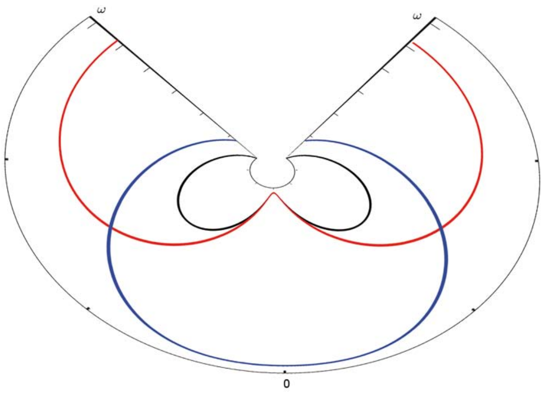
Lehrstuhl für Statistik
Institut für Mathematik
Universität Würzburg

A new biased estimator for multivariate regression models with highly collinear variables



Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius–Maximilians–Universität Würzburg

vorgelegt von

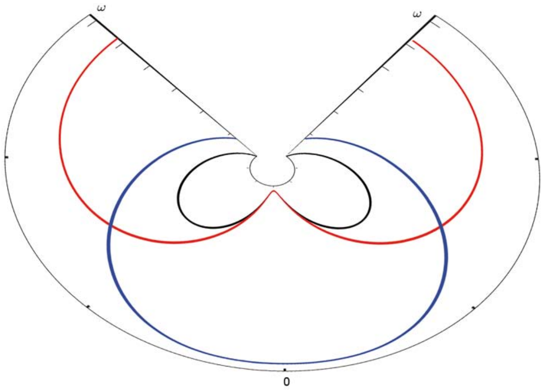
Julia Wissel

aus
Mömbris

Februar 2009

Lehrstuhl für Statistik
Institut für Mathematik
Universität Würzburg

A new biased estimator for multivariate regression models with highly collinear variables



Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius–Maximilians–Universität Würzburg

vorgelegt von

Julia Wissel

aus
Mömbris

Februar 2009

Contents

Introduction

In the scientific fields of astronomy and geodesy scientists have always looked for solutions to ease the navigation on the wide oceans and to allow adventurers to discover new and unknown landscapes during the Age of Exploration. It was a common practice at that time to rely on land sightings to determine the current position of the ship. However, to enable over sea travels the position of the ship had to be determined by means of the celestial bodies. Therefore it was an issue to describe their behavior.

A first major step to describe the movement and hence predict the trajectory of a star was performed by Carl Friedrich Gauss. He used observations and thereby developed the fundamentals for least squares analysis in 1795. However, the honor of presenting the first publication of the least squares method belongs to Legendre. Within his work, published in 1805, he presents the rule as a convenient method only, whereas a first proof of the method was published by Robert Adrain in 1808.

Thus the traditional least squares statistic is probably one of the oldest, but still most popular, technique in modern statistics, where it is often used to estimate the parameters of a linear regression model. Current and also traditional approaches minimize the sum of the squared differences between the predicted and the observed values of the dependent variable, which represents the sum of the squared error terms of the model. As a solution one gets the ordinary least squares estimator, which is also equivalent to the maximum likelihood estimator in case of independent and normally distributed error variables.

The popularity of the least squares estimator can be traced back to the formulation of the Gauss–Markov–Theorem, which states the least squares estimator to be the best linear unbiased estimator. Nevertheless, in many cases there will be rather severe statistical implications of remaining only in the class of unbiased estimators.

In applied work few variables are free of measurement error and/ or are non-stochastic. As a consequence, only few statistical models are correctly specified, and thus these specification errors result in a biased outcome when the least squares estimation is used. In the preface of their book, Vinod and Ullah found the right words to describe the boon and bane of the least squares estimator:

> "Since the exposition of the Gauss–Markov–Theorem [...] practically all fields in the natural and social sciences have widely and sometimes blindly used the Ordinary Least Squares method".

It is therefore not astonishing that the result of Charles Stein in the late 1950's, stating that there exists a better alternative to the least squares estimator under certain conditions, firstly remained unnoticed. Some years later, James and Stein proposed an explicit biased estimator for which the improvement to the least squares estimator in terms of the mean squared error was quite substantial in the usual linear model. Parallel to this, Hoerl and Kennard suggested another biased estimator, the so-called ridge regression estimator, still one of the most popular biased estimator today.

Probably the best motivation for further research in the field of biased estimation was the the problem of multicollinearity. In case of multicollinearity, i.e. if there is a strong dependency between the columns of the design matrix, the least squares estimates tend to be very unstable and unreliable. Although the Gauss-Markov–Theorem assures that the least squares estimator has minimal total variance in the class of the unbiased estimators, it is not guaranteed that the total variance is small. Therefore, it may be more advantageous to accept a slightly biased estimator with smaller variance.

Trenkler (1981,[**58**]) wrote

> "Insisting on least squares which means optimal fit at any price can lead to poor predictive qualities of a correctly specified model. Likewise, trying to remove one or more observation vectors from the sample to improve the bad condition of the regressor matrix is not advisible since relevant information may be thrown away".

Because one often encounters the problem of multicollinearity in applied work, there is almost no way out than using biased estimators. Especially the ridge estimator seems to be applicable for multicollinear data. Therefore, during the past 30 years many different kinds of estimators have been presented as alternatives to least squares estimators for the estimation of the parameters of a linear regression model. Some of these formulations use Bayesian methods, others employ the context of the frequentist point of view. Often different approaches yield the same or estimators with similar mathematical forms. As a result, many

papers and books have been written about ridge and related estimators.
Already in 1981, Trenkler gave a survey of biased estimation and his bibliography contained more than 500 entries.

The goal of this manuscript is to present another biased estimator, with improved mean squared error properties than the least squares estimator, and to compare it with the famous ridge estimator.
The thesis is divided into seven chapters and is organized as follows:
In Chapter 1 the used regression model is specified and the necessary assumptions are explained.
Afterwards an introduction to risk functions and especially to the mean squared error for comparing estimators is given in Chapter 2.
The method of standardization of regression coefficients, often used in regression theory and applied work, is described in detail in Chapter 3. Thereby the discussion about the advantages and disadvantages of standardization in regression theory is not ignored.
After a review of the problem of multicollinearity in Chapter 4, an introduction to ridge estimation will be given in Chapter 5. Because of the exhaustive investigations done in this field, it is intended to give only a rough overview. Thereby the emphasis lies on the ridge estimator of Hoerl and Kennard and its statistical properties, which will be useful for the considerations in the following chapters. For a more detailed information, the reader is recommended to the cited literature.
Finally, the *disturbed least squares estimator* and its theoretical properties for standardized data will be presented in Chapter 6. It is based on adding a small quantity $\omega\psi_j,\ j = 1, \dots, p$ on each regressor. It will be proven that we can always find an ω, such that the mean squared error of the disturbed least squares estimator is smaller than the corresponding one of the least squares estimator. Besides the standardized model, we will also find a solution for unstandardized data and the disturbed least squares estimator will be embedded in the class of the ridge estimators.
We will conclude this thesis by means of a simulation study, which tries to evaluate the performance of the proposed disturbed least squares estimator compared to the least squares and ridge estimator in Chapter 7.

Closing this section a few words regarding our notation: Given a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ we write $X_j,\ j = 1, \dots, p$ for the j–th column and $\boldsymbol{x}_j^T$ for the j–th row of $\boldsymbol{X}$. The mean of the j–th column X_j is denoted by $\bar{X}_j,\ j = 1, \dots, p$.

If a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive semidefinite or positive definite we write $\boldsymbol{A} \geq 0$ or $\boldsymbol{A} > 0$.

If we strike out the i–th row and j–th column of $\boldsymbol{A}$ we write $\boldsymbol{A}_{\{i,j\}}$, whereas (if not defined otherwise) $\boldsymbol{A}_{\{j\}}$ means that the j–th column of $\boldsymbol{A}$ is missing. Furthermore we will often use the vector $\boldsymbol{\beta}_{\{\beta_0\}}$ which is equal to $\boldsymbol{\beta}$, except that the coefficient β_0 is missing. The notation $\boldsymbol{\gamma}_\omega \in \mathbb{R}^{p \times 1}$ should emphasize the dependence of the vector $\boldsymbol{\gamma}$ on ω. If the coefficients of $\boldsymbol{\gamma}_\omega$ are used, we write $\gamma_j^\omega,\ j = 1, \ldots, p$.

Different calculations and plots within this thesis were made either with the software package `SAS` or `MATLAB`.

CHAPTER 1

Specification of the Model

Consider the ordinary linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n,\ n \in \mathbb{N},$$

where $\beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}$ are unknown regression coefficients, y_i are observations of the dependent variable, $x_{i,j},\ j = 1, \ldots p$ are observations of the p non–constant independent variables (or regressors) and ε_i are unknown errors with $\mathrm{E}(\varepsilon_i) = 0$. We may rewrite this model in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.0.1}$$

where $\boldsymbol{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{X} := \begin{bmatrix} \mathbf{1}_n & X_1 & \ldots & X_p \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}$ with $\mathbf{1}_n := [1]_{1\leq i\leq n}$. Hence the first column of the design matrix $\boldsymbol{X}$ consist only of ones and the remaining columns are denoted by $X_j := [x_{i,j}]_{1\leq i\leq n}$, $j = 1, \ldots, p$. We use the following assumptions:

Assumption 1: $\boldsymbol{X}$ is a non–stochastic matrix of regressors,
Assumption 2: $\boldsymbol{X}$ has full column rank, i.e. $\boldsymbol{X}^T\boldsymbol{X}$ has rank $p+1$.
Assumption 3: $n \geq p+1$, i.e. we have at least just as many observations as unknown regression coefficients.
Assumption 4: The vector $\boldsymbol{\varepsilon}$ of the unknown errors ε_i is multivariate normal distributed with covariance matrix $\sigma^2\boldsymbol{I}_n$, i.e. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$.

Thus with Theorem A.9.3 and Assumption 2 the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is positive definite. The least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is derived by minimizing the residual sum of squares (RSS) of $\boldsymbol{\beta}^*$. Thus minimize

$$\begin{aligned} \mathrm{RSS}(\boldsymbol{\beta}^*) &:= \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*\right)^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*) \\ &= \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^{*T}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^* - 2\boldsymbol{\beta}^{*T}\boldsymbol{X}^T\boldsymbol{y} \end{aligned} \tag{1.0.2}$$

by differentiation, where $\boldsymbol{x}_i^T$, $i = 1, \ldots, n$ denotes the i-th row vector of $\boldsymbol{X}$. Then

$$\frac{\partial \mathrm{RSS}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^* - 2\boldsymbol{X}^T\boldsymbol{y}. \tag{1.0.3}$$

Set (1.0.3) equal to zero. Thus the *normal equations*

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}^T\boldsymbol{y} \tag{1.0.4}$$

have the solution

$$\hat{\boldsymbol{\beta}} := \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{1.0.5}$$

because $\boldsymbol{X}^T\boldsymbol{X}$ is invertible. This is the well known *least squares estimator*. Some of the properties of the least squares estimator (1.0.5) are given in the following theorem

THEOREM 1.0.1. *In the ordinary linear regression model (1.0.1) we have*

(1) $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$*, i.e.* $\hat{\boldsymbol{\beta}}$ *is unbiased,*
(2) *the covariance matrix of* $\hat{\boldsymbol{\beta}}$ *is given by* $\Sigma(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$*,*
(3) $\hat{\boldsymbol{\beta}}$ *is the best linear unbiased estimator (BLUE), i.e. for any linear unbiased estimator* $\breve{\boldsymbol{\beta}}$ *we have*

$$\mathrm{var}(\hat{\beta}_j) \le \mathrm{var}(\breve{\beta}_j), \quad j = 0, 1, \dots, p$$

(Gauss–Markov–Theorem).

PROOF. See Falk (2002,[**11**]), p. 118-121.

□

NOTE 1.0.2. Strictly speaking, point (3) of Theorem 1.0.1 is only a consequence of the Gauss–Markov–Theorem, which states that the covariance matrix of any other unbiased estimators exceeds the one of the least squares estimator by a positive semidefinite matrix (see e.g. G. Trenkler (1981,[**58**])).

The following well known lemma will be useful for several examinations within this manuscript.

LEMMA 1.0.3. *In the standard model (1.0.1) we have*

$$\mathrm{E}\left(\mathrm{RSS}(\hat{\boldsymbol{\beta}})\right) = (n-p-1)\sigma^2,$$

where $\mathrm{RSS}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2$.

PROOF. See Falk (2002,[**11**]), p. 122.

□

As a consequence an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}})}{n-p-1}. \tag{1.0.6}$$

We will call (1.0.6) the least squares estimator of σ^2.
The residual vector $\hat{\boldsymbol{\varepsilon}} := \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ has mean zero, i.e.

$$\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$$

and the covariance matrix is given by

$$\Sigma(\hat{\boldsymbol{\varepsilon}}) = \sigma^2 \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \right). \tag{1.0.7}$$

(Proof see Falk (2002,[**11**]), p. 125)

CHAPTER 2

Criteria for Comparing Estimators

The task of a statistician is to estimate the true but unknown vector $\boldsymbol{\beta}$ of the regression coefficients in (1.0.1). It is common to choose an estimator $\boldsymbol{b} \in \mathbb{R}^{(p+1)\times 1}$ which is linear in $\boldsymbol{y}$, i.e.

$$\boldsymbol{b} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{d}.$$

The matrices $\boldsymbol{C} \in \mathbb{R}^{(p+1)\times n}$ and $\boldsymbol{d} \in \mathbb{R}^{(p+1)\times 1}$ are non–stochastic matrices which have to be determined by minimizing a suitably chosen risk function. From (1.0.5) we can see that the least squares estimator is a linear estimator with

$$\boldsymbol{C} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$

and

$$\boldsymbol{d} = \boldsymbol{0}.$$

The following definition gives a distinction within the class of linear estimators.

Definition 2.0.4. *$\boldsymbol{b}$ is called a homogeneous estimator of $\boldsymbol{\beta}$ if $\boldsymbol{d} = \boldsymbol{0}$. Otherwise $\boldsymbol{b}$ is called heterogeneous.*

It is well known, that in the model (1.0.1)

$$\text{Bias}(\boldsymbol{b}) = \text{E}(\boldsymbol{b}) - \boldsymbol{\beta} = \boldsymbol{C}\text{E}(\boldsymbol{y}) + \boldsymbol{d} - \boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{d} - \boldsymbol{\beta}$$

and

$$\Sigma(\boldsymbol{b}) = \boldsymbol{C}\Sigma(\boldsymbol{y})\boldsymbol{C}^T = \sigma^2\boldsymbol{C}\boldsymbol{C}^T. \tag{2.0.1}$$

In Chapter 1 we have measured the goodness of fit by the residual sum of squares RSS. Analogously we define for the random variable $\boldsymbol{b}$ the quadratic loss function

$$\boldsymbol{L}(\boldsymbol{b}) = (\boldsymbol{b} - \boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{b} - \boldsymbol{\beta}), \tag{2.0.2}$$

where $\boldsymbol{W}$ is a symmetric and positive semidefinite $(p + 1) \times (p + 1)$ matrix. Obviously the loss (2.0.2) depends on the sample. Thus we have to consider the average or expected loss over all possible samples, which is called the *risk*.

Definition 2.0.5. *Let $\boldsymbol{W}$ be a symmetric, positive semidefinite $(p+1)\times(p+1)$ matrix. The quadratic risk of an estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is defined as*

$$\boldsymbol{R}(\boldsymbol{b}) = \mathrm{E}\left((\boldsymbol{b}-\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{b}-\boldsymbol{\beta})\right). \tag{2.0.3}$$

For the special case $\boldsymbol{W} = \boldsymbol{I}_{(p+1)}$ we get the well known *(multivariate) mean squared error*, which will be the main criteria for comparison of estimators to be used in the rest of this manuscript.

2.1. Multivariate Mean Squared Error

The (multivariate) mean squared error of an (biased) estimator $\boldsymbol{b}$ of $\boldsymbol{\beta} \in \mathbb{R}^{(p+1)\times 1}$ is defined by

$$\begin{aligned}\mathrm{MSE}(\boldsymbol{b}) &:= \mathrm{E}\left((\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})\right)\\ &= \mathrm{E}\left((\boldsymbol{b}-\mathrm{E}(\boldsymbol{b})+\mathrm{E}(\boldsymbol{b})-\boldsymbol{\beta})^T(\boldsymbol{b}-\mathrm{E}(\boldsymbol{b})+\mathrm{E}(\boldsymbol{b})-\boldsymbol{\beta})\right)\\ &= \mathrm{E}\left((\boldsymbol{b}-\mathrm{E}(\boldsymbol{b}))^T(\boldsymbol{b}-\mathrm{E}(\boldsymbol{b}))\right)+(\mathrm{E}(\boldsymbol{b})-\boldsymbol{\beta})^T(\mathrm{E}(\boldsymbol{b})-\boldsymbol{\beta})\\ &= \mathrm{tr}(\Sigma(\boldsymbol{b}))+\mathrm{Bias}(\boldsymbol{b})^T\mathrm{Bias}(\boldsymbol{b}),\end{aligned} \tag{2.1.4}$$

where tr represents the trace of a matrix (see Appendix A.1). If we define the Euclidean length of a vector $\boldsymbol{v}$ by $\|\boldsymbol{v}\|_2 = \sqrt{\boldsymbol{v}^T\boldsymbol{v}}$, then $\mathrm{MSE}(\boldsymbol{b})$ in (2.1.4) measures the average of the squared Euclidean distance between $\boldsymbol{b}$ and $\boldsymbol{\beta}$. Thus an estimator with small mean squared error will be close to the true parameter. It is well known from the Gauss–Markov–Theorem (see Theorem 1.0.1) that the least squares estimator has the smallest total variance among all unbiased estimators. But this does not imply that there cannot exist any biased estimator with smaller total variance. By allowing a small amount of bias it may be possible to get a biased estimator with smaller mean squared error than the least squares estimator. Some biased estimators will be discussed in detail in Chapter 5.

Because (2.0.3) is a generalization of the multivariate mean squared error, the quadratic risk is also called the *weighted mean squared error* of $\boldsymbol{b}$ or, for short, $\mathrm{WMSE}(\boldsymbol{b})$.

Consider an arbitrary vector $\boldsymbol{w} \in \mathbb{R}^{(p+1)\times 1}$. Then we have

$$\begin{aligned}\mathrm{MSE}\left(\boldsymbol{w}^T\boldsymbol{b}\right) &= \mathrm{E}\left((\boldsymbol{w}^T\boldsymbol{b}-\boldsymbol{w}^T\boldsymbol{\beta})^T(\boldsymbol{w}^T\boldsymbol{b}-\boldsymbol{w}^T\boldsymbol{\beta})\right)\\ &= \mathrm{E}\left((\boldsymbol{b}-\boldsymbol{\beta})^T\boldsymbol{w}\boldsymbol{w}^T(\boldsymbol{b}-\boldsymbol{\beta})\right).\end{aligned} \tag{2.1.5}$$

Thus the mean squared error of a parametric function $\boldsymbol{w}^T\boldsymbol{b}$ of an estimator is equivalent to the $\mathrm{WMSE}(\boldsymbol{b})$ with $\boldsymbol{W} = \boldsymbol{w}\boldsymbol{w}^T$.

2.2. Matrix Mean Squared Error

The weighted mean squared error is closely related to the matrix valued criterion of the mean squared error of an estimator. The *matrix mean squared error* is defined by the $(p+1)\times(p+1)$ matrix

$$\text{MtxMSE}(\boldsymbol{b}) := \text{E}\left((\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})^T\right). \tag{2.2.6}$$

We can write (2.2.6) also as

$$\begin{aligned}\text{MtxMSE}(\boldsymbol{b}) &= \text{E}\left((\boldsymbol{b}-\text{E}(\boldsymbol{b})+\text{E}(\boldsymbol{b})-\boldsymbol{\beta})(\boldsymbol{b}-\text{E}(\boldsymbol{b})+\text{E}(\boldsymbol{b})-\boldsymbol{\beta})^T\right)\\ &= \text{E}\left((\boldsymbol{b}-\text{E}(\boldsymbol{b}))(\boldsymbol{b}-\text{E}(\boldsymbol{b}))^T\right)+(\text{E}(\boldsymbol{b})-\boldsymbol{\beta})\,(\text{E}(\boldsymbol{b})-\boldsymbol{\beta})^T\\ &= \Sigma(\boldsymbol{b})+\text{Bias}(\boldsymbol{b})\text{Bias}(\boldsymbol{b})^T. \end{aligned}\tag{2.2.7}$$

If we take trace on both sides of (2.2.7) we get (2.1.4), that is

$$\text{tr}(\text{MtxMSE}(\boldsymbol{b})) = \text{MSE}(\boldsymbol{b}) = \text{E}\left((\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})\right)$$

and generally we get for any $\boldsymbol{w}\in\mathbb{R}^{(p+1)\times 1}$ and equation (A.1.2)

$$\begin{aligned}\boldsymbol{w}^T(\text{MtxMSE}(\boldsymbol{b}))\boldsymbol{w} &= \text{tr}\left(\boldsymbol{w}^T\text{E}\left((\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})^T\right)\boldsymbol{w}\right)\\ &= \text{E}\left(\text{tr}\left(\boldsymbol{w}^T(\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})^T\boldsymbol{w}\right)\right)\\ &= \text{E}\left(\text{tr}\left((\boldsymbol{b}-\boldsymbol{\beta})^T\boldsymbol{w}\boldsymbol{w}^T(\boldsymbol{b}-\boldsymbol{\beta})\right)\right).\end{aligned}$$

Thus if $\boldsymbol{w}^T(\text{MtxMSE}(\boldsymbol{b}))\boldsymbol{w}\geq 0$, so is the weighted mean squared error for $\boldsymbol{W}=\boldsymbol{w}\boldsymbol{w}^T$. Finally, from (2.1.5) we have $\text{MSE}\left(\boldsymbol{w}^T\boldsymbol{b}\right)\geq 0$.
Consider two competing estimators $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ and

$$\Delta = \text{MtxMSE}(\boldsymbol{b}_2)-\text{MtxMSE}(\boldsymbol{b}_1).$$

The following Theorem of Theobald (1974,[**55**]) states an estimator having a smaller matrix mean squared error than another estimator, iff it has a smaller weighted mean squared error for arbitrary $\boldsymbol{W}$.

THEOREM 2.2.1. *The following conditions are equivalent*

(1) Δ *is positive semidefinite.*
(2) $\text{WMSE}(\boldsymbol{b}_2)-\text{WMSE}(\boldsymbol{b}_1)\geq 0$
for all positive semidefinite matrices $\boldsymbol{W}\in\mathbb{R}^{(p+1)\times(p+1)}$.

PROOF. See Theobald (1974,[**55**]).

□

A similar result may be established for Δ being positive definite, if "positive semidefinite" and "$\geq$" are replaced by "positive definite" and "$>$". If Δ is a positive (semi)definite matrix, $\boldsymbol{b}_1$ is to be preferred to $\boldsymbol{b}_2$. As a consequence of Theorem 2.2.1

$$\delta := \text{MSE}(\boldsymbol{b}_2)-\text{MSE}(\boldsymbol{b}_1)\geq 0,$$

if Δ is a positive (semi)definite matrix. Thus a weaker criterion for $\boldsymbol{b}_1$ to be preferred to $\boldsymbol{b}_2$ is that $\delta\geq 0$.

For any two linear, homogeneous estimators $\boldsymbol{b}_i = \boldsymbol{C}_i\boldsymbol{y},\ i = 1, 2$ we get

$$\begin{aligned}\Delta &= \Sigma(\boldsymbol{b}_2) - \Sigma(\boldsymbol{b}_1) + \text{Bias}(\boldsymbol{b}_2)\text{Bias}^T(\boldsymbol{b}_2) - \text{Bias}(\boldsymbol{b}_1)\text{Bias}^T(\boldsymbol{b}_1)\\ &= \sigma^2\boldsymbol{S} - \text{Bias}(\boldsymbol{b}_1)\text{Bias}^T(\boldsymbol{b}_1) + \text{Bias}(\boldsymbol{b}_2)\text{Bias}^T(\boldsymbol{b}_2),\end{aligned}$$

where $\boldsymbol{S} = \boldsymbol{C}_2\boldsymbol{C}_2^T - \boldsymbol{C}_1\boldsymbol{C}_1^T$. If the matrix

$$\sigma^2\boldsymbol{S} - \text{Bias}(\boldsymbol{b}_1)\text{Bias}^T(\boldsymbol{b}_1) \tag{2.2.8}$$

is positive semidefinite, the matrix Δ can be written as the sum of two positive semidefinite matrices, because from Theorem A.9.2, (5) we know that the matrix $\text{Bias}(\boldsymbol{b}_2)\text{Bias}^T(\boldsymbol{b}_2)$ is positive semidefinite. As a consequence Δ is also a positive semidefinite matrix. To prove the positive semidefiniteness of (2.2.8), we consider the following theorem.

THEOREM 2.2.2. *Let $\boldsymbol{A}$ be a $(p+1) \times (p+1)$ positive definite matrix, let $\boldsymbol{a}$ be a non–zero $(p+1) \times 1$ column vector and let d be a positive scalar. Then $d\boldsymbol{A} - \boldsymbol{a}\boldsymbol{a}^T$ is positive semidefinite, iff $\boldsymbol{a}^T\boldsymbol{A}^{-1}\boldsymbol{a} \leq d$.*

PROOF. See Farebrother (1976,[**12**], in the appendix).

□

It is not difficult to see, that the matrix given in (2.2.8) is of the type $d\boldsymbol{A} - \boldsymbol{a}\boldsymbol{a}^T$. We can write

$$\text{Bias}(\boldsymbol{b}_i) = (\boldsymbol{C}_i\boldsymbol{X} - \boldsymbol{I}_{p+1})\boldsymbol{\beta}, \quad i = 1, 2.$$

From Theorem 2.2.2, Trenkler (1980,[**57**]) obtained the following result.

LEMMA 2.2.3. *Let $\boldsymbol{b}_i = \boldsymbol{C}_i\boldsymbol{y},\ i = 1, 2$ be two homogeneous linear estimators of $\boldsymbol{\beta}$ such that $\boldsymbol{S}$ is a positive definite matrix. Furthermore let the following inequality be valid*

$$\boldsymbol{\beta}^T(\boldsymbol{C}_1\boldsymbol{X} - \boldsymbol{I}_{p+1})^T\boldsymbol{S}^{-1}(\boldsymbol{C}_1\boldsymbol{X} - \boldsymbol{I}_{p+1})\boldsymbol{\beta} < \sigma^2.$$

Then

$$\Delta = \text{MtxMSE}(\boldsymbol{b}_2) - \text{MtxMSE}(\boldsymbol{b}_1) > 0,$$

where $\boldsymbol{S} = \boldsymbol{C}_2\boldsymbol{C}_2^T - \boldsymbol{C}_1\boldsymbol{C}_1^T$.

Within the class of homogeneous linear estimators there is a *best–linear estimator* of $\boldsymbol{\beta}$ with respect to the matrix mean squared error, namely

$$\hat{\boldsymbol{\beta}}_{opt} = \boldsymbol{A}_0\boldsymbol{y}$$

with

$$\boldsymbol{A}_0 = \boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{X}^T + \sigma^2\boldsymbol{I}_n\right)^{-1},$$

see e.g. G. Trenkler (1981,[**58**]).
In Stahlecker and Trenkler (1983,[**49**]) a heterogeneous version of the best–linear estimator is considered. But since both estimators depend on the unknown parameters $\boldsymbol{\beta}$ and σ^2, they are not operational.

