subset \{1,\dots,m\}$,

$$\min\{\xi_S, \xi_T\} \;\leq\; \xi_{S\cup T} \;\leq\; \max\{\xi_S, \xi_T\}.$$

This condition is satisfied, for instance, if the minimizer ξ_S of L_S is unique for any nonvoid set S. It is also satisfied if each function L_j is convex and lower semicontinuous with $L_j(t) \to \infty$ as $|t| \to \infty$, and if ξ_S is always the smallest (or always the largest) minimizer of L_S.

Here one can start the PAVA with the partition of all maximal index intervals S on which $j \mapsto \xi_{\{j\}}$ is non-increasing, because the singletons $\{j\}$, $j \in S$, can be pooled initially.

Explicit versions of the PAVA

The general description of the PAVA is rather vague about where to look for adjacent violators. Here is a more explicit description. Instead of the partition $\mathcal{S}$ we consider a tuple $\boldsymbol{s} = (s_1,\dots,s_m)$ and an integer variable $b \in \{1,\dots,m\}$ such that $0 < s_1 < \cdots < s_b \leq m$. With $s_0 := 0$, this tuple $\boldsymbol{s}$ corresponds to the partition $\mathcal{S}$ consisting of the sets $S_a := \{s_{a-1}+1,\dots,s_a\}$, $1 \leq a \leq b$, and the singletons $\{j\}$, $s_b < j \leq m$. Moreover, let $\boldsymbol{g} = (g_1,\dots,g_m)$ be a tuple such that $g_a = \xi_{S_a}$ for $1 \leq a \leq b$. Note that the components s_j and g_j are irrelevant for $b < j \leq m$, but it is more economical to initialize the vectors $\boldsymbol{s}$ and $\boldsymbol{g}$ just once with a given length m and to keep track

```
b ← 1
s ← (1)_{j=1}^m                          % start with S_1 = {1}
g ← (ξ_{{1}})_{j=1}^m
for k ← 2 to m do
    b ← b + 1
    s_b ← k                              % add new set {k}
    g_b ← ξ_{{k}}
    while b > 1 and g_b ≤ g_{b−1} do
        s_{b−1} ← k                      % pool S_{b−1} and S_b
        g_{b−1} ← ξ_{S_{b−1} ∪ S_b}
        b ← b − 1
    end while
end for
f̂ ← (0)_{j=1}^m                          % initialize f̂ and
for a ← 1 to b do                        % assign values to it
    for j ← s_{a−1} + 1 to s_a do
        f̂_j ← g_a
    end for
end for
```

Table 10.1: Schematic pseudo-code for the PAVA.

of the number b of relevant entries. Table 10.1 contains pseudo-code for the PAVA in terms of these auxiliary objects. Each run of the inner while-loop corrresponds to a pooling of two adjacent violators.

Specifically, if $\xi_S = \bar{y}_S$ as in Examples 10.1, 10.3 and 10.4, there should be another auxiliary vector $\boldsymbol{v} = (v_1, \ldots, v_m)$ such that $v_a = w_{S_a}$ for $1 \le a \le b$. Moreover, here one can use the refined version of the PAVA and start with the partition consisting of all maximal index intervals S on which $j \mapsto \bar{y}_j$ is non-increasing. Then the PAVA reads as in Table 10.2. Note that increasing b and generating the new set S_b requires only $O(\#S_b)$ operations, whereas the pooling of two adjacent violators requires only a fixed number of operations. Consequently, the total running time of the PAVA is of order $O(m)$.

Example 10.6. Figure 10.1 shows a simulated data set with $n = 200$ pairs (X_i, Y_i). In addition to these data points, one sees an isotonic least squares fit $\widehat{\mu}$ as a blue line. It is obtained from $\widehat{\boldsymbol{f}}$ by linear interpolation on $[x_j, x_{j+1}]$, $1 \le j < m$, and constant extrapolation on $(-\infty, x_1]$ and on $[x_m, \infty)$. The data have been generated as $Y_i = \mu(X_i) + \varepsilon_i$ with independent standard Gaussian errors ε_i, and $\mu(x) = \min(\max(x, -1), 1)$. The latter function μ is depicted as a green and dotted line too.

10.2 Further Considerations for Isotonic Least Squares

We consider the minimization of $L(\boldsymbol{f}) = \sum_{j=1}^m L_j(f_j)$ over all $\boldsymbol{f} \in \mathbb{R}^m_{\uparrow}$, where $L_j(t) = w_j(t - \bar{y}_j)^2$. Here the mapping $L : \mathbb{R}^m \to \mathbb{R}$ is strictly convex, so the PAVA yields the unique minimizer

$k \leftarrow 0$
$b \leftarrow 0$
$\boldsymbol{s} \leftarrow (0)_{j=1}^{m}$
$\boldsymbol{g} \leftarrow (0)_{j=1}^{m}$
$\boldsymbol{v} \leftarrow (0)_{j=1}^{m}$
while $k < m$ do
    $k \leftarrow k + 1$
    $v_{\text{new}} \leftarrow w_k$ % start a new set with $\{k\}$
    $G_{\text{new}} \leftarrow w_k \bar{y}_k$
    while $k < m$ and $\bar{y}_{k+1} \leq \bar{y}_k$ do
        $k \leftarrow k + 1$
        $v_{\text{new}} \leftarrow v_{\text{new}} + w_k$ % extend the new set
        $G_{\text{new}} \leftarrow G_{\text{new}} + w_k \bar{y}_k$
    end while
    $b \leftarrow b + 1$
    $s_b \leftarrow k$ % add the new set to $\mathcal{S}$
    $g_b \leftarrow G_{\text{new}} / v_{\text{new}}$
    $v_b \leftarrow v_{\text{new}}$
    while $b > 1$ and $g_b \leq g_{b-1}$ do
        $s_{b-1} \leftarrow k$ % pool S_{b-1} and S_b
        $v_{\text{temp}} \leftarrow v_{b-1} + v_b$
        $g_{b-1} \leftarrow (v_{b-1} g_{b-1} + v_b g_b) / v_{\text{temp}}$
        $v_{b-1} \leftarrow v_{\text{temp}}$
        $b \leftarrow b - 1$
    end while
end for
$\widehat{\boldsymbol{f}} \leftarrow (0)_{j=1}^{m}$ % initialize $\widehat{\boldsymbol{f}}$ and
for $a \leftarrow 1$ to b do % assign values to it
    for $j \leftarrow s_{a-1} + 1$ to s_a do
        $\widehat{f}_j \leftarrow g_a$
    end for
end for

Table 10.2: Pseudo-code for the PAVA with weighted averages.

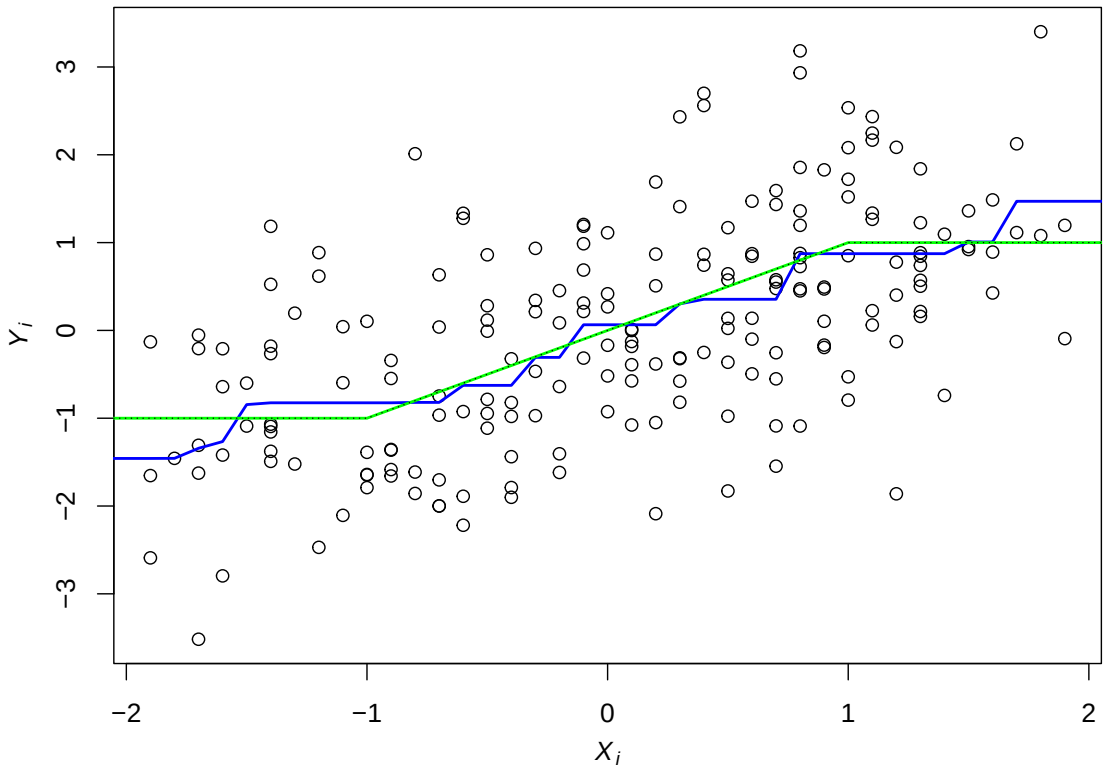

Figure 10.1: Example for isotonic means based on $n = 200$ simulated data.

$\widehat{\boldsymbol{f}}$ of L over $\mathbb{R}^m_{\uparrow}$. In this section we derive various properties and characterizations of $\widehat{\boldsymbol{f}}$.

For notational convenience, for indices $1 \le a \le b \le m$ we use the subscript 'a, b' instead of '$\{a, \ldots, b\}$', so $w_{a,b} = \sum_{j=a}^{b} w_j$, and $\bar{y}_{a,b} = w_{a,b}^{-1} \sum_{j=a}^{b} w_j \bar{y}_j$. In addition we consider the local weighted average

$$\widehat{f}_{a,b} \;:=\; w_{a,b}^{-1} \sum_{j=a}^{b} w_j \widehat{f}_j$$

of $\widehat{\boldsymbol{f}}$. Note that by isotonicity of $j \mapsto \widehat{f}_j$,

$$\widehat{f}_a \;\le\; \widehat{f}_{a,b} \;\le\; \widehat{f}_b.$$

Lemma 10.7 (Inequalities for local weighted averages). *For arbitrary indices $1 \le k \le \ell \le m$,*

$$\widehat{f}_{k,\ell} \;\ge\; \bar{y}_{k,\ell} \quad \textit{if } \widehat{f}_\ell < \widehat{f}_{\ell+1}, \tag{10.1}$$

$$\widehat{f}_{k,\ell} \;\le\; \bar{y}_{k,\ell} \quad \textit{if } \widehat{f}_{k-1} < \widehat{f}_k, \tag{10.2}$$

where $\widehat{f}_0 := -\infty$ and $\widehat{f}_{m+1} := \infty$.

Proof. For any $t \in \mathbb{R}$ let $\boldsymbol{f}(t)$ be given by

$$f_j(t) \;:=\; \widehat{f}_j + t\, 1_{[k \le j \le \ell]}.$$

Note that $\boldsymbol{f}(0) = \widehat{\boldsymbol{f}}$. If $\widehat{f}_\ell < \widehat{f}_{\ell+1}$, then $\boldsymbol{f}(t) \in \mathbb{R}^m_\uparrow$ whenever $0 \le t \le \widehat{f}_{\ell+1} - \widehat{f}_\ell$. Consequently, optimality of $\widehat{\boldsymbol{f}}$ implies that

$$0 \;\le\; \frac{d}{dt}\Big|_{t=0} L(\boldsymbol{f}(t)) \;=\; 2\sum_{j=k}^{\ell} w_j(\widehat{f}_j - \bar{y}_j) \;=\; 2w_{k,\ell}(\widehat{f}_{k,\ell} - \bar{y}_{k,\ell}).$$

If $\widehat{f}_{k-1} < \widehat{f}_k$, then $\boldsymbol{f}(t) \in \mathbb{R}^m_\uparrow$ whenever $\widehat{f}_{k-1} - \widehat{f}_k \le t \le 0$. Consequently,

$$0 \;\ge\; \frac{d}{dt}\Big|_{t=0} L(\boldsymbol{f}(t)) \;=\; 2w_{k,\ell}(\widehat{f}_{k,\ell} - \bar{y}_{k,\ell}).$$

□

The inequalities in Lemma 10.7 lead to an explicit representation of $\widehat{\boldsymbol{f}}$ which is useless algorithmically but very useful for geometrical and theoretical considerations.

Lemma 10.8 (Min-max and max-min formulae). *For any* $k \in \{1, \dots, m\}$,

$$\widehat{f}_k \;=\; \max_{a\le k} \min_{b\ge k} \bar{y}_{a,b} \;=\; \min_{b\ge k} \max_{a\le k} \bar{y}_{a,b}.$$

Proof. To verify these formulae, let $1 \le a(k) \le k \le b(k) \le m$ such that $\widehat{f}_{a(k)-1} < \widehat{f}_{a(k)} = \widehat{f}_{b(k)} < \widehat{f}_{b(k)+1}$.

Proof of the max-min formula. If $a \le k$ and $b(k) < b \le m$, then

$$\bar{y}_{a,b} \;=\; \frac{w_{a,b(k)}\bar{y}_{a,b(k)} + w_{b(k)+1,b}\bar{y}_{b(k)+1,b}}{w_{a,b}} \;>\; \bar{y}_{a,b(k)},$$

because

$$\bar{y}_{a,b(k)} \;\le\; \widehat{f}_{a,b(k)} \;\le\; \widehat{f}_{b(k)} \;=\; \widehat{f}_k \quad\text{and}\quad \bar{y}_{b(k)+1,b} \;\ge\; \widehat{f}_{b(k)+1,b} \;\ge\; \widehat{f}_{b(k)+1} \;>\; \widehat{f}_k,$$

by (10.1) and (10.2), respectively. Consequently,

$$\max_{a\le b} \min_{b\ge k} \bar{y}_{a,b} \;=\; \max_{a\le b} \min_{b=k,\dots,b(k)} \bar{y}_{a,b} \;\le\; \max_{a\le b} \bar{y}_{a,b(k)} \;\le\; \widehat{f}_k,$$

and a second application of (10.2) leads to

$$\max_{a\le b} \min_{b\ge k} \bar{y}_{a,b} \;=\; \max_{a\le b} \min_{b=k,\dots,b(k)} \bar{y}_{a,b} \;\ge\; \min_{b=k,\dots,b(k)} \bar{y}_{a(k),b} \;\ge\; \min_{b=k,\dots,b(k)} \widehat{f}_{a(k),b} \;=\; \widehat{f}_k.$$

Proof of the min-max formula. In principle, this formula follows from the max-min formula after replacing $(w_j, \bar{y}_j)$ with $(w_{m+1-j}, -\bar{y}_{m+1-j})$ for $1 \le j \le m$ and noting that the resulting vector $\widehat{\boldsymbol{f}}$ for the modified data corresponds to the vector $(-\widehat{f}_{m+1-j})_{j=1}^m$ for the original data. But here is a direct argument: If $1 \le a < a(k)$ and $b \ge k$, then

$$\bar{y}_{a,b} \;=\; \frac{w_{a,a(k)-1}\bar{y}_{a,a(k)-1} + w_{a(k),b}\bar{y}_{a(k),b}}{w_{a,b}} \;<\; \bar{y}_{a(k),b},$$

because

$$\bar{y}_{a,a(k)-1} \;\le\; \widehat{f}_{a,a(k)-1} \;\le\; \widehat{f}_{a(k)-1} \;<\; \widehat{f}_k \quad\text{and}\quad \bar{y}_{a(k),b} \;\ge\; \widehat{f}_{a(k),b} \;\ge\; \widehat{f}_{a(k)} \;=\; \widehat{f}_k,$$

by (10.1) and (10.2), respectively. Consequently,

$$\min_{b\ge k}\max_{a\le k}\bar{y}_{a,b} \;=\; \min_{b\ge k}\max_{a=a(k),\dots,k}\bar{y}_{a,b} \;\ge\; \min_{b\le k}\bar{y}_{a(k),b} \;\ge\; \widehat{f}_k,$$

and a second application of (10.1) leads to

$$\min_{b\ge k}\max_{a\le b}\bar{y}_{a,b} \;=\; \min_{b\ge k}\max_{a=a(k),\dots,k}\bar{y}_{a,b} \;\le\; \max_{a=a(k),\dots,k}\bar{y}_{a,b(k)} \;\le\; \max_{a=a(k),\dots,k}\widehat{f}_{a,b(k)} \;=\; \widehat{f}_k.$$

□

Our final result provides a characterization of $\widehat{\boldsymbol{f}}$ in terms of weighted partial sum functions. Let

$$\mathcal{T} \;=\; \{t_0, t_1, \dots, t_m\}$$

with $t_0 := 0$ and $t_k := \sum_{j=1}^{k} w_j = w_{1,k}$ for $1 \le k \le m$. Now we define weighted partial sum functions $S, \widehat{F} : \mathcal{T} \to \mathbb{R}$ of $\boldsymbol{y}$ and $\widehat{\boldsymbol{f}}$, respectively, via $S(0) := 0$, $\widehat{F}(0) := 0$ and

$$\begin{aligned} S(t_k) \;&:=\; \sum_{j=1}^{k} w_j\bar{y}_j \;=\; w_{1,k}\bar{y}_{1,k}, \\ \widehat{F}(t_k) \;&:=\; \sum_{j=1}^{k} w_j\widehat{f}_j \;=\; w_{1,k}\widehat{f}_{1,k}. \end{aligned}$$

Lemma 10.9 (Greatest convex minorant). *The function $\widehat{F}$ is a convex function on $\mathcal{T}$ such that $\widehat{F} \le S$. For any convex function F on $\mathcal{T}$, $F \le S$ implies that $F \le \widehat{F}$.*

Proof. By construction of $\widehat{F}$, the slope

$$\frac{\widehat{F}(t_k) - \widehat{F}(t_{k-1})}{t_k - t_{k-1}} \;=\; \widehat{f}_k$$

is isotonic in $k \in \{1, \dots, m\}$, whence $\widehat{F}$ is convex. Note also that $\widehat{F}(0) = S(0) = 0$, and (10.2) implies that $\widehat{F}(t_k) \le S(t_k)$ for $k = 1, \dots, m$. Moreover, combining this with (10.1) shows that $\widehat{F}(t_k) = S(t_k)$ whenever $\widehat{f}_k < \widehat{f}_{k+1}$.

For fixed $k \in \{1, \dots, m\}$, let $0 \le \alpha(k) < k \le \beta(k) \le m$ such that $\widehat{f}_{\alpha(k)} < \widehat{f}_{\alpha(k)+1} = \widehat{f}_{\beta(k)} < \widehat{f}_{\beta(k)+1}$. Since $\widehat{f}_j = \widehat{f}_k$ for $\alpha(k) < j \le \beta(k)$ and $\widehat{F}(t_{\alpha(k)}) = S(t_{\alpha(k)})$, $\widehat{F}(t_{\beta(k)}) = S(t_{\beta(k)})$,

$$\begin{aligned} \widehat{F}(t_k) \;&=\; \widehat{F}(t_{\alpha(k)}) + (t_k - t_{\alpha(k)})\widehat{f}_k \\ &=\; \widehat{F}(t_{\alpha(k)}) + (t_k - t_{\alpha(k)})\frac{\widehat{F}(t_{\beta(k)}) - \widehat{F}(t_{\alpha(k)})}{t_{\beta(k)} - t_{\alpha(k)}} \\ &=\; (1-\lambda_k)S(t_{\alpha(k)}) + \lambda_k S(t_{\beta(k)}), \end{aligned}$$

where $\lambda_k := (t_k - t_{\alpha(k)})/(t_{\beta(k)} - t_{\alpha(k)}) \in (0, 1]$. But for a convex function $F : \mathcal{T} \to \mathbb{R}$ with $F \le S$,

$$\begin{aligned} F(t_k) \;&=\; F\big((1-\lambda_k)t_{\alpha(k)} + \lambda_k t_{\beta(k)}\big) \\ &\le\; (1-\lambda_k)F(t_{\alpha(k)}) + \lambda_k F(t_{\beta(k)}) \\ &\le\; (1-\lambda_k)S(t_{\alpha(k)}) + \lambda_k S(t_{\beta(k)}) \;=\; \widehat{F}(t_k). \end{aligned}$$

□

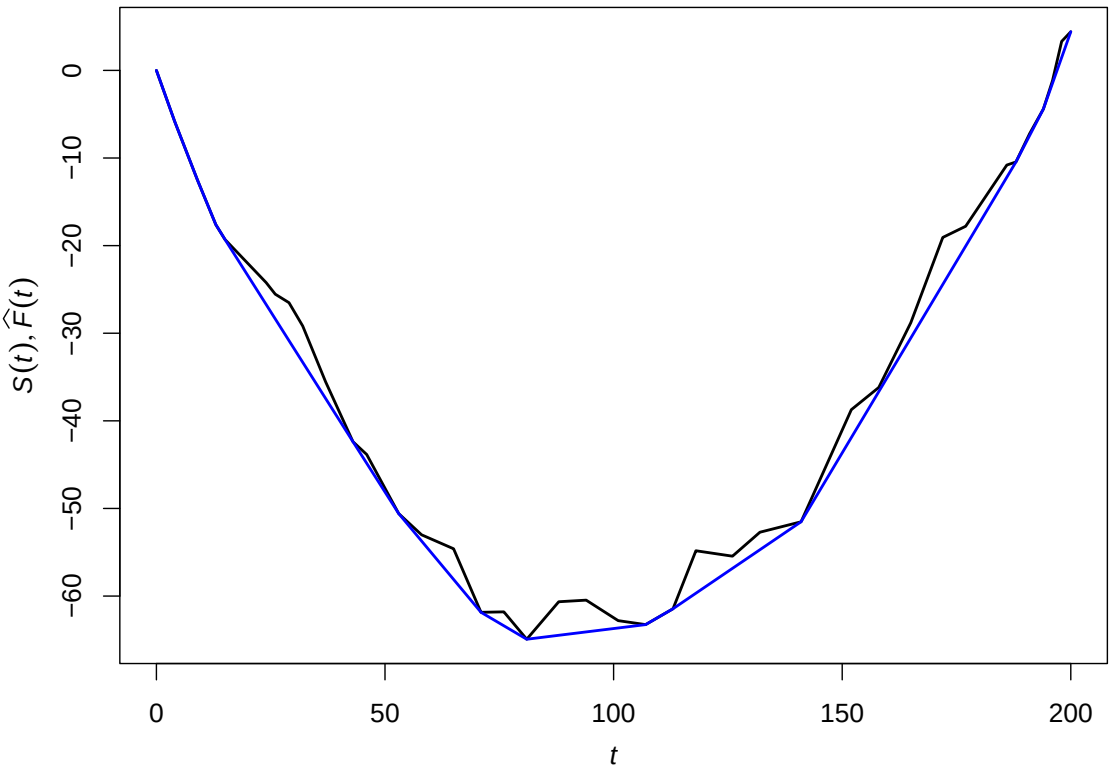

Figure 10.2: Partial sum functions for the data in Figure 10.1.