Additional Reading 2.2.4. In G. Trenkler (1980,[**57**]) a comparison of some biased estimators (see Chapter 5) with respect to the generalized mean squared error is given.
Criteria for comparison of more general estimators is presented by D. Trenkler and G. Trenkler (1983,[**59**]). The interested reader may also consult the referenced article of G. Trenkler and Ihorst (1990,[**63**]).

Note 2.2.5. Within this manuscript we will use the mean squared error (2.1.4) for measuring the performance of different estimators. It should be mentioned that there also exists other loss functions, which may be more appropriate for many given problems.
In Varian (1975,[**66**]) the LINEX (linear–exponential) loss function is introduced. It depends not only upon the second moment of $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$, but also upon the entire sets of moments.
In Zellner (1994,[**72**]) a balanced loss function is proposed, which incorporates a measure for the goodness of fit of the model as well as a measure for the precision of the estimation. A good overview and more advices on literature about the LINEX and balanced loss functions are given in Rao and Toutenbourg (2007,[**45**]).

CHAPTER 3

Standardization of the Regression Coefficients

It is usually difficult to compare regression coefficients, when they are measured in different units of measurement or differ extremely in their magnitude. For this reason it is sometimes helpful to work with scaled regressors. By standardization we mean here changing the origin and also the scale of the data.

3.1. Centering Regression Models

We consider a linear regression model with intercept

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n,\ n \geq p + 1. \tag{3.1.1}$$

We have

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_{i,1} + \ldots + \beta_p \frac{1}{n}\sum_{i=1}^{n} x_{i,p} + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i,$$

or in another notation

$$\bar{y} = \beta_0 + \beta_1 \bar{X}_1 + \ldots + \beta_p \bar{X}_p + \bar{\varepsilon}. \tag{3.1.2}$$

Subtracting (3.1.2) from (3.1.1) implies a centered regression model without intercept

$$y_i^c = \beta_1 x_{i,1}^c + \ldots + \beta_p x_{i,p}^c + \varepsilon_i^c, \quad i = 1, \ldots, n,$$

or in vector notation

$$\boldsymbol{y}^c = \boldsymbol{X}^c \boldsymbol{\beta}_{\{\beta_0\}} + \boldsymbol{\varepsilon}^c, \tag{3.1.3}$$

where

$$\begin{aligned}
\boldsymbol{\beta}_{\{\beta_0\}}^T &:= \begin{bmatrix} \beta_1, & \ldots & , \beta_p \end{bmatrix}, \\
x_{i,j}^c &:= x_{i,j} - \bar{X}_j, \\
y_i^c &:= y_i - \bar{y}, \\
\varepsilon_i^c &:= \varepsilon_i - \bar{\varepsilon}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, p.
\end{aligned} \tag{3.1.4}$$

Lemma 3.1.1. *Let*

$$\boldsymbol{P} := \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \quad \in \mathbb{R}^{n\times n},$$

where $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{1}_n$ an $n \times 1$ vector consisting only of ones. $\boldsymbol{P}$ is a projection matrix and it is

$$\begin{aligned}\boldsymbol{y}^c &= \boldsymbol{P}\boldsymbol{y},\\ \boldsymbol{\varepsilon}^c &= \boldsymbol{P}\boldsymbol{\varepsilon},\\ \boldsymbol{X}^c &= \boldsymbol{P}\boldsymbol{X}_{\{1\}},\end{aligned}$$

with

$$\boldsymbol{X}_{\{1\}} := \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,p}\end{bmatrix} \in \mathbb{R}^{n\times p}.$$

Proof. A symmetric matrix $\boldsymbol{P}$ is a projection matrix, iff $\boldsymbol{P}$ is idempotent, i.e. $\boldsymbol{P}^2 = \boldsymbol{P}$ (see Appendix A.6). It is easy to see, that $\boldsymbol{P}^T = \boldsymbol{P}$ and

$$\begin{aligned}\boldsymbol{P}^2 = \boldsymbol{P}^T\boldsymbol{P} &= (\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)^T(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\\ &= \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^T\mathbf{1}_n\mathbf{1}_n^T\\ &= \boldsymbol{I}_n - \frac{2}{n}\mathbf{1}_n\mathbf{1}_n^T + \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\\ &= \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T = \boldsymbol{P}.\end{aligned}$$

Furthermore it is

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{y} - \begin{bmatrix}\bar{y}\\ \vdots \\ \bar{y}\end{bmatrix} = \boldsymbol{y}^c$$

and analogously $\boldsymbol{\varepsilon}^c = \boldsymbol{P}\boldsymbol{\varepsilon}, \boldsymbol{X}^c = \boldsymbol{P}\boldsymbol{X}_{\{1\}}$.

□

Thus (3.1.3) can be written as

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{X}_{\{1\}}\boldsymbol{\beta}_{\{\beta_0\}} + \boldsymbol{P}\boldsymbol{\varepsilon}. \tag{3.1.5}$$

Minimizing the residual sum of squares of the centered model

$$\begin{aligned}\mathrm{RSS}(\boldsymbol{\beta}^*_{\{\beta_0\}}) &:= \left(\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{X}_{\{1\}}\boldsymbol{\beta}^*_{\{\beta_0\}}\right)^T \left(\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{X}_{\{1\}}\boldsymbol{\beta}^*_{\{\beta_0\}}\right)\\ &= \left(\boldsymbol{y} - \boldsymbol{X}_{\{1\}}\boldsymbol{\beta}^*_{\{\beta_0\}}\right)^T \boldsymbol{P} \left(\boldsymbol{y} - \boldsymbol{X}_{\{1\}}\boldsymbol{\beta}^*_{\{\beta_0\}}\right)\end{aligned}$$

results in the least squares estimator $\hat{\boldsymbol{\beta}}^c := \left[\hat{\beta}_1^c, \quad \dots \quad , \hat{\beta}_p^c\right]$ of the centered model (3.1.3)

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^c &= \left((\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} (\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T \boldsymbol{P}\boldsymbol{y} \\
&= \left((\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} (\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T (\boldsymbol{P}\boldsymbol{X}_{\{1\}}\boldsymbol{\beta}_{\{\beta_0\}} + \boldsymbol{P}\boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta}_{\{\beta_0\}} + \left((\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} (\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T \boldsymbol{P}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_{\{\beta_0\}} + \left(\boldsymbol{X}_{\{1\}}^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} \boldsymbol{X}_{\{1\}}^T \boldsymbol{P}\boldsymbol{\varepsilon}, \qquad (3.1.6)
\end{aligned}$$

where $\boldsymbol{\beta}_{\{\beta_0\}} = \left[\beta_1, \quad \dots \quad , \beta_p\right]^T$. It follows

$$\mathrm{E}(\hat{\boldsymbol{\beta}}^c) = \boldsymbol{\beta}_{\{\beta_0\}} + \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P}\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}_{\{\beta_0\}}$$

and

$$\begin{aligned}
\Sigma(\hat{\boldsymbol{\beta}}^c) &= \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P}\Sigma(\boldsymbol{\varepsilon}) \left(\left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P}\right)^T \\
&= \sigma^2 \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1} = \sigma^2 \left(\boldsymbol{X}^{cT} \boldsymbol{X}^c\right)^{-1}.
\end{aligned}$$

From (3.1.2) we can get an estimator for β_0 using the estimated coefficients of the centered model

$$\hat{\beta}_0^c := \bar{y} - \sum_{i=1}^{p} \hat{\beta}_i^c \bar{X}_i = \bar{y} - \frac{1}{n}\boldsymbol{1}_n^T \boldsymbol{X}_{\{1\}}\hat{\boldsymbol{\beta}}^c. \qquad (3.1.7)$$

Consider now the unstandardized regression model (3.1.1) in vector notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{1}_n & \boldsymbol{X}_{\{1\}} \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_{\{\beta_0\}} \end{bmatrix}.$$

In this case the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, given in (1.0.5), can be written as

$$\begin{aligned}
\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{\{\beta_0\}} \end{bmatrix} &= \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \begin{bmatrix} \boldsymbol{1}_n^T \boldsymbol{1}_n & \boldsymbol{1}_n^T \boldsymbol{X}_{\{1\}} \\ \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{1}_n & \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{X}_{\{1\}} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{1}_n^T \boldsymbol{y} \\ \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \end{bmatrix}
\end{aligned}$$

$$= \begin{bmatrix} n & \mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \\ \boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n & \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{X}_{\{1\}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n^T \boldsymbol{y} \\ \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \end{bmatrix}. \tag{3.1.8}$$

Using Theorem A.5.2 for matrix, we can deduce from (3.1.8)

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \begin{bmatrix} \frac{1}{n} + \frac{1}{n^2}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \boldsymbol{Q}^{-1} \boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n & -\frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \boldsymbol{Q}^{-1} \\ -\frac{1}{n}\boldsymbol{Q}^{-1}\boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n & \boldsymbol{Q}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}_n^T \boldsymbol{y} \\ \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n}\mathbf{1}_n^T \boldsymbol{y} + \frac{1}{n^2}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \boldsymbol{Q}^{-1} \boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n \mathbf{1}_n^T \boldsymbol{y} - \frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \boldsymbol{Q}^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \\ -\frac{1}{n}\boldsymbol{Q}^{-1}\boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n \mathbf{1}_n^T \boldsymbol{y} + \boldsymbol{Q}^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \end{bmatrix}, \end{aligned}$$

with

$$\begin{aligned} \boldsymbol{Q} &= \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{X}_{\{1\}} - \frac{1}{n} (\mathbf{1}_n^T \boldsymbol{X}_{\{1\}})^T (\mathbf{1}_n^T \boldsymbol{X}_{\{1\}}) \right) \\ &= \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} = \left(\boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^T \boldsymbol{P}\boldsymbol{X}_{\{1\}}. \end{aligned}$$

Hence it follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\{\beta_0\}} &= -\frac{1}{n}\boldsymbol{Q}^{-1}\boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n \mathbf{1}_n^T \boldsymbol{y} + \boldsymbol{Q}^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \\ &= -\frac{1}{n} \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n \mathbf{1}_n^T \boldsymbol{y} \\ &\quad + \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \\ &= \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} - \frac{1}{n} \left(\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \right)^T \mathbf{1}_n^T \boldsymbol{y} \right) \\ &= \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \boldsymbol{X}_{\{1\}}^T \boldsymbol{P} \boldsymbol{y} \\ &= \left((\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} (\boldsymbol{P}\boldsymbol{X}_{\{1\}})^T \boldsymbol{P} \boldsymbol{y} \end{aligned}$$

and from the first row of (3.1.6) we see

$$\hat{\boldsymbol{\beta}}_{\{\beta_0\}} = \hat{\boldsymbol{\beta}}^c.$$

Furthermore it follows with Lemma 3.1.1

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{n}\mathbf{1}_n^T \boldsymbol{y} + \frac{1}{n^2}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \boldsymbol{Q}^{-1} \boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n \mathbf{1}_n^T \boldsymbol{y} - \frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \boldsymbol{Q}^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \\ &= \bar{y} + \frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \mathbf{1}_n \bar{y} \\ &\quad - \frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{y} \\ &= \bar{y} - \frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \left(\boldsymbol{X}_{\{1\}}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \boldsymbol{X}_{\{1\}}^T \left(\boldsymbol{y} - \mathbf{1}_n \bar{y} \right) \\ &= \bar{y} - \frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \left(\boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{X}_{\{1\}} \right)^{-1} \boldsymbol{X}_{\{1\}}{}^T \boldsymbol{P} \boldsymbol{y} \\ &= \bar{y} - \frac{1}{n}\mathbf{1}_n^T \boldsymbol{X}_{\{1\}} \hat{\boldsymbol{\beta}}_{\{\beta_0\}} \end{aligned}$$

and thus we have from (3.1.7)

$$\hat{\beta}_0 = \hat{\beta}_0^c.$$

Hence we get exactly the same estimated coefficients by minimizing the residual sum of squares of the uncentered (3.1.1) or of the centered model (3.1.5).

3.2. Scaling Centered Regression Models

The following standardization converts the matrix $\boldsymbol{X}^T\boldsymbol{X}$ into a correlation matrix. Let $\boldsymbol{y}^* := \boldsymbol{y}^c$ denote the centered dependent variable as before and put

$$z_{i,j} = \frac{x_{i,j} - \bar{X}_j}{\sqrt{S_{jj}}}, \quad i = 1, \ldots, n, \; j = 1, \ldots, p$$

with

$$S_{jj} = \sum_{i=1}^{n} (x_{i,j} - \bar{X}_j)^2. \tag{3.2.9}$$

Using these new variables, the regression model (3.1.1) becomes

$$y_i^* = \gamma_1 z_{i,1} + \gamma_2 z_{i,2} + \cdots + \gamma_p z_{i,p} + \varepsilon_i^*, \quad i = 1, \ldots, n,$$

or in vector notation

$$\boldsymbol{y}^* = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}^*, \tag{3.2.10}$$

where

$$\begin{aligned}
\boldsymbol{y}^* &= \boldsymbol{Py}, \\
\boldsymbol{Z} &=: \begin{bmatrix} Z_1 & \ldots & Z_p \end{bmatrix} = \boldsymbol{PX}_{\{1\}}\boldsymbol{D}^{-1}, \\
\boldsymbol{\gamma} &= \boldsymbol{D\beta}_{\{\beta_0\}}, \\
\boldsymbol{\varepsilon}^* &= \boldsymbol{P\varepsilon}
\end{aligned} \tag{3.2.11}$$

and

$$\boldsymbol{D} = \begin{bmatrix} \sqrt{S_{11}} & & \\ & \ddots & \\ & & \sqrt{S_{pp}} \end{bmatrix}. \tag{3.2.12}$$

All of the scaled regressors have sample mean equal to zero and the Euclidean norm of each column $Z_j := [z_{i,j}]_{1 \le i \le n}$, $j = 1, \ldots, p$ of $\boldsymbol{Z}$ is equal to one.
The least squares estimator of (3.2.10) is given by

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T\boldsymbol{y}^* \\
&= \left(\boldsymbol{D}^{-1}\boldsymbol{X}_{\{1\}}^T\boldsymbol{PX}_{\{1\}}\boldsymbol{D}^{-1}\right)^{-1} \boldsymbol{D}^{-1}\boldsymbol{X}_{\{1\}}^T\boldsymbol{Py}
\end{aligned}$$

$$= \boldsymbol{D}\left(\boldsymbol{X}_{\{1\}}^T\boldsymbol{P}\boldsymbol{X}_{\{1\}}\right)^{-1}\boldsymbol{X}_{\{1\}}^T\boldsymbol{P}\boldsymbol{y} = \boldsymbol{D}\hat{\boldsymbol{\beta}}_{\{\beta_0\}} \tag{3.2.13}$$

and thus the relationship between the estimates of the original and standardized regression coefficients is given by

$$\hat{\beta}_j = \hat{\gamma}_j\left(\frac{1}{S_{jj}}\right)^{\frac{1}{2}}, \quad j = 1, \ldots, p$$

and

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p}\hat{\beta}_j\bar{X}_j. \tag{3.2.14}$$

Furthermore we have from (3.2.13) and Theorem 1.0.1

$$\mathrm{E}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{D}\mathrm{E}(\hat{\boldsymbol{\beta}}_{\{\beta_0\}}) = \boldsymbol{D}\boldsymbol{\beta}_{\{\beta_0\}}.$$

With (2.0.1) it follows

$$\begin{aligned}\Sigma(\hat{\boldsymbol{\gamma}}) &= \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\Sigma(\boldsymbol{y}^*)\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{P}\Sigma(\boldsymbol{y})\boldsymbol{P}^T\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\\ &= \sigma^2\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\\ &= \sigma^2\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} - \frac{\sigma^2}{n}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\\ &= \sigma^2\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1},\end{aligned} \tag{3.2.15}$$

because $\boldsymbol{Z}$ is centered and thus $\boldsymbol{Z}^T\mathbf{1}_n = \mathbf{0} \in \mathbb{R}^{p\times 1}$.

Note 3.2.1. Many computer programs use scaling to reduce problems arising from round–off errors in the $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ matrix. But it is up to the decision of the analyst whether to use standardized data or not. For a discussion about standardization in regression theory see Section 5.3 in Chapter 5.

From (3.2.11) it follows

$$\begin{aligned}\mathrm{E}(\boldsymbol{\varepsilon}^*) &= \mathbf{0}\\ \Sigma(\boldsymbol{\varepsilon}^*) &= \boldsymbol{P}\Sigma(\boldsymbol{\varepsilon})\boldsymbol{P}^T = \sigma^2\boldsymbol{P},\end{aligned} \tag{3.2.16}$$

i.e. the covariance matrix of $\boldsymbol{\varepsilon}^*$ differs from the corresponding one of the vector of the error terms $\boldsymbol{\varepsilon}$ in the uncentered model. Consider now the residual sum of squares of $\hat{\boldsymbol{\gamma}}$

$$\begin{aligned}\mathrm{RSS}(\hat{\boldsymbol{\gamma}}) &= (\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})\\ &= \left(\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{X}_{\{1\}}\hat{\boldsymbol{\beta}}_{\{\beta_0\}}\right)^T\left(\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{X}_{\{1\}}\hat{\boldsymbol{\beta}}_{\{\beta_0\}}\right)\\ &= \left(\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{X}_{\{1\}}\hat{\boldsymbol{\beta}}_{\{\beta_0\}} - \boldsymbol{P}\boldsymbol{\beta}_0\right)^T\left(\boldsymbol{P}\boldsymbol{y} - \boldsymbol{P}\boldsymbol{X}_{\{1\}}\hat{\boldsymbol{\beta}}_{\{\beta_0\}} - \boldsymbol{P}\boldsymbol{\beta}_0\right)\end{aligned}$$

with $\boldsymbol{\beta}_0^T := \left[\beta_0, \quad \dots \quad, \beta_0\right] \in \mathbb{R}^{1\times n}$ and $\boldsymbol{P}\boldsymbol{\beta}_0 = \boldsymbol{0}$. It follows

$$\mathrm{RSS}(\hat{\boldsymbol{\gamma}}) = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{P} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \hat{\boldsymbol{\varepsilon}}^T \boldsymbol{P} \hat{\boldsymbol{\varepsilon}}.$$

From (1.0.7) we have

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\varepsilon}}) &= \boldsymbol{0}, \\ \Sigma(\hat{\boldsymbol{\varepsilon}}) &= \sigma^2 \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right) \end{aligned}$$

and we get with Lemma 1.0.3, Theorem A.6.2 and (A.1.2)

$$\begin{aligned} \mathrm{E}\left(\mathrm{RSS}(\hat{\boldsymbol{\gamma}})\right) &= \mathrm{E}\left(\hat{\boldsymbol{\varepsilon}}^T \boldsymbol{P} \hat{\boldsymbol{\varepsilon}}\right) \\ &= \mathrm{E}\left(\hat{\boldsymbol{\varepsilon}}^T (\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^T)\hat{\boldsymbol{\varepsilon}}\right) \\ &= \mathrm{E}\left(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}\right) - \frac{1}{n}\mathrm{E}\left(\hat{\boldsymbol{\varepsilon}}^T\boldsymbol{1}_n\boldsymbol{1}_n^T\hat{\boldsymbol{\varepsilon}}\right) \\ &= (n-p-1)\sigma^2 - \frac{1}{n}\mathrm{tr}(\boldsymbol{1}_n\boldsymbol{1}_n^T\Sigma(\hat{\boldsymbol{\varepsilon}})) \\ &= (n-p-1)\sigma^2 - \frac{\sigma^2}{n}\mathrm{tr}\left(\boldsymbol{1}_n\boldsymbol{1}_n^T(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\right) \\ &= (n-p-2)\sigma^2 + \frac{\sigma^2}{n}\boldsymbol{1}_n^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{1}_n. \qquad (3.2.17) \end{aligned}$$

Hence we have

$$\mathrm{E}\left(\mathrm{RSS}(\hat{\boldsymbol{\gamma}})\right) \neq \mathrm{E}\left(\mathrm{RSS}(\hat{\boldsymbol{\beta}})\right).$$

An unbiased estimator of σ^2 is then given by

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{\gamma}})}{n-p-2+\frac{1}{n}\boldsymbol{1}_n^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{1}_n}.$$

CHAPTER 4

Multicollinearity

Applications of regression models exist in almost every field of research, all of which require estimates of the unknown parameters. Important decisions are often based on the magnitudes of these individual estimates, e.g. tests of significance associated with them. These decisions and inferences can be misleading, even erroneous, when *multicollinearity* is present in the data. The columns $X_1, \ldots, X_p \in \mathbb{R}^{n \times 1}$ of the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ are said to be linearly dependent, if there exists a non–trivial solution $\nu_1, \ldots, \nu_p \in \mathbb{R}$ of the equation

$$\sum_{j=1}^{p} \nu_j X_j = 0. \tag{4.0.1}$$

If (4.0.1) holds for the columns of $\boldsymbol{X}$, multicollinearity is said to exist. In regression theory already near linear dependencies among the regressors, which result in a near singularity of the matrix $\boldsymbol{X}^T\boldsymbol{X}$, are defined as multicollinearity. Thus the question of multicollinearity, as Farrar and Glauber (1967,[**14**]) pointed out, is not one of existence but one of degree. "Multicollinearity" is used in this manuscript when (4.0.1) is approximately true and "exact multicollinearity" when the relationship is exact. Before discussing the effects and detection of multicollinearity in more detail, some sources of the phenomenon are examined.

4.1. Sources of Multicollinearity

Multicollinearity can occur for a variety of reasons, but there are primarily the following three sources

(1) an overdefined model,
(2) sampling techniques and
(3) physical constraints on the model.

An overdefined model has more regressors than observations. This type of model arises frequently in medical research where many pieces of information are taken on each individual in a study. The usual approach of dealing with multicollinearity in this context is to eliminate some of the regressor variables from consideration. With Assumption 3 in Chapter 1 we exclude this situation for the rest of the considerations within this manuscript. The second source of multicollinearity arises when the analyst samples only a subspace of the region of the regressors.

This subspace is approximately a hyperplane defined by one ore more of the relationships of the form (4.0.1). Constraints on the model or the population being sampled can cause multicollinearity. Constraints often occur in problems, where the regressors have to add to a constant.

4.2. Harmful Effects of Multicollinearity

The presence of multicollinearity has a number of potentially serious effects on the least squares estimates of the regression coefficients. These effects can be comprehended easily, if there are only two regressors Z_1 and Z_2 in the standardized model (3.2.10), discussed in Chapter 3. Denote by $\rho_{1,2}$ the correlation coefficient between the two regressors and $\rho_{y,i}$ the correlation coefficient between the centerted dependent variable $\boldsymbol{y}^*$ and $Z_i,\ i = 1, 2$. The least squares estimator $\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^*$ requires the computation of the inverse

$$\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} = \frac{\begin{bmatrix} 1 & -\rho_{1,2} \\ -\rho_{1,2} & 1 \end{bmatrix}}{|\boldsymbol{Z}^T\boldsymbol{Z}|},$$

where $|\boldsymbol{Z}^T\boldsymbol{Z}| = 1 - \rho_{1,2}^2$. As equation (4.0.1) becomes exact, it follows $\rho_{1,2}^2 \to 1$ and $|\boldsymbol{Z}^T\boldsymbol{Z}| \to 0$. As a consequence $\operatorname{var}(\hat{\gamma}_i) \to \infty,\ i = 1, 2$ and $\operatorname{cov}(\hat{\gamma}_i, \hat{\gamma}_j) \to \pm\infty$ for $\rho_{1,2} \to \mp 1$, because we know from (3.2.15) that the covariance matrix of $\hat{\boldsymbol{\gamma}}$ is given by

$$\Sigma(\hat{\boldsymbol{\gamma}}) = \sigma^2 \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}.$$

Thus a strong pairwise linear relationship between Z_1 and Z_2 results in very large variances and covariances for the estimates of the regression coefficients. Consider now the least squares estimator of $\boldsymbol{\gamma}$

$$\hat{\boldsymbol{\gamma}} = \frac{\begin{bmatrix} \rho_{y,1} - \rho_{1,2}\rho_{y,2} \\ \rho_{y,2} - \rho_{1,2}\rho_{y,1} \end{bmatrix}}{|\boldsymbol{Z}^T\boldsymbol{Z}|}.$$

Assume now $\rho_{1,2} = 1$. As a consequence it is $|\boldsymbol{Z}^T\boldsymbol{Z}| = 0$ and thus $\hat{\boldsymbol{\gamma}}$ is not defined. However, note that

$$\hat{\gamma}_1 + \hat{\gamma}_2 = \left(\rho_{y,1}(1 - \rho_{1,2}) + \rho_{y,2}(1 - \rho_{1,2})\right)\left(|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^{-1} = \frac{(\rho_{y,1} + \rho_{y,2})}{(1 + \rho_{1,2})}$$

remains well defined even with $\rho_{1,2} = 1$. By contrast

$$\hat{\gamma}_1 - \hat{\gamma}_2 = \frac{(\rho_{y,2} - \rho_{y,1})}{(1 - \rho_{1,2})}$$

is not defined, i.e. $\pm(\hat{\gamma}_1 + \hat{\gamma}_2)$ is estimable, whereas $\pm(\hat{\gamma}_1 - \hat{\gamma}_2)$ is inestimable. Thus in the presence of exact multicollinearity, there exist linear combinations of the vector $\boldsymbol{\gamma}$ which are inestimable. This danger is particularly present in case of near multicollinearity, but it may not be immediately detected.

In the general case of p regressors it is more difficult to assess the effects of multicollinearity on individual parameter estimates, but some specific comments can be made.

4.2.1. The Variances of $\hat{\gamma}$

The covariance matrix of the least squares estimator $\hat{\boldsymbol{\gamma}}$ of the standardized model (3.2.10) is given by

$$\Sigma(\hat{\boldsymbol{\gamma}}) = \sigma^2 \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}.$$

With the spectral decomposition $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ (see Theorem A.8.1) and (A.1.2) we get

$$\sum_{j=1}^{p} \mathrm{var}(\hat{\gamma}_j) = \sigma^2 \mathrm{tr}(\boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^T) = \sigma^2 \mathrm{tr}(\underbrace{\boldsymbol{V}^T\boldsymbol{V}}_{=\boldsymbol{I}_p}\boldsymbol{\Lambda}^{-1}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}, \tag{4.2.2}$$

where $\lambda_j,\ j = 1, \dots, p$ denote the eigenvalues of $\boldsymbol{Z}^T\boldsymbol{Z}$ in descending order. If there exists strong multicollinearity between the regressors $Z_j,\ j = 1, \dots, p$ at least one eigenvalue will be very small (follows from Theorem A.8.3) and the total variance of $\hat{\boldsymbol{\gamma}}$ will be very large. Consider

$$\mathrm{var}(\hat{\gamma}_j) = \sigma^2 \sum_{k=1}^{p} \frac{v_{j,k}^2}{\lambda_k}, \tag{4.2.3}$$

where $v_{j,k}$ denotes the (j,k)-th element of the matrix $\boldsymbol{V} =: [v_{j,k}]_{1\le j,k\le p}$. Since λ_p is the smallest eigenvalue, it is usually the case that the p-th summand in (4.2.3) is responsible for a large variance. However, sometimes $v_{j,k}$ is also small and the p-th summand in (4.2.3) is small compared to the remaining summands. Then at least one of the remaining eigenvalues can be responsible for a large variance. Theil (1971,[**54**], p. 166) showed that the diagonal elements of $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ can be expressed as

$$r_{j,j} = \frac{1}{1 - R_j^2}, \quad j = 1, \dots, p,$$

where R_j^2 represents the squared multiple correlation coefficient (see Falk (2002,[**11**]), chapter 3), when Z_j is regressed on the remaining $(p-1)$ regressors. In case of multicollinearity, one or some $R_j^2,\ j = 1, \dots, p$ will be close to unity and hence $r_{j,j}$ will be large. Since the variance of $\hat{\gamma}_j$ is

$$\mathrm{var}(\hat{\gamma}_j) = \frac{\sigma^2}{1 - R_j^2}, \quad j = 1, \dots, p,$$

a value of R_j^2 close to unity implies a large variance of the corresponding least squares estimate of γ_j.

4.2.2. Unstable and Large Estimates of γ

When important statistical decisions are based on the results of a regression analysis, the researcher needs a stable or robust result. We consider how the least squares estimates of $\boldsymbol{\gamma}$ can be affected by small changes in the design matrix $\boldsymbol{Z}$ or the dependent variable $\boldsymbol{y}$. Denote the perturbed $\boldsymbol{y}^*$ by $\boldsymbol{y}_p^*$ and perturbed $\hat{\boldsymbol{y}}^* = \boldsymbol{Z}\boldsymbol{\gamma} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^*$ by $\hat{\boldsymbol{y}}_p^* = \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}_p^*$. It can be obtained

$$\frac{\left\|\hat{\gamma} - \boldsymbol{Z}^+\boldsymbol{y}_p^*\right\|_2}{\|\hat{\gamma}\|_2} \leq \operatorname{cond}(\boldsymbol{Z}^T\boldsymbol{Z})\frac{\left\|\hat{\boldsymbol{y}}^* - \hat{\boldsymbol{y}}^*{}_p\right\|_2}{\|\hat{\boldsymbol{y}}^*\|_2}, \tag{4.2.4}$$

where $\boldsymbol{Z}^+ := \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T$ and $\|\cdot\|_2$ denotes the Euclidean norm. Thus the effect of perturbations in $\boldsymbol{y}^*$ on $\hat{\gamma}$ may be amplified by the condition number $\operatorname{cond}(\boldsymbol{Z}^T\boldsymbol{Z})$ of $\boldsymbol{Z}^T\boldsymbol{Z}$, which is usually greater than unity (see Appendix A.8.1). In the presence of multicollinearity $\operatorname{cond}(\boldsymbol{Z}^T\boldsymbol{Z}) \gg 1$ and the effect may be severe. On the other hand it can be stated

$$\begin{aligned}\frac{\left\|\hat{\gamma} - (\boldsymbol{Z}+\boldsymbol{E})^+\boldsymbol{y}_p^*\right\|_2}{\|\hat{\gamma}\|_2} &\leq \operatorname{cond}(\boldsymbol{Z}^T\boldsymbol{Z})\,\|\boldsymbol{E}_1\|_2 \\ &+ \operatorname{cond}(\boldsymbol{Z}^T\boldsymbol{Z})^2\,\|\boldsymbol{E}_2\|_2\,\frac{\|\boldsymbol{y}^* - \hat{\boldsymbol{y}}^*\|_2}{\|\hat{\boldsymbol{y}}^*\|_2} + \operatorname{cond}(\boldsymbol{Z}^T\boldsymbol{Z})^3\,\|\boldsymbol{E}_2\|_2^2\,,\end{aligned} \tag{4.2.5}$$

where $\boldsymbol{E}$ denotes the matrix of perturbations and $\boldsymbol{E} = \boldsymbol{E}_1 + \boldsymbol{E}_2$. $\boldsymbol{E}_1$ is the component of $\boldsymbol{E}$ lying in the column space of $\boldsymbol{Z}$ and $\boldsymbol{E}_2$ is the component orthogonal to the column space of $\boldsymbol{Z}$. The main point of (4.2.5) is that the upper bound can be very large, if $\operatorname{cond}(\boldsymbol{Z}^T\boldsymbol{Z})$ is large. Thus data perturbations in $\boldsymbol{y}^*$ or $\boldsymbol{Z}$ may change $\|\hat{\gamma}\|_2$ anywhere in the interval between 0 and the right hand sides of (4.2.4) and (4.2.5), respectively.

Note, that the upper bounds (4.2.4) and (4.2.5) are often too large compared to what might be expected.

For further explanation and exact expressions see Golub (1996,[**17**], chapter 2.7) and Stewart (1973,[**52**], chapter 4.4) and Stewart (1969,[**51**]).

Consider the squared distance from $\hat{\boldsymbol{\gamma}}$ to the true parameter vector $\boldsymbol{\gamma}$

$$\mathrm{L}^2 := [\hat{\gamma} - \gamma]^T\,[\hat{\gamma} - \gamma]\,.$$

With (4.2.2) we have

$$\mathrm{E}(\mathrm{L}^2) = \mathrm{MSE}(\hat{\boldsymbol{\gamma}}) = \sigma^2\mathrm{tr}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} = \sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j} \tag{4.2.6}$$

and thus the expected squared distance from the least squares estimator to the true parameter $\boldsymbol{\gamma}$ will be large in case of multicollinearity.

Furthermore we get

$$\begin{aligned}\mathrm{E}(\hat{\gamma}^T\hat{\gamma}) &= \mathrm{E}\left(\boldsymbol{y}^{*T}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-2}\boldsymbol{Z}^T\boldsymbol{y}^*\right)\\ &= \mathrm{E}\left((\boldsymbol{Z}\gamma+\boldsymbol{\varepsilon}^*)^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-2}\boldsymbol{Z}^T(\boldsymbol{Z}\gamma+\boldsymbol{\varepsilon}^*)\right)\\ &= \boldsymbol{\gamma}^T\gamma+\mathrm{E}\left(\boldsymbol{\varepsilon}^{*T}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-2}\boldsymbol{Z}^T\boldsymbol{\varepsilon}^*\right),\end{aligned}$$

because $\mathrm{E}(\boldsymbol{\varepsilon}^*) = \mathbf{0}$ from (3.2.16). With Theorem A.6.2, Lemma 3.1.1, (3.2.16) and (A.1.2) it follows

$$\begin{aligned}\mathrm{E}(\hat{\gamma}^T\hat{\gamma}) &= \boldsymbol{\gamma}^T\gamma+\mathrm{tr}\left(\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-2}\boldsymbol{Z}^T\Sigma(\boldsymbol{\varepsilon}^*)\right)\\ &= \boldsymbol{\gamma}^T\gamma+\sigma^2\mathrm{tr}\left((\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\right), \qquad (4.2.7)\end{aligned}$$

because $\boldsymbol{Z}^T\mathbf{1}_n = \mathbf{0}$. As a consequence $\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\gamma}}$ is on the average longer than $\boldsymbol{\gamma}^T\boldsymbol{\gamma}$ and multicollinearity strengthens this effect. From this fact, many authors, e.g. McDonald and Galarneau (1975,[**37**]), Marquardt and Snee (1975,[**33**]) or Hoerl and Kennard (1970,[**24**]) concluded that in case of multicollinearity, the average length of the least square estimator $\hat{\boldsymbol{\gamma}}$ is too large. Brook and Moore (1980,[**5**]) pointed out that this implication is obviously false, but they confirmed the statement that multicollinearity tends to produce least squares estimates $\hat{\gamma}_j$, $j = 1,\ldots,p$, which are too large in absolute value.

Silvey (1969, [**46**]) discussed the effects of multicollinearities on the estimation of parametric functions $\boldsymbol{p}^T\boldsymbol{\gamma}$ and showed that precise estimation is possible when $\boldsymbol{p} \in \mathbb{R}^{p\times 1}$ is a linear combination of eigenvectors corresponding to large eigenvalues of $\boldsymbol{Z}^T\boldsymbol{Z}$, whereas imprecise estimation occurs when $\boldsymbol{p}$ is a linear combination of eigenvectors corresponding to small eigenvalues of $\boldsymbol{Z}^T\boldsymbol{Z}$. It follows from Silvey's result and is proven explicitly by Greenberg (1975,[**18**]) that the linear combination $\boldsymbol{p}^T\boldsymbol{\gamma}$, $\boldsymbol{p}^T\boldsymbol{p} = 1$, for which the least squares estimator has its smallest variance occurs for $\boldsymbol{p}^T = V_1$ (where V_1 denotes the eigenvector associated with the largest eigenvalue of the matrix $\boldsymbol{Z}^T\boldsymbol{Z}$). It follows immediately that $\boldsymbol{p}^T\boldsymbol{\gamma}$ is estimated with maximum variance (using least squares) when $\boldsymbol{p}^T = V_p$ (eigenvector associated with the smallest eigenvalue). Therefore Silvey suggested collecting a new set of values for the regressors in the direction of a eigenvector (associated with a large eigenvalue), in order to combat multicollinearity. However, in practice, a complete freedom of choice of the new values may not be available.

NOTE 4.2.1. For the sake of completeness the following effects of multicollinearity should be mentioned.

- Because of the large variances of the estimates, the corresponding confidence intervals tend to be much wider and this may result in insignificant t-statistics. In contrast, the R^2–value of the model can still be relatively high.

- Farrar and Glauber (1967,[**14**]) proposed that multicollinearity can also result in $\hat{\gamma}_j$ to "have the wrong sign", i.e. opposite to the expectation of the researcher.

4.3. Multicollinearity Diagnostics

Suitable diagnostic procedures should directly reflect the degree of multicollinearity and provide helpful information in determining which regressors are involved. In literature several techniques have been proposed, but we will only discuss and illustrate the most common ones.

4.3.1. The Correlation Matrix and the Variance Inflation Factor

A very simple measure of multicollinearity is the inspection of the off diagonal elements $\rho_{i,j},\ i,j = 1,\ldots,p,\ i \neq j$ of the correlation matrix of $\boldsymbol{X}$. If the regressors X_i and $X_j,\ i,j = 1,\ldots,p,\ i \neq j$ are nearly linearly dependent, the absolute value of $\rho_{i,j}$ will be near to unity. Unfortunately, if more than two regressors are involved in a near linear dependence, there is no assurance that any of the pairwise correlations $\rho_{i,j}$ will be large. Generally, inspection of the $\rho_{i,j}$ is not sufficient for detecting anything more complex than pairwise multicollinearity. Nevertheless, the diagonal elements $r_{j,j},\ j = 1,\ldots,p$ of the inverse of the correlation matrix are very useful in detecting multicollinearity. We have seen in (4.2.3), that

$$\mathrm{VIF}_j := r_{j,j} = \frac{1}{1 - R_j^2} = \frac{\mathrm{var}(\hat{\gamma}_j)}{\sigma^2}, \quad j = 1,\ldots,p.$$

For $\boldsymbol{Z}^T\boldsymbol{Z}$ being an orthogonal matrix we have $r_{j,j} = 1$. Thus the *variance inflation factor* $\mathrm{VIF}_j,\ j = 1,\ldots,p$ can be viewed as the factor by which the variance of $\hat{\gamma}_j$ is increased due to the near linear dependence among the regressors. One or more large variance inflation factors indicate multicollinearity. Various recommendations have been made concerning the magnitudes of variance inflation factors which are indicative for multicollinearity. A variance inflation factor of 10 or greater is usually considered sufficient to indicate a multicollinearity.

4.3.2. The Eigensystem Analysis of $\boldsymbol{Z}^T\boldsymbol{Z}$

The eigenvalues $\lambda_1,\ldots,\lambda_p$ of $\boldsymbol{Z}^T\boldsymbol{Z}$ can be used to measure the extent of multicollinearity in the data. If there are near linear dependencies between the columns of $\boldsymbol{Z}$, one or more eigenvalues will be small. A good diagnostic indicator for multicollinearity is the *condition number* defined in Appendix A.8.1. With (A.8.13) we get

$$\mathrm{cond}(\boldsymbol{Z}^T\boldsymbol{Z}) = \frac{\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})}{\lambda_{\min}(\boldsymbol{Z}^T\boldsymbol{Z})}.$$

The condition number shows the spread in the eigenvalue spectrum of $\boldsymbol{Z}^T\boldsymbol{Z}$. Generally, if the condition number is less than 100, there is no serious problem with multicollinearity. Condition numbers between 100 and 1000 imply moderate to strong multicollinearity and if $\text{cond}(\boldsymbol{Z}^T\boldsymbol{Z})$ exceeds 1000, severe multicollinearity is indicated (see Belsley, Kuh and Welsch (1980,[**2**])). The *condition indices* of $\boldsymbol{Z}^T\boldsymbol{Z}$ are

$$\text{cond}_j(\boldsymbol{Z}^T\boldsymbol{Z}) = \frac{\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})}{\lambda_j(\boldsymbol{Z}^T\boldsymbol{Z})}, \quad j = 1, \dots, p. \tag{4.3.8}$$

Clearly the largest condition index is the condition number. The number of condition indices that are large is a useful measure of the number of near linear dependencies in $\boldsymbol{Z}^T\boldsymbol{Z}$.

Belsley, Kuh and Welsch (1980,[**2**]) proposed an approach based on the condition indices and the singular value decomposition of $\boldsymbol{Z}$ given in Theorem A.8.2. The $n \times p$ matrix $\boldsymbol{Z}$ can be decomposed in

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Theta}\boldsymbol{V}^T,$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times p}, \boldsymbol{V} \in \mathbb{R}^{p\times p}$ are both orthogonal and $\boldsymbol{\Theta} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with the singular values $\theta_k,\ k = 1, \dots, p$ on its diagonal. Multicollinearity between the columns of $\boldsymbol{Z}$ is reflected in the size of the singular values. Analogously to (4.3.8), Belsley, Kuh and Welsch (1980,[**2**]) defined the condition indices of $\boldsymbol{Z}$ by

$$\text{cond}_k(\boldsymbol{Z}) = \frac{\theta_{\max}(\boldsymbol{Z})}{\theta_k(\boldsymbol{Z})}, \quad k = 1, \dots, p.$$

Note, that this approach deals directly with the design matrix $\boldsymbol{Z}$. The covariance matrix of $\hat{\boldsymbol{\gamma}}$ is then given by

$$\Sigma(\hat{\boldsymbol{\gamma}}) = \sigma^2\boldsymbol{V}\boldsymbol{\Theta}^{-2}\boldsymbol{V}^T$$

and the variance of the j-th regression coefficient is given by

$$\text{var}(\hat{\gamma}_j) = \sigma^2 \sum_{k=1}^{p} \frac{v_{j,k}^2}{\theta_k^2} = \sigma^2\text{VIF}_j, \quad j = 1, \dots, p. \tag{4.3.9}$$

Equation (4.3.9) decomposes $\text{var}(\hat{\gamma}_j)$ into a sum of components, each associated with one and only one of the p singular values θ_k. Since these θ_k^2 appear in the denominator, other things being equal, those components associated with near

singular value	$\text{var}(\gamma_1)$	$\text{var}(\gamma_2)$	$\dots$	$\text{var}(\gamma_p)$
θ_1	$\pi_{1,1}$	$\pi_{1,2}$	$\cdots$	$\pi_{1,p}$
θ_2	$\pi_{2,1}$	$\pi_{2,2}$	$\cdots$	$\pi_{2,p}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
θ_p	$\pi_{p,1}$	$\pi_{p,2}$	$\cdots$	$\pi_{p,p}$

TABLE 4.3.1. Variance decomposition matrix

dependencies (small θ_k) will be large relative to the other components. This suggests that an unusually high proportion of the variance of two or more coefficients, concentrated in components associated with the same small singular value, provides evidence that the corresponding near dependency is causing problems. Therefore let

$$\phi_{j,k} := \frac{v_{j,k}^2}{\theta_k^2}$$

and

$$\mathrm{VIF}_j = \sum_{k=1}^{p} \phi_{j,k}, \quad j,k = 1,\dots,p.$$

Then, the variance decomposition proportions are

$$\pi_{j,k} := \frac{\phi_{j,k}}{\mathrm{VIF}_j}.$$

If we array $\pi_{j,k}$ in a *variance decomposition matrix* (see Table 4.3.1), the elements of each column are just the proportions of the variance of each $\hat{\gamma}_j$ contributed to the k-th singular value. If a high proportion of the variance for two or more regression coefficients is associated with one small singular value, multicollinearity is indicated. Belsley, Kuh and Welsch (1980,[**2**]) suggested, that the regressors should be scaled to unit length but not centered, when computing the variance decomposition matrix. Only then the role of the intercept in near linear dependences can be diagnosed. But there is still some controversy about this (therefore see also Section 5.3 in Chapter 5).

Note 4.3.1. The condition number is not invariant to scaling, i.e. the condition number of $\boldsymbol{X}^T\boldsymbol{X}$ is not equal to the condition number of $\boldsymbol{Z}^T\boldsymbol{Z}$. In case of the unstandardized matrix, the scale of the regressors or possible great differences in their magnitude can have an impact on the condition number and thus on the multicollinearity diagnostic. Therefore we suggested calculating the condition number of the standardized matrix $\boldsymbol{Z}^T\boldsymbol{Z}$.

Additional Reading 4.3.2. Most books about regression theory deal with the problem of multicollinearity, e.g. Vinod and Ullah (1981,[**67**]), Theil (1971,[**54**]), Draper and Smith (1981,[**9**]), Montgomery, Peck and Vining (2006,[**38**]) or Chatterjee and Hadi (2006,[**6**]).

For more detailed and not mentioned information about the sources, harmful effects and detection of multicollinearity, the reader is recommended to Mason (1975,[**35**]), Gunst (1983,[**21**]), Farrar and Glauber (1967,[**14**]), Stewart (1987,[**50**]), Willan and Watts (1978,[**70**]) to mention just a few. Another good overview and a more detailed examination of the eigenstructure of the design matrix in case of multicollinearity with detailed simulation studies is given in Belsley, Kuh and Welsch (1980,[**2**]).

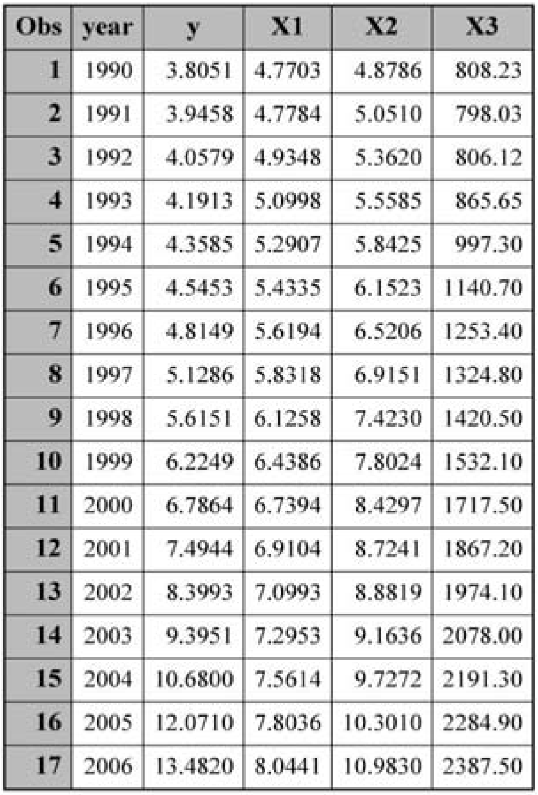
4.4. Example: Multicollinearity of the Economic Data

In the course of the financial crisis in the United States and over the whole world there is a big discussion about the life of the american people on credit. The data in Table 4.4.2 is taken from the Economic Report of the President (2007,[**10**]) and represents the relationship between the dependent variable

$\boldsymbol{y}$: Mortage dept outstanding (in trillions of dollars)

and the three other independent variables

$\boldsymbol{X}_1$: Personal consumption (in trillions of dollars),
$\boldsymbol{X}_2$: Personal income (in trillions of dollars),
$\boldsymbol{X}_3$: Consumer credit outstanding (in trillions of dollars).



Obs	year	y	X1	X2	X3
1	1990	3.8051	4.7703	4.8786	808.23
2	1991	3.9458	4.7784	5.0510	798.03
3	1992	4.0579	4.9348	5.3620	806.12
4	1993	4.1913	5.0998	5.5585	865.65
5	1994	4.3585	5.2907	5.8425	997.30
6	1995	4.5453	5.4335	6.1523	1140.70
7	1996	4.8149	5.6194	6.5206	1253.40
8	1997	5.1286	5.8318	6.9151	1324.80
9	1998	5.6151	6.1258	7.4230	1420.50
10	1999	6.2249	6.4386	7.8024	1532.10
11	2000	6.7864	6.7394	8.4297	1717.50
12	2001	7.4944	6.9104	8.7241	1867.20
13	2002	8.3993	7.0993	8.8819	1974.10
14	2003	9.3951	7.2953	9.1636	2078.00
15	2004	10.6800	7.5614	9.7272	2191.30
16	2005	12.0710	7.8036	10.3010	2284.90
17	2006	13.4820	8.0441	10.9830	2387.50

Table 4.4.2. Economic Data

Within the considered 17 years the mortage dept and consumer credit outstanding have tripled, whereas the personal income only doubled. It is obvious that the correlation between the independent variables and also the correlation between the dependent and the independent variables have to be high. We consider the regression model

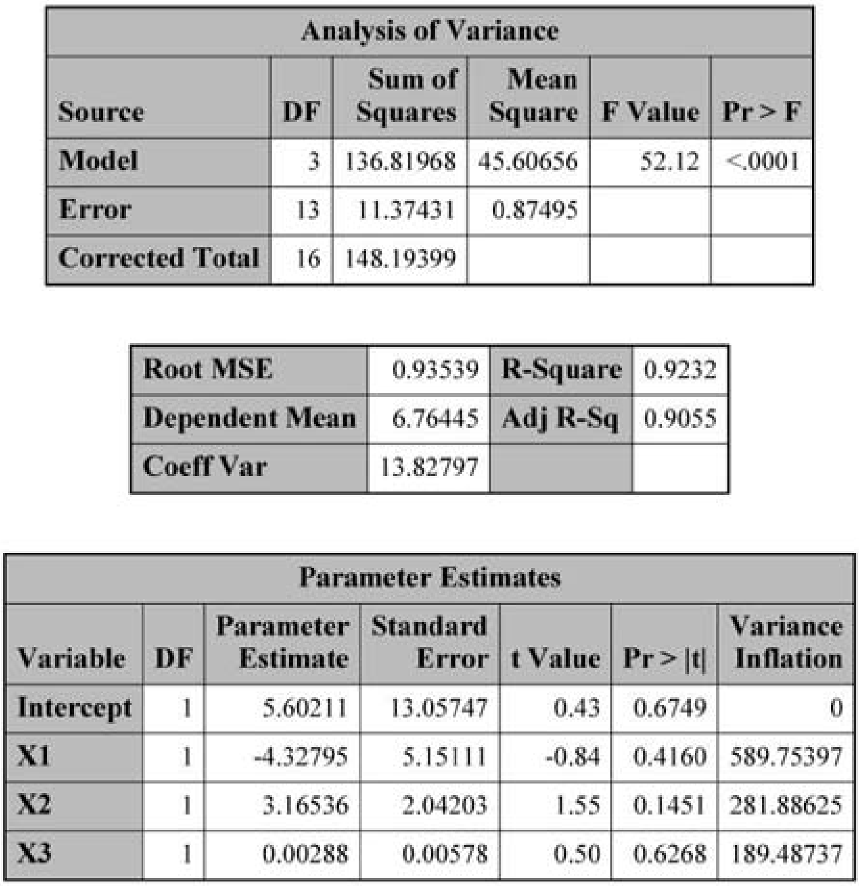
$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \beta_3 \boldsymbol{X}_3 + \boldsymbol{\varepsilon}$$

or in vector notation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{4.4.10}$$

and assume $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{17})$. The output of the `REG`–procedure of `SAS` in Table 4.4.3 shows the least squares estimator $\hat{\boldsymbol{\beta}}^T := \begin{bmatrix}\hat{\beta}_0, & \hat{\beta}_1, & \hat{\beta}_2, & \hat{\beta}_3\end{bmatrix}$ of $\boldsymbol{\beta}^T := \begin{bmatrix}\beta_0, & \beta_1, & \beta_2, & \beta_3\end{bmatrix}$ and the summary statistics of the model.
The following examinations indicate that multicollinearity may cause problems.

- The R^2–value of the model is high, whereas the p–values of the t–tests for the individual parameters tell us, that none parameter is statistical

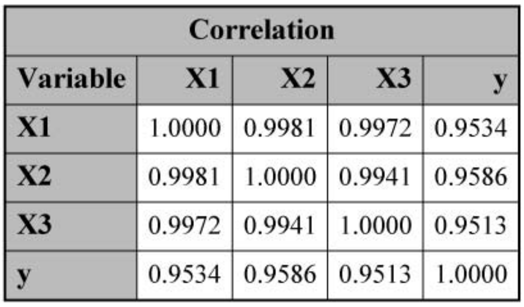


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	136.81968	45.60656	52.12	<.0001
Error	13	11.37431	0.87495		
Corrected Total	16	148.19399			

Root MSE	0.93539	R-Square	0.9232
Dependent Mean	6.76445	Adj R-Sq	0.9055
Coeff Var	13.82797		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	5.60211	13.05747	0.43	0.6749	0
X1	1	-4.32795	5.15111	-0.84	0.4160	589.75397
X2	1	3.16536	2.04203	1.55	0.1451	281.88625
X3	1	0.00288	0.00578	0.50	0.6268	189.48737

TABLE 4.4.3. Analysis of variance and parameter estimates of the Economic Data



Correlation				
Variable	**X1**	**X2**	**X3**	**y**
X1	1.0000	0.9981	0.9972	0.9534
X2	0.9981	1.0000	0.9941	0.9586
X3	0.9972	0.9941	1.0000	0.9513
y	0.9534	0.9586	0.9513	1.0000

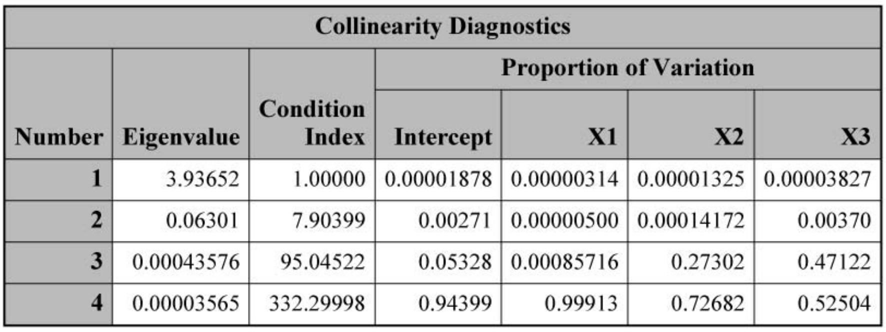
TABLE 4.4.4. Correlation matrix of the Economic Data

significant. Thus the summary statistics says that the three independent variables taken together are important, but any regressor may be deleted from the model provided the others are retained. These results are a characteristical for models, where multicollinearity is present.