Example 10.6 (continued). Figure 10.2 shows for the data in Figure 10.1 the corresponding partial sum function $t \mapsto S(t)$ as a black line and the function $t \mapsto \widehat{F}(t)$ as a blue line. (We used linear interpolation between neighboring points t_{k-1} and t_k, $1 \le k \le m$.)

Consistency. The min-max and max-min formulae for $\widehat{f}$ imply consistency properties of $\widehat{f}$. We present two such results, where for simplicity we restrict our attention to the special where that $X_i = i/n$. Moreover, we assume that

$$Y_i \;=\; \mu(i/n) + \varepsilon_i$$

with independent random errors $\varepsilon_1, \ldots, \varepsilon_n$ such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) \le \sigma^2$ for some fixed $\sigma > 0$.

Theorem 10.10. *Let $\widehat{\mu} \in \mathcal{F}_\uparrow$ minimize $L(f) = \sum_{i=1}^n (Y_i - f(X_i))^2$ over all $f \in \mathcal{F}_\uparrow$.*

(a) *Suppose that the true mean function μ is Lipschitz-continuous on some nondegenerate interval $[z_0, z_1] \subset [0,1]$. Then for each fixed $x \in (z_0, z_1)$,*

$$\left|\widehat{\mu}(x) - \mu(x)\right| \;=\; O_p(n^{-1/3}).$$

Suppose that the true mean function μ is constant on some nondegenerate interval $[z_0, z_1] \subset [0,1]$.

Then for any fixed $\delta \in (0, z_1 - z_0)$,

$$\sup_{x\in[z_0,z_1-\delta]} \big(\widehat{\mu}(x) - \mu(x)\big)^+ \;=\; O_p(n^{-1/2}),$$
$$\sup_{x\in[z_0+\delta,z_1]} \big(\mu(x) - \widehat{\mu}(x)\big)^+ \;=\; O_p(n^{-1/2}).$$

Proof of Theorem 10.10. Note that $m = n$ and $x_j = X_j$, $w_{jk} = k + 1 - j$ for $1 \le j \le k \le n$. In addition to the averages $\bar{y}_{jk}$ we use the averages $\bar{\varepsilon}_{jk} := (k + 1 - j)^{-1} \sum_{i=j}^{k} \varepsilon_i$.

Suppose that μ is Lipschitz-continuous on $[z_0, z_1]$ with Lipschitz constant L. With $\delta_n := n^{-1/3}$, for sufficiently large n, $[x - \delta_n, x] \subset [z_0, x]$ and $\{i : X_i \in [x - \delta_n, x]\}$ is nonvoid. Precisely, the indices $j_n := \min\{i : X_i \ge x - n^{-1/3}\}$ and $k_n := \max\{i : X_i \le x\}$ satisfy

$$k_n + 1 - j_n \;=\; n\delta_n(1 + o(1)) \;=\; n^{2/3}(1 + o(1)).$$

According to the max-min formula,

$$\widehat{\mu}(x) \;\ge\; \widehat{f}_{k_n} \;=\; \max_{a\le k_n} \min_{b\ge k_n} \bar{y}_{ab} \;\ge\; \min_{b\ge k_n} \bar{y}_{j_n b}.$$

But with the Lipschitz constant L of μ on $[z_0, z_1]$,

$$\bar{y}_{j_n b} \;=\; (b + 1 - j_n) \sum_{i=j_n}^{b} (\mu(i/n) + \varepsilon_i) \;\ge\; \mu(x) - L\delta_n + \bar{\varepsilon}_{j_n b},$$

whence

$$\mu(x) - \widehat{\mu}(x) \;\le\; L\delta_n + \max_{b\ge k_n} |\bar{\varepsilon}_{j_n b}|.$$

Now, a well-known inequality of Kolmogorov states that for any $\eta > 0$,

$$\mathbb{P}\Big(\max_{b\ge k_n} |\bar{\varepsilon}_{j_n b}| \ge \eta\Big) \;\le\; \frac{4\sigma^2}{\eta^2(k_n + 1 - j_n)}.$$

Consequently, for any fixed $C > 0$,

$$\mathbb{P}\big(\mu(x) - \widehat{\mu}(x) \ge (L + C)\delta_n\big) \;\le\; \frac{4\sigma^2}{C^2\delta_n^2(k_n + 1 - j_n)} \;=\; \frac{4\sigma(1 + o(1))}{C^2},$$

because $\delta_n^2(k_n + 1 - j_n) = 1 + o(1)$. This shows that $\big(\mu(x) - \widehat{\mu}(x)\big)^+ = O_p(n^{-1/3})$. Analogously one can show that $\big(\widehat{\mu}(x) - \mu(x)\big)^+ = O_p(n^{-1/3})$. This proves part (a).

Now suppose that μ is constant on $[z_0, z_1]$. For sufficiently large n, the index set $\{i : X_i \in [z_0, z_0 + \delta]\}$ is nonvoid. Precisely, $j_n := \min\{i : X_i \ge z_0\}$ and $k_n := \max\{i : X_i \le z_0 + \delta\}$ satisfy

$$k_n + 1 - j_n \;=\; n\delta(1 + o(1)).$$

Now, since $\mu \equiv \mu(z_0)$ and $\widehat{\mu}$ is isotonic on $[z_0, z_1]$,

$$\sup_{x\in[z_0+\delta,z_1]} \big(\mu(x) - \widehat{\mu}(x)\big) \;\le\; \big(\mu(z_0) - \widehat{f}_{k_n}\big),$$

and by means of the max-min formula,

$$\widehat{f}_{k_n} \;\geq\; \min_{b\geq k_n} \bar{y}_{j_n b} \;\geq\; \mu(z_0) + \min_{b\geq k} \bar{\varepsilon}_{j_n b}.$$

Another application of Kolmogorov's inequality leads to

$$\begin{aligned}&\mathbb{P}\Big(\sup_{x\in[z_0+\delta,z_1]} \big(\mu(x) - \widehat{\mu}(x)\big) \geq Cn^{-1/2}\Big) \\ &\quad\leq\; \mathbb{P}\Big(\max_{b\geq k_n} |\bar{\varepsilon}_{j_n b}| \geq\; Cn^{-1/2}\Big) \;\leq\; \frac{4\sigma^2}{C^2 n^{-1}(k_n+1-j_n)} \;=\; \frac{4\sigma^2(1+o(1))}{C^2\delta}.\end{aligned}$$

Consequently, $\sup_{x\in[z_0+\delta,z_1]} \big(\mu(x) - \widehat{\mu}(x)\big)^+$ is of order $O_p(n^{-1/2})$. Analogously one can show that $\sup_{x\in[z_0,z_1-\delta]} \big(\widehat{\mu}(x) - \mu(x)\big)^+$ is of order $O_p(n^{-1/2})$. This concludes the proof of part (b). □

10.3 Isotonic Distributional Regression

The methods mentioned in this section are described an analyzed in detail by Mösching and Dümbgen (2020). Suppose we are interested in the complete distributions $Q(\cdot \,|\, x)$, $x \in \mathcal{X}$, rather than just a specific feature of them. Equivalently, we would like to estimate the entire distribution function $G(\cdot \,|\, x)$ or quantile function $G^{-1}(\cdot \,|\, x)$ of $Q(\cdot \,|\, x)$ for each $x \in \mathcal{X}$. This is possible under the following assumption:

(SO) The mapping $x \mapsto Q(\cdot \,|\, x)$ is isotonic on $\mathcal{X}$ with respect to stochastic order.

This constraint can be reformulated in two equivalent ways:

(SO') For any fixed $y \in \mathbb{R}$, $x \mapsto G(y \,|\, x)$ is antitonic (monotone decreasing) on $\mathcal{X}$;

(SO") For any fixed $\gamma \in (0,1)$, $x \mapsto G^{-1}(\gamma \,|\, x)$ is isotonic on $\mathcal{X}$.

In view of (SO") one could think about estimating $x \mapsto G^{-1}(\gamma \,|\, x)$ for any γ in a sufficiently large finite subset of $(0,1)$ via a suitable variant of the PAVA. But characterization (SO') leads to a more elegant solution: Since $G(y_o \,|\, x) = \int 1_{[y\leq y_o]}\, Q(dy \,|\, x)$, we estimate $G(y_o \,|\, x)$ by $\widehat{G}(y_o \,|\, x) := \widehat{f}(x)$, where $\widehat{f} = \widehat{f}_{y_o}$ is a function in

$$\operatorname*{arg\,min}_{f\in\mathcal{F}_\downarrow} \sum_{i=1}^n \big(1_{[Y_i\leq y_o]} - f(X_i)\big)^2,$$

and $\mathcal{F}_\downarrow$ denotes the family of all antitonic functions on $\mathcal{X}$. The good news is that it suffices to consider all $y_o \in \{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n-1)}\}$, where $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ are the order statistics of $Y_1, \ldots, Y_n$. Hence, the estimation of $G(\cdot \,|\, \cdot)$ amounts to applying a suitable version of the PAVA at most $n-1$ times. Thus, the running time of the whole procedure is only $O(n^2)$. For further details and refinements, we refer also to Henzi et al. (2022).

Example 10.11. In a survey of UK households in 1973, the annual income (X) and the annual expenditures for various commodities such as food or housing (Y) have been determined several thousand households. Figures 10.3 and 10.4 depict the log-transformed observation pairs $(\log_{10}(X_i), \log_{10}(Y_i))$ and estimated γ-quantile curves for $\gamma \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$, resulting from the isotonic estimators $\widehat{G}(\cdot \,|\, x)$. (The observations with the five smallest or largest X- or Y-values are not shown.)

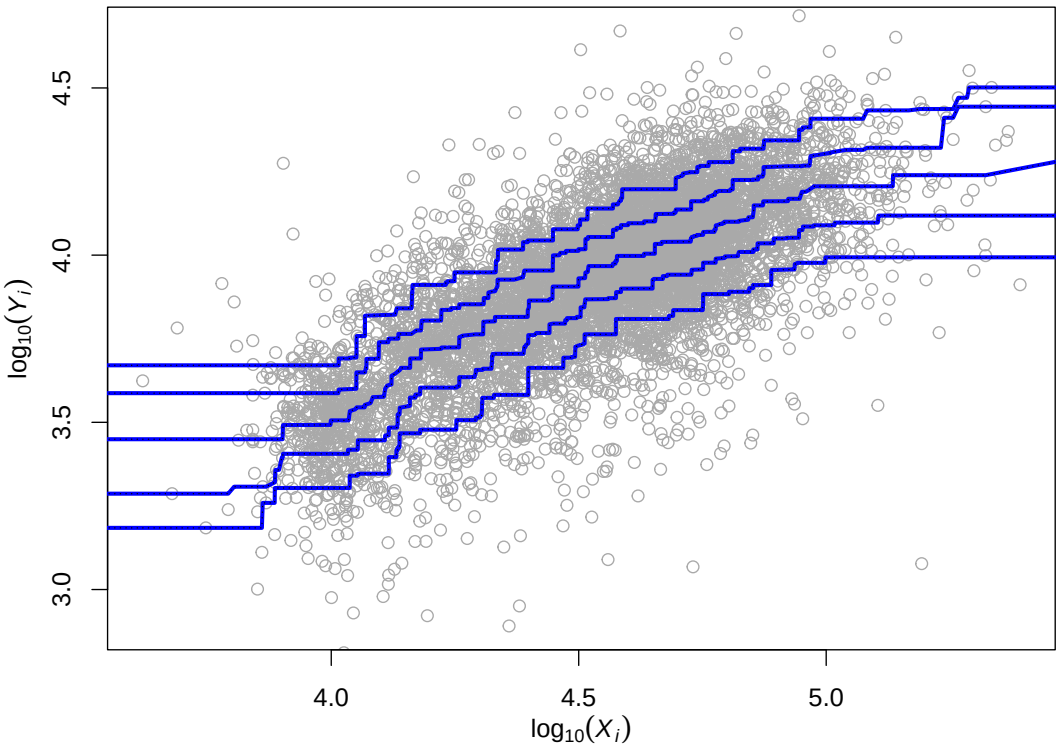

Figure 10.3: Log-transformed annual incomes (X) and expenditures for food (Y) for $n = 7125$ UK households in 1973.

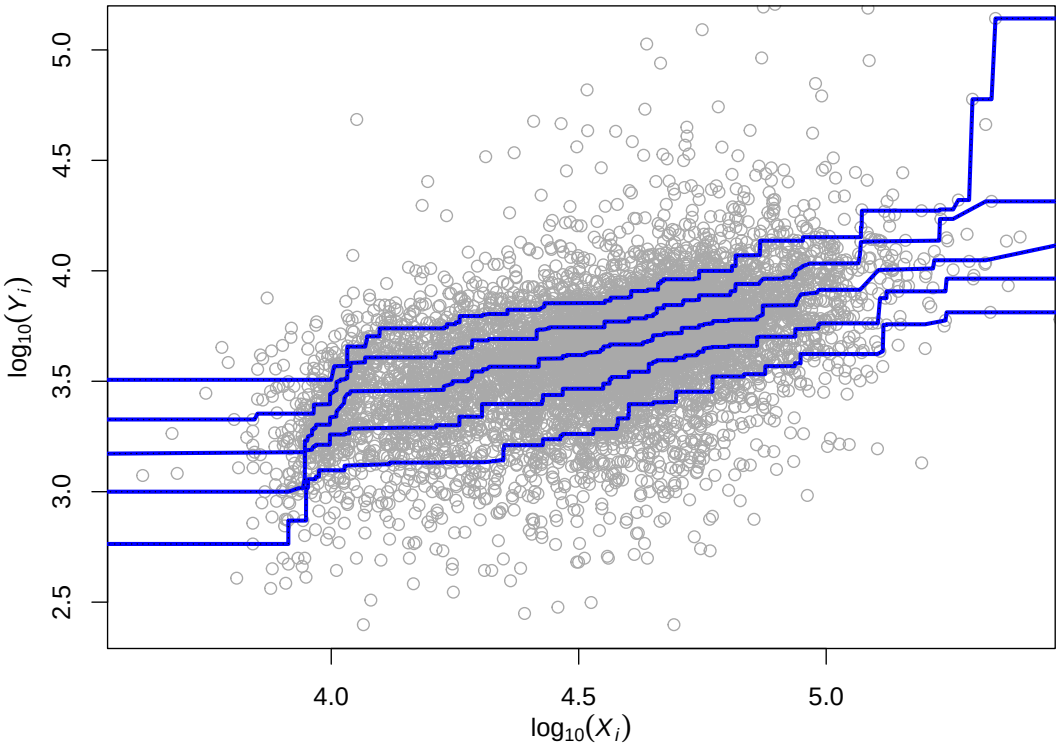

Figure 10.4: Log-transformed annual incomes (X) and expenditures for housing (Y) for $n = 7110$ UK households in 1973.

Appendix A

Miscellaneous Auxiliary Results

A.1 The QR Decomposition

Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ such that its first $\min(n,p)$ columns are linearly independent. Then there exists an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ and an upper triangular matrix $\boldsymbol{R} \in \mathbb{R}^{n\times p}$ with nonzero diagonal entries such that

$$\boldsymbol{A} \;=\; \boldsymbol{Q}\boldsymbol{R}.$$

The conditions on $\boldsymbol{Q}$ and $\boldsymbol{R}$ mean that

$$\boldsymbol{Q}^\top\boldsymbol{Q} \;=\; \boldsymbol{I}_n \quad \text{and} \quad R_{ij} \begin{cases} \neq\; 0 & \text{if } i = j, \\ =\; 0 & \text{if } i > j. \end{cases}$$

In case of $n \geq p$, one could also write

$$\boldsymbol{A} \;=\; \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{R}} \tag{A.1}$$

with $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{n\times p}$ containing the first p columns of $\boldsymbol{Q}$ and $\tilde{\boldsymbol{R}} \in \mathbb{R}^{p\times p}$ consisting of the first p rows of $\boldsymbol{R}$. Then, $\tilde{\boldsymbol{Q}}^\top\tilde{\boldsymbol{Q}} = \boldsymbol{I}_p$, meaning that the columns of $\tilde{\boldsymbol{Q}}$ are orthonormal vectors in $\mathbb{R}^n$, and the square matrix $\tilde{\boldsymbol{R}}$ is upper triangular with nonzero diagonal terms, too. In this setting, suppose we want to compute for $\boldsymbol{y} \in \mathbb{R}^n$ the vector

$$\boldsymbol{x} \;:=\; \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{v}\| \;=\; (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y}.$$

Then, the reduced QR decomposition (A.1) of $\boldsymbol{A}$ yields the equivalent representation

$$\boldsymbol{x} \;=\; \tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{Q}}^\top\boldsymbol{y}.$$

Hence, computing $\boldsymbol{x}$ amounts to solving the linear equation system

$$\tilde{\boldsymbol{R}}\boldsymbol{x} \;=\; \tilde{\boldsymbol{Q}}^\top\boldsymbol{y}.$$

Due to $\tilde{\boldsymbol{R}}$ being upper triangular, the latter task can be accomplished easily via back substitution.

Now, the main question is how to obtain a QR decomposition. In case of $n \geq p$, one could construct a reduced QR decomposition (A.1) by means of the Gram–Schmidt method. A faster and numerically more stable procedure is based on Householder transformations:

Consideration 1 (Householder matrices) If $\boldsymbol{b} \in \mathbb{R}^n$ is a unit vector, then

$$\boldsymbol{H} \;:=\; \boldsymbol{I}_n - 2\boldsymbol{b}\boldsymbol{b}^\top$$

describes the reflection at the hyperplane $\boldsymbol{b}^\perp$. That means,

$$\boldsymbol{H}\boldsymbol{x} \;=\; \begin{cases} -\boldsymbol{x} & \text{if } \boldsymbol{x} \in \operatorname{span}(\boldsymbol{b}), \\ \boldsymbol{x} & \text{if } \boldsymbol{x} \in \boldsymbol{b}^\perp. \end{cases}$$

In particular, $\boldsymbol{H}$ is a symmetric orthogonal matrix, that means,

$$\boldsymbol{H} \;=\; \boldsymbol{H}^\top \quad \text{and} \quad \boldsymbol{H}^2 \;=\; \boldsymbol{I}_n.$$

If $\boldsymbol{a}, \boldsymbol{r}$ are two different vectors in $\mathbb{R}^n$ with $\|\boldsymbol{a}\| = \|\boldsymbol{r}\| > 0$, then

$$\boldsymbol{H} \;=\; \boldsymbol{I}_n - \frac{2}{\|\boldsymbol{a}-\boldsymbol{r}\|^2}(\boldsymbol{a}-\boldsymbol{r})(\boldsymbol{a}-\boldsymbol{r})^\top \tag{A.2}$$

describes a reflection at the the hyperplane $(\boldsymbol{a}-\boldsymbol{r})^\perp$, and

$$\boldsymbol{H}\boldsymbol{a} \;=\; \boldsymbol{r}, \quad \boldsymbol{H}\boldsymbol{r} \;=\; \boldsymbol{a}.$$

Consideration 2 (Start of algorithm) Let $\boldsymbol{a}$ be the first column of $\boldsymbol{A}$, and let

$$\boldsymbol{r} \;:=\; \begin{cases} (-\|\boldsymbol{a}\|, 0, \ldots, 0)^\top & \text{if } a_1 \geq 0, \\ (+\|\boldsymbol{a}\|, 0, \ldots, 0)^\top & \text{if } a_1 < 0. \end{cases}$$

If we define $\boldsymbol{H}$ as in (A.2), then

$$\boldsymbol{A} \;=\; \boldsymbol{Q}\boldsymbol{R},$$

where

$$\boldsymbol{Q} \;:=\; \boldsymbol{H} \quad \text{and} \quad \boldsymbol{R} \;:=\; \boldsymbol{H}\boldsymbol{A} \;=\; \begin{bmatrix} \pm\|\boldsymbol{a}\| & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{A}_o \end{bmatrix}$$

with matrices $\boldsymbol{R}_{12} \in \mathbb{R}^{1\times(p-1)}$ and $\boldsymbol{A}_o \in \mathbb{R}^{(n-1)\times(p-1)}$ such that $\operatorname{rank}(\boldsymbol{A}_o) = \min(n-1, p-1)$.

Consideration 3 (Induction step) Suppose that after $k < \min(n-1, p)$ steps, we have obtained a decomposition

$$\boldsymbol{A} \;=\; \boldsymbol{Q}\boldsymbol{R}$$

with an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ and a matrix $\boldsymbol{R} \in \mathbb{R}^{n\times p}$ of the following form:

$$\boldsymbol{R} \;=\; \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{0} & \boldsymbol{A}_o \end{bmatrix}$$

with an upper triangular matrix $\boldsymbol{R}_{11} \in \mathbb{R}^{k\times k}$, some matrix $\boldsymbol{R}_{12} \in \mathbb{R}^{k\times(p-k)}$ and a matrix $\boldsymbol{A}_o \in \mathbb{R}^{(n-k)\times(p-k)}$ with rank $\min(n-k, p-k)$. Now we construct $\boldsymbol{H}_o$ as $\boldsymbol{H}$ in Consideration 2 with $\boldsymbol{A}_o$ in place of $\boldsymbol{A}$, that means $\boldsymbol{H}_o \in \mathbb{R}^{(n-k)\times(n-k)}$ is orthogonal such that the first column of

$\boldsymbol{H}_o\boldsymbol{A}_o$ is of the form $(r_o, 0, \ldots, 0)^\top \in \mathbb{R}^{n-k}$ for some $r_o \neq 0$. Then the equation $\boldsymbol{A} = \boldsymbol{QR}$ remains valid if we replace $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n]$ with

$$\boldsymbol{Q}\begin{bmatrix}\boldsymbol{I}_k & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{H}_o\end{bmatrix} \;=\; \big[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k,\, [\boldsymbol{q}_{k+1}, \ldots, \boldsymbol{q}_n]\boldsymbol{H}_o\big]$$

and $\boldsymbol{R}$ with

$$\begin{bmatrix}\boldsymbol{I}_k & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{H}_o\end{bmatrix}\boldsymbol{R} \;=\; \begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \boldsymbol{0} & \boldsymbol{H}_o\boldsymbol{A}_o\end{bmatrix} \;=\; \begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{R}_{12}\\ \boldsymbol{0} & \begin{bmatrix} r_o & * \\ \boldsymbol{0} & *\end{bmatrix}\end{bmatrix}.$$

The new matrix $\boldsymbol{Q}$ is orthogonal, too, and the first $k+1$ columns of the new matrix $\boldsymbol{R}$ constitute an upper triangular matrix.

Remark. To multiply a matrix $\boldsymbol{C}$ from the left or from the right with a Householder matrix $\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{b}\boldsymbol{b}^\top$, there is no need to generate the matrix $\boldsymbol{H}$ explicitly. Instead one can compute $\boldsymbol{CH} = \boldsymbol{C} - 2(\boldsymbol{Cb})\boldsymbol{b}^\top$ or $\boldsymbol{HC} = \boldsymbol{C} - 2\boldsymbol{b}(\boldsymbol{b}^\top\boldsymbol{C})$, which is substantially faster.