- $\boldsymbol{y}$ and $\boldsymbol{X}_1$ are positively correlated and thus we would not expect a negative estimate of β_1.
- Table 4.4.3 displays the variance inflation factors of the model, which are the diagonal elements of the inverse of the correlation matrix of $\boldsymbol{X}$. Because all variance inflation factors are greater than 10, multicollinearity is indicated.
- Table 4.4.4 shows, that the pairwise correlation coefficients of the three independent variables are high. Thus there is a strong linear relationship among all pairs of regressors.
- Finally we also consider the examination of the eigensystem of $\boldsymbol{X}$. `SAS` follows the approach of Belsley, Kuh and Welsch (1980,[**2**]), i.e. before calculating the eigenvalues and the variance decomposition matrix the columns of $\boldsymbol{X}$ are scaled to have unit length.
  The analysis in `REG`–procedure is reported to the eigenvalues of the scaled matrix of $\boldsymbol{X}^T\boldsymbol{X}$ rather than the singular values. But from (A.8.13) we know, that the eigenvalues of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ are the squares of the singular values of $\boldsymbol{X}$. The condition indices are the square roots of the ratio of the largest eigenvalue $\lambda_{\max}$ of the scaled matrix $\boldsymbol{X}^T\boldsymbol{X}$ to each individual eigenvalue λ_j, $j = 2, 3$. From Table 4.4.5 the largest condition index, which is the condition number of the scaled matrix $\boldsymbol{X}$ is given by

$$\text{cond}(\boldsymbol{X}) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\frac{3.93652}{0.00003568}} = 332.15196,$$

which implies a strong problem with multicollinearity. The second largest condition index is given by ≈ 95, which indicates another dependency affecting the regression estimates. From Belsley (1991,[**4**]) we get a good advice on interpreting a variance decomposition matrix like given in Table 4.4.5. Here we have "coexisting or simultaneous near dependencies": A high proportion (>0.5) of the variance of two or more regression coefficients associated with a singular value (or here eigenvalue) indicates that the corresponding regressor is involved in "at least one near dependency". But unfortunately these proportions "cannot always be relied upon to determine which regressors are involved in which specific near dependency". A large proportion of the variance is associated with all regressors, i.e. all regressors are involved in the multicollinearities and the variances of all coefficients may be inflated.



Collinearity Diagnostics

			Proportion of Variation			
Number	**Eigenvalue**	**Condition Index**	**Intercept**	**X1**	**X2**	**X3**
1	3.93652	1.00000	0.00001878	0.00000314	0.00001325	0.00003827
2	0.06301	7.90399	0.00271	0.00000500	0.00014172	0.00370
3	0.00043576	95.04522	0.05328	0.00085716	0.27302	0.47122
4	0.00003565	332.29998	0.94399	0.99913	0.72682	0.52504

TABLE 4.4.5. Variance decomposition matrix of the Economic Data

Of course the data is chosen in a way, such that multicollinearity could have been expected. It is the nature of the three regressors that each is determined by and helps to determine the others. It is obvious, that for example the variable "personal consumption" is highly correlated with the variable "personal income". Thus it is not unreasonable to conclude that there are not three variables, but in fact only one.

CHAPTER 5

Ridge Regression

As shown in the previous chapter, multicollinearity can result in very poor estimates of the regression coefficients and the corresponding variances may be considerably inflated. The Gauss–Markov–Theorem 1.0.1, (3) assures that $\hat{\boldsymbol{\beta}}$ (and $\hat{\boldsymbol{\gamma}}$ respectively) has minimum total variance in the class of linear unbiased estimators of $\boldsymbol{\beta}$, but there is no guarantee that this variance is small. One way to overcome this problem is to drop the requirement of having an unbiased estimator of $\boldsymbol{\beta}$ and try to find a biased estimator $\boldsymbol{\beta}^*$ with a smaller mean squared error. Maybe by allowing a small amount of bias, the total variance of $\boldsymbol{\beta}^*$ can be made small, such that the mean squared error of $\boldsymbol{\beta}^*$ is less than the corresponding one of the unbiased least squares estimator.

A number of procedures have been developed for obtaining biased estimators with optimal statistical properties. But the best known and still the most popular technique is the *ridge regression*, originally proposed by Hoerl and Kennard (1970,[**24**]).

5.1. The Ridge Estimator of Hoerl and Kennard

The ridge estimator is found by solving a slightly modified version of the normal equations in the standardized model (3.2.10). Specifically the ridge estimator $\hat{\boldsymbol{\gamma}}_r$ of $\boldsymbol{\gamma}$ is defined as solution to

$$(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)\hat{\boldsymbol{\gamma}}_r = \boldsymbol{Z}^T\boldsymbol{y}^*,$$

i.e.

$$\hat{\boldsymbol{\gamma}}_r = (\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^*, \tag{5.1.1}$$

where $k > 0$ is a constant selected by the analyst. Note that for $k = 0$, the ridge estimator is equal to the least squares estimator. The ridge estimator is a linear transformation of the least squares estimator since

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_r &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^* \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^* \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)\hat{\boldsymbol{\gamma}} \\ &= \left(\boldsymbol{I}_p + k(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\right)^{-1}\hat{\boldsymbol{\gamma}} = \boldsymbol{K}_r\hat{\boldsymbol{\gamma}},\end{aligned} \tag{5.1.2}$$

where $\boldsymbol{K}_r := \left(\boldsymbol{I}_p + k(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\right)^{-1} = \boldsymbol{I}_p - k\left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right)^{-1}$ (see Lemma A.3.4). Therefore, since $\mathrm{E}(\hat{\boldsymbol{\gamma}}_r) = \mathrm{E}(\boldsymbol{K}_r\hat{\boldsymbol{\gamma}}) = \boldsymbol{K}_r\boldsymbol{\gamma}$, the ridge estimator $\hat{\boldsymbol{\gamma}}_r$ is a biased estimator of $\boldsymbol{\gamma}$. The bias is given by

$$\begin{aligned}\mathrm{Bias}(\hat{\boldsymbol{\gamma}}_r) &= \mathrm{E}(\hat{\boldsymbol{\gamma}}_r) - \boldsymbol{\gamma} = (\boldsymbol{K}_r - \boldsymbol{I}_p)\,\boldsymbol{\gamma} \\ &= -k\left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right)^{-1}\boldsymbol{\gamma}\end{aligned}$$

and the squared bias can be written as

$$\mathrm{Bias}^T(\hat{\boldsymbol{\gamma}}_r)\mathrm{Bias}(\hat{\boldsymbol{\gamma}}_r) = k^2\boldsymbol{\gamma}^T\left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right)^{-2}\boldsymbol{\gamma}.$$

With the equations (5.1.1), (2.0.1) and Lemma 3.1.1, the covariance matrix of $\hat{\boldsymbol{\gamma}}_r$ is given by

$$\begin{aligned}\Sigma(\hat{\boldsymbol{\gamma}}_r) &= (\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\Sigma(\boldsymbol{y}^*)\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1} \\ &= (\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\Sigma(\boldsymbol{\varepsilon}^*)\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1} \\ &= (\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\boldsymbol{P}\Sigma(\boldsymbol{\varepsilon})\boldsymbol{P}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1} \\ &= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1} \\ &= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1} \\ &= \sigma^2\boldsymbol{K}_r(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1} =: \sigma^2\boldsymbol{K}_r\boldsymbol{Z}_r, \end{aligned} \tag{5.1.3}$$

because $\boldsymbol{Z}^T\mathbf{1}_n$ is a null matrix due to the centered matrix $\boldsymbol{Z}$. Denote by $\lambda_j,\ j = 1,\dots,p$ the eigenvalues of $\boldsymbol{Z}^T\boldsymbol{Z}$. The spectral decomposition (see (A.8.1)) is given by

$$\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{5.1.4}$$

with

$$\boldsymbol{\Lambda} = \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{bmatrix}$$

and

$$\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_p.$$

The columns V_j of $\boldsymbol{V}$ are the eigenvectors to the eigenvalues $\lambda_j, j = 1,\dots,p$. The matrix $\boldsymbol{Z}^T\boldsymbol{Z}$ is positive definite and thus we get from (5.1.4)

$$(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^T.$$

Hence the inverse $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ of $\boldsymbol{Z}^T\boldsymbol{Z}$ has the eigenvalues $\frac{1}{\lambda_j}$ to the eigenvectors $V_j,\ j = 1,\ldots,p$. To get the eigenvalue decomposition of (5.1.3) we consider the following Lemma.

Lemma 5.1.1. *Let λ_j be the eigenvalues of the positive definite matrix $\boldsymbol{Z}^T\boldsymbol{Z}$ to the eigenvectors $V_j,\ j = 1,\ldots,p$. Then it follows.*

(1) *The matrix $\boldsymbol{Z}_r = (\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)^{-1}$ has the eigenvalues $\xi_j = \frac{1}{\lambda_j+k}$ to the eigenvectors V_j.*

(2) *The matrix $\boldsymbol{K}_r = \left(k(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} + \boldsymbol{I}_p\right)^{-1}$ has the eigenvalues $\mu_j = \frac{\lambda_j}{\lambda_j+k}$ to the eigenvectors V_j.*

Proof. Proof of (1): For $j = 1,\ldots,p$, λ_j is an eigenvalue of $\boldsymbol{Z}^T\boldsymbol{Z}$ to the eigenvector V_j, iff

$$\boldsymbol{Z}^T\boldsymbol{Z}V_j = \lambda_j V_j, \quad j = 1,\ldots,p. \tag{5.1.5}$$

But (5.1.5) implies

$$\left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right) V_j = (\lambda_j + k)\, V_j, \quad j = 1,\ldots,p.$$

Thus the matrix $(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)$ has the eigenvalues $\lambda_j + k$ to the eigenvectors V_j. Because $(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p)$ is a positive definite matrix for $k > 0$, its inverse, defined by $\boldsymbol{K}_r$, has the eigenvalues $\xi_j := \frac{1}{\lambda_j+k}$ to the eigenvectors V_j.
Proof of (2): The eigenvalues of $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ are given by $\frac{1}{\lambda_j}$ to the eigenvectors $V_j,\ j = 1,\ldots,p$. We obtain

$$\begin{aligned} (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}V_j &= \frac{1}{\lambda_j}V_j \\ \Leftrightarrow k(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}V_j &= \frac{k}{\lambda_j}V_j \\ \Leftrightarrow \left(k(\boldsymbol{Z}^T\boldsymbol{Z})^{-1} + \boldsymbol{I}_p\right) V_j &= \left(\frac{k}{\lambda_j} + 1\right) V_j, \quad j = 1,\ldots,p. \end{aligned}$$

Thus the eigenvalues of $\boldsymbol{K}_r$ are given by $\mu_j = \left(\frac{k}{\lambda_j} + 1\right)^{-1} = \frac{\lambda_j}{\lambda_j+k}$ to the eigenvectors V_j.

□

With the help of Lemma 5.1.1 and (A.1.2), the trace of the covariance matrix of $\hat{\gamma}$ (5.1.3) can be written as

$$\begin{aligned} \sum_{j=1}^{p} \operatorname{var}(\hat{\gamma}_j^r) &= \sigma^2 \operatorname{tr}\left(\boldsymbol{V}\begin{bmatrix}\mu_1 & & \\ & \ddots & \\ & & \mu_p\end{bmatrix}\boldsymbol{V}^T\boldsymbol{V}\begin{bmatrix}\xi_1 & & \\ & \ddots & \\ & & \xi_p\end{bmatrix}\boldsymbol{V}^T\right) \\ &= \sigma^2 \operatorname{tr}\left(\begin{bmatrix}\mu_1 & & \\ & \ddots & \\ & & \mu_p\end{bmatrix}\begin{bmatrix}\xi_1 & & \\ & \ddots & \\ & & \xi_p\end{bmatrix}\right) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}. \end{aligned} \tag{5.1.6}$$

From (5.1.6) it follows, that the total variance of $\hat{\boldsymbol{\gamma}}_r$ is a decreasing function of k. The mean squared error of $\hat{\boldsymbol{\gamma}}_r$ is given by

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\gamma}}_r) &= \sum_{j=1}^{p} \mathrm{var}(\hat{\gamma}_j^r) + \mathrm{bias}^T(\hat{\boldsymbol{\gamma}}_r)\mathrm{bias}(\hat{\boldsymbol{\gamma}}_r) \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \boldsymbol{\gamma}^T(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I})^{-2}\boldsymbol{\gamma}. \qquad (5.1.7)\end{aligned}$$

Hoerl and Kennard (1970,[**24**]) showed that the bias of $\hat{\boldsymbol{\gamma}}_r$ is an increasing function of k. But the aim of using ridge regression is to find a value for k, such that the reduction in the total variance is greater than the increase in the squared bias. As a consequence the mean squared error of the ridge estimator $\hat{\boldsymbol{\gamma}}_r$ will be smaller than that of the least squares estimator $\hat{\boldsymbol{\gamma}}$.
The following theorem shows, that it is always possible to reduce the mean squared error of the least squares estimator.

THEOREM 5.1.2 (Existence theorem). *There always exists a $k > 0$ such that*

$$\mathrm{MSE}(\hat{\boldsymbol{\gamma}}_r) < \mathrm{MSE}(\hat{\boldsymbol{\gamma}}).$$

PROOF. See Hoerl and Kennard (1970,[**24**]), Theorem 4.3.

□

The residual sum of squares of $\hat{\boldsymbol{\gamma}}_r$ is given by

$$\begin{aligned}\mathrm{RSS}(\hat{\boldsymbol{\gamma}}_r) &= (\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_r)^T(\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_r) \\ &= ((\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}) + (\boldsymbol{Z}\hat{\boldsymbol{\gamma}} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_r))^T((\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}) + (\boldsymbol{Z}\hat{\boldsymbol{\gamma}} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_r)) \\ &= (\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}) + 2(\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{Z}\hat{\boldsymbol{\gamma}} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_r) \\ &\qquad + (\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_r)^T\boldsymbol{Z}^T\boldsymbol{Z}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_r) \\ &= (\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}) + (\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_r)^T\boldsymbol{Z}^T\boldsymbol{Z}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_r) \\ &= \mathrm{RSS}(\hat{\boldsymbol{\gamma}}) + (\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_r)^T\boldsymbol{Z}^T\boldsymbol{Z}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_r), \qquad (5.1.8)\end{aligned}$$

because with $\boldsymbol{Z}\hat{\boldsymbol{\gamma}} = \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^*$ we obtain

$$2(\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{Z}\hat{\boldsymbol{\gamma}} - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}_r) = 2\boldsymbol{y}^{*T}\left(\boldsymbol{I}_n - \boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\right)\boldsymbol{Z}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_r) = 0.$$

The first term on the right hand side of (5.1.8) is the residual sum of squares of $\hat{\boldsymbol{\gamma}}$. As a consequence the residual sum of squares of the ridge estimator will be bigger than the corresponding one of the least squares estimator and will thus not necessarily provide the best "fit" to the data.

NOTE 5.1.3. Hoerl and Kennard (1970,[**24**]) originally proposed the ridge estimator for the standardized model, but the calculations also remain true for the unstandardized matrix $\boldsymbol{X}$.

5.1.1 Another Approach to Ridge Regression

The residual sum of squares of an arbitrary estimator $\boldsymbol{g}$ of the model (3.2.10) is given by

$$\begin{aligned}\mathrm{RSS}(\boldsymbol{g}) &= (\boldsymbol{y}^* - \boldsymbol{Z}\boldsymbol{g})^T(\boldsymbol{y}^* - \boldsymbol{Z}\boldsymbol{g}) \\ &= (\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T(\boldsymbol{y}^* - \boldsymbol{Z}\hat{\boldsymbol{\gamma}}) + (\boldsymbol{g} - \hat{\boldsymbol{\gamma}})^T\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{g} - \hat{\boldsymbol{\gamma}}) \\ &= \mathrm{RSS}(\hat{\boldsymbol{\gamma}}) + \phi(\boldsymbol{g}). \qquad (5.1.9)\end{aligned}$$

Contours of constant $\mathrm{RSS}(\boldsymbol{g})$ are the surfaces of hyperellipsoids centered at $\hat{\boldsymbol{\gamma}}$. The value of $\mathrm{RSS}(\boldsymbol{g})$ is the minimum value $\mathrm{RSS}(\hat{\boldsymbol{\gamma}})$ plus the value of the quadratic form in $(\boldsymbol{g} - \hat{\boldsymbol{\gamma}})$. There is a continuum of values $\boldsymbol{g}_0$, that will satisfy the relationship $\mathrm{RSS}(\boldsymbol{g}_0) = \mathrm{RSS}(\hat{\boldsymbol{\gamma}}) + \phi_0$, where $\phi_0 > 0$ is a fixed increment. However, equation (4.2.6) shows, that on average the distance from $\hat{\boldsymbol{\gamma}}$ to $\boldsymbol{\gamma}$ will tend to be large if there is a small eigenvalue of $\boldsymbol{Z}^T\boldsymbol{Z}$. In particular, the worse the conditioning of $\boldsymbol{Z}^T\boldsymbol{Z}$, the more $\hat{\boldsymbol{\gamma}}$ can be expected to be too "long" (see (4.2.7)). On the other hand, the worse the conditioning, the further one can move from $\hat{\boldsymbol{\gamma}}$ without an appreciable increase in the residual sum of squares. In view of (4.2.7) it seems reasonable that if one moves away from $\hat{\boldsymbol{\gamma}}$, the movement should be in a direction which will shorten the length of the regression vector. This implies

$$\min \boldsymbol{g}^T\boldsymbol{g}$$

subject to

$$(\boldsymbol{g} - \hat{\boldsymbol{\gamma}})^T\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{g} - \hat{\boldsymbol{\gamma}}) = \phi_0. \qquad (5.1.10)$$

In mathematical optimization the problem of finding a minimum of a function subject to a constraint like in (5.1.10) is solved with the method of Lagrange multipliers (see e.g. Thomas and Finney (1998,[**56**]), p. 980). This implies minimizing the function

$$\boldsymbol{F} := \boldsymbol{g}^T\boldsymbol{g} + \frac{1}{k}\left((\boldsymbol{g} - \hat{\boldsymbol{\gamma}})^T\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{g} - \hat{\boldsymbol{\gamma}}) - \phi_0\right),$$

where $\frac{1}{k}$ is the multiplier. Then

$$\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{g}} = 2\boldsymbol{g} + \frac{1}{k}\left(2\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{g} - 2\boldsymbol{Z}^T\boldsymbol{Z}\hat{\boldsymbol{\gamma}}\right) = \boldsymbol{0}.$$

Thus the solution is given by

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_r &:= \left(\boldsymbol{I}_p + \frac{1}{k}\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\frac{1}{k}\boldsymbol{Z}^T\boldsymbol{Z}\hat{\boldsymbol{\gamma}} \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + k\boldsymbol{I}_p\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^*,\end{aligned}$$

where k is chosen to satisfy the constraint (5.1.10). This is the ridge estimator.

NOTE 5.1.4. Various interpretations of $\hat{\boldsymbol{\gamma}}_r$ have been advanced in literature. A Bayesian formulation is given by Goldstein and Smith (1974,[**16**]) and also in the original paper of Hoerl and Kennard.

5.2. Estimating The Biasing Factor k

The previous section dealt with the properties of the ridge estimator for non–stochastic k. In this section we will consider different methods for choosing k, because much of the controversy concerning ridge regression centers around the choice of the biasing parameter k.

5.2.1. The Ridge Trace

Hoerl and Kennard (1970,[**24**]) have suggested that an appropriate value of k may be determined by inspecting the *ridge trace*. The ridge trace is a plot of the coefficients of $\hat{\boldsymbol{\gamma}}_r$ versus k for values of k usually in the intervall $(0, 1]$. If multicollinearity is severe, the instability in the regression coefficients will be obvious from the ridge trace. As k is increased, some of the ridge estimates will vary dramatically. At some value of k, the ridge estimates of $\hat{\boldsymbol{\gamma}}_r$ will stabilize. The objective is to select a reasonably small value of k at which the ridge estimates $\hat{\boldsymbol{\gamma}}_r$ are stable. Hopefully this will produce a set of estimates with smaller mean squared error than the least squares estimates. Of course this method of choosing k is somewhat subjective, because two people examining the plot might have different opinions as to where the regression coefficients stabilize.

To simplify this decision, D. Trenkler and G. Trenkler (1995,[**64**]) introduced a global criterion for the degree of stability of the ridge trace. Therefore they measured the (weighted) squared Euclidean distance between the least squares and the ridge estimator and regarded the first derivative with respect to k as "velocity of change".

5.2.2. Estimation Procedures Based on Using Sample Statistics

A number of different formulae have been proposed in literature for estimating the biasing factor k (see Note 5.2.2). We will only consider a few of them, listed below.

(1) Hoerl, Kennard and Baldwin (1975,[**26**]) have suggested that an appropriate choice of k in the standardized model is

$$\hat{k} = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\gamma}}}, \tag{5.2.11}$$

where $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}^2$ define the least squares estimates of $\boldsymbol{\gamma}$ and σ^2 in the standardized model. They argued that this estimator is a reasonable choice, because the minimum of the mean squared error is obtained for $k = \frac{p\sigma^2}{\boldsymbol{\gamma}^T\boldsymbol{\gamma}}$, if $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}_p$ (see Hoerl and Kennard (1970,[**24**])).

In the same paper they showed via simulations that the resulting ridge estimator has a significant improvement in the mean squared error over the least squares estimator for (5.2.11). However, the data used for the simulation study was computed for a standardized model (3.2.10), but with $\boldsymbol{\varepsilon}^* \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. Thus

$$\mathrm{E}\left(\mathrm{RSS}(\hat{\boldsymbol{\gamma}})\right) = \sigma^2(n-p). \tag{5.2.12}$$

This justifies using $\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{\gamma}})}{n-p}$ as estimator for σ^2 in (5.2.11).
But usually in applied work unstandardized data is given, which is subsequently standardized by the analyst. Then as shown in (3.2.17), the equation (5.2.12) is not valid any more after having standardized the data.

(2) In a subsequent paper, Hoerl and Kennard (1976,[**27**]) proposed the following iterative procedure for selecting k: Start with the initial estimate of k, given in (5.2.11). Denote this value by $\hat{k}_0$. Then calculate

$$\hat{k}_i = \frac{p\hat{\sigma}^2}{\sum_{j=1}^p (\hat{\gamma}_j(\hat{k}_{i-1}))^2}, \quad i \geq 1,$$

until the difference between the successive estimates $\hat{k}_i$ of k is negligible.

(3) McDonald and Galarneau (1975,[**37**]) suggested the following method: Let $\boldsymbol{Q}$ be defined by

$$\boldsymbol{Q} := \hat{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\gamma}} - \hat{\sigma}^2 \sum_{j=1}^p \frac{1}{\lambda_j},$$

where λ_j denote the eigenvalues of the matrix $\boldsymbol{Z}^T\boldsymbol{Z}$. Then an estimator $\hat{k}$ is given by solving the equation

$$\hat{\boldsymbol{\gamma}}_r(k)^T \hat{\boldsymbol{\gamma}}_r(k) = \boldsymbol{Q}, \tag{5.2.13}$$

if $\boldsymbol{Q} > 0$, otherwise $k = 0$ or $k = \infty$.
k put in paranthesis should emphasize the dependence of $\hat{\boldsymbol{\gamma}}_r$ to k and the fact that k is determined in a way such that (5.2.13) is fulfilled. In (4.2.7) we showed that

$$\mathrm{E}(\hat{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\gamma}}) = \boldsymbol{\gamma}^T \boldsymbol{\gamma} + \sigma^2 \mathrm{tr}\left((\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\right).$$

Thus an estimator of k calculated by (5.2.13) leads to an unbiased estimator of $\boldsymbol{\gamma}^T\boldsymbol{\gamma}$, because

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\gamma}}_r(k)^T \hat{\boldsymbol{\gamma}}_r(k)) &= \mathrm{E}(\hat{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\gamma}}) - \hat{\sigma}^2 \sum_{j=1}^p \frac{1}{\lambda_j} \\ &= \boldsymbol{\gamma}^T \boldsymbol{\gamma}.\end{aligned}$$

A disadvantage of this method is, that $\boldsymbol{Q}$ may be negative. In this case $k = 0$ leads to the least squares estimator and $k = \infty$ to the zero vector. G. Trenkler (1981,[**58**]) proposed a modification of (5.2.13) by choosing k such that

$$\hat{\gamma}_r(k)^T \hat{\gamma}_r(k) = \text{abs}(\boldsymbol{Q}) \tag{5.2.14}$$

is fulfilled, where abs($\boldsymbol{Q}$) denotes the absolute value of $\boldsymbol{Q}$.
In their simulation study, McDonald and Galarneau computed an unstandardized regression model (including an intercept)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.2.15}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$. Afterwards they standardized this model and used

$$\hat{\sigma}^2 = \frac{\text{RSS}(\hat{\gamma})}{n-p}$$

as estimator for σ^2 in the original model (5.2.15). Hence they also assumed (5.2.12) to be valid for standardized data. In subsequent papers many authors adopted this assumption (e.g. Wichern and Churchill (1978,[**69**]) or Clark and Troskie (2006,[**7**])).

Note 5.2.1. There is no guarantee that these methods are superior to the straightforward inspection of the ridge trace.

Additional Reading 5.2.2. Different researcher concerned themselves with ridge regression and the number of published articles and book is hardly manageable. As a consequence many approaches have been suggested including different techniques for estimating the biasing factor k.

- Marquardt (1970,[**32**]) proposed using a value of k such that the "maximum variance inflation factor should be larger than 1.0 but certainly not as large as 10". Hoerl and Kennard (1970,[**24**]) proposed an extension of the ridge estimator, where an arbitrary diagonal matrix with positive diagonal elements is considered instead of $k\boldsymbol{I}$. This estimator is called the *generalized ridge estimator of Hoerl and Kennard* and will be considered in Section 5.4. Guilkey and Murphy (1975,[**20**]) modificated the above mentioned procedures of Hoerl, Kennard and Baldwin (1975,[**26**]) and McDonald and Galarneau (1975,[**37**]) in a way that results in estimated coefficients with a more moderate increase in the bias. Kibria (2003,[**28**]) proposed a new method using the geometric mean and median of the coefficients of the least squares estimates.

A comparison of some mentioned estimators is given in Wichern and Churchill (1978,[**69**]) and Clark and Troskie (2006,[**7**]).

- There are also several resampling schemes like cross validation, Bootstrap or Jackknife, which can be used for the estimation of k. A good overview of the existing methods and another bootstrap approach, including a simulation study, is given by Delaney and Chatterjee (1986,[**8**]).
- In Note 2.2.5 we have referred to the LINEX and balanced loss functions. The performance of ridge estimators under the LINEX and balanced loss function are examined in Ohtani (1995,[**41**]) or Wan (2002,[**68**]).

NOTE 5.2.3. Many of the properties of the ridge estimator follow from the assumption that the value of k is fixed. In practice k is stochastic since it is estimated from the data. It is of interest to ask if the optimality properties shown above also hold for stochastic k. It has been shown via simulations that ridge regression generally offers improvement in mean squared error over least squares even if k is estimated from the data. Newhouse and Oman (1971,[**39**]) generalized the conditions under which ridge regression leads to a smaller mean squared error than the least squares method. The expected improvement depends on the orientation of $\boldsymbol{\gamma}$ relative to the eigenvectors of $\boldsymbol{Z}^T\boldsymbol{Z}$. The greatest (fewest) expected improvement will be obtained, if $\boldsymbol{\gamma}$ coincides with the eigenvector associated with the smallest (largest) eigenvalue of $\boldsymbol{Z}^T\boldsymbol{Z}$.