Remark. In R, the QR decomposition is at the core of the procedure $\mathsf{qr.solve}(\boldsymbol{A}, \boldsymbol{B})$ with input arguments $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ such that $\mathrm{rank}(\boldsymbol{A}) = p$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q] \in \mathbb{R}^{p\times q}$. It returns a matrix $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_q] \in \mathbb{R}^{p\times q}$ such that

$$\boldsymbol{c}_k \;=\; \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathbb{R}^p} \|\boldsymbol{A}\boldsymbol{v} - \boldsymbol{b}_k\|.$$

To do so, the first p steps of the QR algorithm are applied to the matrix $[\boldsymbol{A}, \boldsymbol{B}] \in \mathbb{R}^{n\times(p+q)}$ in place of $\boldsymbol{A}$, yielding

$$[\boldsymbol{A}, \boldsymbol{B}] \;=\; \boldsymbol{QR}$$

with $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ orthogonal and

$$\boldsymbol{R} \;=\; \begin{cases} [\boldsymbol{R}_{11}, \boldsymbol{D}] & \text{if } n = p,\\ \begin{bmatrix}\boldsymbol{R}_{11} & \boldsymbol{D}\\ \boldsymbol{0} & \boldsymbol{R}_{22}\end{bmatrix} & \text{if } n > p.\end{cases}$$

Here $\boldsymbol{R}_{11} \in \mathbb{R}^{p\times p}$ is upper triangular with nonzero diagonal, $\boldsymbol{R}_{22} \in \mathbb{R}^{(n-p)\times q}$, and $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_q] \in \mathbb{R}^{p\times q}$. Then the desired matrix $\boldsymbol{C}$ solves the linear equation system

$$\boldsymbol{R}_{11}\boldsymbol{C} \;=\; \boldsymbol{D}$$

and is computed via backsubstitution.

A.2 Expected Values and Covariances

The distribution of a real-valued random variables X is roughly characterized by its expected value $\mathbb{E}(X)$ and its variance $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$. Recall the definition of the covariance of

two real-valued random variables X, Y with $\mathbb{E}(X^2), \mathbb{E}(Y^2) < \infty$:

$$\begin{aligned}\operatorname{Cov}(X,Y) &:= \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big)\\ &= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)\\ &= \operatorname{Cov}(Y,X).\end{aligned}$$

For an additional random variable Z with $\mathbb{E}(Z^2) < \infty$ and fixed numbers $\alpha, \beta \in \mathbb{R}$ the following rules are useful:

$$\begin{aligned}\mathbb{E}(\alpha+\beta X) &= \alpha + \beta\,\mathbb{E}(X),\\ \operatorname{Cov}(\alpha+\beta X, Y) &= \beta\operatorname{Cov}(X,Y),\\ \operatorname{Cov}(X+Y,Z) &= \operatorname{Cov}(X,Z)+\operatorname{Cov}(Y,Z).\end{aligned}$$

Furthermore,

$$\operatorname{Cov}(X,Y) = 0 \quad \text{if } X \text{ and } Y \text{ are stochastically independent.}$$

Now we generalize these quantities to random matrices and vectors.

Definition A.1. **(a)** Let $\boldsymbol{Q} = (Q_{ij})_{i\le p, j\le q} \in \mathbb{R}^{p\times q}$ be a random matrix. The expected value of $\boldsymbol{Q}$ is defined componentwise,

$$\mathbb{E}(\boldsymbol{Q}) := \begin{bmatrix}\mathbb{E}(Q_{11}) & \cdots & \mathbb{E}(Q_{1q})\\ \vdots & & \vdots\\ \mathbb{E}(Q_{p1}) & \cdots & \mathbb{E}(Q_{pq})\end{bmatrix} \in \mathbb{R}^{p\times q}.$$

Here we assume that all expected values $\mathbb{E}\,|Q_{ij}|$ are finite.

(b) Let $\boldsymbol{X} = (X_i)_{i=1}^p \in \mathbb{R}^p$ and $\boldsymbol{Y} = (Y_j)_{j=1}^q \in \mathbb{R}^q$ be random vectors with finite expected values $\mathbb{E}(\|\boldsymbol{X}\|^2)$ and $\mathbb{E}(\|\boldsymbol{Y}\|^2)$. The *covariance (matrix)* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined to be the matrix

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y}) &:= \mathbb{E}\big((\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathbb{E}(\boldsymbol{Y}))^\top\big)\\ &= \mathbb{E}(\boldsymbol{X}\boldsymbol{Y}^\top) - \mathbb{E}(\boldsymbol{X})\,\mathbb{E}(\boldsymbol{Y})^\top\\ &= \big(\operatorname{Cov}(X_i,Y_j)\big)_{i\le p, j\le q} \in \mathbb{R}^{p\times q},\end{aligned}$$

and the *covariance (matrix)* of $\boldsymbol{X}$ is the symmetric matrix

$$\operatorname{Var}(\boldsymbol{X}) := \operatorname{Cov}(\boldsymbol{X},\boldsymbol{X}) \in \mathbb{R}^{p\times p}.$$

Its diagonal contains the variances $\operatorname{Var}(X_1), \ldots, \operatorname{Var}(X_p)$.

To what extend are these definitions of $\mathbb{E}(\boldsymbol{X})$, $\operatorname{Var}(\boldsymbol{X})$ and $\operatorname{Cov}(\boldsymbol{X},\boldsymbol{Y})$ meaningful? On the one hand, for arbitrary fised numbers $a \in \mathbb{R}$ and vectors $\boldsymbol{b} \in \mathbb{R}^p$ one can express expected value and variance of $a + \boldsymbol{b}^\top\boldsymbol{X}$ as follows:

$$\begin{aligned}\mathbb{E}(a+\boldsymbol{b}^\top\boldsymbol{X}) &= \mathbb{E}\Big(a+\sum_{i=1}^p b_i X_i\Big) = a + \sum_{i=1}^p b_i\,\mathbb{E}(X_i)\\ &= a + \boldsymbol{b}^\top\,\mathbb{E}(\boldsymbol{X}),\\ \operatorname{Var}(a+\boldsymbol{b}^\top\boldsymbol{X}) &= \operatorname{Var}\Big(a+\sum_{i=1}^k b_i X_i\Big) = \sum_{i,j=1}^p b_i b_j \operatorname{Cov}(X_i,X_j)\\ &= \boldsymbol{b}^\top \operatorname{Var}(\boldsymbol{X})\boldsymbol{b}.\end{aligned}$$

Since $\mathrm{Var}(a + \boldsymbol{b}^\top \boldsymbol{X})$ is always non-negative, this representation proves the following property:

$$\mathrm{Var}(\boldsymbol{X}) \text{ is symmetric and positive semidefinite.}$$

In case of a unit vector $\boldsymbol{b}$, the vector $(\boldsymbol{b}^\top \boldsymbol{X})\boldsymbol{b}$ is the orthogonal projection of $\boldsymbol{X}$ onto the line $\mathrm{span}(\boldsymbol{b})$; see the left panel of Figure A.1. Then $\mathrm{Var}(\boldsymbol{b}^\top \boldsymbol{X})$ quantifies the random fluctuations of $\boldsymbol{X}$ in the direction of $\boldsymbol{b}$. The matrix $\mathrm{Var}(\boldsymbol{X})$ is singular if and only if $\boldsymbol{b}^\top \mathrm{Var}(\boldsymbol{X})\boldsymbol{b} = \mathrm{Var}(\boldsymbol{b}^\top \boldsymbol{X})$ equals 0 for some unit vector $\boldsymbol{b}$. But this means that $\boldsymbol{X}$ lies almost surely on the hyperplane

$$\mathbb{H} \;:=\; \big\{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{b}^\top \boldsymbol{x} = \boldsymbol{b}^\top \mathbb{E}(\boldsymbol{X})\big\}.$$

This hyperplane contains the vector $\mathbb{E}(\boldsymbol{X})$ and is perpendicular to the vector $\boldsymbol{b}$; see the right panel of Figure A.1.

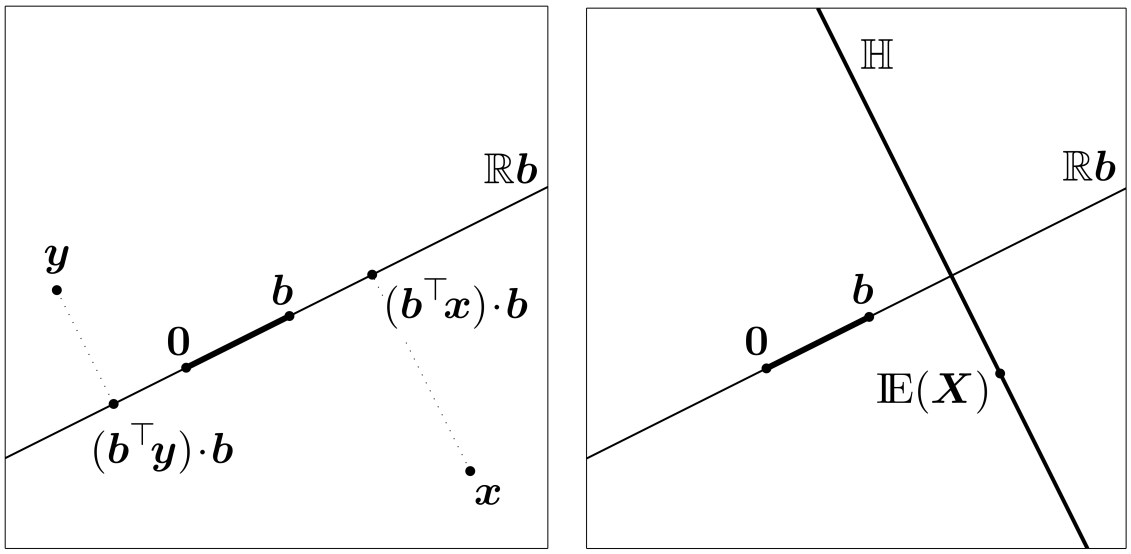

Figure A.1: Geometric interpretation of $\boldsymbol{b}^\top \boldsymbol{X}$ (left panel). Support $\mathbb{H}$ of $\boldsymbol{X}$ in case of $\boldsymbol{b}^\top \mathrm{Var}(\boldsymbol{X})\boldsymbol{b} = 0$ (right panel).

The aforementioned formulae for the expected value and variance of affine functions of $\boldsymbol{X}$ may be generalized to arbitrary affine mappings. The proof of the following lemma is left to the reader as an exercise.

Lemma A.2. *(a) Let $\boldsymbol{Q}, \tilde{\boldsymbol{Q}} \in \mathbb{R}^{p\times q}$ be random matrices with finite expected values $\mathbb{E}\,|Q_{ij}|$ and $\mathbb{E}\,|\widetilde{Q}_{ij}|$. Then*

$$\mathbb{E}(\boldsymbol{Q}^\top) \;=\; \mathbb{E}(\boldsymbol{Q})^\top \quad \textit{and} \quad \mathbb{E}(\boldsymbol{Q} + \tilde{\boldsymbol{Q}}) \;=\; \mathbb{E}(\boldsymbol{Q}) + \mathbb{E}(\tilde{\boldsymbol{Q}}).$$

Moreover, for fixed matrices $\boldsymbol{A} \in \mathbb{R}^{s\times t}$, $\boldsymbol{B} \in \mathbb{R}^{s\times p}$ and $\boldsymbol{C} \in \mathbb{R}^{q\times t}$,

$$\mathbb{E}(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{Q}\boldsymbol{C}) \;=\; \boldsymbol{A} + \boldsymbol{B}\,\mathbb{E}(\boldsymbol{Q})\boldsymbol{C}.$$

(b) Let $\boldsymbol{R} \in \mathbb{R}^{q\times r}$ be an additional random matrix with finite expected values $\mathbb{E}\,|R_{jk}|$. If $\boldsymbol{Q}$ and $\boldsymbol{R}$ are stochastically independent, then

$$\mathbb{E}(\boldsymbol{Q}\boldsymbol{R}) \;=\; \mathbb{E}(\boldsymbol{Q})\,\mathbb{E}(\boldsymbol{R}).$$

(c) *Let $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$ be random vectors with finite expected values $\mathbb{E}(\|\boldsymbol{X}\|^2)$, $\mathbb{E}(\|\tilde{\boldsymbol{X}}\|^2)$ and $\mathbb{E}(\|\boldsymbol{Y}\|^2)$. Then*

$$\mathrm{Cov}(\boldsymbol{Y}, \boldsymbol{X}) \;=\; \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})^\top \quad \textit{and} \quad \mathrm{Cov}(\boldsymbol{X} + \tilde{\boldsymbol{X}}, \boldsymbol{Y}) \;=\; \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{Cov}(\tilde{\boldsymbol{X}}, \boldsymbol{Y}).$$

Moreover, for fixed vectors $\boldsymbol{a} \in \mathbb{R}^s$ and matrices $\boldsymbol{B} \in \mathbb{R}^{s \times p}$,

$$\begin{aligned} \mathbb{E}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}) \;&=\; \boldsymbol{a} + \boldsymbol{B}\,\mathbb{E}(\boldsymbol{X}), \\ \mathrm{Cov}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}, \boldsymbol{Y}) \;&=\; \boldsymbol{B}\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}), \\ \mathrm{Var}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}) \;&=\; \boldsymbol{B}\,\mathrm{Var}(\boldsymbol{X})\boldsymbol{B}^\top. \end{aligned}$$

A.3 Monte Carlo Estimators for Tukey's Method

In Section 3.6 we considered random variables of the following type: For a given set $\mathcal{B}$ of unit vectors $\boldsymbol{b} \in \mathbb{R}^p$,

$$T \;:=\; \frac{W}{\sqrt{S^2/\ell}}$$

with $W := \sup_{\boldsymbol{b} \in \mathcal{B}} |\boldsymbol{b}^\top \boldsymbol{Z}|$, where $\boldsymbol{Z} \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I})$ and $S^2 \sim \chi^2_\ell$ are stochastically independent. Now our goal is to estimate the survival function $\bar{F} : [0, \infty) \to [0, 1]$ of T, i.e.

$$\bar{F}(x) \;:=\; \mathbb{P}(T \geq x).$$

An obvious way to approximate the unknown $(1-\alpha)$-quantile $\kappa_{1-\alpha}$ of T is to simulate independent copies $T_1, T_2, \ldots, T_B$ of T and to set

$$\widehat{\kappa}_{1-\alpha} \;:=\; T_{(m)} \quad \text{with} \quad m \;:=\; \lceil (B+1)(1-\alpha) \rceil,$$

where $T_{(1)}, T_{(2)}, \ldots, T_{(B)}$ are the order statistics of $T_1, T_2, \ldots, T_B$. The choice of m is justified in Exercise A.3 below. However, in the present setting one can generate a refined Monte Carlo approximation of the survival function $\bar{F}$.

Integrating out S^2. We may write

$$\begin{aligned} \bar{F}(x) \;&=\; \mathbb{P}\big(W \geq x\sqrt{S^2/\ell}\big) \\ &=\; \mathbb{E}\,\mathbb{P}\big(S^2 \leq \ell W^2/x^2 \,\big|\, M\big) \\ &=\; \mathbb{E}\, G_\ell(\ell W^2/x^2) \end{aligned}$$

for arbitrary $x \geq 0$, where G_ℓ stands for the distribution function of χ^2_ℓ. Hence a Monte Carlo estimator of $\bar{F}(x)$ may be constructed as follows: We simulate a large number of stochastically independent copies $W_1, W_2, \ldots, W_B$ of M. Then we estimate $\bar{F}(x)$ by

$$\widehat{\bar{F}}(x) \;:=\; \frac{1}{B} \sum_{s=1}^{B} G_\ell(\ell W_s^2/x^2).$$

Efficient simulation of W. In the special case $\mathcal{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2\}$, note that

$$\boldsymbol{W} \sim \mathrm{N}_2\Big(\boldsymbol{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\Big)$$

with $\rho := \boldsymbol{b}_1^\top \boldsymbol{b}_2 \in [-1, 1]$. Here W has the same distribution as

$$\max\big\{|Z_1|, |\rho Z_1 + \bar{\rho} Z_2|\big\}$$

with $\bar{\rho} := \sqrt{1 - \rho^2}$ and $\boldsymbol{Z} \sim \mathrm{N}_2(\boldsymbol{0}, \boldsymbol{I}_2)$.

In the general setting, we considered $\mathcal{B} = \big\{\|\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi}\|^{-1}\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi} \colon \boldsymbol{\psi} \in \mathcal{P}\big\}$ for some set $\mathcal{P} \subset \mathbb{R}^p \setminus \{\boldsymbol{0}\}$, where $\boldsymbol{\Gamma} = \boldsymbol{D}^\top \boldsymbol{D}$. If we determine a QR decomposition of $\boldsymbol{D}$, that is, $\boldsymbol{D} = \boldsymbol{Q}\boldsymbol{R}$ with a matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times p}$ such that $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_p$ and a nonsingular, upper triangular matrix $\boldsymbol{R}$, then $\boldsymbol{\Gamma} = \boldsymbol{R}^\top \boldsymbol{R}$. Now let $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^\top$ with orthogonal matrices $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{p \times p}$ and a diagonal matrix $\boldsymbol{L}$ with strictly positive diagonal entries (SVD decomposition). Then

$$\boldsymbol{\Gamma} \;=\; \boldsymbol{R}^\top \boldsymbol{Q}^\top \boldsymbol{Q} \boldsymbol{R} \;=\; \boldsymbol{R}^\top \boldsymbol{R} \;=\; \boldsymbol{V}\boldsymbol{L}^2\boldsymbol{V}^\top,$$

whence $\boldsymbol{\Gamma}^{-1/2} = \boldsymbol{V}\boldsymbol{L}^{-1}\boldsymbol{V}^\top = \boldsymbol{T}\boldsymbol{R}^{-\top}$ with the orthogonal matrix $\boldsymbol{T} := \boldsymbol{V}\boldsymbol{U}^\top$, and $\boldsymbol{R}^{-\top} := (\boldsymbol{R}^\top)^{-1} = (\boldsymbol{R}^{-1})^\top$. Consequently,

$$\begin{aligned} W \;&=\; \sup_{\boldsymbol{b} \in \mathcal{B}} |\boldsymbol{b}^\top \boldsymbol{Z}| \\ &=\; \sup_{\boldsymbol{\psi} \in \mathcal{P}} \|\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi}\|^{-1} \big|(\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\psi})^\top \boldsymbol{Z}\big| \\ &=\; \sup_{\boldsymbol{\psi} \in \mathcal{P}} \|\boldsymbol{R}^{-\top}\boldsymbol{\psi}\|^{-1} \big|(\boldsymbol{R}^{-\top}\boldsymbol{\psi})^\top \boldsymbol{T}^\top \boldsymbol{Z}\big| \end{aligned}$$

has the same distribution as

$$\sup_{\boldsymbol{\psi} \in \mathcal{P}} \|\boldsymbol{R}^{-\top}\boldsymbol{\psi}\|^{-1} \big|(\boldsymbol{R}^{-\top}\boldsymbol{\psi})^\top \boldsymbol{Z}\big|.$$

Exercise A.3 (Monte Carlo confidence bounds with pivotal statistics). Let $Y \in \mathcal{Y}$ be a random variable with distribution depending on an unknown parameter $\theta \in \Theta$. Furthermore, let $T : \mathcal{Y} \times \Theta \to \mathbb{R}$ be measurable in its first argument such that $T(Y, \theta)$ has a known continuous distribution P_o. Next, let $T_1, T_2, \ldots, T_B$ be stochastically independent random variables with distribution P_o, also independent from Y, and let $T_{(1)} < T_{(2)} < \cdots < T_{(B)}$ be their order statistics. Show that for for any $m \in \{1, 2, \ldots, B\}$,

$$C_m = C_m(Y, T_1, \ldots, T_B) \;:=\; \big\{\eta \in \Theta : T(Y, \eta) \le T_{(m)}\big\}$$

satisfies the equation

$$\mathbb{P}(C_m \ni \theta) \;=\; \frac{m}{B+1}.$$

A.4 B Splines

For $m, d \in \mathbb{N}$ and real numbers $t_0 < t_1 < \ldots < t_m$, let $\mathcal{S}_d(t_0, t_1, \ldots, t_m)$ be the set of all splines of order d with knots $t_0, t_1, \ldots, t_m$, as introduced in Section 5.1. We have seen already that it is a real vector space of dimension $m + d$.

Now we consider temporarily an infinite-dimensional variant. Namely, let $\boldsymbol{t} = (t_z)_{z\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ be a double sequence such that

$$\cdots < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < \cdots \quad \text{and} \quad \lim_{z\to\pm\infty} t_z \;=\; \pm\infty.$$

With $\mathcal{S}_d(\boldsymbol{t})$ we denote the set of all functions $f : \mathbb{R} \to \mathbb{R}$ with the following properties:

- On each interval $[t_{z-1}, t_z]$, $z \in \mathbb{Z}$, f is a polynomial of order d;
- f is $d-1$ times continuously differentiable.

Each function $f \in \mathcal{S}_d(\boldsymbol{t})$ may be represented as

$$f(x) \;=\; \sum_{j=0}^{d} a_j (x-t_0)^j + \sum_{k\geq 1} b_k \max(x-t_k,0)^d + \sum_{\ell\leq 0} c_\ell \max(t_\ell - x, 0)^d$$

with unique real coefficients a_j ($0 \leq j \leq d$), b_k ($k \geq 1$) and c_ℓ ($\ell \leq 0$). If x is restricted to $[t_{-(m+1)}, t_{m+1}]$ for some integer $m \geq 1$, one may write

$$f(x) \;=\; \sum_{j=0}^{d} a_j (x-t_0)^j + \sum_{k=1}^{m} b_k \max(x-t_k,0)^d - \sum_{\ell=-m}^{0} c_\ell \max(t_\ell - x, 0)^d$$

For various reasons, the basis functions $(\cdot - t_0)^j$, $\max(\cdot - t_k, 0)^d$ and $\max(t_\ell - \cdot, 0)^d$ are problematic, and one would rather use basis functions with compact support. To this end, there exists a special construction due to Carl de Boor:

Definition A.4 (B Splines). For $z \in \mathbb{Z}$ let

$$b_{z,0}(x) \;:=\; 1_{[t_z \leq x < t_{z+1}]}.$$

For $d = 1, 2, 3, \ldots$ define recursively

$$b_{z,d}(x) \;:=\; \frac{x - t_z}{\delta_{z,d}}\, b_{z,d-1}(x) + \frac{t_{z+1+d} - x}{\delta_{z+1,d}}\, b_{z+1,d-1}(x), \tag{A.3}$$

where $\delta_{y,d} := t_{y+d} - t_y$.