5.3. Standardization in Ridge Regression

There is a big controversy in literature about the standardization of data in regression and ridge regression methods. In opposition to Smith and Campbell (1980,[**48**]), Marquardt (1980,[**34**]) pointed out that

(1) the quality of the predictor variable (here regressors) structure of a data set can be assessed properly only in terms of a standardized scale. This applies to both, least squares and ridge estimation,
(2) the interpretability of a model equation is enhanced by expression in standardized form, no matter how the model was estimated.

Furthermore Marquardt and Snee (1975,[**33**]) argued, that "the ill conditioning that results from failure to standardize is all the more insidious because it is not due to any real defect in the data, but only the arbitrary origins of the scales on which the predictor variables are expressed". That is why they recommend standardizing whenever a constant term is present in the model. Belsley, Kuh and Welsch (1984,[**3**]), by contrast, indicated that "mean centering typically masks the role of the constant term in any underlying near dependencies and produces

misreadingly favorable conditioning diagnostics". Especially for a ridge regression model Vinod (1981,[**67**],p. 180) pointed out, that the appearance of a ridge trace that does not plot standardized regression coefficients may be dramatically changed by a simple translation of the origin and scale transformation of the variables. In this case, there is the danger of naively misinterpreting the meaning of the plot.
In summary most authors recommend standardizing the data as we do, so that $\boldsymbol{Z}^T\boldsymbol{Z}$ is in the form of a correlation matrix. For further information see also Stewart (1987,[**52**]) and King (1986,[**29**]).

5.4. A General Form of the Ridge Estimator

In Section 5.1 we introduced the ridge estimator of Hoerl and Kennard for standardized data. In literature the ridge estimator is called the *original ridge estimator* if unstandardized data is used. It differs from the least squares estimator because of the addition of the matrix $k\boldsymbol{I}_p$ to $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{(p+1)\times(p+1)}$ ($\boldsymbol{X}$ may include an intercept).
Actually this matrix could take a number of different forms and thus different kinds of ridge estimators can be formulated. These different kinds of ridge estimators may be derived if the argumentation of Hoerl and Kennard (1970,[**24**]), that was given in Section 5.1.1, is generalized slightly. Instead of minimizing $\boldsymbol{b}^T\boldsymbol{b}$ we minimize the weighted distance

$$\min \boldsymbol{b}^T\boldsymbol{H}\boldsymbol{b},$$

subject to the side condition that $\boldsymbol{b}$ lies on the ellipsoid

$$(\boldsymbol{b}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{b}-\hat{\boldsymbol{\beta}}) = \phi_0.$$

We assume $\boldsymbol{H} \in \mathbb{R}^{(p+1)\times(p+1)}$ to be a symmetric, positive semidefinite matrix. A slightly more general optimization problem than that of (5.1.10) may now be solved by minimizing the Lagrangian function

$$\boldsymbol{F}_g := \boldsymbol{b}^T\boldsymbol{H}\boldsymbol{b} + \frac{1}{k}\left((\boldsymbol{b}-\hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{b}-\hat{\boldsymbol{\beta}}) - \phi_0\right).$$

The solution for $\boldsymbol{X}^T\boldsymbol{X}$ having full rank is given by

$$\hat{\boldsymbol{\beta}}_g = \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{5.4.16}$$

There are four special cases that are of interest

(1) With $\boldsymbol{H} = \boldsymbol{I}_{p+1}$, the solution (5.4.16) reduces to the ridge estimator of Hoerl and Kennard, given in (5.1.1).

(2) For $\boldsymbol{H} = \frac{1}{k}\boldsymbol{G}$, where $\boldsymbol{G} \in \mathbb{R}^{(p+1)\times(p+1)}$ is any positive semidefinite matrix, we get

$$\hat{\boldsymbol{\beta}}_g = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{G}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}. \tag{5.4.17}$$

This is the generalized ridge regression estimator that was proposed by C.R. Rao (1975,[**43**]).

(3) With Theorem A.8.1 we can write $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$, where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues and $\boldsymbol{V}$ is the orthogonal matrix of eigenvectors. Let $\boldsymbol{H} = \frac{1}{k}\boldsymbol{V}\boldsymbol{K}\boldsymbol{V}^T$, where $\boldsymbol{K}$ is a positive definite diagonal matrix. The ridge estimator takes the form

$$\hat{\boldsymbol{\beta}}_g = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}\boldsymbol{K}\boldsymbol{V}^T\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}. \tag{5.4.18}$$

This estimator was also proposed by Hoerl and Kennard (1970,[**24**]). It is known in literature as the *generalized ridge estimator of Hoerl and Kennard.*

(4) By setting $\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{X}$ we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}_g &= \left((1+k)\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y} \\ &= \frac{1}{1+k}\hat{\boldsymbol{\beta}} = \tau\hat{\boldsymbol{\beta}}, \quad \tau \in [0,1]. \end{aligned} \tag{5.4.19}$$

This estimator was originally proposed by Mayer and Willke (1973,[**36**], Proposition 1).

NOTE 5.4.1. Estimators of the form (5.4.19) define a whole class of biased estimators, so called *shrinkage estimators.* They are obtained by shrinking the least squares estimator towards the origin. The parameter τ can be chosen to be deterministic or stochastic (see e.g. Mayer and Willke (1973,[**36**])). One of the most famous shrinkage estimators is the James–Stein–estimator for the standardized model

$$\hat{\boldsymbol{\gamma}}_s := \left(1 - \frac{c\sigma^2}{\boldsymbol{\gamma}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\gamma}}\right)\hat{\boldsymbol{\gamma}},$$

where $c > 0$ is an arbitrary constant. In this case τ contains, besides $c \in \mathbb{R}$, the unknown parameters σ^2 and $\boldsymbol{\gamma}$, which have to be estimated. From (5.1.2) it follows that even the ridge estimator is of the type of a shrinkage estimator. But in case of the ridge estimator, k is the only unknown parameter. For further information on shrinkage estimation see Gruber (1998,[**19**]), Ohtani (2000,[**42**]) or Farebrother (1977, [**13**]).

The covariance matrix of (5.4.16) is given by

$$\Sigma(\hat{\boldsymbol{\beta}}_g) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right)\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}$$

and the bias by

$$\begin{aligned}\operatorname{Bias}(\hat{\boldsymbol{\beta}}_g) &= \operatorname{E}(\hat{\boldsymbol{\beta}}_g) - \boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= \left(\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right) - \boldsymbol{I}_p\right)\boldsymbol{\beta}.\end{aligned}$$

With Lemma A.3.4 we have

$$\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right) - \boldsymbol{I}_p = -\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1} k\boldsymbol{H}$$

and it follows for the mean squared error matrix

$$\operatorname{MtxMSE}(\hat{\boldsymbol{\beta}}_g) = \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}\left(k^2\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{H} + \sigma^2\boldsymbol{X}^T\boldsymbol{X}\right)\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}.$$

From Chapter 2 we know that the generalized ridge estimator $\hat{\boldsymbol{\beta}}_g$ is preferred to the least squares estimator $\hat{\boldsymbol{\beta}}$, if

$$\Delta := \operatorname{MtxMSE}(\hat{\boldsymbol{\beta}}) - \operatorname{MtxMSE}(\hat{\boldsymbol{\beta}}_g) \tag{5.4.20}$$

is a positive semidefinite matrix.

The following Theorem contains the main result about the matrix mean squared error of the generalized form of the ridge estimator given in (5.4.16).

Theorem 5.4.2. *The matrix Δ, given in (5.4.20), is positive semidefinite, iff*

$$\boldsymbol{\beta}^T\left(\frac{2}{k}\boldsymbol{H}^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)^{-1}\boldsymbol{\beta} \le \sigma^2, \tag{5.4.21}$$

for $\boldsymbol{H}$ being a symmetric, positive definite matrix.

Proof. We have to show that

$$\begin{aligned}\Delta = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1}&\left(k^2\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{H}\right. \\ &\left. + \sigma^2\boldsymbol{X}^T\boldsymbol{X}\right)\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)^{-1} \ge 0.\end{aligned} \tag{5.4.22}$$

Therefore suppose that Δ can be written as $\Delta = \boldsymbol{U}^T\boldsymbol{U}$ for any matrix $\boldsymbol{U}$ having p columns (see Theorem A.9.3). Because $\boldsymbol{H}$ is symmetric and positive definite by assumption, the matrix $\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)$ is also symmetric and positive definite. Thus we get

$$\begin{aligned}&\boldsymbol{p}^T\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)\Delta\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)\boldsymbol{p} \\ &\qquad = \boldsymbol{p}^T\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)\boldsymbol{U}^T\boldsymbol{U}\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)\boldsymbol{p} \\ &\qquad\quad = \boldsymbol{p}^T\left(\boldsymbol{U}\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)\right)^T\left(\boldsymbol{U}\left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right)\right)\boldsymbol{p} \ge 0,\end{aligned}$$

for an arbitrary vector $\boldsymbol{p} \in \mathbb{R}^{(p+1)\times 1}$. As a consequence multiplication of both sides of inequality (5.4.22) by $(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H})$ preserves positiv semidefinitness of the difference. Thus (5.4.22) is equivalent to

$$\sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right) \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right) - k^2\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{H} - \sigma^2\boldsymbol{X}^T\boldsymbol{X} \geq 0. \tag{5.4.23}$$

It is

$$\sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right) \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{H}\right) = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X} + 2k\boldsymbol{H} + k^2\boldsymbol{H}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{H}\right)$$

and we get for (5.4.23)

$$\sigma^2 \left(2k\boldsymbol{H} + k^2\boldsymbol{H}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{H}\right) - k^2\boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{H} \geq 0. \tag{5.4.24}$$

Multiplication of both sides of (5.4.24) by $\frac{1}{k}\boldsymbol{H}^{-1}$ yields

$$\sigma^2 \left(\frac{2}{k}\boldsymbol{H}^{-1} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) - \boldsymbol{\beta}\boldsymbol{\beta}^T \geq 0. \tag{5.4.25}$$

Inequality (5.4.25) is equivalent to (5.4.21) by virtue of Theorem 2.2.2 in Chapter 2.

□

As a consequence of Chapter 2 all parametric functions $\boldsymbol{p}^T\hat{\boldsymbol{\beta}}_g,\ \boldsymbol{p} \in \mathbb{R}^{(p+1)\times 1}$ of the generalized ridge estimator have a mean squared error that is less than or equal to that of the least squares estimator, iff (5.4.21) is fulfilled. The following corollary, given in Gruber (1998,[**19**], p. 125), specializes the result of Theorem 5.4.2 to the ridge estimators of Hoerl, Kennard and Mayer, Willke.

Corollary 5.4.3. *The matrix* Δ*, given in (5.4.20), is positive semidefinite for*

(1) *the ordinary ridge regression estimator (5.1.1), iff*

$$\boldsymbol{\beta}^T \left(\frac{2}{k}\boldsymbol{I}_p + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)\boldsymbol{\beta} \leq \sigma^2,$$

(2) *the generalized ridge estimator of C. R. Rao (5.4.17), iff*

$$\boldsymbol{\beta}^T \left(2\boldsymbol{G}^{-1} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)\boldsymbol{\beta} \leq \sigma^2$$

and $\boldsymbol{G}$ *is a symmetric, positive definite matrix.*

(3) *the generalized ridge estimator of Hoerl, Kennard (5.4.18), iff*

$$\boldsymbol{\beta}^T \left(2\boldsymbol{K}^{-1} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)\boldsymbol{\beta} \leq \sigma^2,$$

(4) *the estimator of Mayer and Wilke (5.4.19), iff*

$$\boldsymbol{\beta}^T \left(\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{\beta} \leq \frac{k+2}{k}\sigma^2.$$

NOTE 5.4.4. For the sake of completeness consider the following two remarks.

- Already Marquardt (1970,[**32**]) proposed estimators of the form

  $$\hat{\boldsymbol{b}}_g = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}\right)^+ \boldsymbol{X}^T\boldsymbol{y}, \tag{5.4.26}$$

  where $\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}\right)^+$ denotes the Moore–Penrose inverse of the matrix $\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}\right)^+$ (definition see e.g. Harville (1997,[**23**])) and $\boldsymbol{C}$ is any symmetric matrix commuting with $\boldsymbol{X}^T\boldsymbol{X}$. It is not difficult to see that the generalized ridge estimator of Hoerl and Kennard in (5.4.18) is of the form of (5.4.26).
  Lowerre (1974,[**31**]) developed conditions on the matrix $\boldsymbol{C}$, under which each coefficient of the estimator (5.4.26) has a smaller mean squared error than the corresponding ones of the least squares estimator.
- Most theoretical examinations on ridge type estimators are done using the singular value decomposition of $\boldsymbol{Z}$ (see Theorem A.8.2)

  $$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Theta}\boldsymbol{V}^T.$$

  Then the least squares estimator can be written as

  $$\begin{aligned}\hat{\gamma} &= \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{y}^* = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{y}^*\\ &= \boldsymbol{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{y}^*,\end{aligned}$$

  where $\boldsymbol{\Lambda}$ is the diagonal matrix containing the eigenvalues of $\boldsymbol{Z}^T\boldsymbol{Z}$. Obenchain (1978,[**40**]) considered generalized ridge estimators of the form

  $$\hat{\gamma}_r^* = \boldsymbol{V}\boldsymbol{\Xi}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{y}^*,$$

  where $\boldsymbol{\Xi}$ is a diagonal matrix with the non–stochastic "ridge factors" $\xi_1, \ldots, \xi_p \in \mathbb{R}$. He found ridge factors, which achieve either minimum mean squared error parallel to an arbitrary direction in the coefficient space or minimum weighted mean squared for an arbitrary positive definite matrix $\boldsymbol{W}$, defined like in Chapter 2. Note that because of (5.1.2) and Lemma 5.1.1, we get the ridge estimator of Hoerl and Kennard for $\xi_j = \frac{\lambda_j}{\lambda_j + k}$, $j = 1, \ldots, p$.
  D. Trenkler and G. Trenkler (1984,[**61**]) extended the examinations of Obenchain to an imhomogeneous estimator of a linear transform $\boldsymbol{B}\boldsymbol{\beta}$ with a known matrix $\boldsymbol{B}$ (which may be inestimable). In contrast to Obenchain they did not claim a positive definite matrix $\boldsymbol{W}$ and a regular design matrix. The main problem of both, the estimator of Obenchain and D. Trenkler and G. Trenkler is, that they depend on unknown parameters, which have to be estimated.

Additional Reading 5.4.5. Besides ridge type estimators or shrinkage estimators, mentioned in Note 5.4.1, several alternatives to the least squares estimator like principle components, Bayes or minimax estimators have been introduced. G. Trenkler (1980,[**57**]) also introduced an iteration and inversion estimator, which have similar properties as the ridge and shrinkage estimators. In D. Trenkler and G. Trenkler (1984,[**62**]) they showed that the ridge and iteration estimator can be made very close to the principal component estimator. G. Trenkler (1981,[**58**]) and Rao and Toutenbourg (2007,[**45**]) give a good overview of some alternatives to least squares.

5.5. Example: Ridge Regression of the Economic Data

To calculate the ridge estimator for the Economic Data of Section 4.4 in dependence of different k, we can use the `RIDGE` option of the `REG`–procedure of `SAS`.Within this option the following calculations are done.

(1) The regression model (4.4.10) of the Economic Data with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{17})$ is considered. With (1.0.6) and Table 4.4.3 an estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\text{RSS}(\hat{\boldsymbol{\beta}})}{n-p-1} = \frac{11.369}{17-3-1} = 0.87453. \tag{5.5.27}$$

First of all the design matrix is centered and scaled analogous to Chapter 3 and the ridge estimator $\hat{\boldsymbol{\gamma}}_r$ of $\boldsymbol{\gamma}$, corresponding to the biasing factor k, is given by

$$\hat{\boldsymbol{\gamma}}_r = \left(\boldsymbol{Z}^T \boldsymbol{Z} + k\boldsymbol{I}_3\right)^{-1} \boldsymbol{Z}^T \boldsymbol{y}^*. \tag{5.5.28}$$

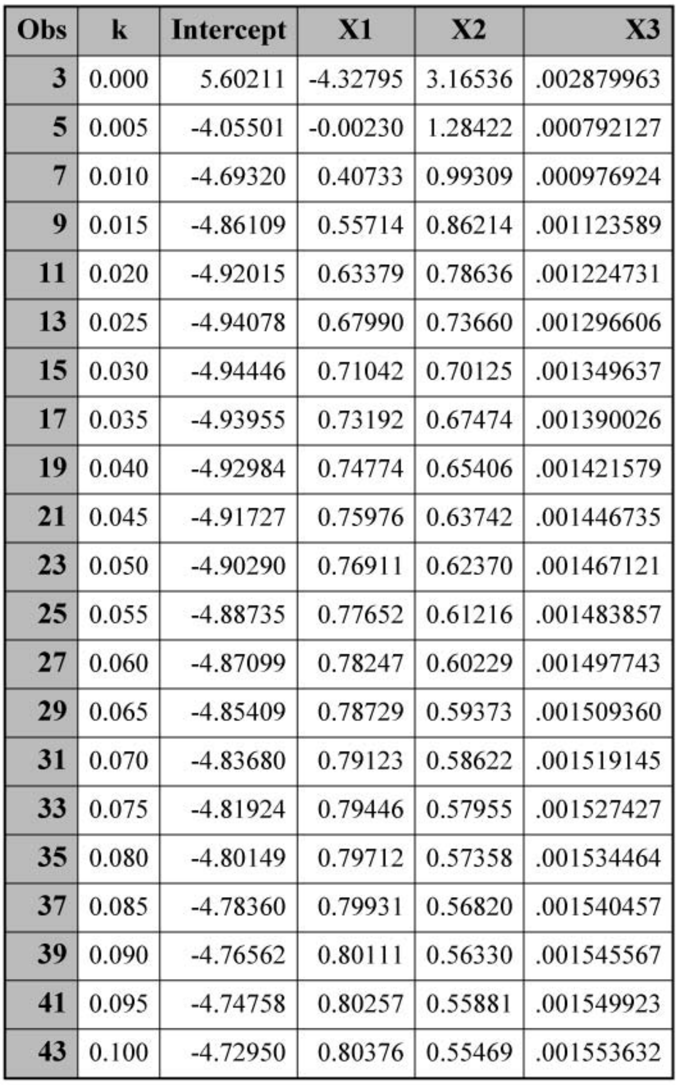
Hence `SAS` follows Hoerl and Kennard (1970,[**24**]) by performing the ridge regression in the standardized model (see Note 5.1.3).

(2) Afterwards the ridge estimator (5.5.28) is transformed back with the help of the relationship (3.2.13), in order to get the ridge estimator of the original, unstandardized model.
Hence the ridge estimator $\hat{\boldsymbol{\beta}}^r_{\{\beta_0\}}{}^T := \begin{bmatrix}\hat{\beta}^r_1, & \hat{\beta}^r_2, & \hat{\beta}^r_3\end{bmatrix}$ of $\boldsymbol{\beta}_{\{\beta_0\}}{}^T = \begin{bmatrix}\beta_1, & \beta_2, & \beta_3\end{bmatrix}$ corresponding to the biasing factor k is computed by

$$\hat{\boldsymbol{\beta}}^r_{\{\beta_0\}} = \boldsymbol{D}^{-1}\hat{\boldsymbol{\gamma}}_r = \boldsymbol{D}^{-1}\left(\boldsymbol{Z}^T \boldsymbol{Z} + k\boldsymbol{I}_3\right)^{-1} \boldsymbol{Z}^T \boldsymbol{y}^*,$$

where $\boldsymbol{D}$ is given in (3.2.12). With (3.2.14) the ridge estimator of the intercept is given by

$$\hat{\beta}^r_0 = \bar{y} - \sum_{j=1}^{3} \hat{\beta}^r_j \bar{X}_j.$$



Obs	k	Intercept	X1	X2	X3
3	0.000	5.60211	-4.32795	3.16536	.002879963
5	0.005	-4.05501	-0.00230	1.28422	.000792127
7	0.010	-4.69320	0.40733	0.99309	.000976924
9	0.015	-4.86109	0.55714	0.86214	.001123589
11	0.020	-4.92015	0.63379	0.78636	.001224731
13	0.025	-4.94078	0.67990	0.73660	.001296606
15	0.030	-4.94446	0.71042	0.70125	.001349637
17	0.035	-4.93955	0.73192	0.67474	.001390026
19	0.040	-4.92984	0.74774	0.65406	.001421579
21	0.045	-4.91727	0.75976	0.63742	.001446735
23	0.050	-4.90290	0.76911	0.62370	.001467121
25	0.055	-4.88735	0.77652	0.61216	.001483857
27	0.060	-4.87099	0.78247	0.60229	.001497743
29	0.065	-4.85409	0.78729	0.59373	.001509360
31	0.070	-4.83680	0.79123	0.58622	.001519145
33	0.075	-4.81924	0.79446	0.57955	.001527427
35	0.080	-4.80149	0.79712	0.57358	.001534464
37	0.085	-4.78360	0.79931	0.56820	.001540457
39	0.090	-4.76562	0.80111	0.56330	.001545567
41	0.095	-4.74758	0.80257	0.55881	.001549923
43	0.100	-4.72950	0.80376	0.55469	.001553632

TABLE 5.5.1. Regression estimates in dependence of k

Table 5.5.1 shows the estimates of the regression coefficients in dependence of k. To find an optimal k we will apply some of the techniques mentioned in Section 5.2.

- With the help of Table 5.5.1 we get the ridge trace of the Economic Data, shown in Figure 5.5.1. The ridge trace illustrates the instability of the least squares estimates as there is a large change in the regression coefficients for small k. The coefficient $\hat{\beta}_1^r$ even changes sign. As mentioned in Section 4.4, the negative sign of $\hat{\beta}_1$ is not expected and thus probably due to multicollinearity. However, the coefficients seem

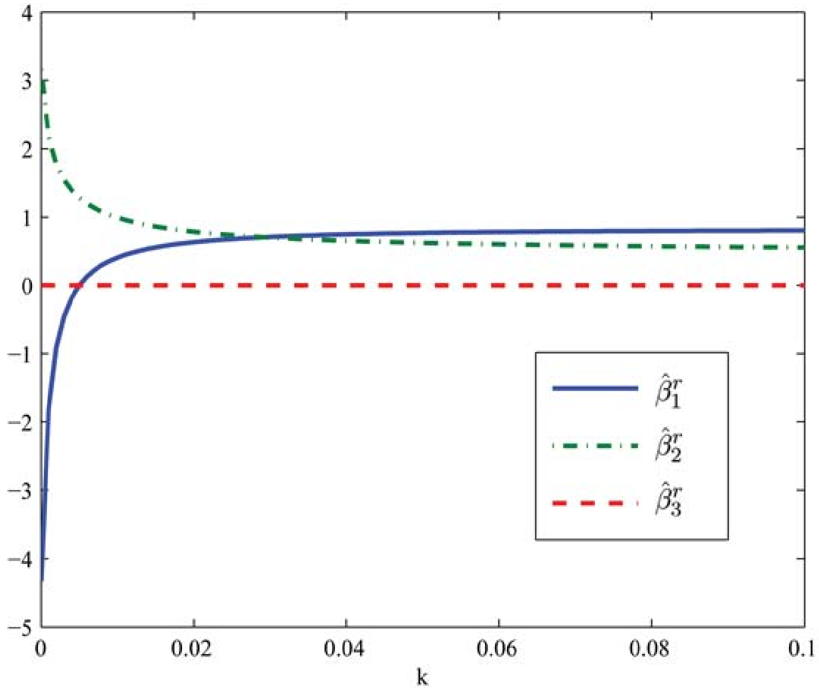


FIGURE 5.5.1. Ridge trace of the Economic Data

to stabilize as k increases. We want to choose k large enough to provide stable coefficients, but not unnecessarily large as this introduces additional bias. Hence choosing k around 0.05 seems to be suitable.

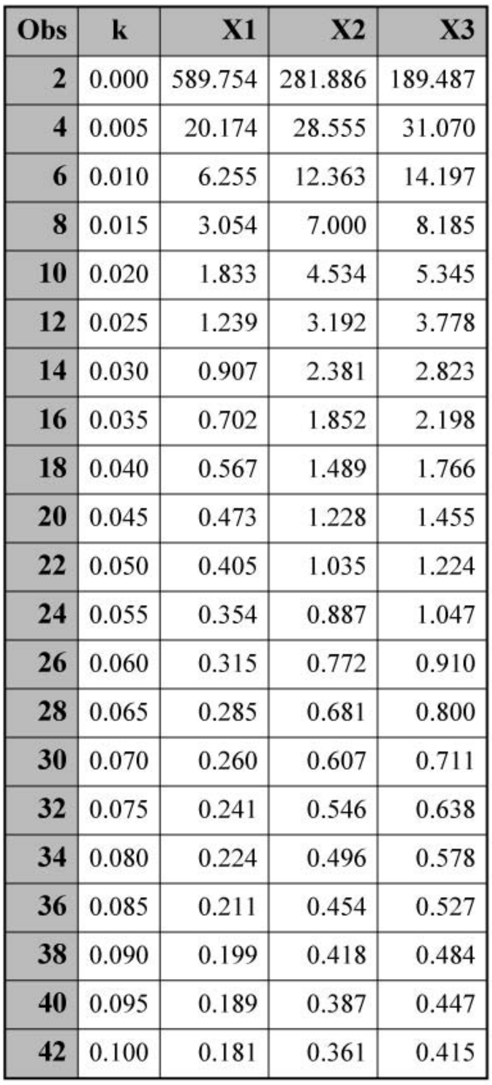
- SAS also provides an option, which calculates the variance inflation factors of the regression estimates in dependence of k. These are given in Table 5.5.2. Marquardt (1970,[**32**]) proposed using the variance inflation factors for getting an optimal k. He recommended choosing a k, for which the variance inflation factors are bigger than one, but smaller than 10 (see Note 5.2.2). Thus from Table 5.5.2 we also have to choose k around 0.05.