Remark A.5 (The special case $d = 1$). The space $\mathcal{S}_1(\boldsymbol{t})$ consists of all continuous functions $f : \mathbb{R} \to \mathbb{R}$ such that f is affine on each interval $[t_z, t_{z+1}]$, $z \in \mathbb{Z}$. The basis functions $b_{z,1}$ belong to this space and are given by

$$b_{z,1}(t_y) \;=\; 1_{[y=z+1]} \quad \text{for } y, z \in \mathbb{Z}.$$

In particular, $b_{z,1} > 0$ on (t_z, t_{z+2}) and $b_{z,1} \equiv 0$ on $(-\infty, t_z] \cup [t_{z+2}, \infty)$, and

$$b'_{z,1} \;= \frac{b_{z,0}}{\delta_{z,1}} - \frac{b_{z+1,0}}{\delta_{z+1,1}} \quad \text{on } \mathbb{R} \setminus \{t_z, t_{z+1}, t_{z+2}\}.$$

Moreover,

$$\int_{\mathbb{R}} b_{z,1}(x)\, dx \;=\; \frac{\delta_{z,2}}{2}.$$

An arbitrary function $f \in \mathcal{S}_1(\boldsymbol{t})$ may be written as

$$f(x) \;=\; \sum_{z\in\mathbb{Z}} b_{z,1}(x) f(t_{z+1}).$$

In particular, $\sum_{z\in\mathbb{Z}} b_{z,1} \equiv 1$.

Theorem A.6. *For $d \geq 1$, the functions $b_{z,d}$, $z \in \mathbb{Z}$, have the following properties:*

(i)
$$b_{z,d} \begin{cases} > 0 & \text{on } (t_z, t_{z+d+1}), \\ = 0 & \text{on } (-\infty, t_z] \cup [t_{z+d+1}, \infty). \end{cases}$$

(ii)
$$\sum_{z \in \mathbb{Z}} b_{z,d} \equiv 1.$$

(iii) $b_{z,d} \in \mathcal{S}_d(\boldsymbol{t})$, *and*
$$b'_{z,d} = d\Big(\frac{b_{z,d-1}}{\delta_{z,d}} - \frac{b_{z+1,d-1}}{\delta_{z+1,d}}\Big) \quad \text{on } \begin{cases} \mathbb{R} \setminus \{t_z : z \in \mathbb{Z}\} & \text{if } d = 1, \\ \mathbb{R} & \text{if } d \geq 2. \end{cases}$$

(iv)
$$\int_{\mathbb{R}} b_{z,d}(x)\, dx = \frac{\delta_{z,d+1}}{d+1}.$$

(v) *The set $\mathcal{S}_d(\boldsymbol{t})$ coincides with the set of all functions*
$$f = \sum_{z \in \mathbb{Z}} \lambda_z b_{z,d}$$
with $\boldsymbol{\lambda} = (\lambda_z)_{z \in \mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$. The parameter sequence $\boldsymbol{\lambda}$ is uniquely determined by the function $f = \sum_z \lambda_z b_{z,d}$.

Proof of Theorem A.6. In case of $d = 1$, properties (i–v) follow directly from Remark A.5. Now we assume that properties (i–v) are satisfied for a fixed $d \geq 1$, and we verify them for $e := d + 1$ in place of d. We write "property $(\cdot)_{(e)}$" to indicate this replacement.

First of all, let us rewrite the recursion formula (A.3) as

$$b_{z,e}(x) = \frac{x - t_z}{\delta_{z,e}}\, b_{z,d}(x) + \frac{t_{z+e+1} - x}{\delta_{z+1,e}}\, b_{z+1,d}(x). \tag{A.4}$$

Property (i) implies that for any $z \in \mathbb{Z}$ and $y \in \{z-1, z\}$, $b_{y,d} > 0$ on (t_z, t_{y+d+1}) and $b_{y,d} = 0$ on $(-\infty, t_y] \cup [t_{y+d+1}, \infty)$. Since

$$(t_z, t_{z+d+1}) \cup (t_{z+1}, t_{z+1+d+1}) = (t_z, t_{z+e+1}),$$

property (i) and (A.4) imply property (i)$_{(e)}$: Outside of the interval (t_z, t_{z+e+1}), both functions $b_{z,d}$ and $b_{z+1,d}$ vanish. And for x within this interval, both factors $x - t_z$ and $t_{z+e+1} - x$ are strictly positive, both factors $b_{z,d}(x)$ and $b_{z+1,d}(x)$ are nonnegative with at least one of them being strictly positive.

Property (ii)$_{(e)}$ is also a consequence of property (ii) and (A.4). For any $x \in \mathbb{R}$,

$$\begin{aligned} \sum_{z \in \mathbb{Z}} b_{z,e}(x) &= \sum_{z \in \mathbb{Z}} \Big(\frac{x - t_z}{\delta_{z,e}}\, b_{z,d}(x) + \frac{t_{z+e+1} - x}{\delta_{z+1,e}}\, b_{z+1,d}(x) \Big) \\ &= \sum_{z \in \mathbb{Z}} \frac{x - t_z}{\delta_{z,e}}\, b_{z,d}(x) + \sum_{y \in \mathbb{Z}} \frac{t_{y+e} - x}{\delta_{y,e}}\, b_{y,d}(x) \\ &= \sum_{z \in \mathbb{Z}} \Big(\frac{x - t_z}{\delta_{z,e}} + \frac{t_{z+e} - x}{\delta_{z,e}} \Big)\, b_{z,d}(x) \\ &= \sum_{z \in \mathbb{Z}} b_{z,d}(x) = 1. \end{aligned}$$

Note that in the previous manipulations of infinite sums we rely on property (i), implying that $b_{z,d}(x) \neq 0$ for only finitely many $z \in \mathbb{Z}$.

Property (iii) implies that all functions $b_{z,d}$, $z \in \mathbb{Z}$, belong to $\mathcal{S}_d(\boldsymbol{t})$ and satisfy

$$b'_{z,d} \;=\; d\Big(\frac{b_{z,d-1}}{\delta_{z,d}} - \frac{b_{z+1,d-1}}{\delta_{z+1,d}}\Big) \quad \text{on } \; U := \mathbb{R} \setminus \{t_z : z \in \mathbb{Z}\}.$$

Then for $x \in U$,

$$\begin{aligned} b'_{z,e}(x) \;&=\; \frac{\partial}{\partial x}\Big(\frac{x - t_z}{\delta_{z,e}}\, b_{z,d}(x) + \frac{t_{z+e+1} - x}{\delta_{z+1,e}}\, b_{z+1,d}(x)\Big) \\ &=\; \frac{b_{z,e-1}(x)}{\delta_{z,e}} - \frac{b_{z+1,e-1}(x)}{\delta_{z+1,e}} + \frac{x - t_z}{\delta_{z,e}}\, b'_{z-1,d}(x) + \frac{t_{z+e+1} - x}{\delta_{z+1,e}}\, b'_{z+1,d}(x). \end{aligned} \tag{A.5}$$

The latter two summands may be written as

$$\begin{aligned} \frac{x - t_z}{\delta_{z,e}}\, b'_{z,d}(x) \;&=\; \frac{x - t_z}{\delta_{z,e}}\, d\Big(\frac{b_{z,d-1}(x)}{\delta_{z,d}} - \frac{b_{z+1,d-1}(x)}{\delta_{z+1,d}}\Big) \\ &=\; d\,\frac{1}{\delta_{z,e}}\Big(\frac{x - t_z}{\delta_{z,d}}\, b_{z,d-1}(x) + \frac{t_z - x}{\delta_{z+1,d}}\, b_{z+1,d-1}(x)\Big) \\ &=\; d\,\frac{1}{\delta_{z,e}}\Big(\frac{x - t_z}{\delta_{z,d}}\, b_{z,d-1}(x) - \frac{t_{z+d+1} - x}{\delta_{z+1,d}}\, b_{z+1,d-1}(x)\Big) \\ &\qquad\qquad -\; d\,\frac{\delta_{z,d+1}}{\delta_{z,e}\delta_{z+1,d}}\, b_{z+1,d-1}(x) \\ &=\; d\,\frac{b_{z,d}(x)}{\delta_{z,e}} - d\,\frac{b_{z+1,d-1}(x)}{\delta_{z+1,d}}, \end{aligned}$$

and

$$\begin{aligned} \frac{t_{z+e+1} - x}{\delta_{z+1,e}}\, b'_{z+1,d}(x) \;&=\; \frac{t_{z+e+1} - x}{\delta_{z+1,e}}\, d\Big(\frac{b_{z+1,d-1}(x)}{\delta_{z+1,d}} - \frac{b_{z+2,d-1}(x)}{\delta_{z+2,d}}\Big) \\ &=\; -\,d\,\frac{1}{\delta_{z+1,e}}\Big(\frac{x - t_{z+d+2}}{\delta_{z+1,d}}\, b_{z+1,d-1}(x) + \frac{t_{z+d+2} - x}{\delta_{z+2,d}}\, b_{z+2,d-1}(x)\Big) \\ &=\; -\,d\,\frac{1}{\delta_{z+1,e}}\Big(\frac{x - t_{z+1}}{\delta_{z+1,d}}\, b_{z+1,d-1}(x) + \frac{t_{z+d+2} - x}{\delta_{z+2,d}}\, b_{z+2,d-1}(x)\Big) \\ &\qquad\qquad +\; d\,\frac{\delta_{z+1,d+1}}{\delta_{z+1,e}\delta_{z+1,d}}\, b_{z+1,d-1}(x) \\ &=\; -\,d\,\frac{b_{z+1,d}(x)}{\delta_{z,e}} +\; d\,\frac{b_{z+1,d-1}(x)}{\delta_{z+1,d}}. \end{aligned}$$

Plugging-in these two equations in (A.5) shows that

$$b'_{z,e} \;=\; \frac{e}{\delta_{z-1,e}}\, b_{z-1,d} - \frac{e}{\delta_{z,e}}\, b_{z,d} \quad \text{on } U.$$

But both sides of the latter equation are continuous functions on $\mathbb{R}$, and $b_{z,e}$ is continuous by (A.4). Hence one can deduce from the mean value theorem that even

$$b'_{z,e} \;=\; \frac{e}{\delta_{z,e}}\, b_{z,d} - \frac{e}{\delta_{z+1,e}}\, b_{z+1,d} \quad \text{on } \mathbb{R}.$$

In particular, $b'_{z,e} \in \mathcal{S}_d(\boldsymbol{t})$, so the function $b_{z,e}$ itself is a spline function in $\mathcal{S}_{d+1}(\boldsymbol{t}) = \mathcal{S}_e(\boldsymbol{t})$. This concludes the proof of property $(\text{iii})_{(e)}$.

As to property (iv)$_{(e)}$, by partial integration, $b_{z,e}(\pm\infty) = 0$ and property (iii)$_{(e)}$,

$$\int_{\mathbb{R}} b_{z,e}(x)\,dx \;=\; -\int_{\mathbb{R}} x b'_{z,e}(x)\,dx \;=\; -e\int_{\mathbb{R}} x\Big(\frac{b_{z,d}(x)}{\delta_{z,e}} - \frac{b_{z+1,d}(x)}{\delta_{z+1,e}}\Big)\,dx.$$

The integrand may be rewritten as

$$\begin{aligned} x\Big(\frac{b_{z,d}(x)}{\delta_{z,e}} - \frac{b_{z+1,d}(x)}{\delta_{z+1,e}}\Big) &= \frac{x-t_z}{\delta_{z,e}}\,b_{z,d}(x) + \frac{t_{z+d+1}-x}{\delta_{z+1,e}}\,b_{z+1,d}(x) \\ &\quad + \frac{t_z}{\delta_{z,e}}\,b_{z,d}(x) - \frac{t_{z+d+1}}{\delta_{z+1,e}}\,b_{z+1,d}(x) \\ &= b_{z,e}(x) + \frac{t_z}{\delta_{z-1,e}}\,b_{z,d}(x) - \frac{t_{z+d+1}}{\delta_{z+1,e}}\,b_{z+1,d}(x). \end{aligned}$$

Hence property (iv) implies that

$$\begin{aligned} \int_{\mathbb{R}} b_{z,e}(x)\,dx \;&=\; -e\int_{\mathbb{R}} b_{z,e}(x)\,dx - \frac{et_z}{\delta_{z,e}}\int_{\mathbb{R}} b_{z,d}(x)\,dx + \frac{et_{z+d+1}}{\delta_{z+1,e}}\int_{\mathbb{R}} b_{z+1,d}(x)\,dx \\ &=\; -e\int_{\mathbb{R}} b_{z,e}(x)\,dx - t_z + t_{z+d+1} \\ &=\; -e\int_{\mathbb{R}} b_{z,e}(x)\,dx + \delta_{z,e+1}, \end{aligned}$$

which implies property (iv)$_{(e)}$.

It remains to verify property (v)$_{(e)}$. Any function $f \in \mathcal{S}_e(\boldsymbol{t})$ satisfies $f' \in \mathcal{S}_d(\boldsymbol{t})$. Hence, there exists a double sequence $\boldsymbol{\lambda} \in \mathbb{R}^{\mathbb{Z}}$ such that $f' = \sum_{z\in\mathbb{Z}} \lambda_z b_{z,d}$. Now let

$$\mu_z \;:=\; \frac{\delta_{z,e}\lambda_z}{e}, \quad h_z \;:=\; \frac{e}{\delta_{z,e}}\,b_{z,d}$$

and

$$\nu_z \;:=\; \begin{cases} \sum_{k=1}^{z} \mu_k & \text{if } z > 0, \\ 0 & \text{if } z = 0, \\ -\sum_{k=z+1}^{0} \mu_k & \text{if } z < 0. \end{cases}$$

Then it follows from property (iii)$_{(e)}$ that

$$f' \;=\; \sum_{z\in\mathbb{Z}} \mu_z h_z \;=\; \sum_{z\in\mathbb{Z}} (\nu_z - \nu_{z-1}) h_z \;=\; \sum_{z\in\mathbb{Z}} \nu_z (h_z - h_{z+1}) \;=\; \sum_{z\in\mathbb{Z}} \nu_z b'_{z,e}.$$

Consequently,

$$f \;=\; C + \sum_{z\in\mathbb{Z}} \nu_z b_{z,e} \;=\; \sum_{z\in\mathbb{Z}} (C + \nu_z) b_{z,e},$$

because of property (ii)$_{(e)}$. Finally, to verify uniqueness of the series representation of functions in $\mathcal{S}_e(\boldsymbol{t})$, let $\boldsymbol{\lambda} \in \mathbb{R}^{\mathbb{Z}}$ such that

$$\sum_{z\in\mathbb{Z}} \lambda_z b_{z,e} \;\equiv\; 0.$$

Then, by property (iii)$_{(e)}$,

$$\begin{aligned} 0 \;&\equiv\; \sum_{z\in\mathbb{Z}} \lambda_z b'_{z,e} \\ &\equiv\; \sum_{z\in\mathbb{Z}} \lambda_z \Big(\frac{e}{\delta_{z,e}}\, b_{z,d} - \frac{e}{\delta_{z+1,e}}\, b_{z+1,d} \Big) \\ &\equiv\; \sum_{z\in\mathbb{Z}} \lambda_z \frac{e}{\delta_{z,e}}\, b_{z,d} - \sum_{z\in\mathbb{Z}} \lambda_{z-1} \frac{e}{\delta_{z,e}}\, b_{z,d} \\ &\equiv\; \sum_{z\in\mathbb{Z}} (\lambda_z - \lambda_{z-1}) \frac{e}{\delta_{z,e}}\, b_{z,d}. \end{aligned}$$

By property (v), $\boldsymbol{\lambda} = (c)_{z\in\mathbb{Z}}$ for some constant $c \in \mathbb{R}$. But then property (ii)$_{(e)}$ implies that

$$0 \;\equiv\; \sum_{z\in\mathbb{Z}} \lambda_z b_{z,e} \;\equiv\; c \sum_{z\in\mathbb{Z}} b_{z,e} \;\equiv\; c,$$

whence $\boldsymbol{\lambda} = \mathbf{0}$. □

Remark A.7 (B-Splines for $\mathcal{S}_d(t_0, t_1, \ldots, t_m)$). It follows from Theorem A.6 that the functions $b_{-d,d}, b_{-d+1,d}, \ldots, b_{m-1,d}$ constitute a basis of $\mathcal{S}_d(t_0, t_1, \ldots, t_m)$. Indeed, any function $f : [t_0, t_m] \to \mathbb{R}$ in the latter space may by extended to a function $f \in \mathcal{S}_d(\boldsymbol{t})$ as follows:

$$f(x) \;:=\; \begin{cases} \sum_{i=0}^{d} \dfrac{f^{(i)}(t_0+)}{i!}\, (x - t_0)^i & \text{for } x < t_0, \\ \sum_{i=0}^{d} \dfrac{f^{(i)}(t_m-)}{i!}\, (x - t_m)^i & \text{for } x > t_m. \end{cases}$$

Then $f \equiv \sum_{z\in\mathbb{Z}} \lambda_z b_{z,d}$ with certain coefficients $\lambda_z \in \mathbb{R}$, $z \in \mathbb{Z}$. But $b_{z,d} \equiv 0$ on $[t_0, t_m]$ whenever $z < -d$ or $z \geq m$. Hence

$$f \;\equiv\; \sum_{z=-d}^{m-1} \lambda_z b_{z,d} \quad \text{on } [t_0, t_m].$$

If one is only interested in the functions $b_{-d,d}, \ldots, b_{m-1,d}$ as a basis for $\mathcal{S}_d(t_0, t_1, \ldots, t_m)$, they may be computed via the following recursion: For $x \in [t_0, t_m]$,

$$b_{z,d}(x) \;=\; \begin{cases} \dfrac{t_1 - x}{\delta_{-d+1,d}}\, b_{-d+1,d-1}(x) & \text{if } z = -d, \\ \dfrac{x - t_z}{\delta_{z,d}}\, b_{z,d-1}(x) + \dfrac{t_{z+d+1} - x}{\delta_{z+1,d}}\, b_{z+1,d-1}(x) & \text{if } -d < z < m-1, \\ \dfrac{x - t_{m-1}}{\delta_{m-1,d}}\, b_{m-1,d-1}(x) & \text{if } z = m-1. \end{cases}$$

Moreover,

$$b'_{z,d} \;=\; d \begin{cases} \dfrac{b_{-d+1,d-1}}{\delta_{-d+1,d}} & \text{if } z = 1 \\ \dfrac{b_{z,d-1}}{\delta_{z,d}} - \dfrac{b_{z+1,d-1}}{\delta_{z+1,d}} & \text{if } -d < z < m-1 \\ \dfrac{b_{m-1,d-1}}{\delta_{m-1,d}} & \text{if } z = m-1 \end{cases}$$

on

$$\begin{cases} [t_0, t_m] \setminus \{t_1, \ldots, t_{m-1}\} & \text{if } d = 1, \\ [t_0, t_m] & \text{if } d \geq 2, \end{cases}$$

Here, one may even set $t_z = t_0$ for $-d \leq z < 0$ and $t_z = t_m$ for $m < z \leq m+d$: If we let $t_z \to t_0$ for $-d \leq z < 0$ and $t_z \to t_m$ for $m < z \leq m + d$, the basis functions $b_{y,c}$, $1 \leq y \leq m + c$, $0 \leq c \leq d$, converge to basis functions satisfying the same recursion formulae with

$$\delta_{y,c} \;:=\; t_{\min(y+c,m)} - t_{\max(y,0)}.$$

That means, we don't have to specify t_z for $z \notin \{0, 1, \ldots, m\}$. And with that modification,

$$\int_{t_0}^{t_m} b_{z,d}(x)\, dx \;=\; \frac{\delta_{z,d+1}}{d+1} \quad \text{for } -d \leq z < m.$$

A.5 Weak Convergence of Distributions

For $n = 1, 2, 3, \ldots$ let X_n be a random variable with distribution P_n on a metric space $(\mathcal{X}, d)$ (equipped with the σ-field of its Borel sets). Let X be an additional random variable with distribution P on $\mathcal{X}$. In what follows, asymptotic statements refer to $n \to \infty$, unless stated otherwise.

Convergence in distribution. One says that X_n *converges in distribution to* X and writes

$$X_n \;\to_{\mathcal{L}}\; X,$$

if

$$\lim_{n\to\infty} \mathbb{E}\, f(X_n) \;=\; \mathbb{E}\, f(X)$$

for arbitrary bounded and continuous functions $f : \mathcal{X} \to \mathbb{R}$.

In statistics, people often rephrase this statement as "X_n *is asymptotically distributed as* X" or "X_n *has asymptotic distribution* P". Sometimes we abuse notation slightly and write

$$X_n \;\to_{\mathcal{L}}\; P.$$

Weak convergence. Convergence in distribution in the sense above is equivalent to the following statement about the distributions P_n: One says that P_n *converges weakly to* P and writes

$$P_n \;\to_{w}\; P,$$

if

$$\lim_{n\to\infty} \int f\, dP_n \;=\; \int f\, dP$$

for arbitrary bounded and continuous functions $f : \mathcal{X} \to \mathbb{R}$.

Theorem A.8 (Portmanteau-Theorem). *The following statements about P and $(P_n)_n$ are equivalent:*

(i) *P_n converges weakly to P.*
(ii) *For arbitrary bounded and Lipschitz-continuous functions $f : (\mathcal{X}, d) \to [0, 1]$,*

$$\lim_{n\to\infty} \int f \, dP_n \;=\; \int f \, dP.$$

(iii$_a$) *For arbitrary open sets $U \subset \mathcal{X}$,*

$$\liminf_{n\to\infty} P_n(U) \;\geq\; P(U).$$

(iii$_b$) *For arbitrary closed sets $A \subset \mathcal{X}$,*

$$\limsup_{n\to\infty} P_n(A) \;\leq\; P(A).$$

(iv) *For arbitrary Borel sets $B \subset \mathcal{X}$,*

$$\lim_{n\to\infty} P_n(B) \;=\; P(B) \quad \textit{whenever } P(\partial B) = 0.$$

For the reader's convenience, a proof of the Portmanteau-Theorem is given at the end of this subsection.

The fact that the second statement implies both the first and third statements may be verified by means of the following exercise, a variation of Exercise 4.7.

Exercise A.9 (Approximation by Lipschitz-continuous functions). Let f be a function on a metric space $(\mathcal{X}, d)$ with values in $[a, \infty)$ for some $a \in \mathbb{R}$. For $L > 0$ and $x \in \mathcal{X}$ let

$$f_L(x) \;:=\; \inf_{y\in\mathcal{X}} \big(f(y) + L\, d(x, y)\big).$$

Show that $f_L(x)$ has the following properties:

(i) $a \leq f_L(x) \leq f(x)$.

(ii) $\big|f_L(x) - f_L(x')\big| \leq L\, d(x, x')$ for arbitrary $x, x' \in \mathcal{X}$.

(iii) $f_L(x)$ is non-decreasing in $L \geq 0$, and

$$f_\infty(x) := \lim_{L\uparrow\infty} f_L(x) \;=\; \liminf_{x'\to x} f(x').$$

(iv) Show that $\tilde{f}_L := \min(f_L, a + L)$ also has the properties (i-iii).

Exercise A.10 (Continuous Mapping Theorem). Let X and X_n ($n \in \mathbb{N}$) be random variables with values in a metric space $(\mathcal{X}, d_\mathcal{X})$ such that $X_n \to_\mathcal{L} X$ as $n \to \infty$. Further let H be a continuous mapping from $(\mathcal{X}, d_\mathcal{X})$ into another metric space $(\mathcal{Y}, d_\mathcal{Y})$. Show that

$$H(X_n) \;\to_\mathcal{L}\; H(X) \quad \text{as } n \to \infty.$$

Exercise A.11 (Fatou's Lemma and Scheffé's Theorem for weak convergence). Let P and P_n ($n \in \mathbb{N}$) be probability distributions on a metric space $(\mathcal{X}, d)$ such that $P_n \to_w P$ as $n \to \infty$.

(a) Let $h : \mathcal{X} \to [0, \infty)$ be lower semicontinuous. That means, $f(x) = \liminf_{x' \to x} f(x')$ for arbitrary $x \in \mathcal{X}$. Show that

$$\int h \, dP \;\leq\; \liminf_{n\to\infty} \int h \, dP_n.$$

Hint: Exercise A.9.

(b) Let $h : \mathcal{X} \to [0, \infty)$ be a lower semicontinuous and unbounded functions. Suppose that

$$\limsup_{n\to\infty} \int h \, dP_n \;\leq\; \int h \, dP \;<\; \infty.$$

Show that

$$\lim_{n\to\infty} \int g \, dP_n \;=\; \int g \, dP$$

for any continuous functions $g : \mathcal{X} \to [0, \infty)$ such that $\sup_{x\in\mathcal{X}} |g(x)|/(1 + h(x)) < \infty$.

Exercise A.12 (Cartesian products). Let $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ be metric spaces. Further, let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R} \times \mathbb{R}$ such that $\|(v_1, v_2)\|$ is non-decreasing in $|v_1|$ and $|v_2|$. Show that

$$d\big((x_1, y_1), (x_2, y_2)\big) \;:=\; \big\|\big(d_\mathcal{X}(x_1, x_2), d_\mathcal{Y}(y_1, y_2)\big)\big\|$$

defines a metric on $\mathcal{X} \times \mathcal{Y}$, and that that the resulting topology does not depend on the particular norm $\|\cdot\|$.

Exercise A.13 (Slutsky's Lemma).

(a) Consider metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$. The Cartesian product $\mathcal{X} \times \mathcal{Y}$ is equipped with the metric $d\big((x, y), (x', y')\big) := \max\big\{d_\mathcal{X}(x, x'), d_\mathcal{Y}(y, y')\big\}$. For $n = 1, 2, 3, \ldots$ let (X_n, Y_n) be a random variable with values in $\mathcal{X} \times \mathcal{Y}$ such that

$$X_n \;\to_\mathcal{L}\; X \quad \text{and} \quad Y_n \;\to_p\; y_o$$

for a random variable $X \in \mathcal{X}$ and a fixed point $y_o \in \mathcal{Y}$. Show that

$$(X_n, Y_n) \;\to_\mathcal{L}\; (X, y_o).$$

(b) For $n = 1, 2, 3, \ldots$ let (A_n, B_n, X_n) be a random variable with values in $\mathbb{R}^q \times \mathbb{R}^{q\times p} \times \mathbb{R}^p$ such that

$$A_n \;\to_p\; a, \quad B_n \;\to_p\; B \quad \text{and} \quad X_n \;\to_\mathcal{L}\; X$$

for a fixed vector $a \in \mathbb{R}^q$, a fixed Matrix $B \in \mathbb{R}^{q\times p}$ and a random variable $X \in \mathbb{R}^p$. Show that

$$A_n + B_n X_n \;\to_\mathcal{L}\; a + BX.$$

Weak convergence in $\mathbb{R}^d$. In the important special case of $\mathcal{X} = \mathbb{R}^d$ with standard Euclidean distance, the following statements are equivalent:

- The sequence $(P_n)_n$ converges weakly to P.
- For arbitrary *infinitely often differentiable* functions $f : \mathbb{R}^d \to \mathbb{R}$ such that f *and all its partial derivatives are bounded*,
$$\lim_{n\to\infty} \int f\, dP_n \;=\; \int f\, dP.$$
- For arbitrary $\boldsymbol{t} \in \mathbb{R}^d$,
$$\lim_{n\to\infty} \int \exp(\boldsymbol{i}\boldsymbol{t}^\top \boldsymbol{x})\, P_n(d\boldsymbol{x}) \;=\; \int \exp(\boldsymbol{i}\boldsymbol{t}^\top \boldsymbol{x})\, P(d\boldsymbol{x}).$$

To verify equivalence of the first two statements, note that if $f : \mathbb{R}^d \to \mathbb{R}$ is bounded and Lipschitz-continuous, then $\sup_{\boldsymbol{x}\in\mathbb{R}^d} \big|f_\epsilon(\boldsymbol{x}) - f(\boldsymbol{x})\big| \to 0$ as $\epsilon \downarrow 0$, where

$$f_\epsilon(\boldsymbol{x}) \;:=\; \mathbb{E}\, f(\boldsymbol{x} + \epsilon \boldsymbol{Z}), \quad \boldsymbol{Z} \sim \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{I}).$$

Indeed, if L is the Lipschitz constant of f, then

$$\big|f_\epsilon(\boldsymbol{x}) - f(\boldsymbol{x})\big| \;\le\; L\epsilon\, \mathbb{E}\, \|\boldsymbol{Z}\| \;\le\; L\epsilon\sqrt{d}.$$

On the other hand, with $C_d := (2\pi)^{-d/2}$,

$$\begin{aligned} f_\epsilon(\boldsymbol{x}) \;&=\; C_d \int_{\mathbb{R}^d} f(\boldsymbol{x} + \epsilon\boldsymbol{z}) \exp(-\|\boldsymbol{z}\|^2/2)\, d\boldsymbol{z} \\ &=\; C_d \epsilon^{-d} \int_{\mathbb{R}^d} f(\boldsymbol{y}) \exp\big(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/(2\epsilon^2)\big)\, d\boldsymbol{y}. \end{aligned}$$

By induction, one can show that for any tuple $\alpha \in \mathbb{N}_0^d$, there exists a d-variate polynomial $p_{\epsilon,\alpha} : \mathbb{R}^d \to \mathbb{R}$ of degree $\alpha_1 + \cdots + \alpha_d$ such that

$$\frac{\partial^{\alpha_1+\cdots+\alpha_d}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \exp\big(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/(2\epsilon^2)\big) \;=\; p_{\epsilon,\alpha}(\boldsymbol{x} - \boldsymbol{y}) \exp\big(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/(2\epsilon^2)\big),$$

and since the latter function is integrable, dominated convergence and an induction argument show that

$$\begin{aligned} \frac{\partial^{\alpha_1+\cdots+\alpha_d}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} f_\epsilon(\boldsymbol{x}) \;&=\; C_d \epsilon^{-d} \int_{\mathbb{R}^d} f(\boldsymbol{y}) p_{\epsilon,\alpha}(\boldsymbol{x} - \boldsymbol{y}) \exp\big(-\|\boldsymbol{x} - \boldsymbol{y}\|^2/(2\epsilon^2)\big)\, d\boldsymbol{y} \\ &=\; C_d \epsilon^{-d} \int_{\mathbb{R}^d} f(\boldsymbol{x} - \boldsymbol{v}) p_{\epsilon,\alpha}(\boldsymbol{v}) \exp\big(-\|\boldsymbol{v}\|^2/(2\epsilon^2)\big)\, d\boldsymbol{v}, \end{aligned}$$

obviously a bounded function of $\boldsymbol{x}$.

Weak convergence in $\mathbb{R}$. In case of $\mathcal{X} = \mathbb{R}$, convergence in distribution and weak convergence may be characterized in terms of the distribution functions F_n and F of X_n and X, respectively: For arbitrary $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F_n(x) \;=\; F(x) \quad \text{if } F(x-) = F(x). \tag{A.6}$$

If the limiting distribution function F is continuous, statement (A.6) is even equivalent to a uniform convergence:

$$\lim_{n\to\infty} \sup_{\text{intervals } B\subset\mathbb{R}} \big|P_n(B) - P(B)\big| \;=\; 0.$$

Proof of Theorem A.8. Obviously, statement (i) implies (ii).

To show that (ii) implies (iii$_a$), let $U \subset \mathcal{X}$ be an open set, and for $L > 0$ let

$$f_L(x) \;:=\; \min\{L\, d(x, \mathcal{X} \setminus U), 1\}$$

with $d(x, \mathcal{X} \setminus U) := \inf_{y \in \mathcal{X}\setminus U} d(x, y)$. Then $0 \le f_L \uparrow 1_U$ as $L \uparrow \infty$, and f_L is Lipschitz-continuous with constant L. Hence, (ii) implies that for any fixed L,

$$\liminf_{n\to\infty} P_n(U) \;\ge\; \liminf_{n\to\infty} \int f_L \, dP_n \;=\; \int f_L \, dP.$$

But $\int f_L \, dP \uparrow P(U)$ as $L \uparrow \infty$, so (iii$_a$) is satisfied as well.

That (iii$_a$) and (iii$_b$) are equivalent follows from the fact that the complement of an open set is closed and the complement of a closed set is open.

Suppose that (iii$_a$) and (iii$_b$) are satisfied. For any Borel set $B \subset \mathcal{X}$, note that its interior B^o, its closure $\overline{B}$ and its boundary ∂B satisfy the relations $B^o \subset B \subset \overline{B} = B^o \cup \partial B$. Consequently, if $P(\partial B) = 0$, then

$$\liminf_{n\to\infty} P_n(B) \;\ge\; \liminf_{n\to\infty} P_n(B^o) \;\overset{(iii_a)}{\ge}\; P(B^o) = P(B)$$

and

$$\limsup_{n\to\infty} P_n(B) \;\le\; \liminf_{n\to\infty} P_n(\overline{B}) \;\overset{(iii_b)}{\le}\; P(\overline{B}) = P(B),$$

whence (iv) is satisfied as well.

Finally suppose that property (iv) is satisfied. For a continuous function $f : \mathcal{X} \to [a, b]$,

$$\int f \, dP_{(n)} \; a + \int_a^b P_{(n)}(\{f \ge t\})\, dt.$$

Continuity of f implies that $\{f \ge t\}$ is closed and $\partial\{f \ge t\} \subset \{f = t\}$. But $P(\{f = t\}) > 0$ for at most countably many $t \in [a, b]$. Consequently, $P_n(\{f \ge t\}) \in [0, 1]$ converges to $P(\{f \ge t\})$ for all but at most countably many $t \in [a, b]$. Hence, by dominated convergence,

$$\int f \, dP_n = a + \int_a^b P_n(\{f \ge t\})\, dt \;\to\; a + \int_a^b P(\{f \ge t\})\, dt = \int f \, dP.$$

Thus, (i) is satisfied too. □

A.6 Lindeberg's Central Limit Theorem

The key message of Lindeberg's Central Limit Theorem is that the distribution of a sum of stochastically independent random variables is approximately Gaussian, if each summand has only little impact on the total sum. In what follows we provide precise statements in this vein.

A.6.1 The Univariate Case

Theorem A.14. *Let $Y_1, Y_2, \ldots, Y_n$ be stochastically independent random variables with mean 0 and $\sum_{i=1}^n \mathbb{E}(Y_i^2) = 1$. Then*

$$\sup_{x\in\mathbb{R}} \Big| \mathbb{P}\Big\{ \sum_{i=1}^n Y_i \le x \Big\} - \Phi(x) \Big| \;\le\; C\Lambda^{1/2}.$$

Here, C is a universal constant, and

$$\Lambda \;:=\; \sum_{i=1}^n \mathbb{E}\big(Y_i^2 \min(|Y_i|, 1)\big).$$

The quantity Λ measures the influence of the single summands Y_i on the total sum. Obviously, $\Lambda \le 1$. If, for instance, the modulus of each summand Y_i is bounded from above by some constant $\kappa \le 1$, then

$$\Lambda \;\le\; \sum_{i=1}^n \mathbb{E}\left(|Y_i|^3\right) \;\le\; \sum_{i=1}^n \mathbb{E}\left(\kappa Y_i^2\right) \;=\; \kappa.$$

Example A.15 (Binomial distributions). Let $X \sim \mathrm{Bin}(n,p)$ for some $n \in \mathbb{N}$ and $p \in (0,1)$. Then X has the same distribution as $\sum_{i=1}^n X_i$ with independent random variables $X_1, \ldots, X_n \in \{0,1\}$ such that $\mathbb{E}(X_i) = \mathbb{P}(X_i = 1) = p$. Here

$$\frac{X - np}{\sqrt{np(1-p)}} \;=_{\mathcal{L}}\; \sum_{i=1} Y_i \quad \text{with} \quad Y_i \;:=\; \frac{X_i - p}{\sqrt{np(1-p)}},$$

and the summands $Y_1, \ldots, Y_n$ satisfy the assumptions of Theorem A.14. In particular,

$$|Y_i| \;\le\; \frac{1}{\sqrt{np(1-p)}},$$

whence

$$\Lambda \;\le\; \frac{1}{\sqrt{np(1-p)}}.$$

Consequently, the distribution of X is approximately Gaussian if its variance, $\mathrm{Var}(X) = np(1-p)$, is large.

A proof of Theorem A.14 is given, for instance, in the monograph of Barbour and Chen (2005), utilizing a very intriguing technique of Charles Stein. As shown later, the original proof of Lindeberg leads to the bound $C\Lambda^{1/4}$ instead of $C\Lambda^{1/2}$. Lindeberg's idea is to replace the summands Y_i successively by independent, Gaussian summands. Roughly saying, $\mathbb{E}\big(Y_i^2 \min(1, |Y_i|)\big)$ bounds the approximation error resulting from replacing Y_i with a random variable with distribution $\mathrm{N}(0, \mathrm{Var}(Y_i))$.

A.6.2 The Multivariate Case

For the multivariate case we need a so-called triangular scheme of random vectors $\boldsymbol{Y}_{ni} \in \mathbb{R}^d$, $n \in \mathbb{N}$, $1 \le i \le n$. Again, asymptotic statements are meant as $n \to \infty$.

Theorem A.16. *For $n \in \mathbb{N}$ let $\boldsymbol{Y}_{n1}, \boldsymbol{Y}_{n2}, \dots, \boldsymbol{Y}_{nn}$ be stochastically independent random vectors in $\mathbb{R}^d$ such that*

$$\mathbb{E}(\boldsymbol{Y}_{ni}) \;=\; \mathbf{0} \quad \textit{and} \quad \mathbb{E}(\|\boldsymbol{Y}_{ni}\|^2) \;<\; \infty.$$

Further, suppose that

$$\boldsymbol{\Sigma}_n := \sum_{i=1}^n \mathbb{E}\big(\boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top\big) \;\to\; \boldsymbol{\Sigma},$$

$$\Lambda_n := \sum_{i=1}^n \mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^2 \min(1, \|\boldsymbol{Y}_{ni}\|)\big) \;\to\; 0.$$

This implies that

$$\sum_{i=1}^n \boldsymbol{Y}_{ni} \;\to_{\mathcal{L}}\; \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma}).$$

Moreover,

$$\mathbb{E}\Big\|\sum_{i=1}^n \boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top - \boldsymbol{\Sigma}_n\Big\|_F \;\to\; 0 \quad \textit{and} \quad \mathbb{E}\Big(\max_{i=1,2,\dots,n} \|\boldsymbol{Y}_{ni}\|^2\Big) \;\to\; 0.$$

Here $\|A\|_F$ denotes the Frobenius norm $\sqrt{\operatorname{trace}(A^\top A)}$ of a matrix A.

Again, Λ_n measures the influence of the single summands $\boldsymbol{Y}_{ni}$ on their sum. For instance, if

$$\|\boldsymbol{Y}_{ni}\| \;\le\; \kappa_n \quad \text{for } 1 \le i \le n,$$

then

$$\Lambda_n \;\le\; \kappa_n \sum_{i=1}^n \mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^2\big) \;=\; \kappa_n \operatorname{trace}(\boldsymbol{\Sigma}_n) \;=\; \kappa_n \cdot O(1).$$

Exercise A.17. For $n \in \mathbb{N}$ let $W_{n1}, \dots, W_{nn}$ be random variables with values in $[0, \infty)$ such that

$$\sum_{i=1}^n \mathbb{E}(W_{ni}) \;=\; O(1).$$

Show that the following three conditions are equivalent:

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}\big(W_{ni} \min(W_{ni}, 1)\big) \;=\; 0; \tag{i}$$

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}\big(W_{ni} \min(W_{ni}^\delta, 1)\big) \;=\; 0 \quad \text{for any fixed } \delta > 0; \tag{ii}$$

$$\lim_{n\to\infty} \sum_{i=1}^n \mathbb{E}(W_{ni} 1_{[W_{ni} \ge \varepsilon]}) \;=\; 0 \quad \text{for any fixed } \varepsilon > 0. \tag{iii}$$

Example A.18 (Multinomial distributions). For $n \in \mathbb{N}$ let $\boldsymbol{H}_n \sim \mathrm{Mult}(n, \boldsymbol{p}_n)$ with a probability vector $\boldsymbol{p}_n = (p_n(j))_{j=1}^K \in (0,1)^K$. Suppose that

$$\boldsymbol{p}_n \;\to\; \boldsymbol{p} \quad \text{and} \quad \min_{j=1,\dots,K} np_n(j) \;\to\; \infty$$

for a fixed probability vector $\boldsymbol{p} \in [0,1]^K$. It is possible that some components of $\boldsymbol{p}$ are zero, as long as the corresponding components of $\boldsymbol{p}_n$ converge sufficiently slowly to zero. These two assumptions imply that

$$n^{-1/2}(\boldsymbol{H}_n - n\boldsymbol{p}_n) \;\to_{\mathcal{L}}\; \mathrm{N}_K\big(\boldsymbol{0}, \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\big)$$

and

$$n^{-1/2}\,\mathrm{diag}(\boldsymbol{p}_n)^{-1/2}(\boldsymbol{H}_n - n\boldsymbol{p}_n) \;\to_{\mathcal{L}}\; \mathrm{N}_K\Big(\boldsymbol{0}, \boldsymbol{I} - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top\Big).$$

To verify this, we represent $\boldsymbol{H}_n$ as $\sum_{i=1}^n \boldsymbol{H}_{ni}$ with random vectors $\boldsymbol{H}_{ni} \in \{0,1\}^K$, where

$$\boldsymbol{H}_{ni} \;=\; \begin{cases} (1,0,\ldots,0,0)^\top & \text{with prob. } p_n(1), \\ (0,1,0,\ldots,0)^\top & \text{with prob. } p_n(2), \\ \vdots & \\ \big(0,0,\ldots,0,1\big)^\top & \text{with prob. } p_n(K). \end{cases}$$

But then $n^{-1/2}\boldsymbol{A}_n(\boldsymbol{H}_n - n\boldsymbol{p}_n) = \sum_{i=1}^n \boldsymbol{Y}_{ni}$ with

$$\boldsymbol{Y}_{ni} \;:=\; n^{-1/2}\boldsymbol{A}_n(\boldsymbol{H}_{ni} - \boldsymbol{p}_n),$$

where $\boldsymbol{A}_n = \boldsymbol{I}$ or $\boldsymbol{A}_n = \mathrm{diag}(\boldsymbol{p}_n)^{-1/2}$. One can easily verify that the assumptions of Theorem A.16 are satisfied with $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$ or $\boldsymbol{\Sigma} = \boldsymbol{I} - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top$, respectively. In particular,

$$\|\boldsymbol{Y}_{ni}\| \;\le\; \sqrt{2/n} \quad \text{or} \quad \|\boldsymbol{Y}_{ni}\| \;\le\; \Big(\min_{j=1,\ldots,K} np_n(j)\Big)^{-1/2},$$

because $\|\boldsymbol{H}_{ni} - \boldsymbol{p}_n\|^2 = 2 - 2\boldsymbol{H}_{ni}^\top\boldsymbol{p}_n \le 2$ and

$$\begin{aligned} &\big\|\,\mathrm{diag}(\boldsymbol{p}_n)^{-1/2}(\boldsymbol{H}_{ni} - \boldsymbol{p}_n)\big\|^2 \\ &\quad= \sum_{j=1}^K \Big(\frac{H_{ni}(j)^2}{p_n(j)} - 2H_{ni}(j) + p_n(j)\Big) \;=\; \sum_{j=1}^K \frac{H_{ni}(j)}{p_n(j)} - 1 \;\le\; \Big(\min_{j=1,\ldots,K} p_n(j)\Big)^{-1}. \end{aligned}$$

Exercise A.19. Let $\boldsymbol{p}$ and $\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \ldots$ be probability vectors in $\mathbb{R}^K$ (i.e. vectors with nonnegative entries summing to one). Further let $(\boldsymbol{M}_n)_{n\in\mathbb{N}}$ be a sequence of random vectors such that $\boldsymbol{M}_n \sim \mathrm{Mult}(n, \boldsymbol{p}_n)$, where

$$\boldsymbol{p}_n \to \boldsymbol{p} \quad \text{and} \quad \min_{1\le j\le K} np_n(j) \to \infty.$$

(a) Consider Pearson's chi-squared statistic

$$T_n \;:=\; n\sum_{j=1}^K \frac{(\widehat{p}_n(j) - p_n(j))^2}{p_n(j)}.$$

Show that

$$T_n \;\to_{\mathcal{L}}\; \chi^2_{K-1}.$$

(b) We define the Hellinger statistic as $H_n := 4n\big\|\sqrt{\widehat{\boldsymbol{p}}_n} - \sqrt{\boldsymbol{p}_n}\big\|^2$. Show that

(b.1) $$\max_{1\le j\le K}\left|\frac{\widehat{\boldsymbol{p}}_n(j)}{\boldsymbol{p}_n(j)} - 1\right| \;\to_p\; 0;$$

(b.2) $$H_n - T_n \;\to_p\; 0;$$

(b.3) $$2\sqrt{n}\big(\sqrt{\widehat{\boldsymbol{p}}_n} - \sqrt{\boldsymbol{p}_n}\big) \;\to_{\mathcal{L}}\; \mathrm{N}_K(\boldsymbol{0}, \boldsymbol{I} - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top).$$

Interpret the result (b.3) geometrically: Show that $\big\{\sqrt{\boldsymbol{p}} : \boldsymbol{p} \in [0,1]^K, \sum_{j=1}^K p(j) = 1\big\}$ is a certain subset of the unit sphere $\mathbb{R}^K$, and determine a linear subspace of $\mathbb{R}^K$ on which the limiting distribution $\mathrm{N}_K(\boldsymbol{0}, \boldsymbol{I} - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top)$ is concentrated.

Example A.20 (Sample means). Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots$ be independent, identically distributed random vectors in $\mathbb{R}^d$ such that $\mathbb{E}(\|\boldsymbol{X}_1\|^2) < \infty$. Let $\boldsymbol{\mu} := \mathbb{E}(\boldsymbol{X}_1)$ and $\boldsymbol{\Sigma} := \mathrm{Var}(\boldsymbol{X}_1)$. Then the sample means $\bar{\boldsymbol{X}}_n := n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$ satisfy the following limit theorem:

$$\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \;\to_{\mathcal{L}}\; \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

The reason is that $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) = \sum_{i=1}^n \boldsymbol{Y}_{ni}$ with

$$\boldsymbol{Y}_{ni} \;:=\; n^{-1/2}(\boldsymbol{X}_i - \boldsymbol{\mu}),$$

so $\mathbb{E}(\boldsymbol{Y}_{ni}) = \boldsymbol{0}$, $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}$, and

$$\Lambda_n \;=\; \mathbb{E}\big(\|\boldsymbol{X}_1 - \boldsymbol{\mu}\|^2 \min(n^{-1/2}\|\boldsymbol{X}_1 - \boldsymbol{\mu}\|, 1)\big) \;\to\; 0$$

by dominated convergence.