Unfortunately other methods for finding an optimal estimate of k are not implemented in SAS.
Denote by

$$\boldsymbol{y}^* = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^* \tag{5.5.29}$$

the standardized model of (4.4.10) of the Economic Data. The standardized data is given in Table 5.5.3 and with the help of Table 5.5.4, which shows the summary statistics of the model (5.5.29), we get the least squares estimator of $\boldsymbol{\gamma}$

$$\hat{\boldsymbol{\gamma}}^T = \begin{bmatrix} -19.106, & 24.356, & 6.4207 \end{bmatrix}. \tag{5.5.30}$$



Obs	k	X1	X2	X3
2	0.000	589.754	281.886	189.487
4	0.005	20.174	28.555	31.070
6	0.010	6.255	12.363	14.197
8	0.015	3.054	7.000	8.185
10	0.020	1.833	4.534	5.345
12	0.025	1.239	3.192	3.778
14	0.030	0.907	2.381	2.823
16	0.035	0.702	1.852	2.198
18	0.040	0.567	1.489	1.766
20	0.045	0.473	1.228	1.455
22	0.050	0.405	1.035	1.224
24	0.055	0.354	0.887	1.047
26	0.060	0.315	0.772	0.910
28	0.065	0.285	0.681	0.800
30	0.070	0.260	0.607	0.711
32	0.075	0.241	0.546	0.638
34	0.080	0.224	0.496	0.578
36	0.085	0.211	0.454	0.527
38	0.090	0.199	0.418	0.484
40	0.095	0.189	0.387	0.447
42	0.100	0.181	0.361	0.415

TABLE 5.5.2. Variance inflation factors in dependence of k

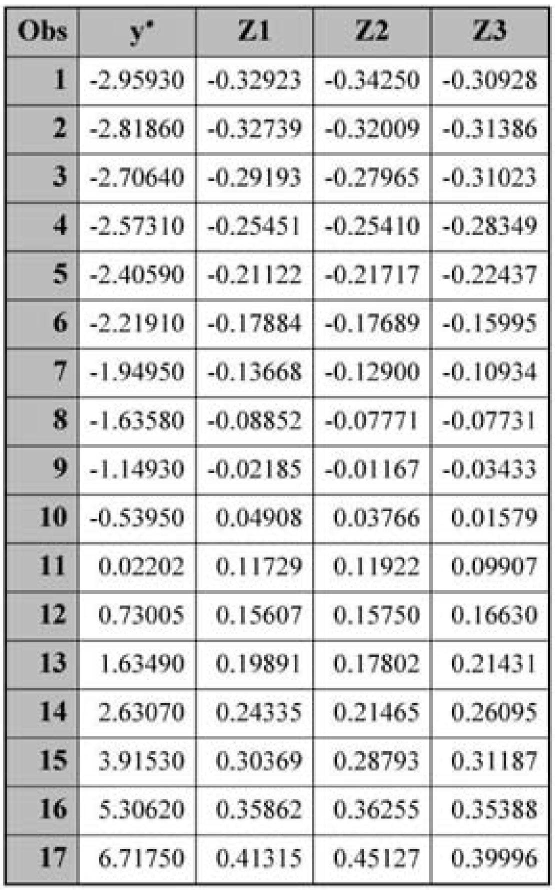
We can calculate the following estimates for k:

- $\hat{k} = \frac{4\hat{\sigma}_Z^2}{\hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\gamma}}}$, suggested by Hoerl and Kennard and given in (5.2.11). As described in point (1) of Section 5.2.2,

$$\hat{\sigma}_Z^2 = \frac{\text{RSS}(\hat{\boldsymbol{\gamma}})}{n-p} = \frac{11.368}{14} = 0.8120, \tag{5.5.31}$$

is chosen as an estimator of σ^2. RSS($\hat{\boldsymbol{\gamma}}$) denotes the residual sum of squares of $\hat{\boldsymbol{\gamma}}$ and is given in Table 5.5.4. Thus it follows

$$\hat{k} = 0.00325.$$



Obs	y*	Z1	Z2	Z3
1	-2.95930	-0.32923	-0.34250	-0.30928
2	-2.81860	-0.32739	-0.32009	-0.31386
3	-2.70640	-0.29193	-0.27965	-0.31023
4	-2.57310	-0.25451	-0.25410	-0.28349
5	-2.40590	-0.21122	-0.21717	-0.22437
6	-2.21910	-0.17884	-0.17689	-0.15995
7	-1.94950	-0.13668	-0.12900	-0.10934
8	-1.63580	-0.08852	-0.07771	-0.07731
9	-1.14930	-0.02185	-0.01167	-0.03433
10	-0.53950	0.04908	0.03766	0.01579
11	0.02202	0.11729	0.11922	0.09907
12	0.73005	0.15607	0.15750	0.16630
13	1.63490	0.19891	0.17802	0.21431
14	2.63070	0.24335	0.21465	0.26095
15	3.91530	0.30369	0.28793	0.31187
16	5.30620	0.35862	0.36255	0.35388
17	6.71750	0.41315	0.45127	0.39996

TABLE 5.5.3. Data of the standardized model

After transforming back the ridge estimates for $\hat{k}$ we get the ridge estimator for the Economic Data

$$\hat{\boldsymbol{\beta}}^T(\hat{k}) = \Big[-3.3497, \quad -0.3787, \quad 1.5042, \quad 7.8142 \cdot 10{-}4\Big]^T .$$

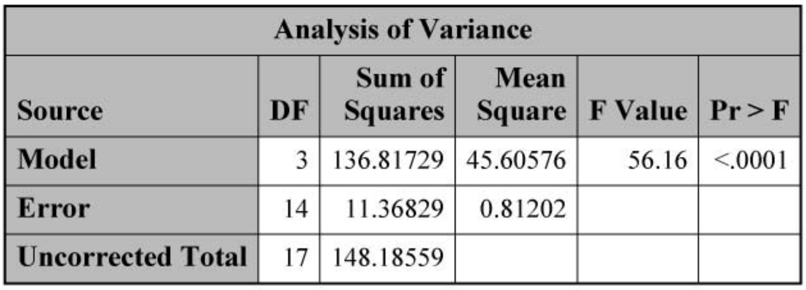
- To get an estimator of k with the method proposed by McDonald and Galarneau in Section 5.2.2, (3) we use `MATLAB` for finding a solution of the equation

$$\hat{\boldsymbol{\gamma}}_r(k)^T \hat{\boldsymbol{\gamma}}_r(k) \approx \hat{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\gamma}} - \hat{\sigma}^2 \mathrm{tr}\left((\boldsymbol{Z}^T \boldsymbol{Z})^{-1}\right). \qquad (5.5.32)$$

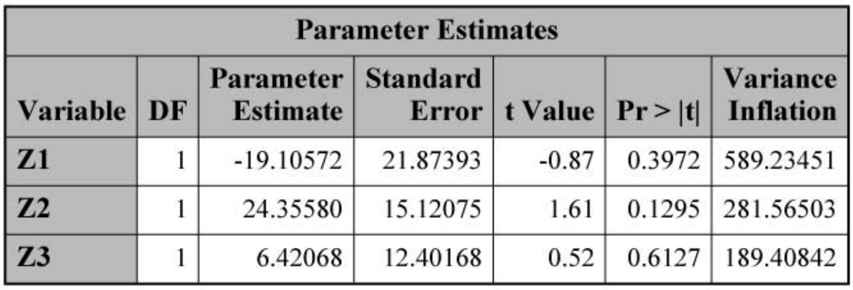
Since equation (5.2.13) is usually solved with the help of numerical methods, the solution will not be exact. Therefore we write "≈" in (5.5.32).

If we take, as proposed by McDonald and Galarneau, $\hat{\sigma}^2_Z$ as estimator of σ^2 we get

$$\boldsymbol{Q}_Z := \hat{\boldsymbol{\gamma}}^T \hat{\boldsymbol{\gamma}} - \hat{\sigma}^2_Z \mathrm{tr}\left((\boldsymbol{Z}^T \boldsymbol{Z})^{-1}\right) = 138.0$$



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	136.81729	45.60576	56.16	<.0001
Error	14	11.36829	0.81202		
Uncorrected Total	17	148.18559			



Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Z1	1	-19.10572	21.87393	-0.87	0.3972	589.23451
Z2	1	24.35580	15.12075	1.61	0.1295	281.56503
Z3	1	6.42068	12.40168	0.52	0.6127	189.40842

TABLE 5.5.4. Analysis of variance and parameter estimates of the standardized Economic Data

and the solution is given by

$$\hat{k}_Z = 0.0033.$$

This yields to

$$\hat{\boldsymbol{\beta}}^r(\hat{k}) = \begin{bmatrix} -3.3781, & -0.3642, & 1.4962, & 7.80 \cdot 10^{-4} \end{bmatrix}^T.$$

- The ridge estimator is designed to have a smaller mean squared error than the least squares estimator. Therefore it is obvious to choose k in a way, such that the mean squared error, given in (5.1.7), is minimized. Of course, the mean squared error has to be estimated, because the parameters σ^2 and $\boldsymbol{\gamma}$ are unknown.

  With $\hat{\sigma}^2 = \frac{\text{RSS}(\hat{\boldsymbol{\beta}})}{17-3-1} = 0.87453$ and $\hat{\boldsymbol{\gamma}}$, given in (5.5.30), we can calculate the estimated mean squared error in dependence of k by

$$\widehat{\text{MSE}}_k(\hat{\boldsymbol{\gamma}}_r) = \hat{\sigma}^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \hat{\boldsymbol{\gamma}}^T (\boldsymbol{Z}^T \boldsymbol{Z} + k\boldsymbol{I})^{-2} \hat{\boldsymbol{\gamma}}. \qquad (5.5.33)$$

  Figure 6.8.91 displays the estimated total variance, the estimated squared bias and the estimated mean squared error of $\hat{\boldsymbol{\gamma}}_r$ in dependence of k. The minimum of the estimated mean squared error is given for

$$k_{\text{min}} = 0.0012$$

and the function value of the estimated mean squared error for k_{min} is

$$\widehat{\text{MSE}}(\hat{\boldsymbol{\gamma}}_{k_{\text{min}}}) = 517.06. \tag{5.5.34}$$

After transforming back the ridge estimator for $k_{\text{min}} = 0.0012$ we get

$$\hat{\boldsymbol{\beta}}^r(\hat{k}) = \Big[-0.7410, \quad -1.6031, \quad 2.0893, \quad 0.0012\Big]^T.$$

NOTE 5.5.1. More information about the used procedure in `MATLAB` for solving (5.5.32) is given in Chapter 7.

The considered methods result in estimates for k, differing enormously in magnitude. For the ridge trace and with the help of the variance inflation factors we would choose a k, 10 times larger than for the remaining methods. Maybe this example can illustrate the difficulty of finding an optimal estimator of k.

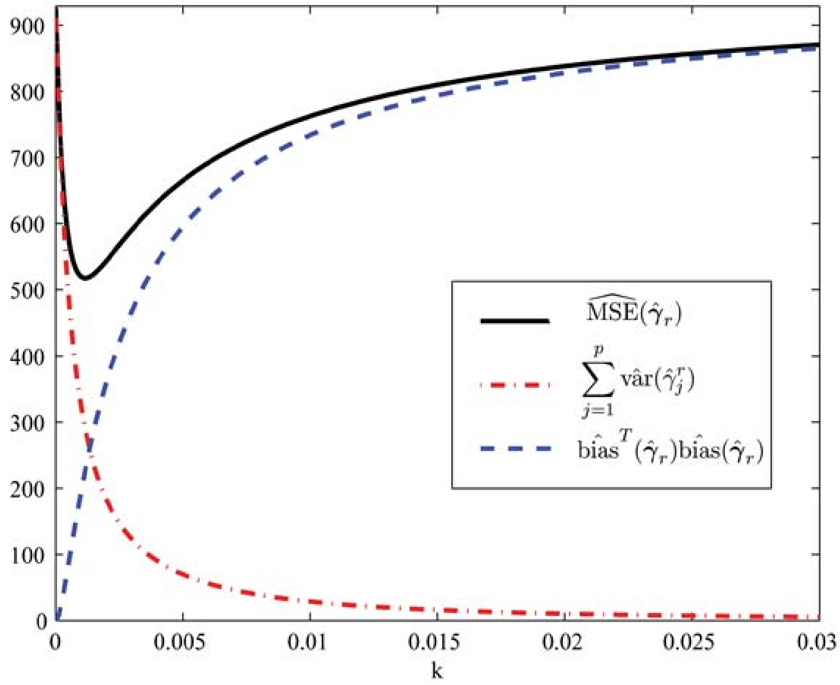


FIGURE 5.5.2. Estimated mean squared error of the ridge estimator

CHAPTER 6

The Disturbed Linear Regression Model

Often in the application of statistics when a linear regression model is fitted, some of the independent variables are highly correlated and thus might have a linear relationship between them. When this happens the least squares estimates will be very imprecise, because their covariance matrix is nearly singular. A possible solution to this problem was the formulation of the ridge estimators, we considered in Chapter 5. Ridge estimators have been shown to give more precise estimates of the regression coefficients and as a consequence, they have found diverse applications in dealing with multicollinear data.

We will now propose another biased estimator, which we call the *disturbed least squares estimator*. It will be derived by minimizing a slightly changed version of the residual sum of squares. This is based on adding a small quantity $\omega\psi_j$ on the standardized regressors $Z_j,\ j = 1, \dots, p$. The resulting biased estimator is described in dependence of ω and it will be shown that its mean squared error is smaller than the corresponding one of the least squares estimator for suitably chosen ω.

Furthermore we will also consider the matrix mean squared error of the disturbed least squares estimator and finally we will show, that the disturbed regression estimator can be embedded in the class of ridge estimators.

6.1. Derivation of the Disturbed Linear Regression Model

We assume $\boldsymbol{Z}$ to be a standardized $n \times p$ matrix, $n > p$, of full column rank, containing the non–constant columns $Z_j,\ j = 1, \dots, p$. Instead of minimizing the usual residual sum of squares like in (1.0.2), we minimize

$$\min_{\gamma} \sum_{i=1}^{n} (y_i^* - \boldsymbol{z}_i^T \boldsymbol{\gamma} - \omega \boldsymbol{\psi}^T \boldsymbol{\gamma})^2, \tag{6.1.1}$$

where $\boldsymbol{z}_i^T$ denotes the i-th row of $\boldsymbol{Z}$. The vector $\boldsymbol{\psi}^T = \begin{bmatrix} \psi_1, & \dots & , \psi_p \end{bmatrix} \neq \boldsymbol{0}$ is assumed to be fixed and $\omega \in \mathbb{R}$ is chosen arbitrarily. We assume ω to be very small, so that the disturbance remains small. (6.1.1) is equivalent to

$$\min_{\gamma} \left(\boldsymbol{y}^* - (\boldsymbol{Z} + \omega\boldsymbol{\Psi})\boldsymbol{\gamma}\right)^T \left(\boldsymbol{y}^* - (\boldsymbol{Z} + \omega\boldsymbol{\Psi})\boldsymbol{\gamma}\right),$$

with

$$\boldsymbol{\Psi} := \begin{bmatrix} \psi_1 & \dots & \psi_p \\ \vdots & & \vdots \\ \psi_1 & \dots & \psi_p \end{bmatrix} \in \mathbb{R}^{n \times p}, \ n > p. \tag{6.1.2}$$

The normal equations are given by

$$(\boldsymbol{Z} + \omega\boldsymbol{\Psi})^T (\boldsymbol{Z} + \omega\boldsymbol{\Psi}) \boldsymbol{\gamma} = (\boldsymbol{Z} + \omega\boldsymbol{\Psi})^T \boldsymbol{y}^*. \tag{6.1.3}$$

Because $\boldsymbol{y}^*$ is centered we have $\sum_{i=1}^n y_i^* = 0$ and thus

$$\boldsymbol{\Psi}^T \boldsymbol{y}^* = \begin{bmatrix} \psi_1 \sum_{i=1}^n y_i^* \\ \vdots \\ \psi_p \sum_{i=1}^n y_i^* \end{bmatrix} = \mathbf{0} \quad \in \mathbb{R}^{p \times 1}.$$

Because $\boldsymbol{Z}$ is also centered we get in the same way

$$\omega \boldsymbol{\Psi}^T \boldsymbol{Z} = \omega \boldsymbol{Z}^T \boldsymbol{\Psi} = \mathbf{0}.$$

Thus we have

$$\boldsymbol{M} = [m_{u,v}]_{1 \le u,v \le p} := \big(\boldsymbol{Z} + \omega\boldsymbol{\Psi}\big)^T \big(\boldsymbol{Z} + \omega\boldsymbol{\Psi}\big) = \boldsymbol{Z}^T\boldsymbol{Z} + \omega^2 \boldsymbol{\Psi}^T\boldsymbol{\Psi}. \tag{6.1.4}$$

$\boldsymbol{M}$ is a positive definite matrix, because $\boldsymbol{Z}^T\boldsymbol{Z}$ is positive definite by assumption and $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ is positive semidefinite and thus

$$\boldsymbol{p}^T \boldsymbol{M} \boldsymbol{p} = \boldsymbol{p}^T \left(\boldsymbol{Z}^T\boldsymbol{Z}\right) \boldsymbol{p} + \omega^2 \boldsymbol{p}^T \left(\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right) \boldsymbol{p} > 0$$

is fulfilled for any $\boldsymbol{p} \in \mathbb{R}^{p \times 1}$. The least squares estimator of (6.1.1), which we call the *disturbed least squares estimator (DLSE)* can then be calculated by

$$\begin{aligned} \tilde{\boldsymbol{\gamma}}_\omega &= \Big((\boldsymbol{Z} + \omega\boldsymbol{\Psi})^T (\boldsymbol{Z} + \omega\boldsymbol{\Psi})\Big)^{-1} (\boldsymbol{Z} + \omega\boldsymbol{\Psi})^T \boldsymbol{y}^* \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2 \boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^{-1} \boldsymbol{Z}^T \boldsymbol{y}^*. \end{aligned} \tag{6.1.5}$$

Because of the additional matrix $\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}$, the disturbed least squares estimator (6.1.5) will not be unbiased any more and its covariance matrix will differ from the corresponding one of the least squares estimator, given in (3.2.15). To get the bias and the variances of the coefficients of $\tilde{\boldsymbol{\gamma}}_\omega$ and thus the mean squared error of $\tilde{\boldsymbol{\gamma}}_\omega$ in dependence of ω, we have to examine the inverse of the matrix $\boldsymbol{M}$, given in (6.1.4). Therefore let $\boldsymbol{M}_{\{u,v\}}$ represent the $(p-1) \times (p-1)$ submatrix of $\boldsymbol{M}$ obtained by striking out the u-th row and v-th column, i.e.

$$\boldsymbol{M}_{\{u,v\}} = \begin{bmatrix} m_{1,1} & \dots & m_{1,v-1} & m_{1,v+1} & \dots & m_{1,p} \\ \vdots & & \vdots & \vdots & & \vdots \\ m_{u-1,1} & \dots & m_{u-1,v-1} & m_{u-1,v+1} & \dots & m_{u-1,p} \\ m_{u+1,1} & \dots & m_{u+1,v-1} & m_{u+1,v+1} & \dots & m_{u+1,p} \\ \vdots & & \vdots & \vdots & & \vdots \\ m_{p,1} & \dots & m_{p,v-1} & m_{p,v+1} & \dots & m_{p,p} \end{bmatrix}, \ u,v = 1,\dots,p.$$

Then with Corollary A.3.3 the inverse of $\boldsymbol{M}$ is expressible as

$$\boldsymbol{M}^{-1} = \frac{\tilde{\boldsymbol{M}}}{|\boldsymbol{M}|}, \tag{6.1.6}$$

where $\tilde{\boldsymbol{M}} := [\tilde{m}_{u,v}]_{1\le u,v\le p} \in \mathbb{R}^{p\times p}$ denotes the adjoint matrix of $\boldsymbol{M}$, i.e.

$$\tilde{m}_{u,v} := (-1)^{u+v} \left|\boldsymbol{M}_{\{v,u\}}\right|.$$

Before describing (6.1.6) in dependence of ω, we drop the assumption of having a standardized design matrix and try to find an expression for $\boldsymbol{M}_x^{-1} := \big((\boldsymbol{X}+\omega\boldsymbol{\Psi})^T(\boldsymbol{X}+\omega\boldsymbol{\Psi})\big)^{-1}$ with an arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. Therefore we establish the following lemma.

Lemma 6.1.1. *Let $\boldsymbol{X}$ represent an arbitrary $n\times p$ matrix, $n \ge p$ and $\boldsymbol{\Psi}$ an $n\times p$ matrix, $n \ge p$, whose columns are all constant to ψ_j, $j = 1,\dots,p$. Then the determinant of the matrix $\boldsymbol{M}_x = (\boldsymbol{X}+\omega\boldsymbol{\Psi})^T(\boldsymbol{X}+\omega\boldsymbol{\Psi})$ is expressible as*

$$|\boldsymbol{M}_x| = \omega^2\boldsymbol{\psi}^T\boldsymbol{A}_x\boldsymbol{\psi} + 2\omega{\boldsymbol{b}_x}^T\boldsymbol{\psi} + |\boldsymbol{X}^T\boldsymbol{X}|, \tag{6.1.7}$$

with

$$\boldsymbol{\psi}^T := \Big[\psi_1, \ \dots \ ,\psi_p\Big],$$
$${\boldsymbol{b}_x}^T := \Big[|\boldsymbol{X}^T\boldsymbol{X}_{[1]}|, \ \dots \ ,|\boldsymbol{X}^T\boldsymbol{X}_{[p]}|\Big]$$

and

$$\boldsymbol{A}_x := \begin{bmatrix} |{\boldsymbol{X}_{[1]}}^T\boldsymbol{X}_{[1]}| & \dots & |{\boldsymbol{X}_{[1]}}^T\boldsymbol{X}_{[p]}| \\ \vdots & \ddots & \vdots \\ |{\boldsymbol{X}_{[p]}}^T\boldsymbol{X}_{[1]}| & \dots & |{\boldsymbol{X}_{[p]}}^T\boldsymbol{X}_{[p]}| \end{bmatrix},$$

where $\boldsymbol{X}_{[j]}$, $j = 1,\dots,p$ is identical to the $n\times p$ matrix $\boldsymbol{X}$, except that the j-th column of $\boldsymbol{X}$ is replaced by a column of ones.

Proof. Application of the Cauchy–Binet formula (A.2.6) implies

$$\begin{aligned}|\boldsymbol{M}_x| &= \sum_{1\le j_1\le\ldots\le j_p\le n}\left|(\boldsymbol{X}+\omega\boldsymbol{\Psi})^T\begin{bmatrix}1 & \ldots & p\\ j_1 & \ldots & j_p\end{bmatrix}\right|^2\\ &= \sum_{1\le j_1\le\ldots\le j_p\le n}\left|\boldsymbol{X}^T\begin{bmatrix}1 & \ldots & p\\ j_1 & \ldots & j_p\end{bmatrix}+\omega\boldsymbol{\Psi}^T\begin{bmatrix}1 & \ldots & p\\ j_1 & \ldots & j_p\end{bmatrix}\right|^2 . \qquad (6.1.8)\end{aligned}$$

The summands in (6.1.8) are determinants of the sum of the two $p\times p$ matrices $\boldsymbol{X}^T\begin{bmatrix}1 & \ldots & p\\ j_1 & \ldots & j_p\end{bmatrix}$ and $\omega\boldsymbol{\Psi}^T\begin{bmatrix}1 & \ldots & p\\ j_1 & \ldots & j_p\end{bmatrix}$ and the summation is over all $s=\binom{n}{p}$ subsets $\{j_1,\ldots,j_p\}$, $1\le j_1\le\ldots\le j_p\le n$ of cardinality p of $\{1,\ldots,n\}$, which we define by $\{j_{1,k},\ldots,j_{p,k}\},\ k=1,\ldots,s$.
Let

$$\boldsymbol{X}^T\langle k\rangle := \boldsymbol{X}^T\begin{bmatrix}1 & \ldots & p\\ j_{1,k} & \ldots & j_{p,k}\end{bmatrix}$$

and

$$\boldsymbol{\Psi}^T\langle k\rangle := \boldsymbol{\Psi}^T\begin{bmatrix}1 & \ldots & p\\ j_{1,k} & \ldots & j_{p,k}\end{bmatrix}$$

represent the matrices of the k-th subset $\{j_{1,k},\ldots,j_{p,k}\}$ of $\{1,\ldots,n\}$ in (6.1.8). To get an expression for each summand $\left|\boldsymbol{X}^T\langle k\rangle+\omega\boldsymbol{\Psi}^T\langle k\rangle\right|,\ k=1,\ldots,s$, in (6.1.8) we use Corollary A.4.2:
Let $\{i_1,\ldots,i_r\}\subseteq\{1,\ldots,n\}$ and define by $|\boldsymbol{C}_p^{\{i_1,\ldots,i_r\}}|_k$ the determinant of the $p\times p$ matrix, whose $i_1,\ldots,i_r$-th rows are identical to the $i_1,\ldots,i_r$-th rows of $\omega\boldsymbol{\Psi}^T\langle k\rangle$ and whose remaining rows are identical to the rows of $\boldsymbol{X}^T\langle k\rangle$, $k=1,\ldots,s$. For $r\ge 2$ there are at least two rows of $\boldsymbol{C}_p^{\{i_1,\ldots,i_r\}}$ identical to the corresponding rows of $\omega\boldsymbol{\Psi}^T\langle k\rangle$ and thus linearly dependent. Consequently $|\boldsymbol{C}_p^{\{i_1,\ldots,i_r\}}|_k=0,\ r\ge 2$. For $r=1$ we can write $i_r=r$ and we get

$$\left|\boldsymbol{C}_p^{\{r\}}\right|_k=\omega\psi_r|\boldsymbol{X}_{[r]}{}^T\langle k\rangle|,$$

where $\boldsymbol{X}_{[r]}{}^T\langle k\rangle$ is identical to $\boldsymbol{X}^T\langle k\rangle$, except that the r-th row of $\boldsymbol{X}^T\langle k\rangle$ is replaced by a row of ones. Finally we have $|\boldsymbol{C}_p^{\{\}}|_k=|\boldsymbol{X}^T\langle k\rangle|$ for the null set. Hence it follows

$$\begin{aligned}\left|\boldsymbol{X}^T\langle k\rangle+\omega\boldsymbol{\Psi}^T\langle k\rangle\right| &= \sum_{\{i_1,\ldots,i_r\}}\left|\boldsymbol{C}_p^{\{i_1,\ldots,i_r\}}\right|_k = |\boldsymbol{X}^T\langle k\rangle|+\sum_{r=1}^{p}\left|\boldsymbol{C}_p^{\{r\}}\right|_k\\ &= |\boldsymbol{X}^T\langle k\rangle|+\omega\psi_1|\boldsymbol{X}_{[1]}{}^T\langle k\rangle|+\ldots+\omega\psi_p|\boldsymbol{X}_{[p]}{}^T\langle k\rangle|, \qquad (6.1.9)\end{aligned}$$

where $\boldsymbol{X}_{[j]}{}^T\langle k\rangle,\ j=1,\ldots,p$ is, as described above, identical to $\boldsymbol{X}^T\langle k\rangle$, except that each entry of the j-th row of $\boldsymbol{X}^T\langle k\rangle$ is replaced by a one. Finally we obtain for (6.1.8)

$$
\begin{aligned}
|\boldsymbol{M}_x| &= \sum_{k=1}^{s} \left(|\boldsymbol{X}^T \langle k \rangle| + \sum_{r=1}^{p} \left| \boldsymbol{C}_p^{\{r\}} \right|_k \right)^2 \\
&= \sum_{k=1}^{s} \Bigg(|\boldsymbol{X}^T \langle k \rangle|^2 + 2\omega |\boldsymbol{X}^T \langle k \rangle| \sum_{r=1}^{p} \psi_r |\boldsymbol{X}_{[r]}{}^T \langle k \rangle| \\
&\qquad + \omega^2 \sum_{t=1}^{p} \sum_{r=1}^{p} \psi_r \psi_t |\boldsymbol{X}_{[r]}{}^T \langle k \rangle| |\boldsymbol{X}_{[t]}{}^T \langle k \rangle| \Bigg) \\
&= \sum_{k=1}^{s} |\boldsymbol{X}^T \langle k \rangle|^2 + 2\omega \sum_{r=1}^{p} \psi_r \sum_{k=1}^{s} |\boldsymbol{X}^T \langle k \rangle| |\boldsymbol{X}_{[r]}{}^T \langle k \rangle| \\
&\qquad + \omega^2 \sum_{t=1}^{p} \sum_{r=1}^{p} \psi_r \psi_t \sum_{k=1}^{s} |\boldsymbol{X}_{[r]}{}^T \langle k \rangle| |\boldsymbol{X}_{[t]}{}^T \langle k \rangle|. \qquad (6.1.10)
\end{aligned}
$$

We have from the Cauchy–Binet formula

$$
\begin{aligned}
|\boldsymbol{X}^T \boldsymbol{X}| &= \sum_{k=1}^{s} |\boldsymbol{X}^T \langle k \rangle|^2, \\
|\boldsymbol{X}^T \boldsymbol{X}_{[r]}| &= \sum_{k=1}^{s} |\boldsymbol{X}^T \langle k \rangle| |\boldsymbol{X}_{[r]}{}^T \langle k \rangle|, \\
|\boldsymbol{X}_{[r]}{}^T \boldsymbol{X}_{[t]}| &= \sum_{k=1}^{s} |\boldsymbol{X}_{[r]}{}^T \langle k \rangle| |\boldsymbol{X}_{[t]}{}^T \langle k \rangle|
\end{aligned}
$$

and thus it follows for (6.1.10)

$$
|\boldsymbol{M}_x| = \omega^2 \sum_{t=1}^{p} \sum_{r=1}^{p} \psi_r \psi_t |\boldsymbol{X}_{[r]}{}^T \boldsymbol{X}_{[t]}| + 2\omega \sum_{r=1}^{p} \psi_r |\boldsymbol{X}^T \boldsymbol{X}_{[r]}| + |\boldsymbol{X}^T \boldsymbol{X}|. \qquad (6.1.11)
$$

We conclude that (6.1.11) can be written as

$$
|\boldsymbol{M}_x| = \omega^2 \boldsymbol{\psi}^T \boldsymbol{A}_x \boldsymbol{\psi} + 2\omega \boldsymbol{b}_x{}^T \boldsymbol{\psi} + |\boldsymbol{X}^T \boldsymbol{X}|.
$$

□

Notation 6.1.2. For convenience we will often use the notation of Theorem 6.1.1 within this chapter, i.e. for $\boldsymbol{A} \in \mathbb{R}^{p \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times p}$, $n \geq p$ we write for the Cauchy–Binet formula given in Theorem A.2.6

$$
|\boldsymbol{AB}| = \sum_{1 \leq j_1 \leq \ldots \leq j_p \leq n} \left| \boldsymbol{A} \begin{bmatrix} 1 & \ldots & p \\ j_1 & \ldots & j_p \end{bmatrix} \right| \left| \boldsymbol{B} \begin{bmatrix} j_1 & \ldots & j_p \\ 1 & \ldots & p \end{bmatrix} \right|
$$

$$= \sum_{1 \le j_1 \le \ldots \le j_p \le n} \left| \boldsymbol{A} \begin{bmatrix} 1 & \ldots & p \\ j_1 & \ldots & j_p \end{bmatrix} \right| \left| \boldsymbol{B}^T \begin{bmatrix} 1 & \ldots & p \\ j_1 & \ldots & j_p \end{bmatrix} \right|$$

$$=: \sum_{k=1}^{s} |\boldsymbol{A}\langle k \rangle| \, |\boldsymbol{B}\langle k \rangle| ,$$

where

$$\boldsymbol{A}\langle k \rangle := \boldsymbol{A} \begin{bmatrix} 1 & \ldots & p \\ j_{1,k} & \ldots & j_{p,k} \end{bmatrix} \in \mathbb{R}^{p \times p}, \quad k = 1, \ldots, s,$$

and the summation is over all $s = \binom{n}{p}$ subsets $\{j_{1,k}, \ldots, j_{p,k}\}$ of $\{1, \ldots, n\}$ of cardinality p.

The following example should illustrate the proof of Lemma 6.1.1.

Example 6.1.3. Consider the matrices

$$\boldsymbol{X}^T = \begin{bmatrix} -1 & 2 & 2 \\ 1 & 5 & 3 \end{bmatrix},$$

and

$$\boldsymbol{\Psi}^T = \begin{bmatrix} \psi_1 & \psi_1 & \psi_1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \in \mathrm{R}^{2 \times 3}.$$

The determinant of the matrix $\boldsymbol{M}_x = (\boldsymbol{X} + \omega\boldsymbol{\Psi})^T (\boldsymbol{X} + \omega\boldsymbol{\Psi})$ is given by

$$|\boldsymbol{M}_x| = \sum_{\{j_1, j_2\}} \left| (\boldsymbol{X} + \omega\boldsymbol{\Psi})^T \begin{bmatrix} 1 & 2 \\ j_1 & j_2 \end{bmatrix} \right|^2$$

$$= \sum_{\{j_1, j_2\}} \left| \boldsymbol{X}^T \begin{bmatrix} 1 & 2 \\ j_1 & j_2 \end{bmatrix} + \omega\boldsymbol{\Psi}^T \begin{bmatrix} 1 & 2 \\ j_1 & j_2 \end{bmatrix} \right|^2, \quad 1 \le j_1 \le j_2 \le 3$$

where the summation is over the $s = \binom{3}{2} = 3$ subsets of $\{1, 2, 3\}$ of cardinality 2. Thus

$$\{j_{1,1}, j_{2,1}\} = \{1, 2\}$$
$$\{j_{1,2}, j_{2,2}\} = \{1, 3\}$$
$$\{j_{1,3}, j_{2,3}\} = \{2, 3\}$$

and we get

$$
\begin{aligned}
|\boldsymbol{M}_x| &= \sum_{k=1}^{3} |\boldsymbol{X}^T \langle k \rangle + \omega \boldsymbol{\Psi}^T \langle k \rangle|^2 \\
&= \left| \boldsymbol{X}^T \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} + \omega \boldsymbol{\Psi}^T \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \right|^2 + \left| \boldsymbol{X}^T \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} + \omega \boldsymbol{\Psi}^T \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} \right|^2 \\
&\qquad + \left| \boldsymbol{X}^T \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} + \omega \boldsymbol{\Psi}^T \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \right|^2 \\
&= \left| \begin{bmatrix} -1 & 2 \\ 1 & 5 \end{bmatrix} + \omega \begin{bmatrix} \psi_1 & \psi_1 \\ 0 & 0 \end{bmatrix} \right|^2 + \left| \begin{bmatrix} -1 & 2 \\ 1 & 3 \end{bmatrix} + \omega \begin{bmatrix} \psi_1 & \psi_1 \\ 0 & 0 \end{bmatrix} \right|^2 \\
&\qquad + \left| \begin{bmatrix} 2 & 2 \\ 5 & 3 \end{bmatrix} + \omega \begin{bmatrix} \psi_1 & \psi_1 \\ 0 & 0 \end{bmatrix} \right|^2
\end{aligned}
\tag{6.1.12}
$$

The summands in (6.1.12) are determinants of the sum of two matrices. Thus we can apply Corollary A.4.2

$$
\left| \boldsymbol{X}^T \langle k \rangle + \omega \boldsymbol{\Psi}^T \langle k \rangle \right| = \sum_{\{i_1,\ldots,i_r\}} \left| \boldsymbol{C}_p^{\{i_1,\ldots,i_r\}} \right|_k ,
$$

where the summation is over all subsets of $\{1, 2\}$, i.e.

$$
\begin{aligned}
&\{\} \\
&\{1\},\ \{2\} \\
&\{1, 2\}.
\end{aligned}
$$

$\boldsymbol{C}_p^{\{i_1,\ldots,i_r\}}$ is a 2×2 matrix whose $i_1, \ldots, i_r$-th rows are identical to the $i_1, \ldots, i_r$-th rows of $\omega \boldsymbol{\Psi}^T \langle k \rangle$ and whose remaining rows are identical to the remaining rows of $\boldsymbol{X}^T \langle k \rangle$. Thus we get for the first summand in (6.1.12)

$$
\begin{aligned}
\left| \begin{bmatrix} -1 & 2 \\ 1 & 5 \end{bmatrix} + \omega \begin{bmatrix} \psi_1 & \psi_1 \\ 0 & 0 \end{bmatrix} \right| &= \left| \begin{bmatrix} -1 & 2 \\ 1 & 5 \end{bmatrix} \right| + \omega \left| \begin{bmatrix} \psi_1 & \psi_1 \\ 1 & 5 \end{bmatrix} \right| \\
&\qquad + \omega \left| \begin{bmatrix} -1 & 2 \\ 0 & 0 \end{bmatrix} \right| + \omega \left| \begin{bmatrix} \psi_1 & \psi_1 \\ 0 & 0 \end{bmatrix} \right|
\end{aligned}
\tag{6.1.13}
$$

It follows

$$
\begin{aligned}
\left| \begin{bmatrix} -1 & 2 \\ 1 & 5 \end{bmatrix} + \omega \begin{bmatrix} \psi_1 & \psi_1 \\ 0 & 0 \end{bmatrix} \right| &= \left| \begin{bmatrix} -1 & 2 \\ 1 & 5 \end{bmatrix} \right| + \omega\psi_1 \left| \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix} \right| \\
\left| \begin{bmatrix} -1 & 2 \\ 1 & 3 \end{bmatrix} + \omega \begin{bmatrix} \psi_1 & \psi_1 \\ 0 & 0 \end{bmatrix} \right| &= \left| \begin{bmatrix} -1 & 2 \\ 1 & 3 \end{bmatrix} \right| + \omega\psi_1 \left| \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \right|
\end{aligned}
$$

$$\left|\begin{bmatrix}2 & 2\\ 5 & 3\end{bmatrix} + \omega\begin{bmatrix}\psi_1 & \psi_1\\ 0 & 0\end{bmatrix}\right| = \left|\begin{bmatrix}2 & 2\\ 5 & 3\end{bmatrix}\right| + \omega\psi_1\left|\begin{bmatrix}1 & 1\\ 5 & 3\end{bmatrix}\right|$$

and thus

$$|\boldsymbol{M}_x| = \left(\left|\begin{bmatrix}-1 & 2\\ 1 & 5\end{bmatrix}\right| + \omega\psi_1\left|\begin{bmatrix}1 & 1\\ 1 & 5\end{bmatrix}\right|\right)^2 + \left(\left|\begin{bmatrix}-1 & 2\\ 1 & 3\end{bmatrix}\right| + \omega\psi_1\left|\begin{bmatrix}1 & 1\\ 1 & 3\end{bmatrix}\right|\right)^2 \\ + \left(\left|\begin{bmatrix}2 & 2\\ 5 & 3\end{bmatrix}\right| + \omega\psi_1\left|\begin{bmatrix}1 & 1\\ 5 & 3\end{bmatrix}\right|\right)^2.$$

From the definition of $\boldsymbol{X}_{[1]}$ in Theorem 6.1.1 we have

$$\boldsymbol{X}^T\boldsymbol{X}_{[1]} = \begin{bmatrix}-1 & 2 & 2\\ 1 & 5 & 3\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 5\\ 1 & 3\end{bmatrix}$$

and

$$\boldsymbol{X}_{[1]}{}^T\boldsymbol{X}_{[1]} = \begin{bmatrix}1 & 1 & 1\\ 1 & 5 & 3\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 5\\ 1 & 3\end{bmatrix}.$$

With the help of the Cauchy–Binet formula we can write

$$|\boldsymbol{X}^T\boldsymbol{X}| = \left|\begin{bmatrix}-1 & 2\\ 1 & 5\end{bmatrix}\right|^2 + \left|\begin{bmatrix}-1 & 2\\ 1 & 3\end{bmatrix}\right|^2 + \left|\begin{bmatrix}2 & 2\\ 5 & 3\end{bmatrix}\right|^2,$$

$$|\boldsymbol{X}^T\boldsymbol{X}_{[1]}| = \left|\begin{bmatrix}-1 & 2\\ 1 & 5\end{bmatrix}\right|\left|\begin{bmatrix}1 & 1\\ 1 & 5\end{bmatrix}\right| + \left|\begin{bmatrix}-1 & 2\\ 1 & 3\end{bmatrix}\right|\left|\begin{bmatrix}1 & 1\\ 1 & 3\end{bmatrix}\right| + \left|\begin{bmatrix}2 & 2\\ 5 & 3\end{bmatrix}\right|\left|\begin{bmatrix}1 & 1\\ 5 & 3\end{bmatrix}\right|,$$

$$|\boldsymbol{X}_{[1]}{}^T\boldsymbol{X}_{[1]}| = \left|\begin{bmatrix}1 & 1\\ 1 & 5\end{bmatrix}\right|^2 + \left|\begin{bmatrix}1 & 1\\ 1 & 3\end{bmatrix}\right|^2 + \left|\begin{bmatrix}1 & 1\\ 5 & 3\end{bmatrix}\right|^2$$

and thus

$$|\boldsymbol{M}_x| = \omega^2\psi_1^2|\boldsymbol{X}_{[1]}{}^T\boldsymbol{X}_{[1]}| + 2\omega\psi_1|\boldsymbol{X}^T\boldsymbol{X}_{[1]}| + |\boldsymbol{X}^T\boldsymbol{X}|.$$

•

The determinant of $\boldsymbol{M}_x$ is a parabola in ω. W.l.o.g. let $\psi_i \neq 0,\ 1 \leq i \leq q$, $1 \leq q \leq p$ and $\psi_i = 0,\ i > q$. Then we have with Lemma 6.1.1

$$\boldsymbol{\psi}^T\boldsymbol{A}_x\boldsymbol{\psi} = \sum_{s=1}^{q}\sum_{r=1}^{q}\psi_r\psi_s|\boldsymbol{X}_{[r]}{}^T\boldsymbol{X}_{[s]}| = \boldsymbol{\psi}_q^T\boldsymbol{A}_x^q\boldsymbol{\psi}_q,$$

$$\boldsymbol{b}_x^T\boldsymbol{\psi} = \sum_{r=1}^{q}\psi_r|\boldsymbol{X}^T\boldsymbol{X}_{[r]}| = \boldsymbol{b}_x^{q\,T}\boldsymbol{\psi}_q$$

with

$$\begin{aligned}
\boldsymbol{\psi}_q^T &:= \Big[\psi_1, \quad \dots \quad , \psi_q\Big] , \\
\boldsymbol{b}_x^{qT} &:= \Big[|\boldsymbol{X}^T\boldsymbol{X}_{[1]}|, \quad \dots \quad , |\boldsymbol{X}^T\boldsymbol{X}_{[q]}|\Big] , \\
\boldsymbol{A}_x^q &:= \begin{bmatrix} |{\boldsymbol{X}_{[1]}}^T\boldsymbol{X}_{[1]}| & \dots & |{\boldsymbol{X}_{[1]}}^T\boldsymbol{X}_{[q]}| \\ \vdots & \ddots & \vdots \\ |{\boldsymbol{X}_{[q]}}^T\boldsymbol{X}_{[1]}| & \dots & |{\boldsymbol{X}_{[q]}}^T\boldsymbol{X}_{[q]}| \end{bmatrix} .
\end{aligned} \tag{6.1.14}$$

The roots of $|\boldsymbol{M}_x|$ can be calculated by solving the quadratic equation

$$\omega^2\boldsymbol{\psi}_q^T\boldsymbol{A}_x^q\boldsymbol{\psi}_q + 2\omega\boldsymbol{b}_x^{qT}\boldsymbol{\psi} + |\boldsymbol{X}^T\boldsymbol{X}| = 0. \tag{6.1.15}$$

The solutions ω_1, ω_2 are given by

$$\omega_{1/2} = \frac{-2\boldsymbol{b}_x^{qT}\boldsymbol{\psi}_q \pm \sqrt{4\left((\boldsymbol{b}_x^{qT}\boldsymbol{\psi}_q)^2 - \boldsymbol{\psi}_q^T\boldsymbol{A}_x^q\boldsymbol{\psi}_q|\boldsymbol{X}^T\boldsymbol{X}|\right)}}{2\boldsymbol{\psi}_q^T\boldsymbol{A}_x^q\boldsymbol{\psi}_q}$$

with the discriminant

$$D := 4\left((\boldsymbol{b}_x^{qT}\boldsymbol{\psi}_q)^2 - \boldsymbol{\psi}_q^T\boldsymbol{A}_x^q\boldsymbol{\psi}_q|\boldsymbol{X}^T\boldsymbol{X}|\right). \tag{6.1.16}$$

For purposes of proving that $|\boldsymbol{M}_x|$ has at most one root, it is convenient to prove the following lemma which states the positive semidefiniteness of the matrix $\boldsymbol{A}_x^q$.

Lemma 6.1.4. *Define by $\boldsymbol{R}_u$, $u = 1, \dots, m$ arbitrary $n \times p$ matrices with $n \geq p$. A matrix of the form*

$$\boldsymbol{R} = \begin{bmatrix} |\boldsymbol{R}_1^T\boldsymbol{R}_1| & |\boldsymbol{R}_1^T\boldsymbol{R}_2| & \dots & |\boldsymbol{R}_1^T\boldsymbol{R}_m| \\ |\boldsymbol{R}_2^T\boldsymbol{R}_1| & |\boldsymbol{R}_2^T\boldsymbol{R}_2| & \dots & |\boldsymbol{R}_2^T\boldsymbol{R}_m| \\ \vdots & \vdots & \ddots & \vdots \\ |\boldsymbol{R}_m^T\boldsymbol{R}_1| & |\boldsymbol{R}_m^T\boldsymbol{R}_2| & \dots & |\boldsymbol{R}_m^T\boldsymbol{R}_m| \end{bmatrix}$$

can be written as

$$\boldsymbol{R} = \tilde{\boldsymbol{R}}^T\tilde{\boldsymbol{R}},$$

where

$$\tilde{\boldsymbol{R}}^T := \begin{bmatrix} \left|\boldsymbol{R}_1^T\langle 1\rangle\right| & \dots & \left|\boldsymbol{R}_1^T\langle s\rangle\right| \\ \vdots & & \vdots \\ \left|\boldsymbol{R}_m^T\langle 1\rangle\right| & \dots & \left|\boldsymbol{R}_m^T\langle s\rangle\right| \end{bmatrix} \in \mathbb{R}^{m\times s}$$

and $s = \dbinom{n}{p}$.
As a consequence $\boldsymbol{R}$ is a positive semidefinite matrix.

PROOF. From Theorem A.2.3 we have $|\boldsymbol{R}_u^T\boldsymbol{R}_v| = |\boldsymbol{R}_v^T\boldsymbol{R}_u|$, $u, v = 1, \dots, m$. Applying the Cauchy–Binet formula, Corollary A.2.7 and Notation 6.1.2 yields

$$|\boldsymbol{R}_u^T\boldsymbol{R}_u| = \sum_{1\leq j_1\leq\ldots\leq j_p\leq n} \left|\boldsymbol{R}_u^T \begin{bmatrix} 1 & \dots & p \\ j_1 & \dots & j_p \end{bmatrix}\right|^2 = \sum_{k=1}^{s} \left|\boldsymbol{R}_u^T\langle k\rangle\right|^2, \ u = 1, \dots, m$$

and

$$\begin{aligned}|\boldsymbol{R}_u^T\boldsymbol{R}_v| &= \sum_{1\leq j_1\leq\ldots\leq j_p\leq n} \left|\boldsymbol{R}_u^T \begin{bmatrix} 1 & \dots & p \\ j_1 & \dots & j_p \end{bmatrix}\right| \left|\boldsymbol{R}_v^T \begin{bmatrix} 1 & \dots & p \\ j_1 & \dots & j_p \end{bmatrix}\right| \\ &= \sum_{k=1}^{s} \left|\boldsymbol{R}_u^T\langle k\rangle\right| \left|\boldsymbol{R}_v^T\langle k\rangle\right| \\ &\qquad u, v = 1, \dots, m, \ u \neq v. \quad (6.1.17)\end{aligned}$$

The summation is over all $s = \binom{n}{p}$ subsets $\{j_{1,k}, \dots, j_{p,k}\}$, $k = 1, \dots, s$ of $\{1, \dots, n\}$.
Thus the decomposition $\boldsymbol{R} = \tilde{\boldsymbol{R}}^T\tilde{\boldsymbol{R}}$ of the matrix $\boldsymbol{R}$ follows directly from (6.1.17). With Theorem A.9.3, $\boldsymbol{R}$ is positive semidefinite.

□

As a consequence of Lemma 6.1.4 we can write

$$\boldsymbol{A}_x^q = \breve{\boldsymbol{A}}^T\breve{\boldsymbol{A}}$$

with

$$\breve{\boldsymbol{A}}^T = \begin{bmatrix} \left|\boldsymbol{X}_{[1]}{}^T\langle 1\rangle\right| & \dots & \left|\boldsymbol{X}_{[1]}{}^T\langle s\rangle\right| \\ \vdots & & \vdots \\ \left|\boldsymbol{X}_{[q]}{}^T\langle 1\rangle\right| & \dots & \left|\boldsymbol{X}_{[q]}{}^T\langle s\rangle\right| \end{bmatrix} \in \mathbb{R}^{q\times s},$$

and $\boldsymbol{A}_x^q$ is positive semidefinite.
Thus $\boldsymbol{\psi}_q^T\boldsymbol{A}_x^q\boldsymbol{\psi}_q \geq 0$ and for $\boldsymbol{\psi}_q^T\boldsymbol{A}_x^q\boldsymbol{\psi}_q > 0$, the determinant $|\boldsymbol{M}_x|$ is an opened upward parabola.
With the help of Lemma 6.1.4 we can prove the following lemma.

LEMMA 6.1.5. *For $\boldsymbol{A}_x^q$ and $\boldsymbol{b}_x^q$ defined like in (6.1.14) and $|\boldsymbol{X}^T\boldsymbol{X}| \neq 0$ the matrix*

$$\boldsymbol{B} := \boldsymbol{A}_x^q - \frac{\boldsymbol{b}_x^q\boldsymbol{b}_x^{q\,T}}{|\boldsymbol{X}^T\boldsymbol{X}|}$$

is positive semidefinite, i.e. $\boldsymbol{B} \geq 0$.

Proof. Let

$$\boldsymbol{G} := \begin{bmatrix} \boldsymbol{A}_x^q & \boldsymbol{b}_x^q \\ \boldsymbol{b}_x^{qT} & |\boldsymbol{X}^T\boldsymbol{X}| \end{bmatrix} = \left[\begin{array}{ccc|c} |{\boldsymbol{X}_{[1]}}^T\boldsymbol{X}_{[1]}| & \ldots & |{\boldsymbol{X}_{[1]}}^T\boldsymbol{X}_{[q]}| & |\boldsymbol{X}^T\boldsymbol{X}_{[1]}| \\ \vdots & \ddots & \vdots & \vdots \\ |{\boldsymbol{X}_{[q]}}^T\boldsymbol{X}_{[1]}| & \ldots & |{\boldsymbol{X}_{[q]}}^T\boldsymbol{X}_{[q]}| & |\boldsymbol{X}^T\boldsymbol{X}_{[q]}| \\ \hline |\boldsymbol{X}^T\boldsymbol{X}_{[1]}| & \ldots & |\boldsymbol{X}^T\boldsymbol{X}_{[q]}| & |\boldsymbol{X}^T\boldsymbol{X}| \end{array}\right], \\ \in \mathbb{R}^{(q+1)\times(q+1)}.$$

With Corollary A.2.7 we have

$$|\boldsymbol{X}^T\boldsymbol{X}| = \sum_{1\le j_1\le\ldots\le j_p\le n} \left|\boldsymbol{X}^T \begin{bmatrix} 1 & \ldots & p \\ j_1 & \ldots & j_p \end{bmatrix}\right|^2 = \sum_{k=1}^{s} \left|\boldsymbol{X}^T\langle k\rangle\right|^2 = \boldsymbol{X}'^T\boldsymbol{X}',$$

where

$$\boldsymbol{X}'^T := \left[\left|\boldsymbol{X}^T\langle 1\rangle\right|, \quad \ldots \quad, \left|\boldsymbol{X}^T\langle s\rangle\right|\right] \in \mathbb{R}^{1\times s}.$$

It follows again with the help of the Cauchy–Binet formula

$$\begin{aligned} \breve{\boldsymbol{A}}^T\boldsymbol{X}' &= \begin{bmatrix} |{\boldsymbol{X}_{[1]}}^T\langle 1\rangle| & \ldots & |{\boldsymbol{X}_{[1]}}^T\langle s\rangle| \\ \vdots & & \vdots \\ |{\boldsymbol{X}_{[q]}}^T\langle 1\rangle| & \ldots & |{\boldsymbol{X}_{[q]}}^T\langle s\rangle| \end{bmatrix} \begin{bmatrix} |\boldsymbol{X}^T\langle 1\rangle| \\ \vdots \\ |\boldsymbol{X}^T\langle s\rangle| \end{bmatrix} \\ &= \begin{bmatrix} |{\boldsymbol{X}_{[1]}}^T\boldsymbol{X}| \\ \vdots \\ |{\boldsymbol{X}_{[q]}}^T\boldsymbol{X}| \end{bmatrix} = \begin{bmatrix} |\boldsymbol{X}^T\boldsymbol{X}_{[1]}| \\ \vdots \\ |\boldsymbol{X}^T\boldsymbol{X}_{[q]}| \end{bmatrix} = \boldsymbol{b}_x^q, \end{aligned}$$

because

$$\sum_{k=1}^{s} |{\boldsymbol{X}_{[i]}}^T\langle k\rangle|\,|\boldsymbol{X}^T\langle k\rangle| = |{\boldsymbol{X}_{[i]}}^T\boldsymbol{X}|, \quad i = 1,\ldots,q,$$

with the notation given in Notation 6.1.2. Hence we have

$$\boldsymbol{G} = \begin{bmatrix} \breve{\boldsymbol{A}}^T\breve{\boldsymbol{A}} & \breve{\boldsymbol{A}}^T\boldsymbol{X}' \\ \boldsymbol{X}'^T\breve{\boldsymbol{A}} & \boldsymbol{X}'^T\boldsymbol{X}' \end{bmatrix} = \begin{bmatrix} \breve{\boldsymbol{A}}^T \\ \boldsymbol{X}'^T \end{bmatrix} \begin{bmatrix} \breve{\boldsymbol{A}} & \boldsymbol{X}' \end{bmatrix} \in \mathbb{R}^{(q+1)\times(q+1)} \tag{6.1.18}$$

and it follows with Theorem A.9.3, that $\boldsymbol{G}$ is a positive semidefinite matrix. As an immediate consequence of Theorem A.9.4 we have

$$\left(\boldsymbol{A}_x^q - \frac{\boldsymbol{b}_x^q\boldsymbol{b}_x^{qT}}{|\boldsymbol{X}^T\boldsymbol{X}|}\right) \ge 0.$$

□

We can deduce from Lemma 6.1.5

$$
\begin{aligned}
0 \le \boldsymbol{\psi}_q^T \left(\boldsymbol{A}_x^q - \frac{\boldsymbol{b}_x^q \boldsymbol{b}_x^{qT}}{|\boldsymbol{X}^T\boldsymbol{X}|} \right) \boldsymbol{\psi}_q &= \boldsymbol{\psi}_q^T \boldsymbol{A}_x^q \boldsymbol{\psi}_q - \frac{(\boldsymbol{\psi}_q^T \boldsymbol{b}_x^q)(\boldsymbol{b}_x^{qT} \boldsymbol{\psi}_q)}{|\boldsymbol{X}^T\boldsymbol{X}|} \\
&= \boldsymbol{\psi}_q^T \boldsymbol{A}_x^q \boldsymbol{\psi}_q - \frac{(\boldsymbol{\psi}_q^T \boldsymbol{b}_x^q)^2}{|\boldsymbol{X}^T\boldsymbol{X}|} \\
&\qquad\qquad \Leftrightarrow \left(\boldsymbol{b}_x^{qT} \boldsymbol{\psi}\right)^2 - \boldsymbol{\psi}^T \boldsymbol{A}_x^q \boldsymbol{\psi} |\boldsymbol{X}^T\boldsymbol{X}| \le 0,
\end{aligned}
$$

i.e. the discriminant D of $|\boldsymbol{M}_x|$, given in (6.1.16), is smaller than or equal to zero. Thus the quadratic equation $|\boldsymbol{M}_x| = 0$, given in (6.1.15), has no or only one real solution. As a consequence $|\boldsymbol{M}_x|$ has at most one root.
It is

$$
s = \binom{n}{p} = \frac{n!}{(n-p)!p!} = n \frac{(n-1)\cdot(n-2)\cdot\ldots\cdot(p+1)}{p\cdot(p-1)\cdot\ldots\cdot 2\cdot 1} \ge n, \quad p < n
$$

and thus $\boldsymbol{G}$ is positive definite, iff the matrix $\begin{bmatrix}\breve{\boldsymbol{A}} & \boldsymbol{X}'\end{bmatrix} \in \mathbb{R}^{s\times(q+1)}$, $q \le p$ has full column rank, i.e. rank $\left(\begin{bmatrix}\breve{\boldsymbol{A}} & \boldsymbol{X}'\end{bmatrix}\right) = q+1$ (see Theorem A.9.3).
The matrix $\boldsymbol{G}$ is only positive semidefinite if

(1) $q = p = n$, because in this case $s = 1$ and thus rank $\left(\begin{bmatrix}\breve{\boldsymbol{A}} & \boldsymbol{X}'\end{bmatrix}\right) = 1$, but this was excluded by Assumption 3 in Chapter 1,

(2) $\boldsymbol{X}$ does not have full column rank. Then it follows $\boldsymbol{X}' = \boldsymbol{0}$ and thus rank $\left(\begin{bmatrix}\breve{\boldsymbol{A}} & \boldsymbol{X}'\end{bmatrix}\right) < q+1$. But this situation is excluded by Assumption 2 in Chapter 1,

(3) each entry of the j–th column of $\boldsymbol{X}$ equals any constant $c \in \mathbb{R}$. Then $\boldsymbol{X}'$ is a multiple of the j-th column of $\breve{\boldsymbol{A}}$ and it follows again rank $\left(\begin{bmatrix}\breve{\boldsymbol{A}} & \boldsymbol{X}'\end{bmatrix}\right) < q+1$.