Proof of Theorem A.16. Let $\mathcal{G}$ be the family of all twice differentiable functions $g : \mathbb{R}^d \to \mathbb{R}$ such that its second derivative (Hessian matrix) $D^2 g$ satisfies the following conditions: For arbitrary $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\|D^2 g(\boldsymbol{x})\| \;\le\; 1 \quad \text{and} \quad \|D^2 g(\boldsymbol{x}) - D^2 g(\boldsymbol{y})\| \;\le\; 6\|\boldsymbol{x} - \boldsymbol{y}\|.$$

Furthermore, we extend the underlying probability space(s) for each n such that in addition to the random vectors $\boldsymbol{Y}_{ni}$ there exist random vectors $\boldsymbol{Z}_{n1}, \ldots, \boldsymbol{Z}_{nn}$ with distribution $\mathcal{L}(\boldsymbol{Z}_{ni}) = \mathrm{N}_d(\boldsymbol{0}, \mathrm{Var}(\boldsymbol{Y}_{ni}))$ such that all $2n$ random vectors $\boldsymbol{Y}_{ni}$ and $\boldsymbol{Z}_{ni}$ are stochastically independent. With these $2n$ random vectors we define the random sums

$$\boldsymbol{S}_n \;:=\; \sum_{i=1}^n \boldsymbol{Y}_{ni} \quad \text{and} \quad \boldsymbol{T}_n \;:=\; \sum_{i=1}^n \boldsymbol{Z}_{ni} \;\sim\; \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}_n).$$

In what follows, we shall prove that

(A.7) $$\big|\mathbb{E}\, g(\boldsymbol{S}_n) - \mathbb{E}\, g(\boldsymbol{T}_n)\big| \;\le\; K(d)\Lambda_n \quad \text{for all } g \in \mathcal{G}, \quad \text{if } \Lambda_n \le 1,$$

where $K(d) := 2^{3/2}(d^2 + 2d)^{3/4} + 1$. This finding, together with our general considerations about weak convergence, show that $\boldsymbol{S}_n$ converges in distribution to $\mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$.

To verify (A.7), we utilize Lindeberg's trick and define

$$\boldsymbol{U}_{nk} \;:=\; \sum_{i<k} \boldsymbol{Z}_{ni} + \sum_{i>k} \boldsymbol{Y}_{ni}$$

for $1 \le k \le n$. Then

$$\begin{aligned} \boldsymbol{S}_n &= \boldsymbol{U}_{n1} + \boldsymbol{Y}_{n1}, \\ \boldsymbol{U}_{nk} + \boldsymbol{Z}_{nk} &= \boldsymbol{U}_{n,k+1} + \boldsymbol{Y}_{n,k+1} \quad \text{for } 1 \le k < n, \\ \boldsymbol{T}_n &= \boldsymbol{U}_{nn} + \boldsymbol{Z}_{nn}. \end{aligned}$$

Consequently,

$$\begin{aligned} g(\boldsymbol{S}_n) - g(\boldsymbol{T}_n) &= \big(g(\boldsymbol{U}_{n1} + \boldsymbol{Y}_{n1}) - g(\boldsymbol{U}_{n1} + \boldsymbol{Z}_{n1})\big) \\ &\quad + \; g(\boldsymbol{U}_{n2} + \boldsymbol{Y}_{n2}) - g(\boldsymbol{U}_{nn} + \boldsymbol{Z}_{nn}) \\ &= \sum_{k=1}^{2} \big(g(\boldsymbol{U}_{nk} + \boldsymbol{Y}_{nk}) - g(\boldsymbol{U}_{nk} + \boldsymbol{Z}_{nk})\big) \\ &\quad + \; g(\boldsymbol{U}_{n3} + \boldsymbol{Y}_{n3}) - g(\boldsymbol{U}_{nn} + \boldsymbol{Z}_{nn}) \\ &= \ldots \\ &= \sum_{k=1}^{n} \big(g(\boldsymbol{U}_{nk} + \boldsymbol{Y}_{nk}) - g(\boldsymbol{U}_{nk} + \boldsymbol{Z}_{nk})\big), \end{aligned}$$

and it suffices to show that for $1 \le k \le n$,

$$\big|\mathbb{E}\big(g(\boldsymbol{U}_{nk} + \boldsymbol{Y}_{nk}) - g(\boldsymbol{U}_{nk} + \boldsymbol{Z}_{nk})\big)\big| \;\le\; K(d)\,\mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2 \min(1, \|\boldsymbol{Y}_{nk}\|)\big), \tag{A.8}$$

provided that $\Lambda_n \le 1$. To this end, we employ the following Taylor expansion of g: For arbitrary $\boldsymbol{u}, \boldsymbol{x} \in \mathbb{R}^d$,

$$g(\boldsymbol{u} + \boldsymbol{x}) \;=\; g(\boldsymbol{u}) + \nabla g(\boldsymbol{u})^\top \boldsymbol{x} + 2^{-1} \boldsymbol{x}^\top D^2 g(\boldsymbol{u}) \boldsymbol{x} + r(\boldsymbol{u}, \boldsymbol{x})$$

with

$$\begin{aligned} |r(\boldsymbol{u}, \boldsymbol{x})| &= \Big| \int_0^1 (1-t) \boldsymbol{x}^\top \big(D^2 g(\boldsymbol{u} + t\boldsymbol{x}) - D^2 g(\boldsymbol{u})\big) \boldsymbol{x}\, dt \Big| \\ &\le \|\boldsymbol{x}\|^2 \int_0^1 (1-t) \big\| D^2 g(\boldsymbol{u} + t\boldsymbol{x}) - D^2 g(\boldsymbol{u}) \big\|\, dt \\ &\le \|\boldsymbol{x}\|^2 \int_0^1 (1-t) \min(2, 6t\|\boldsymbol{x}\|)\, dt \\ &\le \|\boldsymbol{x}\|^2 \min(1, \|\boldsymbol{x}\|). \end{aligned}$$

Combining this expansion with stochastic independence of $\boldsymbol{U}_{nk}$ and $(\boldsymbol{Y}_{nk}, \boldsymbol{Z}_{nk})$, we obtain the

equation

$$\begin{aligned}
&\mathbb{E}\big(g(\boldsymbol{U}_{nk}+\boldsymbol{Y}_{nk})-g(\boldsymbol{U}_{nk}+\boldsymbol{Z}_{nk})\big)\\
&\quad= \mathbb{E}\big(\nabla g(\boldsymbol{U}_{nk})^\top(\boldsymbol{Y}_{nk}-\boldsymbol{Z}_{nk})\big)+2^{-1}\mathrm{trace}\big(\mathbb{E}\big(D^2 g(\boldsymbol{U}_{nk})(\boldsymbol{Y}_{nk}\boldsymbol{Y}_{nk}^\top-\boldsymbol{Z}_{nk}\boldsymbol{Z}_{nk}^\top)\big)\big)\\
&\qquad+\ \mathbb{E}\,r(\boldsymbol{U}_{nk},\boldsymbol{Y}_{nk})+\mathbb{E}\,r(\boldsymbol{U}_{nk},\boldsymbol{Z}_{nk})\\
&\quad= \mathbb{E}\big(\nabla g(\boldsymbol{U}_{nk})\big)^\top\underbrace{\mathbb{E}(\boldsymbol{Y}_{nk}-\boldsymbol{Z}_{nk})}_{=\boldsymbol{0}}+2^{-1}\mathrm{trace}\big(\mathbb{E}\big(D^2 g(\boldsymbol{U}_{nk})\big)\underbrace{\mathbb{E}(\boldsymbol{Y}_{nk}\boldsymbol{Y}_{nk}^\top-\boldsymbol{Z}_{nk}\boldsymbol{Z}_{nk}^\top)}_{=\boldsymbol{0}}\big)\\
&\qquad+\ \mathbb{E}\,r(\boldsymbol{U}_{nk},\boldsymbol{Y}_{nk})+\mathbb{E}\,r(\boldsymbol{U}_{nk},\boldsymbol{Z}_{nk})\\
&\quad= \mathbb{E}\,r(\boldsymbol{U}_{nk},\boldsymbol{Y}_{nk})+\mathbb{E}\,r(\boldsymbol{U}_{nk},\boldsymbol{Z}_{nk}).
\end{aligned}$$

Consequently,

$$\big|\mathbb{E}\big(g(\boldsymbol{U}_{nk}+\boldsymbol{Y}_{nk})-g(\boldsymbol{U}_{nk}+\boldsymbol{Z}_{nk})\big)\big|\ \le\ \mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2\min(1,\|\boldsymbol{Y}_{nk}\|)\big)+\mathbb{E}\big(\|\boldsymbol{Z}_{nk}\|^3\big),$$

and it suffices to show that

$$\mathbb{E}\big(\|\boldsymbol{Z}_{nk}\|^3\big)\ \le\ 2^{3/2}(d^2+2d)^{3/4}\,\mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2\min(1,\|\boldsymbol{Y}_{nk}\|)\big).$$

With $\boldsymbol{\Sigma}_{nk}:=\mathrm{Var}(\boldsymbol{Y}_{nk})=\mathrm{Var}(\boldsymbol{Z}_{nk})$ and $\boldsymbol{Z}\sim\mathrm{N}_d(\boldsymbol{0},\boldsymbol{I})$,

$$\begin{aligned}
\mathbb{E}\big(\|\boldsymbol{Z}_{nk}\|^3\big)\ &=\ \mathbb{E}\big(\|\boldsymbol{\Sigma}_{nk}^{1/2}\boldsymbol{Z}\|^3\big)\ \le\ \|\boldsymbol{\Sigma}_{nk}\|^{3/2}\,\mathbb{E}\big(\|\boldsymbol{Z}\|^3\big)\\
&\le\ \|\boldsymbol{\Sigma}_{nk}\|^{3/2}\,\mathbb{E}\big(\|\boldsymbol{Z}\|^4\big)^{3/4}\ =\ \|\boldsymbol{\Sigma}_{nk}\|^{3/2}(d^2+2d)^{3/4}.
\end{aligned}$$

Moreover, $\|\boldsymbol{\Sigma}_{nk}\|=\big\|\mathbb{E}(\boldsymbol{Y}_{nk}\boldsymbol{Y}_{nk}^\top)\big\|\le\mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2\big)$, and for arbitrary $\epsilon\in(0,1]$, the latter expectation is not greater than

$$\epsilon^2+\mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2 1_{[\|\boldsymbol{Y}_{nk}\|>\epsilon]}\big)\ \le\ \epsilon^2+\mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2\min(1,\|\boldsymbol{Y}_{nk}\|)\big)/\epsilon.$$

Setting $\epsilon:=\mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2\min(1,\|\boldsymbol{Y}_{nk}\|)\big)^{1/3}\le\Lambda_n^{1/2}$, then we obtain the upper bound

$$2\,\mathbb{E}\big(\|\boldsymbol{Y}_{nk}\|^2\min(1,\|\boldsymbol{Y}_{nk}\|)\big)^{2/3}$$

for $\|\boldsymbol{\Sigma}_{nk}\|$, provided that $\Lambda_n\le 1$.

Now we show that the expected value of $\big\|\sum_{i=1}^n(\boldsymbol{M}_{ni}-\mathbb{E}\,\boldsymbol{M}_{ni})\big\|_F$ converges to 0, where $\boldsymbol{M}_{ni}:=\boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top$. For this purpose we write $\boldsymbol{M}_{ni}=\boldsymbol{M}'_{ni}+\boldsymbol{M}''_{ni}$ with

$$\boldsymbol{M}'_{ni}\ :=\ \max(1-\|\boldsymbol{Y}_{ni}\|,0)\,\boldsymbol{M}_{ni}\quad\text{and}\quad\boldsymbol{M}''_{ni}\ :=\ \min(\|\boldsymbol{Y}_{ni}\|,1)\,\boldsymbol{M}_{ni}.$$

Then

$$\begin{aligned}
\mathbb{E}\Big\|\sum_{i=1}^n(\boldsymbol{M}_{ni}-\mathbb{E}\,\boldsymbol{M}_{ni})\Big\|_F\ &\le\ \mathbb{E}\Big\|\sum_{i=1}^n(\boldsymbol{M}'_{ni}-\mathbb{E}\,\boldsymbol{M}'_{ni})\Big\|_F+\mathbb{E}\Big\|\sum_{i=1}^n(\boldsymbol{M}''_{ni}-\mathbb{E}\,\boldsymbol{M}''_{ni})\Big\|_F\\
&\le\ \Big(\mathbb{E}\Big(\Big\|\sum_{i=1}^n(\boldsymbol{M}'_{ni}-\mathbb{E}\,\boldsymbol{M}'_{ni})\Big\|_F^2\Big)\Big)^{1/2}+2\sum_{i=1}^n\mathbb{E}\,\|\boldsymbol{M}''_{ni}\|_F\\
&\le\ \Big(\sum_{i=1}^n\mathbb{E}\big(\|\boldsymbol{M}'_{ni}\|_F^2\big)\Big)^{1/2}+2\sum_{i=1}^n\mathbb{E}\,\|\boldsymbol{M}''_{ni}\|_F.
\end{aligned}$$

But

$$\begin{aligned}\|\boldsymbol{M}'_{ni}\|_F^2 &= \max(1-\|\boldsymbol{Y}_{ni}\|,0)^2\|\boldsymbol{Y}_{ni}\|^4 \le \min(\|\boldsymbol{Y}_{ni}\|,1)\|\boldsymbol{Y}_{ni}\|^2/4,\\ \|\boldsymbol{M}''_{ni}\|_F &= \min(\|\boldsymbol{Y}_{ni}\|,1)\|\boldsymbol{Y}_{ni}\|^2,\end{aligned}$$

whence $\mathbb{E}\big\|\sum_{i=1}^n(\boldsymbol{M}_{ni}-\mathbb{E}\,\boldsymbol{M}_{ni})\big\|_F \le \Lambda_n^{1/2}/2+2\Lambda_n \to 0$.

It remains to show that that the expected value of $\max_{i=1,\dots,n}\|\boldsymbol{Y}_{ni}\|^2$ converges to 0. To this end, we choose an arbitrary number $\varepsilon_n \in (0,1]$ and write

$$\mathbb{E}\Big(\max_{i=1,\dots,n}\|\boldsymbol{Y}_{ni}\|^2\Big) \le \mathbb{E}\Big(\varepsilon_n^2+\sum_{i=1}^n\|\boldsymbol{Y}_{ni}\|^2\min(\|\boldsymbol{Y}_{ni}\|,1)/\varepsilon_n\Big) = \varepsilon_n^2+\Lambda_n/\varepsilon_n.$$

Setting $\varepsilon_n=\Lambda_n^{1/3}$ in case of $\Lambda_n\le 1$ leads to the bound $2\Lambda_n^{2/3}$ for the expected value in question. □

Proof of a weaker version of Theorem A.14. The proof of Theorem A.16 reveals that in the setting of Theorem A.14,

$$\big|\mathbb{E}\,g(S)-\mathbb{E}\,g(Z)\big| \le K(1)\Lambda \quad \text{for all } g\in\mathcal{G}, \tag{A.9}$$

where $S:=\sum_{i=1}^n Y_i$, $Z\sim\mathrm{N}(0,1)$, and $\mathcal{G}$ is the set of all twice differentiable functions $g:\mathbb{R}\to\mathbb{R}$ such that for arbitrary $x,y\in\mathbb{R}$,

$$|g(x)| \le 1 \quad\text{and}\quad \big|g''(x)-g''(y)\big| \le 6|x-y|.$$

Now let $H:\mathbb{R}\to[0,1]$ be three times differentiable with bounded third derivative such that $H=1$ on $(-\infty,0]$ and $H=0$ on $[1,\infty)$. Then $cH\in\mathcal{G}$ for a suitable constant $c>0$. For $x\in\mathbb{R}$ and $\varepsilon>0$,

$$\begin{aligned}\mathbb{P}(S\le x) &\le \mathbb{E}\,H((S-x)/\varepsilon)\\ &\le \big|\mathbb{E}\,H((S-x)/\varepsilon)-\mathbb{E}\,H((Z-x)/\varepsilon)\big|+\mathbb{P}(Z\le x+\varepsilon)\\ &\le c^{-1}\varepsilon^{-3}K(1)\Lambda+\Phi(x+\varepsilon)\\ &\le c^{-1}\varepsilon^{-3}K(1)\Lambda+(2\pi)^{-1/2}\varepsilon+\Phi(x),\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(S\le x) &\ge \mathbb{E}\,H((S+\varepsilon-x)/\varepsilon)\\ &\ge -\big|\mathbb{E}\,H((S+\varepsilon-x)/\varepsilon)-\mathbb{E}\,H((Z+\varepsilon-x)/\varepsilon)\big|+\mathbb{P}(Z\le x-\varepsilon)\\ &\ge -c^{-1}\varepsilon^{-3}K(1)\Lambda\Phi(x-\varepsilon)\\ &\ge -c^{-1}\varepsilon^{-3}K(1)\Lambda-(2\pi)^{-1/2}\varepsilon+\Phi(x),\end{aligned}$$

because $c\varepsilon^3H((\cdot-a)/\varepsilon)\in\mathcal{G}$ for arbitrary $a\in\mathbb{R}$, and $0<\Phi'\le\Phi'(0)=(2\pi)^{-1/2}$. Consequently,

$$\sup_{x\in\mathbb{R}}\big|\mathbb{P}(S\le x)-\Phi(x)\big| \le c^{-1}\varepsilon^{-3}K(1)\Lambda+(2\pi)^{-1/2}\varepsilon.$$

If we set $\varepsilon:=D\lambda^{1/4}$ for some $D>0$, the latter bound equals $C\Lambda^{1/4}$ for some $C>0$. □

Exercise A.21. Let Z be a real-valued random variable with $\mathbb{E}(Z) = 0$ and $\mathbb{E}(Z^2) = 1$. Show that for any constant $\sigma \geq 0$ and any twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ with bounded second derivative f'',

$$\big|\mathbb{E}\, f(\sigma Z) - \mathbb{E}\, f(Z)\big| \;\leq\; |\sigma^2 - 1| \|f''\|_\infty / 2.$$

A.7 Iteratively Reweighted Least Squares

The QR decomposition and least squares methods are essential ingredients in various regression methods. The general framework is as follows: Let $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n]^\top \in \mathbb{R}^{n \times p}$ with $\operatorname{rank}(\boldsymbol{D}) = p$. For $i = 1, 2, \ldots, n$ let $H_i : \mathbb{R} \to \mathbb{R}$ be a twice continuously differentiable function such that $H_i'' > 0$. Now we consider the target function $L : \mathbb{R}^p \to \mathbb{R}$ given by

$$L(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^{n} H_i(\boldsymbol{d}_i^\top \boldsymbol{\theta}). \tag{A.10}$$

Example A.22 (Least squares). For a vector $\boldsymbol{Y} \in \mathbb{R}^n$ let

$$H_i(x) \;:=\; (Y_i - x)^2.$$

Then

$$H_i'(x) \;=\; 2(x - Y_i) \quad \text{and} \quad H_i''(x) \;=\; 2.$$

Example A.23 (Logistic regression). For a vector $\boldsymbol{Y} \in [0,1]^n$ let

$$H_i(x) \;:=\; -Y_i x + \log(1 + e^x).$$

Then

$$H_i'(x) \;=\; \ell(x) - Y_i \quad \text{and} \quad H_i''(x) \;=\; \ell(x)(1 - \ell(x))$$

with the logistic function

$$\ell(x) \;:=\; \frac{e^x}{1 + e^x} \;=\; \frac{1}{e^{-x} + 1} \;\in\; (0, 1).$$

The target function L in (A.10) is twice continuously differentiable with gradient

$$\nabla L(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} H_i'(\boldsymbol{d}_i^\top \boldsymbol{\theta}) \boldsymbol{d}_i,$$

and Hessian matrix

$$D^2 L(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} H_i''(\boldsymbol{d}_i^\top \boldsymbol{\theta}) \boldsymbol{d}_i \boldsymbol{d}_i^\top.$$

Hence, the second order Taylor expansion of $L(\boldsymbol{\theta}+\cdot)$ reads

$$\begin{aligned} L(\boldsymbol{\theta}+\boldsymbol{v}) \;&\approx\; L(\boldsymbol{\theta}) + \nabla L(\boldsymbol{\theta})^\top \boldsymbol{v} + \frac{1}{2}\,\boldsymbol{v}^\top D^2 L(\boldsymbol{\theta})\boldsymbol{v} \\ &=\; L(\boldsymbol{\theta}) + \sum_{i=1}^{n}\Big(H_i'(\boldsymbol{d}_i^\top\boldsymbol{\theta})\boldsymbol{d}_i^\top\boldsymbol{v} + \frac{1}{2}\,H_i''(\boldsymbol{d}_i^\top\boldsymbol{\theta})(\boldsymbol{d}_i^\top\boldsymbol{v})^2\Big) \\ &=\; L(\boldsymbol{\theta}) - \frac{1}{2}\sum_{i=1}^{n}\frac{H_i'(\boldsymbol{d}_i^\top\boldsymbol{\theta})^2}{H_i''(\boldsymbol{d}_i^\top\boldsymbol{\theta})} + \frac{1}{2}\sum_{i=1}^{n}\Big(\frac{H_i'(\boldsymbol{d}_i^\top\boldsymbol{\theta})}{\sqrt{H_i''(\boldsymbol{d}_i^\top\boldsymbol{\theta})}} + \sqrt{H_i''(\boldsymbol{d}_i^\top\boldsymbol{\theta})}\,\boldsymbol{d}_i^\top\boldsymbol{v}\Big)^2 \\ &=\; c(\boldsymbol{\theta}) + \frac{1}{2}\big\|\boldsymbol{Y}(\boldsymbol{\theta}) - \boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{v}\big\|^2, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{Y}(\boldsymbol{\theta}) \;&:=\; \Big(\frac{-H_1'(\boldsymbol{d}_1^\top\boldsymbol{\theta})}{\sqrt{H_1''(\boldsymbol{d}_1^\top\boldsymbol{\theta})}}, \frac{-H_2'(\boldsymbol{d}_2^\top\boldsymbol{\theta})}{\sqrt{H_2''(\boldsymbol{d}_2^\top\boldsymbol{\theta})}}, \ldots, \frac{-H_n'(\boldsymbol{d}_n^\top\boldsymbol{\theta})}{\sqrt{H_n''(\boldsymbol{d}_n^\top\boldsymbol{\theta})}}\Big)^\top \;\in\; \mathbb{R}^n, \\ \boldsymbol{D}(\boldsymbol{\theta}) \;&:=\; \Big[\sqrt{H_1''(\boldsymbol{d}_1^\top\boldsymbol{\theta})}\,\boldsymbol{d}_1, \sqrt{H_2''(\boldsymbol{d}_2^\top\boldsymbol{\theta})}\,\boldsymbol{d}_2, \ldots, \sqrt{H_n''(\boldsymbol{d}_n^\top\boldsymbol{\theta})}\,\boldsymbol{d}_n\Big]^\top \;\in\; \mathbb{R}^{n\times p}. \end{aligned}$$

Consequently, one step of the Newton–Raphson procedure is given by

$$\boldsymbol{\theta} \;\mapsto\; \boldsymbol{\theta} + \boldsymbol{v}(\boldsymbol{\theta}) \quad \text{with} \quad \boldsymbol{v}(\boldsymbol{\theta}) \;:=\; \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathbb{R}^p}\big\|\boldsymbol{Y}(\boldsymbol{\theta}) - \boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{v}\big\|^2,$$

and $\boldsymbol{v}(\boldsymbol{\theta})$ may be computed by means of the QR decomposition. Furthermore, the directional derivative $\nabla L(\boldsymbol{\theta})^\top\boldsymbol{v}(\boldsymbol{\theta})$ equals

$$\nabla L(\boldsymbol{\theta})^\top\boldsymbol{v}(\boldsymbol{\theta}) \;=\; -\boldsymbol{Y}(\boldsymbol{\theta})^\top\boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{v}(\boldsymbol{\theta}) \;=\; -\|\boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{v}(\boldsymbol{\theta})\|^2.$$

A.8 Couplings and Mallows Distances

In the context of bootstrap procedures and elsewhere, measures of distance between probability distributions are useful. In the present section we consider a general class of such distance measures which are based on *couplings* and related to weak convergence. Throughout this section let $(\mathcal{X}, d)$ be a complete and separable metric space, equipped with its Borel σ-field.

A.8.1 Optimal Transport

We use repeatedly the fact that $\mathcal{X}\times\mathcal{X}$ is also a complete, separable metric space if equipped with the metric

$$d\big((x_1,y_1),(x_2,y_2)\big) \;:=\; \max\big(d(x_1,x_2), d(y_1,y_2)\big)$$

or

$$d\big((x_1,y_1),(x_2,y_2)\big) \;:=\; \sqrt{d(x_1,x_2)^2 + d(y_1,y_2)^2},$$

for instance. And then

$$\mathrm{Borel}(\mathcal{X}\times\mathcal{X}) \;=\; \mathrm{Borel}(\mathcal{X}) \otimes \mathrm{Borel}(\mathcal{X}).$$

Exercise A.24 (Cartesian products). As in Exercise A.12, let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be metric spaces, and for $(x_1, y_1), (x_2, y_2) \in \mathcal{X} \times \mathcal{Y}$ set

$$d\big((x_1, y_1), (x_2, y_2)\big) \;:=\; \big\|\big(d_{\mathcal{X}}(x_1, x_2), d_{\mathcal{Y}}(y_1, y_2)\big)\big\|$$

with an arbitrary norm $\|\cdot\|$ on $\mathbb{R} \times \mathbb{R}$ such that $\|(v_1, v_2)\|$ is non-decreasing in $|v_1|$ and $|v_2|$.

(a) Suppose that both spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are separable. Show that $(\mathcal{X} \times \mathcal{Y}, d)$ is separable, too, and that

$$\mathrm{Borel}(\mathcal{X} \times \mathcal{X}, d) \;=\; \mathrm{Borel}(\mathcal{X}, d_{\mathcal{X}}) \otimes \mathrm{Borel}(\mathcal{Y}, d_{\mathcal{Y}}).$$

(b) Show that $(\mathcal{X} \times \mathcal{Y}, d)$ is complete, provided that both spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ are complete.

For two probability distributions P and Q on $\mathcal{X}$, let $\mathcal{R}(P, Q)$ be the set of all probability distributions R on $\mathcal{X} \times \mathcal{X}$ such that for $(X, Y) \sim R$,

$$X \sim P \;\text{ and }\; Y \sim Q.$$

A distribution $R \in \mathcal{R}(P, Q)$ is a *coupling* of P and Q. A simple but not too useful coupling is the product measure

$$P \otimes Q.$$

Now we consider a continuous cost function $C : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ and seek to minimize $\int C\, dR$ over all couplings $R \in \mathcal{R}(P, Q)$. This is a so-called *transport problem*: We may interpret P and Q as mass distributions on $\mathcal{X}$, and $C(x, y)$ are the costs of transferring a unit mass from point x to point y. A distribution $R \in \mathcal{R}(P, Q)$ may be interpreted as a transport plan: For Borel sets $A, B \subset \mathcal{X}$, $R(A \times B)$ specifies which part of the mass $P(A)$ is transported from the set A into the set B.

The following theorem shows that there is always an optimal coupling of P and Q:

Theorem A.25 (Optimal transport). *There exists a distribution $R_o \in \mathcal{R}(P, Q)$ such that*

$$\int C\, dR_o \;=\; \inf_{R \in \mathcal{R}(P,Q)} \int C\, dR.$$

To prove Theorem A.25, we use Prohorov's Theorem:

Theorem A.26 (Prohorov).

(i) *Let P be a probability measure on $\mathcal{X}$. Then for any $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $P(K) \geq 1 - \varepsilon$.*

(ii) *Let $(P_n)_n$ be a sequence of probability distributions on $\mathcal{X}$. Suppose that for any $\varepsilon > 0$ there exists a compact set $K \subset \mathcal{X}$ such that $P_n(K) \geq 1 - \varepsilon$ for all $n \in \mathbb{N}$. Then there exists a subsequence $(P_{n(k)})_k$ of $(P_n)_n$ which converges weakly to some probability distribution on $\mathcal{X}$.*

Proof of Theorem A.25. Let c_* be the infimum of $\int C\, dR$ over all $R \in \mathcal{R}(P, Q)$. In case of $c_* = \infty$, the assertion is trivial; we could just choose $R_o = P \otimes Q$. Hence let $c_* < \infty$, and let

$(R_n)_n$ be a sequence of distributions in $\mathcal{R}(P,Q)$ such that $\int C\,dR_n \to c_*$. According to part (i) of Theorem A.26, for any $\varepsilon > 0$ there exist compact sets $K_1, K_2 \subset \mathcal{X}$ such that $P(K_1), Q(K_2) \geq 1 - \varepsilon/2$. But then, $K := K_1 \times K_2$ is a compact subset of $\mathcal{X} \times \mathcal{X}$ such that

$$R(K) \;\geq\; 1 - P(\mathcal{X} \setminus K_1) - Q(\mathcal{X} \setminus K_2) \;\geq\; 1 - \varepsilon$$

for all $R \in \mathcal{R}(P,Q)$. According to part (ii) of Theorem A.26 (applied to $\mathcal{X} \times \mathcal{X}$ instead of $\mathcal{X}$), we may replace the sequence $(R_n)_n$ with a subsequence, if necessary, such that it converges weakly to some distribution R_o on $\mathcal{X} \times \mathcal{X}$. The Continuous Mapping Theorem, applied to the continuous projections $\mathcal{X} \times \mathcal{X} \ni (x,y) \mapsto x \in \mathcal{X}$ and $\mathcal{X} \times \mathcal{X} \ni (x,y) \mapsto y \in \mathcal{X}$, implies that $R_o \in \mathcal{R}(P,Q)$, too. Moreover, Exercise A.11 implies that

$$\int C\,dR_o \;\leq\; \liminf_{n\to\infty} \int C\,dR_n \;=\; c_*,$$

whence $c_* = \int C\,dR_o$. □

A.8.2 Optimal Transport on the Real Line

The proof of Theorem A.25 does not provide an explicit optimal coupling. Indeed, the determination of optimal couplings is still an active area of research. But in the special case of $\mathcal{X} = \mathbb{R}$ and

$$C(x,y) \;=\; \phi(x-y)$$

with a convex function $\phi : \mathbb{R} \to [0,\infty)$, there is an explicit version of the optimal coupling based on quantile functions:

Theorem A.27 (Optimal coupling in $\mathbb{R}$). *Let P and Q be probability measures on $\mathbb{R}$ with distribution functions F and G, respectively. Then*

$$\int \phi(x-y)\,R(dx,dy) \;\geq\; \int_0^1 \phi\big(F^{-1}(u) - G^{-1}(u)\big)\,du$$

for arbitrary couplings $R \in \mathcal{R}(P,Q)$. Here, F^{-1} and G^{-1} are the quantile functions of P and Q, respectively.

Theorem A.27 shows that an optimal coupling of the distributions P and Q is given by

$$R_o \;:=\; \mathcal{L}\big(F^{-1}(U), G^{-1}(U)\big) \quad \text{with } U \sim \text{ Unif}(0,1).$$

As a preparation of the proof of Theorem A.27, we pose two exercises:

Exercise A.28. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function.

(a) Show that for arbitrary numbers $a_1 < a_2$ and $b_1 < b_2$,

$$\phi(a_1 - b_2) + \phi(a_2 - b_1) \;\geq\; \phi(a_1 - b_1) + \phi(a_2 - b_2).$$

(b) Show that for arbitrary integers $n \geq 2$, real numbers $a_1 \leq \cdots \leq a_n$ and $b_1 \leq \cdots \leq b_n$, and permutations σ of $\{1, 2, \ldots, n\}$,

$$\sum_{i=1}^n \phi(a_i - b_{\sigma(i)}) \;\geq\; \sum_{i=1}^n \phi(a_i - b_i).$$

Exercise A.29. Let F be a distribution function on the real line. Show that F^{-1} is continuous at $u \in (0,1)$ if and only if $F < u$ on $(-\infty, F^{-1}(u))$ and $F > u$ on $(F^{-1}(u), \infty)$.

Proof of Theorem A.27. Let R be an arbitrary coupling of P and Q such that $\int C\,dR < \infty$. Now consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with independent random variables

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots \sim R \quad \text{and} \quad U \sim \mathrm{Unif}((0,1)).$$

For any integer $n \geq 2$, consider the order statistics $X_{n:1} \leq \cdots \leq X_{n:n}$ of $X_1, \ldots, X_n$, and $Y_{n:1} \leq \cdots \leq Y_{n:n}$ of $Y_1, \ldots, Y_n$. Then it follows from Exercise A.28, Fubini's theorem and Fatou's lemma that

$$\begin{aligned}\int C\,dR \;=\; \lim_{n\to\infty} \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} \phi(X_i - Y_i)\Big) \;&\geq\; \liminf_{n\to\infty} \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} \phi(X_{n:i} - Y_{n:i})\Big) \\ &=\; \liminf_{n\to\infty} \mathbb{E}\big(\phi(X_{n:\lceil nU\rceil} - Y_{n:\lceil nU\rceil})\big) \\ &\geq\; \mathbb{E}\Big(\liminf_{n\to\infty}\, \phi\big(X_{n:\lceil nU\rceil} - Y_{n:\lceil nU\rceil}\big)\Big).\end{aligned}$$

Hence, since ϕ is continuous, it suffices to show that

$$\lim_{n\to\infty} X_{n:\lceil nU\rceil} \;=\; F^{-1}(U) \quad \text{and} \quad \lim_{n\to\infty} Y_{n:\lceil nU\rceil} \;=\; G^{-1}(U)$$

almost surely. It suffices to verify the former statement; the latter one follows analogously. Since F^{-1} is monotone increasing on $(0,1)$, the set of points $u \in (0,1)$ at which F^{-1} is discontinuous is at most countably infinite. Hence, with probability one, U is a continuity point of F^{-1}. This is equivalent to saying that with probability one,

$$F \begin{cases} < U & \text{on } (-\infty, F^{-1}(U)), \\ > U & \text{on } (F^{-1}(U), \infty). \end{cases}$$

Moreover, if $\widehat{F}_n$ denotes the empirical distribution function of $X_1, \ldots, X_n$, it is well-known that

$$\lim_{n\to\infty} \|\widehat{F}_n - F\|_\infty \;=\; 0$$

almost surely. Thus, with probability one, for any $\varepsilon > 0$ there exists a random index $N(\varepsilon)$ such that

$$\widehat{F}_n(F^{-1}(U) - \varepsilon) \;<\; U \;<\; \widehat{F}_n(F^{-1}(U) + \varepsilon) \quad \text{for all } n \geq N(\varepsilon).$$

But with $K := \lceil nU \rceil$,

$$\widehat{F}_n(X_{n:K} -) \;\leq\; (K-1)/n \;<\; U \quad \text{and} \quad \widehat{F}_n(X_{n:K}) \;\geq\; K/n \;\geq\; U,$$

so the inequalites $\widehat{F}_n(F^{-1}(U) - \varepsilon) < U < \widehat{F}_n(F^{-1}(U) + \varepsilon)$ imply that

$$F^{-1}(U) - \varepsilon \;<\; X_{n:\lceil nU\rceil} \;\leq\; F^{-1}(U) + \varepsilon.$$

□

A.8.3 Mallows Distances

An important special case of the cost function $C : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is

$$C(x, y) \ := \ d(x, y)^k$$

for some $k \geq 1$.

Theorem A.30 (Mallows distance). *Let P and Q be probability measures on $\mathcal{X}$. For $k \geq 1$, the minimum*

$$d_{M,k}(P, Q) \ := \ \min_{R \in \mathcal{R}(P,Q)} \Big(\int d(x, y)^k \, R(dx, dy) \Big)^{1/k} \ \in \ [0, \infty]$$

exists. It defines a metric $d_{M,k}(\cdot, \cdot)$, the Mallows distance with exponent k, on the space of all probability measures on $\mathcal{X}$.

In case of $\mathcal{X} = \mathbb{R}$ and $d(x, y) := |x - y|$, one can apply Theorem A.27 and obtains the explicit formula

$$d_{M,k}(P, Q) \ = \ \left(\int_0^1 \left| F^{-1}(u) - G^{-1}(u) \right|^k du \right)^{1/k},$$

where F and G are the distribution functions of P and Q, respectively.

The next theorem establishes for two special cases a connection between weak convergence and convergence with respect to Mallows distances.

Theorem A.31. *Let $P_1, P_2, P_3, \ldots$ and P be probability distributions on $(\mathcal{X}, d)$ and k an arbitrary number in $[1, \infty)$.*

***(a)** In case of a bounded metric $d(\cdot, \cdot)$,*

$$\lim_{n \to \infty} \ d_{M,k}(P_n, P) \ = \ 0$$

if and only if

$$P_n \ \to_w \ P.$$

***(b)** Let $(\mathcal{X}, \| \cdot \|)$ be a separable Banach space with corresponding metric $d(x, y) := \|x - y\|$. Further let $\int \|x\|^k \, P(dx) < \infty$ and $\int \|x\|^k \, P_n(dx) < \infty$ for all $n \geq 1$. Then the following three statements are equivalent:*

$$\lim_{n \to \infty} \ d_{M,k}(P_n, P) \ = \ 0; \tag{b.1}$$

$$P_n \ \to_w \ P \quad \text{and} \quad \int \|x\|^k \, P_n(dx) \ \to \ \int \|x\|^k \, P(dx); \tag{b.2}$$

$$\lim_{n \to \infty} \int f \, dP_n \ = \ \int f \, dP \quad \text{for all } f \in \mathcal{C}(\mathcal{X}) \text{ such that } \sup_{x \in \mathcal{X}} \frac{|f(x)|}{1 + \|x\|^k} < \infty. \tag{b.3}$$

Remark. The function $(x, y) \mapsto \min(d(x, y), 1)$ defines another metric on $\mathcal{X}$, and the topologies generated by these metrics coincide; see Exercise A.32. Hence, part (a) of Theorem A.31 shows that weak convergence of probability measures is equivalent to convergence with respect to the metric

$$(P, Q) \ \mapsto \ \min_{R \in \mathcal{R}(P,Q)} \Big(\int \min(d(x, y), 1)^k \, R(dx, dy) \Big)^{1/k}.$$

Exercise A.32. Let $(\mathcal{X}, d)$ be a metric space, and let $\tau : [0, \infty] \to [0, \infty]$ be a monotone increasing function such that $\lim_{r \to 0} \tau(r) = 0$, but $\tau(r) > 0$ for arbitrary $r > 0$. Furthermore, let τ be subadditive, i.e.

$$\tau(r + s) \ \leq \ \tau(r) + \tau(s) \quad \text{for arbitrary } r, s \geq 0.$$

Show that $\tau \circ d$ defines a metric on $\mathcal{X}$ which induces the same topology as d. That means, a set $U \subset \mathcal{X}$ is open with respect to d if and only if it is open with respect to $\tau \circ d$.

Proof of Theorem A.30. The existence of an optimal coupling of P and Q is a consequence of Theorem A.25. It remains to show that $d_{M,k}(\cdot, \cdot)$ is a metric.

Obviously, $d_{M,k}(P, Q) = d_{M,k}(Q, P) \geq 0$. And it follows from $d_{M,k}(P, Q) = 0$ that there exists a random variable $(X, Y) \in \mathcal{X} \times \mathcal{X}$ such that $X \sim P$, $Y \sim Q$ and $X = Y$ almost surely. But this implies that $P = Q$. Hence, it remains to show that $d_{M,k}(\cdot, \cdot)$ satisfies the triangle inequality, i.e.

$$d_{M,k}(P_1, P_2) \ \leq \ d_{M,k}(P_1, P_0) + d_{M,k}(P_2, P_0)$$

for arbitrary probability distributions P_0, P_1, P_2 on $\mathcal{X}$. To this end, let $R_1 \in \mathcal{R}(P_0, P_1)$ and $R_2 \in \mathcal{R}(P_0, P_2)$ such that

$$\int d(x, y)^k \, R_j(dx, dy) \ = \ d_{M,k}(P_0, P_j)^k \quad \text{for } j = 1, 2.$$

Now we choose Markov kernels $K_j : \mathcal{X} \times \mathrm{Borel}(\mathcal{X}) \to [0, 1]$ such that

$$R_j(B \times C) \ = \ \int_B K_j(x, C) \, P_0(dx) \quad \text{for all } B, C \in \mathrm{Borel}(\mathcal{X}).$$

Then we define a probability distribution $\bar{R}$ on $\mathcal{X} \times \mathcal{X} \times \mathcal{X}$ via

$$\bar{R}(B_0 \times B_1 \times B_2) \ := \ \int_{B_0} K_1(x, B_1) K_2(x, B_2) \, P_0(dx).$$

With the random variables $X_s(\omega_0, \omega_1, \omega_2) := \omega_s$ on $\mathcal{X} \times \mathcal{X} \times \mathcal{X}$, the pair (X_0, X_j) has distribution R_j. Then the triangle inequalities for $d(\cdot, \cdot)$ and for L^k-spaces yield that

$$\begin{aligned} d_{M,k}(P_1, P_2) \ &\leq \ \Big(\mathbb{E}\big[d(X_1, X_2)^k \big] \Big)^{1/k} \\ &\leq \ \Big(\mathbb{E}\big[\big(d(X_0, X_1) + d(X_0, X_2) \big)^k \big] \Big)^{1/k} \\ &\leq \ \Big(\mathbb{E}\big[d(X_0, X_1)^k \big] \Big)^{1/k} + \Big(\mathbb{E}\big[d(X_0, X_2)^k \big] \Big)^{1/k} \\ &= \ d_{M,k}(P_0, P_1) + d_{M,k}(P_0, P_2). \end{aligned}$$

□

Proof of Theorem A.31. The proof uses Skorohod's Theorem: The sequence $(P_n)_n$ converges weakly to P if and only if there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ carrying $\mathcal{X}$-valued random variables $X_1, X_2, X_3, \ldots$ and X such that $X \sim P$, $X_n \sim P_n$ for all n, and

$$\lim_{n \to \infty} X_n \ = \ X \quad \text{almost surely.}$$

In the special case of $\mathcal{X} = \mathbb{R}$ with the usual topology, one may work with $X := F^{-1}(U)$ and $X_n := F_n^{-1}(U)$, where F and F_n are the distribution functions of P and P_n, respectively, while $U \sim \mathrm{Unif}(0,1)$.

Proof of Part (a). Let $0 \leq d(\cdot,\cdot) \leq d_o < \infty$. Suppose that $(P_n)_n$ converges weakly to P. With the random variables X and X_n above,

$$\begin{aligned} \limsup_{n\to\infty} d_{M,k}(P_n,P)^k &\leq \limsup_{n\to\infty} \mathbb{E}\big[d(X_n,X)^k\big] \\ &= 0 \end{aligned}$$

by dominated convergence, because $0 \leq d(X_n,X)^k \leq d_o^k$ and $d(X_n,X)^k \to 0$ almost surely.

Suppose that $d_{M,k}(P_n,P) \to 0$. For each $n \geq 1$ let $R_n \in \mathcal{R}(P_n,P)$ be an optimal coupling of P_n and P. Then for arbitrary bounded and Lipschitz-continuous functions $f : \mathcal{X} \to \mathbb{R}$ with Lipschitz constant L,

$$\begin{aligned} \Big|\int f\,dP_n - \int f\,dP\Big| &= \Big|\int \big(f(x)-f(y)\big)\,R_n(dx,dy)\Big| \\ &\leq \int \big|f(x)-f(y)\big|\,R_n(dx,dy) \\ &\leq L\int d(x,y)\,R_n(dx,dy) \\ &\leq L\Big(\int d(x,y)^k\,R_n(dx,dy)\Big)^{1/k} \\ &= L\,d_{M,k}(P_n,P) \;\to\; 0. \end{aligned}$$

Hence, $(P_n)_n$ converges weakly to P.

Proof of Part (b). Suppose that (b.1) holds true. For each $n \geq 1$ let $R_n \in \mathcal{R}(P_n,P)$ be an optimal coupling of P_n and P. Then one can argue as above that

$$\lim_{n\to\infty}\int f\,dP_n = \int f\,dP$$

for bounded, Lipschitz-continuous functions $f : \mathcal{X} \to \mathbb{R}$. Furthermore, since $\|x\| \leq \|x-y\|+\|y\|$ for arbitrary $x, y \in \mathbb{R}^k$,

$$\begin{aligned} \Big(\int \|x\|^k\,P_n(dx)\Big)^{1/k} &= \Big(\int \|x\|^k\,R_n(dx,dy)\Big)^{1/k} \\ &\leq \Big(\int \|x-y\|^k\,R_n(dx,dy)\Big)^{1/k} + \Big(\int \|y\|^k\,R_n(dx,dy)\Big)^{1/k} \\ &= d_{M,k}(P_n,P) + \Big(\int \|x\|^k\,P(dx)\Big)^{1/k}, \end{aligned}$$

and interchanging the roles of P_n and P leads to the inequality

$$\Big|\Big(\int \|x\|^k\,P_n(dx)\Big)^{1/k} - \Big(\int \|x\|^k\,P(dx)\Big)^{1/k}\Big| \leq d_{M,k}(P_n,P).$$

Hence, $\lim_{n\to\infty}\int \|x\|^k\,P_n(dx) = \int \|x\|^k\,P(dx)$. Thus, (b.2) is satisfied as well.

Suppose that (b.2) holds true. We resort to the random variables X and X_n on $(\Omega, \mathcal{A}, \mathbb{P})$ again. Obviously,

$$d_{M,k}(P_n, P)^k \;\le\; \int \|X_n - X\|^k \, d\,\mathbb{P},$$

and now we show that the expected value on the right hand side converges to 0. For this purpose, we write $X_n = \|X_n\| \cdot U_n$ with a random unit vector $U_n \in \mathcal{X}$ and define

$$\tilde{X}_n \;:=\; \min(\|X_n\|, \|X\|) \cdot U_n.$$

Then

$$\begin{aligned}\Big(\int \|X_n - X\|^k \, d\,\mathbb{P}\Big)^{1/k} \;&\le\; \Big(\int \big(\|\tilde{X}_n - X\| + \|X_n - \tilde{X}_n\|\big)^k \, d\,\mathbb{P}\Big)^{1/k} \\ &\le\; \Big(\int \|\tilde{X}_n - X\|^k \, d\,\mathbb{P}\Big)^{1/k} + \Big(\int \|X_n - \tilde{X}_n\|^k \, d\,\mathbb{P}\Big)^{1/k}.\end{aligned}$$

The construction of $\tilde{X}_n$ implies that $\|\tilde{X}_n\| \le \|X\|$ and $\lim_{n\to\infty} \tilde{X}_n = X$ almost surely. Consequently, $\lim_{n\to\infty} \|\tilde{X}_n - X\|^k = 0$ almost surely, and $\|\tilde{X}_n - X\|^k$ is bounded from above by $2^k\|X\|^k$ with $\int 2^k\|X\|^k \, d\,\mathbb{P} < \infty$. Thus, by dominated convergence,

$$\lim_{n\to\infty} \int \|\tilde{X}_n - X\|^k \, d\,\mathbb{P} \;=\; 0. \tag{A.11}$$

In particular,

$$\lim_{n\to\infty} \int \|\tilde{X}_n\|^k \, d\,\mathbb{P} \;=\; \int \|X\|^k \, d\,\mathbb{P}\,.$$

Another implication of the construction of $\tilde{X}_n$ is that $\|X_n - \tilde{X}_n\| = \|X_n\| - \|\tilde{X}_n\|$. Since $f(t) := t^k$ is convex in $t \ge 0$, $\|X_n - \tilde{X}_n\|^k = f(\|X_n\| - \|\tilde{X}_n\|) - f(0)$ is not larger than $f(\|X_n\|) - f(\|\tilde{X}_n\|) = \|X_n\|^k - \|\tilde{X}_n\|^k$. Consequently,

$$\int \|X_n - \tilde{X}_n\|^k \, d\,\mathbb{P} \;\le\; \int \|X_n\|^k \, d\,\mathbb{P} - \int \|\tilde{X}_n\|^k \, d\,\mathbb{P} \;\to\; 0.$$

The equivalence of (b.2) and (b.3) results from Exercise A.11. □

Exercise A.33. **(a)** Show that for arbitrary distribution functions F and G on the real line,

$$d_{M,1}(F, G) \;=\; \int_{\mathbb{R}} \big|F(x) - G(x)\big| \, dx.$$

(b) Let $Y_1, \ldots, Y_n$ be independent random variables with distribution function F. Show that for the empirical distribution function $\widehat{F}_n$ of $(Y_i)_{i=1}^n$,

$$\sqrt{n}\,\mathbb{E}\, d_{M,1}(\widehat{F}_n, F) \;\to\; \sqrt{\frac{2}{\pi}} \int_{\mathbb{R}} \sqrt{F(y)(1 - F(y))}\, dy.$$

Exercise A.34. Consider vectors $\boldsymbol{\mu}, \tilde{\boldsymbol{\mu}}$ in $\mathbb{R}^p$ and symmetric, positive semidefinite matrices $\boldsymbol{\Sigma}, \tilde{\boldsymbol{\Sigma}}$ in $\mathbb{R}^{p\times p}$. Show that

$$d_{M,2}\big(N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), N_p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})\big)^2 \;\le\; \big\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\big\|^2 + \big\|\boldsymbol{\Sigma}^{1/2} - \tilde{\boldsymbol{\Sigma}}^{1/2}\big\|_F^2.$$

Here $\|\boldsymbol{M}\|_F := \mathrm{trace}(\boldsymbol{M}^\top \boldsymbol{M})^{1/2} = \Big(\sum_{i,j} M_{ij}^2\Big)^{1/2}$ is the Frobenius norm of an arbitrary matrix $\boldsymbol{M}$.

Additional task: Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p\times p}$ be symmetric and positive semidefinite. Show that

$$\big\|A^{1/2} - B^{1/2}\big\|_F^2 \;\le\; p^{1/2}\|A - B\|_F.$$

Lindeberg's CLT revisited. The Mallows distance $d_{M,2}(\cdot,\cdot)$ of distributions on $\mathbb{R}^d$, where $d(x,y)$ is the usual Euclidean norm $\|x-y\|$, is particularly useful in connection with the Central Limit Theorem. Here is an elegant extension of Theorem A.16:

Corollary A.35. *For $n \in \mathbb{N}$ let $\boldsymbol{Y}_{n1}, \boldsymbol{Y}_{n2}, \ldots, \boldsymbol{Y}_{nn}$ be stochastically independent random vectors in $\mathbb{R}^d$ such that*

$$\mathbb{E}(\boldsymbol{Y}_{ni}) \;=\; \boldsymbol{0} \quad \textit{and} \quad \mathbb{E}(\|\boldsymbol{Y}_{ni}\|^2) \;<\; \infty.$$

Further, suppose that

$$\boldsymbol{\Sigma}_n := \sum_{i=1}^n \mathbb{E}\big(\boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top\big) \;=\; O(1) \quad \textit{and} \quad \Lambda_n := \sum_{i=1}^n \mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^2 \min(1, \|\boldsymbol{Y}_{ni}\|)\big) \;\to\; 0.$$

Then

$$d_{M,2}\Big(\mathcal{L}\Big(\sum_{i=1}^n \boldsymbol{Y}_{ni}\Big), \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}_n)\Big) \;\to\; 0$$

and

$$\mathbb{E}\Big\|\sum_{i=1}^n \boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top - \boldsymbol{\Sigma}_n\Big\|_F \;\to\; 0, \quad \mathbb{E}\Big(\max_{i=1,2,\ldots,n} \|\boldsymbol{Y}_{ni}\|^2\Big) \;\to\; 0.$$

Proof. Let $\delta \geq 0$ be the limes superior of

$$\delta_n \;:=\; d_{M,2}\Big(\mathcal{L}\Big(\sum_{i=1}^n \boldsymbol{Y}_{ni}\Big), \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}_n)\Big) + \mathbb{E}\Big\|\sum_{i=1}^n \boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top - \boldsymbol{\Sigma}_n\Big\|_F + \mathbb{E}\Big(\max_{i=1,2,\ldots,n} \|\boldsymbol{Y}_{ni}\|^2\Big)$$

as $n \to \infty$. We want to show that $\delta = 0$. To this end, let $n(1) < n(2) < n(3) < \cdots$ be indices such that $\lim_{k\to\infty} \delta_{n(k)} = \delta$. Since $(\boldsymbol{\Sigma}_n)_n$ is bounded, we may even assume that

$$\boldsymbol{\Sigma} \;:=\; \lim_{k\to\infty} \boldsymbol{\Sigma}_{n(k)}$$

exists. But now we could replace the original triangular array $(\boldsymbol{Y}_{ni})_{n\geq 1, 1\leq i\leq n}$ with the following array $(\tilde{\boldsymbol{Y}}_{ni})_{n\geq 1, 1\leq i\leq n}$ without changing δ: For $n < n(1)$ we set $\tilde{\boldsymbol{Y}}_{ni} := \boldsymbol{0}$, $1 \leq i \leq n$, and for $n(k) \leq n < n(k+1)$, $k \geq 1$,

$$\tilde{\boldsymbol{Y}}_{ni} \;:=\; \begin{cases} \boldsymbol{Y}_{n(k)i} & \text{if } 1 \leq i \leq n(k), \\ \boldsymbol{0} & \text{if } n(k) < i \leq n. \end{cases}$$

In other words, we may assume without loss of generality that the assumptions of Theorem A.16 are satisfied. But then it follows from Theorems A.16 and A.31 that

$$d_{M,2}\Big(\mathcal{L}\Big(\sum_{i=1}^n \boldsymbol{Y}_{ni}\Big), \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})\Big) \;\to\; 0$$

and

$$\mathbb{E}\Big\|\sum_{i=1}^n \boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top - \boldsymbol{\Sigma}_n\Big\|_F \;\to\; 0, \quad \mathbb{E}\Big(\max_{i=1,2,\ldots,n} \|\boldsymbol{Y}_{ni}\|^2\Big) \;\to\; 0.$$

Moreover, Exercise A.34 implies that

$$d_{M,2}\big(\mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}), \mathrm{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}_n)\big) \;\to\; 0,$$

so the triangular inequality for $d_{M,2}(\cdot,\cdot)$ implies that

$$d_{M,2}\Big(\mathcal{L}\Big(\sum_{i=1}^{n} \boldsymbol{Y}_{ni}\Big), \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma}_n)\Big) \;\to\; 0.$$

Hence, the limes superior δ is equal to 0. □

A.9 An Inequality for Sums of Independent Random Vectors

At various places in these lecture notes, inequalities for sums of independent random variables are useful. Let $X_1, X_2, \ldots, X_n$ be independent random variables with values in $\mathbb{R}^d$, and let $\|\cdot\|$ be the usual Euclidean norm on $\mathbb{R}^d$. We assume that $\mathbb{E}\,\|X_i\| < \infty$ for all i and consider

$$S \;:=\; \sum_{i=1}^{n} X_i$$

as well as $\bar{X} := n^{-1}S$.

Let us start with two special cases: By the triangle inequality,

$$\mathbb{E}\,\|S - \mathbb{E}(S)\| \;\le\; \sum_{i=1}^{n} \mathbb{E}\,\|X_i - \mathbb{E}(X_i)\| \;\le\; 2\sum_{i=1}^{n} \mathbb{E}\,\|X_i\|.$$

Moreover,

$$\begin{aligned}
\mathbb{E}\big(\|S - \mathbb{E}(S)\|^2\big) \;&=\; \sum_{i,j=1}^{n} \mathbb{E}\big((X_i - \mathbb{E}(X_i))^\top (X_j - \mathbb{E}(X_j))\big) \\
&=\; \sum_{i=1}^{n} \Big(\mathbb{E}\big(\|X_i\|^2\big) - \big\|\mathbb{E}(X_i)\big\|^2\Big) \\
&\le\; \sum_{i=1}^{n} \mathbb{E}\big(\|X_i\|^2\big),
\end{aligned}$$

provided that all second moments $\mathbb{E}\big(\|X_i\|^2\big)$ are finite. Both inequalities may be generalized to the situation that for a fixed $r \ge 1$, all r-th moments $\mathbb{E}(\|X_i\|^r)$ are finite.

Theorem A.36. *For each $r \ge 1$ there exists a universal constant $C_{r,d}$ such that*

$$\mathbb{E}\big(\|S - \mathbb{E}(S)\|^r\big) \;\le\; C_{r,d}\,\mathbb{E}\bigg(\Big(\sum_{i=1}^{n} \|X_i\|^2\Big)^{r/2}\bigg) \;\le\; C_{r,d} n^{\max(r/2-1,0)} \sum_{i=1}^{n} \mathbb{E}(\|X_i\|^r).$$

In particular,

$$\mathbb{E}\big(\|\bar{X} - \mathbb{E}(\bar{X})\|^r\big) \;\le\; C_{r,d} n^{-\min(r-1,r/2)}\, \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\|X_i\|^r).$$

Remark. The proof presented here yields the constant

$$C_{r,d} \;=\; (2\pi)^{r/2}\, \mathbb{E}(\|G\|^r) d^{-\min(r/2,1)}$$

with a standard Gaussian random vector $G \in \mathbb{R}^d$.

Proof of Theorem A.36. We complement the underlying probability space such that it carries stochastically independent random variables $X_1, \ldots, X_n, X_1', \ldots, X_n'$ and $G_1, \ldots, G_{\max(n,d)}$ such that $\mathcal{L}(X_i') = \mathcal{L}(X_i)$ and $\mathcal{L}(G_j) = \mathrm{N}(0,1)$. Now we apply Jensen's inequality repeatedly. At first, with $\boldsymbol{X} = (X_i)_{i=1}^n$ and $\mathbb{E}_o(\cdot) := \mathbb{E}(\cdot \,|\, \boldsymbol{X})$ one may write

$$\begin{aligned}
\mathbb{E}\big(\|S - \mathbb{E}(S)\|^r\big) \;&=\; \mathbb{E}\Big(\Big\|\mathbb{E}_o\Big(\sum_{i=1}^n (X_i - X_i')\Big)\Big\|^r\Big) \\
&\leq\; \mathbb{E}\Big(\mathbb{E}_o\Big(\Big\|\sum_{i=1}^n (X_i - X_i')\Big\|^r\Big)\Big) \\
&=\; \mathbb{E}\Big(\Big\|\sum_{i=1}^n (X_i - X_i')\Big\|^r\Big).
\end{aligned}$$

For symmetry reasons, $\sum_i (X_i - X_i')$ has the same distribution as $\sum_i \operatorname{sign}(G_i)(X_i - X_i')$, so

$$\begin{aligned}
\mathbb{E}\Big(\Big\|\sum_{i=1}^n (X_i - X_i')\Big\|^r\Big) \;&=\; \mathbb{E}\Big(\Big\|\sum_{i=1}^n \operatorname{sign}(G_i)(X_i - X_i')\Big\|^r\Big) \\
&=\; \mathbb{E}\Big(\Big\|\sum_{i=1}^n \operatorname{sign}(G_i) X_i - \sum_{i=1}^n \operatorname{sign}(G_i) X_i'\Big\|^r\Big) \\
&\leq\; 2^r\, \mathbb{E}\Big(\Big\|\sum_{i=1}^n \operatorname{sign}(G_i) X_i\Big\|^r\Big),
\end{aligned}$$

because

$$\|A \pm B\|^r \;\leq\; 2^r \Big(\frac{\|A\| + \|B\|}{2}\Big)^r \;\leq\; 2^r\, \frac{\|A\|^r + \|B\|^r}{2}.$$

But

$$\sum_{i=1}^n \operatorname{sign}(G_i) X_i \;=\; \sqrt{\pi/2}\, \mathbb{E}\Big(\sum_{i=1}^n G_i X_i \,\Big|\, \big(\operatorname{sign}(G_i)\big)_{i=1}^n, \boldsymbol{X}\Big),$$

because $|G_i|$ and $\operatorname{sign}(G_i)$ are stochastically independent with $\mathbb{E}(|G_i|) = \sqrt{2/\pi}$. Consequently,

$$\begin{aligned}
\mathbb{E}\Big(\Big\|\sum_{i=1}^n \operatorname{sign}(G_i) X_i\Big\|^r\Big) \;&=\; (\pi/2)^{r/2}\, \mathbb{E}\Big(\Big\|\mathbb{E}\Big(\sum_{i=1}^n G_i X_i \,\Big|\, \big(\operatorname{sign}(G_i)\big)_{i=1}^n, \boldsymbol{X}\Big)\Big\|^r\Big) \\
&\leq\; (\pi/2)^{r/2}\, \mathbb{E}\Big(\mathbb{E}\Big(\Big\|\sum_{i=1}^n G_i X_i\Big\|^r \,\Big|\, \big(\operatorname{sign}(G_i)\big)_{i=1}^n, \boldsymbol{X}\Big)\Big) \\
&=\; (\pi/2)^{r/2}\, \mathbb{E}\Big(\Big\|\sum_{i=1}^n G_i X_i\Big\|^r\Big).
\end{aligned}$$

The conditional distribution of $\sum_{i=1}^n G_i X_i$, given $\boldsymbol{X}$, is a centered d-variate Gaussian distribution with covariance matrix

$$\Sigma = \Sigma(\boldsymbol{X}) \;:=\; \sum_{i=1}^n X_i X_i^\top.$$

Hence,

$$\begin{aligned}
\mathbb{E}\Big(\Big\|\sum_{i=1}^n G_i X_i\Big\|^r\Big) \;&=\; \mathbb{E}\,\mathbb{E}_o\Big(\Big\|\sum_{i=1}^n G_i X_i\Big\|^r\Big) \\
&=\; \mathbb{E}\,\mathbb{E}_o\Big(\big\|\Sigma^{1/2} G\big\|^r\Big),
\end{aligned}$$

with $G := (G_j)_{j=1}^d \sim \mathrm{N}_d(0, I_d)$. If we write

$$G \;=\; \|G\|U,$$

then $\|G\|$, the unit vector U and $\boldsymbol{X}$ are stochastically independent, and

$$\mathbb{E}_o\big(\|\Sigma^{1/2}G\|^r\big) \;=\; \mathbb{E}\big(\|G\|^r\big)\,\mathbb{E}_o\big(\|\Sigma^{1/2}U\|^r\big) \;=\; \mathbb{E}\big(\|G\|^r\big)\,\mathbb{E}_o\big((U^\top\Sigma U)^{r/2}\big).$$

Consequently, defining

$$D = D(\boldsymbol{X}) \;:=\; \mathrm{trace}(\Sigma)^{1/2} \;=\; \Big(\sum_{i=1}^n \|X_i\|^2\Big)^{1/2},$$

it suffices to show that

$$\mathbb{E}_o\big((U^\top\Sigma U)^{r/2}\big) \;\le\; \tilde{C}_{r,d}D^r$$

for some constant $\tilde{C}_{r,d}$.

In case of $1 \le r \le 2$, we have the inequality

$$\mathbb{E}_o\big((U^\top\Sigma U)^{r/2}\big) \;\le\; \big(\mathbb{E}_o(U^\top\Sigma U)\big)^{r/2} \;=\; \big(d^{-1}D^2\big)^{r/2} \;=\; d^{-r/2}D^r,$$

because $U^\top\Sigma U =_{\mathcal{L}} \sum_{j=1}^d U_j^2\lambda_j(\Sigma)$, and a symmetry consideration shows that $\mathbb{E}(U_j^2) = 1/d$ for all j.

In case of $r \ge 2$, we use the fact that $\sum_{j=1}^d U_j^2 = 1$:

$$\begin{aligned}
\mathbb{E}_o\big((U^\top\Sigma U)^{r/2}\big) \;&=\; \mathbb{E}_o\Big(\Big(\sum_{j=1}^d U_j^2\lambda_j(\Sigma)\Big)^{r/2}\Big) \;\le\; \mathbb{E}_o\Big(\sum_{j=1}^d U_j^2\lambda_j(\Sigma)^{r/2}\Big)\\
&=\; d^{-1}\sum_{j=1}^d \lambda_j(\Sigma)^{r/2} \;\le\; d^{-1}\sum_{j=1}^d \lambda_j(\Sigma)\mathrm{trace}(\Sigma)^{r/2-1}\\
&=\; d^{-1}\mathrm{trace}(\Sigma)^{r/2} \;=\; d^{-1}D^r.
\end{aligned}$$

All in all, this yields the first asserted inequality, where $C_{r,d} = (2\pi)^{r/2}\,\mathbb{E}(\|G\|^r)d^{-\min(r/2,1)}$.

Since $\|X_i\|^2/D^2 \in [0,1]$, we may conclude that for $1 \le r \le 2$,

$$\sum_{i=1}^n \|X_i\|^r/D^r \;=\; \sum_{i=1}^n \big(\|X_i\|^2/D^2\big)^{r/2} \;\ge\; \sum_{i=1}^n \|X_i\|^2/D^2 \;=\; 1,$$

because $0 < r/2 \le 1$, so $D^r \le \sum_{i=1}^n \|X_i\|^r$. In case of $r > 2$, we apply Jensen's inequality to the convex function $0 \le t \mapsto t^{r/2}$ and obtain the inequality

$$D^r \;=\; n^{r/2}\Big(\frac{1}{n}\sum_{i=1}^n \|X_i\|^2\Big)^{r/2} \;\le\; n^{r/2}\,\frac{1}{n}\sum_{i=1}^n \|X_i\|^r \;=\; n^{r/2-1}\sum_{i=1}^n \|X_i\|^r.$$

□

A.10 Stochastic Landau Symbols

Let $(X_n)_n$ be a sequence or real- or vector-valued random variables, and let $(R_n)_n$ be a sequence of random variables $R_n > 0$. One writes

$$X_n \;=\; O_{\mathrm{p}}(R_n)$$

and says that "X_n is of stochastic order R_n" if for any $\epsilon > 0$ there exists a constant C such that

$$\limsup_{n\to\infty} \mathbb{P}(|X_n| \geq CR_n) \;\leq\; \varepsilon.$$

One writes

$$X_n \;=\; o_{\mathrm{p}}(r_n)$$

and says that "X_n is of smaller stochastic order than R_n" if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n| \geq \epsilon R_n) \;=\; 0.$$

The symbols $O_{\mathrm{p}}(R_n)$ and $o_{\mathrm{p}}(R_n)$ are also used directly as placeholders for sequences of random variables with the stated properties. Often one uses a deterministic sequence $(r_n)_n$ of numbers $r_n > 0$ rather than a random sequence $(R_n)_n$. Important special case are $O_{\mathrm{p}}(1)$ and $o_{\mathrm{p}}(1)$, corresponding to the constant sequence $(1)_n$. Indeed, $X_n = O_{\mathrm{p}}(R_n)$ if and only if $R_n^{-1}X_n = O_{\mathrm{p}}(1)$, and $X_n = o_{\mathrm{p}}(R_n)$ if and only if $R_n^{-1}X_n = o_{\mathrm{p}}(1)$. The statement $X_n = O_{\mathrm{p}}(1)$ is also phrased as "X_n is bounded in probability", and $X_n = o_{\mathrm{p}}(1)$ means that X_n converges to zero in probability.

Exercise A.37. Prove the following rules for stochastic Landau-symbols:

(a) $o_{\mathrm{p}}(R_n) = O_{\mathrm{p}}(R_n)$

(b) $O_{\mathrm{p}}(R_n) + O_{\mathrm{p}}(S_n) = O_{\mathrm{p}}\big(\max(R_n, S_n)\big)$

(c) $o_{\mathrm{p}}(R_n) + o_{\mathrm{p}}(S_n) = o_{\mathrm{p}}\big(\max(R_n, S_n)\big)$

(d) $O_{\mathrm{p}}(R_n)O_{\mathrm{p}}(S_n) = O_{\mathrm{p}}(R_nS_n)$

(e) $O_{\mathrm{p}}(R_n)o_{\mathrm{p}}(S_n) = o_{\mathrm{p}}(R_nS_n)$

(f) $f(a + o_{\mathrm{p}}(1)) = f(a) + o_{\mathrm{p}}(1)$ whenever f is continuous at a.

(g) $f(a + O_{\mathrm{p}}(R_n)) = f(a) + O_{\mathrm{p}}(R_n)$ whenever f is differentiable at a.

These rules are meant to be read from left to right. For example, (e) means that if X_n is of stochastic order R_n and Y_n is of smaller stochastic order than S_n, then X_nY_n is of smaller stochastic order than R_nS_n.

Exercise A.38. Prove the following rules for (stochastic) Landau-symbols:

(a) If X_n converges in distribution to X, then $X_n = O_{\mathrm{p}}(1)$.

(b) If $\mathbb{E}\left(|X_n|^k\right) = O(a_n)$ for some $k > 0$, then $X_n = O_{\mathrm{p}}(a_n^{1/k})$.

(c) If $\mathbb{E}\left(|X_n|^k\right) = o(a_n)$ for some $k > 0$, then $X_n = o_{\mathrm{p}}(a_n^{1/k})$.

In parts (b-c), $(a_n)_n$ is a deterministic sequence of numbers $a_n > 0$.