Otherwise $\boldsymbol{G}$ will usually be positive definite.
Now suppose $\boldsymbol{G}$ is positive definite. As a consequence $|\boldsymbol{M}_x|$ has no roots. Because $|\boldsymbol{M}_x|$ is an opened upward parabola it follows $|\boldsymbol{M}_x| > 0$.

Note 6.1.6. From Theorem A.9.3 it follows directly that $\boldsymbol{M}_x$ is a positive semidefinite matrix and thus $|\boldsymbol{M}_x| \ge 0$. $\boldsymbol{M}_x$ is positive definite, iff $(\boldsymbol{X} + \omega\boldsymbol{\Psi})$ has full column rank p.

As a straightforward consequence of Lemma 6.1.1, we obtain the result expressed in the following Lemma.

LEMMA 6.1.7. *Let $\boldsymbol{X}$ represent an arbitrary $n \times p$ matrix and $\boldsymbol{\Psi}$ an $n \times p$ matrix, $n \geq p$, whose columns are all constant to ψ_j, $j = 1, \ldots, p$. If the inverse of $\boldsymbol{M}_x = \left(\boldsymbol{X} + \omega\boldsymbol{\Psi}\right)^T \left(\boldsymbol{X} + \omega\boldsymbol{\Psi}\right)$ exists, it is expressible as*

$$\boldsymbol{M}_x^{-1} = \frac{\tilde{\boldsymbol{M}}_x}{|\boldsymbol{M}_x|} = \frac{\omega^2 \tilde{\boldsymbol{M}}_x^{quad} + \omega \tilde{\boldsymbol{M}}_x^{lin} + \tilde{\boldsymbol{M}}_x^{const}}{\omega^2 \boldsymbol{\psi}^T \boldsymbol{A}_x \boldsymbol{\psi} + 2\omega \boldsymbol{b}_x^T \boldsymbol{\psi} + |\boldsymbol{X}^T \boldsymbol{X}|}, \tag{6.1.19}$$

with the symmetric matrices

$$\begin{aligned}
\tilde{\boldsymbol{M}}_x^{quad} &:= \left[\tilde{m}_x^{quad}{}_{u,v}\right]_{1 \leq u,v \leq p} = \left[(-1)^{u+v} {\boldsymbol{\psi}_{\{u\}}}^T \tilde{\boldsymbol{A}}_x^{(uv)} \boldsymbol{\psi}_{\{v\}}\right]_{1 \leq u,v \leq p}, \\
\tilde{\boldsymbol{M}}_x^{lin} &:= \left[\tilde{m}_x^{lin}{}_{u,v}\right]_{1 \leq u,v \leq p} = \left[(-1)^{u+v} \left(\tilde{\boldsymbol{b}}_x^{(uv)T} \boldsymbol{\psi}_{\{v\}} + \tilde{\boldsymbol{b}}_x^{(vu)T} \boldsymbol{\psi}_{\{u\}}\right)\right]_{1 \leq u,v \leq p}, \\
\tilde{\boldsymbol{M}}_x^{const} &:= \left[\tilde{m}_x^{const}{}_{u,v}\right]_{1 \leq u,v \leq p} = \left[(-1)^{u+v} \left|{\boldsymbol{X}_{\{u\}}}^T \boldsymbol{X}_{\{v\}}\right|\right]_{1 \leq u,v \leq p}
\end{aligned} \tag{6.1.20}$$

and

$$\begin{aligned}
\boldsymbol{\psi}_{\{u\}} &:= \left[\psi_r\right]_{\substack{1 \leq r \leq p \\ r \neq u}} \in \mathbb{R}^{(p-1)\times 1}, \\
\tilde{\boldsymbol{b}}_x^{(uv)} &:= \left[\left|{\boldsymbol{X}_{\{u\}}}^T \boldsymbol{X}_{\{v\}[r]}\right|\right]_{\substack{1 \leq r \leq p \\ r \neq v}} \in \mathbb{R}^{(p-1)\times 1}, \\
\tilde{\boldsymbol{A}}_x^{(uv)} &:= \left[\left|{\boldsymbol{X}_{\{u\}[r]}}^T \boldsymbol{X}_{\{v\}[s]}\right|\right]_{\substack{1 \leq r,s \leq p \\ u \neq r; v \neq s}} \in \mathbb{R}^{(p-1)\times(p-1)}.
\end{aligned}$$

$\boldsymbol{X}_{\{u\}} \in \mathbb{R}^{n \times (p-1)}$ means that the u–th column of $\boldsymbol{X}$ is missing and $\boldsymbol{X}_{\{u\}[r]}$ is formed out of $\boldsymbol{X}$ by replacing the r–th column of $\boldsymbol{X}$ by a column of ones and afterwards striking out the u–th column.

PROOF. The denominator of (6.1.19) is given by Lemma 6.1.1. For the examination of the adjoint matrix $\tilde{\boldsymbol{M}}_x =: \left[\tilde{m}_{u,v}^x\right]_{1 \leq u,v \leq p}$ of $\boldsymbol{M}_x$, we will use the same arguments as in Lemma 6.1.1.

Let $\boldsymbol{X}_{\{u\}}, \boldsymbol{\Psi}_{\{u\}}$ and respectively $\boldsymbol{X}_{\{v\}}, \boldsymbol{\Psi}_{\{v\}}$ represent the four $n \times (p-1)$ matrices, obtained from $\boldsymbol{X}$ or $\boldsymbol{\Psi}$ by striking out the u-th or v-th column. Applying the Cauchy–Binet formula (A.2.6) implies

$$\begin{aligned}
\tilde{m}_{v,u}^x = \tilde{m}_{u,v}^x &= (-1)^{u+v} \left|\left(\boldsymbol{X}_{\{u\}} + \omega \boldsymbol{\Psi}_{\{u\}}\right)^T \left(\boldsymbol{X}_{\{v\}} + \omega \boldsymbol{\Psi}_{\{v\}}\right)\right| \\
&= (-1)^{u+v} \sum_{k=1}^{\tilde{s}} \left|{\boldsymbol{X}_{\{u\}}}^T \langle k \rangle + \omega {\boldsymbol{\Psi}_{\{u\}}}^T \langle k \rangle\right| \left|{\boldsymbol{X}_{\{v\}}}^T \langle k \rangle + \omega {\boldsymbol{\Psi}_{\{v\}}}^T \langle k \rangle\right|,
\end{aligned}$$

where the summation is over all $\tilde{s} := \binom{n}{p-1}$ subsets of cardinality $(p-1)$ of $\{1, \ldots, n\}$. Set for the k-th subset, $k = 1, \ldots, \tilde{s}$

$$\begin{aligned}
&\left|\boldsymbol{X}_{\{u\}}{}^T\langle k\rangle + \omega\boldsymbol{\Psi}_{\{u\}}{}^T\langle k\rangle\right| \\
&\qquad := \left|\boldsymbol{X}_{\{u\}}{}^T \begin{bmatrix} 1 & \dots & p-1 \\ j_{1,k} & \dots & j_{p-1,k} \end{bmatrix} + \omega\boldsymbol{\Psi}_{\{u\}}{}^T \begin{bmatrix} 1 & \dots & p-1 \\ j_{1,k} & \dots & j_{p-1,k} \end{bmatrix}\right| .
\end{aligned}$$

With Corollary A.4.2 it follows with the same arguments as in (6.1.9)

$$\left|\boldsymbol{X}_{\{u\}}{}^T\langle k\rangle + \omega\boldsymbol{\Psi}_{\{u\}}{}^T\langle k\rangle\right| = |\boldsymbol{X}_{\{u\}}{}^T\langle k\rangle| + \omega\sum_{\substack{r=1\\ r\neq u}}^{p} \psi_r |\boldsymbol{X}_{\{u\}[r]}{}^T\langle k\rangle|$$

and

$$\left|\boldsymbol{X}_{\{v\}}{}^T\langle k\rangle + \omega\boldsymbol{\Psi}_{\{v\}}{}^T\langle k\rangle\right| = |\boldsymbol{X}_{\{v\}}{}^T\langle k\rangle| + \omega\sum_{\substack{s=1\\ s\neq v}}^{p} \psi_s |\boldsymbol{X}_{\{v\}[s]}{}^T\langle k\rangle| ,$$

where $\boldsymbol{X}_{\{u\}[r]}$, $r = 1, \dots, p$, $r \neq u$ is identical to the matrix $\boldsymbol{X}_{\{u\}}$, except that the r-th column of $\boldsymbol{X}$ in $\boldsymbol{X}_{\{u\}}$ is replaced by a column of ones. Hence it follows

$$\begin{aligned}
\tilde{m}^x_{u,v} = (-1)^{u+v} \sum_{k=1}^{\tilde{s}} \Bigg(& |\boldsymbol{X}_{\{u\}}{}^T\langle k\rangle||\boldsymbol{X}_{\{v\}}{}^T\langle k\rangle| \\
& +\omega\sum_{\substack{s=1\\ s\neq v}}^{p}\psi_s|\boldsymbol{X}_{\{u\}}{}^T\langle k\rangle||\boldsymbol{X}_{\{v\}[s]}{}^T\langle k\rangle| + \omega\sum_{\substack{r=1\\ r\neq u}}^{p}\psi_r|\boldsymbol{X}_{\{v\}}{}^T\langle k\rangle||\boldsymbol{X}_{\{u\}[r]}{}^T\langle k\rangle| \\
& +\omega^2\sum_{\substack{s=1\\ s\neq v}}^{p}\sum_{\substack{r=1\\ r\neq u}}^{p}\psi_r\psi_s|\boldsymbol{X}_{\{u\}[r]}{}^T\langle k\rangle||\boldsymbol{X}_{\{v\}[s]}{}^T\langle k\rangle| \Bigg) \\
= (-1)^{u+v} \Bigg(& \omega^2\sum_{\substack{s=1\\ s\neq v}}^{p}\sum_{\substack{r=1\\ r\neq u}}^{p}\psi_r\psi_s|\boldsymbol{X}_{\{u\}[r]}{}^T\boldsymbol{X}_{\{v\}[s]}| + \omega\sum_{\substack{s=1\\ s\neq v}}^{p}\psi_s|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{v\}[s]}| \\
& +\omega\sum_{\substack{r=1\\ r\neq u}}^{p}\psi_r|\boldsymbol{X}_{\{v\}}{}^T\boldsymbol{X}_{\{u\}[r]}| + |\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{v\}}| \Bigg) \\
& = \omega^2\tilde{m}_x^{quad}{}_{u,v} + \omega\tilde{m}_x^{lin}{}_{u,v} + \tilde{m}_x^{const}{}_{u,v} \qquad (6.1.21)
\end{aligned}$$

and this is equivalent to the result to be proved. It is not difficult to see that $\tilde{\boldsymbol{M}}_x^{quad}$,$\tilde{\boldsymbol{M}}_x^{lin}$ and $\tilde{\boldsymbol{M}}_x^{const}$ are symmetric $p \times p$ matrices.

□

NOTE 6.1.8. The matrix $\tilde{\boldsymbol{M}}_x^{const}$ is equivalent to the adjoint matrix of $\boldsymbol{X}^T\boldsymbol{X}$ and for $\omega = 0$ (6.1.19) results in the formula for the calculation of the inverse of $\boldsymbol{X}^T\boldsymbol{X}$ given in Corollary A.3.3.

We showed above that $|\boldsymbol{M}_x|$ is usually bigger than zero and thus the inverse of $\boldsymbol{M}_x$ will usually exist.
Now we concentrate our attention again on the standardized design matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times p},\ n > p$, which is assumed to have full column rank. To describe the inverse of

$$\boldsymbol{M} = \left(\boldsymbol{Z} + \omega\boldsymbol{\Psi}\right)^T \left(\boldsymbol{Z} + \omega\boldsymbol{\Psi}\right) = \boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi},$$

introduced in (6.1.4), in dependence of ω, we can apply Lemma 6.1.7. Because $\boldsymbol{Z}$ is a centered matrix it follows

$$\boldsymbol{Z}^T\boldsymbol{Z}_{[v]} \quad = \quad \begin{bmatrix} \|Z_1\|^2 & \dots & {Z_1}^T Z_{v-1} & 0 & {Z_1}^T Z_{v+1} & \dots & {Z_1}^T Z_p \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ {Z_p}^T Z_1 & \dots & {Z_p}^T Z_{v-1} & 0 & {Z_p}^T Z_{v+1} & \dots & \|Z_p\|^2 \end{bmatrix}$$

and thus $|\boldsymbol{Z}^T\boldsymbol{Z}_{[v]}| = 0,\ v = 1, \dots, p.$
As a consequence

(1) $\boldsymbol{b}_x = \boldsymbol{0}$

in Lemma 6.1.1.
In the same way it is easy to see that

$$\boldsymbol{Z}_{[u]}{}^T\boldsymbol{Z}_{[v]} = \begin{bmatrix} \|Z_1\|^2 & \dots & {Z_1}^T Z_{v-1} & 0 & {Z_1}^T Z_{v+1} & \dots & {Z_1}^T Z_p \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ {Z_{u-1}}^T Z_1 & \dots & {Z_{u-1}}^T Z_{v-1} & 0 & {Z_{u-1}}^T Z_{v+1} & \dots & {Z_{u-1}}^T Z_p \\ 0 & \dots & 0 & n & 0 & \dots & 0 \\ {Z_{u+1}}^T Z_1 & \dots & {Z_{u+1}}^T Z_{v-1} & 0 & {Z_{u+1}}^T Z_{v+1} & \dots & {Z_{u+1}}^T Z_p \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ {Z_p}^T Z_1 & \dots & {Z_p}^T Z_{v-1} & 0 & {Z_p}^T Z_{v+1} & \dots & \|Z_p\|^2 \end{bmatrix}$$

and thus

$$|\boldsymbol{Z}_{[u]}{}^T\boldsymbol{Z}_{[v]}| = (-1)^{u+v} n |\boldsymbol{Z}_{\{u\}}{}^T\boldsymbol{Z}_{\{v\}}|, \qquad p \geq 2,\ u, v = 1, \dots, p. \tag{6.1.22}$$

Hence in case of a standardized and thus centered matrix $\boldsymbol{Z}$, the matrix $\boldsymbol{A}_x$ of Lemma 6.1.1 is identical to n times the adjoint matrix of $|\boldsymbol{Z}^T\boldsymbol{Z}|$, i.e.

(2) $\boldsymbol{A}_x = n \left[(-1)^{u+v} \left|\boldsymbol{Z}_{\{u\}}{}^T\boldsymbol{Z}_{\{v\}}\right|\right]_{1\leq u,v\leq p}.$

With the same argumentation we get for $r \neq u$ and $s \neq v$

$$|\boldsymbol{Z}_{\{u\}}{}^T\boldsymbol{Z}_{\{v\}[s]}| = 0$$

and

$$|\boldsymbol{Z}_{\{u\}[r]}{}^T\boldsymbol{Z}_{\{v\}[s]}| = (-1)^{r+s} n |\boldsymbol{Z}_{\{ur\}}{}^T\boldsymbol{Z}_{\{vs\}}|, \quad p \geq 3, \tag{6.1.23}$$

where $\boldsymbol{Z}_{\{ur\}},\ r = 1,\ldots,p$ means that both, the u-th and r-th column of $\boldsymbol{Z}$ are striked out. Equation (6.1.23) is not valid in case of only two regressors, because it is

$$\boldsymbol{Z}_{\{1\}[2]} = \boldsymbol{Z}_{\{2\}[1]} = \mathbf{1}_n$$

and thus it follows for $p = 2$

$$|{\boldsymbol{Z}_{\{u\}[r]}}^T \boldsymbol{Z}_{\{v\}[s]}| = n, \quad u, v, r, s = 1, 2;\ r \neq u; s \neq v.$$

As a consequence we have

(3) $\tilde{\boldsymbol{b}}_x^{(uv)} = 0$ and

(4) $\tilde{\boldsymbol{A}}_x^{(uv)} = \begin{cases} n & ,p = 2 \\ \left[(-1)^{r+s} n |{\boldsymbol{Z}_{\{ur\}}}^T \boldsymbol{Z}_{\{vs\}}|\right]_{\substack{1\leq r,s\leq p \\ u\neq r; v\neq s}} & ,p \geq 3 \end{cases}.$

With (3),(4) and (6.1.21) it follows

$$(5)\ \tilde{m}_{u,v} = (-1)^{u+v}\left(n\omega^2 \sum_{\substack{r=1\\r\neq u}}^{p} \sum_{\substack{s=1\\s\neq v}}^{p} (-1)^{r+s}\psi_r\psi_s |{\boldsymbol{Z}_{\{ur\}}}^T \boldsymbol{Z}_{\{vs\}}| + |{\boldsymbol{Z}_{\{u\}}}^T \boldsymbol{Z}_{\{v\}}|\right),$$
$$1 \leq u, v \leq p,\ p \geq 3$$

and for $p = 2$ we get

$$\tilde{m}_{u,v} = (-1)^{u+v}\left(n\omega^2 \sum_{\substack{r=1\\r\neq u}}^{2} \sum_{\substack{s=1\\s\neq v}}^{2} \psi_r\psi_s + |{\boldsymbol{Z}_{\{u\}}}^T \boldsymbol{Z}_{\{v\}}|\right), \quad u, v = 1, 2.$$

Thus from (1), (2) and (5), Lemma 6.1.7 has the following implication for the standardized matrix $\boldsymbol{Z}$.

Corollary 6.1.9. *Let $\boldsymbol{Z}$ represent an arbitrary standardized (especially centered) $n \times p$ matrix, $n > p,\ p \geq 2$ and $\boldsymbol{\Psi}$ an $n \times p$ matrix, whose columns are all constant to $\psi_j,\ j = 1,\ldots,p$. Then the inverse of the matrix $\boldsymbol{M} = (\boldsymbol{Z} + \omega\boldsymbol{\Psi})^T(\boldsymbol{Z} + \omega\boldsymbol{\Psi}) = \boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ is expressible as*

$$\boldsymbol{M}^{-1} = \frac{\tilde{\boldsymbol{M}}}{|\boldsymbol{M}|} = \frac{n\omega^2 \tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|},$$

with

$$\tilde{\boldsymbol{M}}^{quad} := \left[\tilde{m}_{u,v}^{quad}\right]_{1\leq u,v\leq p} := \left[(-1)^{u+v} {\boldsymbol{\psi}_{\{u\}}}^T \tilde{\boldsymbol{A}}^{(uv)} \boldsymbol{\psi}_{\{v\}}\right]_{1\leq u,v\leq p},$$

$$\tilde{\boldsymbol{M}}^{const} := \left[\tilde{m}_{u,v}^{const}\right]_{1\leq u,v\leq p} := \left[(-1)^{u+v}\, |{\boldsymbol{Z}_{\{u\}}}^T \boldsymbol{Z}_{\{v\}}|\right]_{1\leq u,v\leq p},$$

and

$$\boldsymbol{\psi}_{\{u\}} := \begin{bmatrix}\psi_r\end{bmatrix}_{\substack{1\le r\le p\\ r\neq u}} \in \mathbb{R}^{(p-1)\times 1},$$

$$\tilde{\boldsymbol{A}}^{(uv)} := \begin{cases} 1 & ,p=2 \\ \left[(-1)^{r+s}|\boldsymbol{Z}_{\{ur\}}{}^T\boldsymbol{Z}_{\{vs\}}|\right]_{\substack{1\le r,s\le p\\ u\neq r; v\neq s}} & ,p\ge 3 \end{cases} \in \mathbb{R}^{(p-1)\times(p-1)}.$$

The matrix $\tilde{\boldsymbol{M}}^{const}$ is equivalent to the adjoint matrix of $\boldsymbol{Z}^T\boldsymbol{Z}$. From Corollary A.3.3 we have

$$\tilde{\boldsymbol{M}}^{const} = |\boldsymbol{Z}^T\boldsymbol{Z}|\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \in \mathbb{R}^{p\times p}. \tag{6.1.24}$$

From Theorem A.9.2, (6) we know that the inverse of $\boldsymbol{Z}^T\boldsymbol{Z}$ is also positive definite. Thus the adjoint matrix of $\boldsymbol{Z}^T\boldsymbol{Z}$ is positive definite, because $|\boldsymbol{Z}^T\boldsymbol{Z}| > 0$ by assumption.

Of course this can also be established with the help of Lemma 6.1.4: From (6.1.22) it follows directly with Lemma 6.1.4 that $\tilde{\boldsymbol{M}}^{const}$ is positive semidefinite and thus $n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} \ge 0$. With $|\boldsymbol{Z}^T\boldsymbol{Z}| > 0$ for $\boldsymbol{Z}$ having full rank it follows $|\boldsymbol{M}| > 0$.

As a consequence the inverse of $\boldsymbol{M}$ always exists for $\boldsymbol{Z}$ having full rank.

Example 6.1.10. Consider

$$\boldsymbol{Z}^T = \begin{bmatrix} -2 & 1 & 1 \\ -2 & 2 & 0 \end{bmatrix},$$

which is the transpose of the mean centered matrix of $\boldsymbol{X}$ in Example 6.1.3 and

$$\boldsymbol{\Psi}^T = \begin{bmatrix} \psi_1 & \psi_1 & \psi_1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Thus we have

$$\boldsymbol{Z}^T\boldsymbol{Z} = \begin{bmatrix} 6 & 6 \\ 6 & 8 \end{bmatrix}$$

and

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 & 0 \end{bmatrix}.$$

From Corollary 6.1.9 it is

$$\begin{aligned} \tilde{\boldsymbol{A}}^{(11)} &= \tilde{\boldsymbol{A}}^{(12)} = \tilde{\boldsymbol{A}}^{(21)} = \tilde{\boldsymbol{A}}^{(22)} = 1, \\ \boldsymbol{\psi}_{\{1\}} &= 0, \\ \boldsymbol{\psi}_{\{2\}} &= \psi_1 \end{aligned}$$

and with

$$\tilde{\boldsymbol{M}}_{quad} = \begin{bmatrix} 0 & 0 \\ 0 & \psi_1^2 \end{bmatrix},$$

$$\tilde{\boldsymbol{M}}^{const} = \begin{bmatrix} 8 & -6 \\ -6 & 6 \end{bmatrix}$$

it follows

$$\boldsymbol{M}^{-1} = \frac{n\omega^2 \begin{bmatrix} 0 & 0 \\ 0 & \psi_1^2 \end{bmatrix} + \begin{bmatrix} 8 & -6 \\ -6 & 6 \end{bmatrix}}{n\omega^2\psi_1^2 8 + \left| \begin{bmatrix} 6 & 6 \\ 6 & 8 \end{bmatrix} \right|} \tag{6.1.25}$$

$$= \frac{1}{24\omega^2\psi_1^2 + 12} \left(3\omega^2\psi_1^2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 8 & -6 \\ -6 & 6 \end{bmatrix} \right). \tag{6.1.26}$$

●

6.2. The Disturbed Least Squares Estimator

In the last chapter we found a decomposition of the determinant $|\boldsymbol{M}|$ and of all entries of the adjoint matrix $\tilde{\boldsymbol{M}}$ in dependence of ω. So we are in a position to describe the matrix $\boldsymbol{M}^{-1}$ and thus the disturbed least squares estimator $\tilde{\gamma}_\omega$ in dependence of ω. With the help of Corollary 6.1.9, it follows for (6.1.5)

$$\tilde{\gamma}_\omega = (\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\boldsymbol{y}^* \tag{6.2.27}$$

$$= \frac{n\omega^2\tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|}\boldsymbol{Z}^T\boldsymbol{y}^*,$$

with

$$\tilde{\boldsymbol{M}}^{quad} = \left[\tilde{m}_{u,v}^{quad}\right]_{1\le u,v\le p} = \left[(-1)^{u+v}\boldsymbol{\psi}_{\{u\}}{}^T\tilde{\boldsymbol{A}}^{(uv)}\boldsymbol{\psi}_{\{v\}}\right]_{1\le u,v\le p},$$

$$\tilde{\boldsymbol{M}}^{const} = \left[\tilde{m}_{u,v}^{const}\right]_{1\le u,v\le p} = \left[(-1)^{u+v}\left|\boldsymbol{Z}_{\{u\}}{}^T\boldsymbol{Z}_{\{v\}}\right|\right]_{1\le u,v\le p}. \tag{6.2.28}$$

For $\omega = 0$, we get the least squares estimator of the standardized model

$$\tilde{\gamma}_0 = \frac{\tilde{\boldsymbol{M}}^{const}}{|\boldsymbol{Z}^T\boldsymbol{Z}|}\boldsymbol{Z}^T\boldsymbol{y}^* = \hat{\gamma}.$$

EXAMPLE 6.2.1. Consider $\boldsymbol{Z}$ and $\boldsymbol{\Psi}$ of Example 6.1.10 and let

$$y^* = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

With (6.1.25) of Example 6.1.10, the disturbed least squares estimator is given by

$$\begin{aligned}\tilde{\gamma}_\omega &= \frac{1}{24\omega^2\psi_1^2+12}\left(3\omega^2\psi_1^2\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}+\begin{bmatrix}8 & -6\\ -6 & 6\end{bmatrix}\right)\begin{bmatrix}-2 & 1 & 1\\ -2 & 2 & 0\end{bmatrix}\begin{bmatrix}-1\\ 0\\ 1\end{bmatrix}\\ &= \frac{1}{2\omega^2\psi_1^2+1}\begin{bmatrix}1\\ 0.5(\omega^2\psi_1^2-1)\end{bmatrix}.\end{aligned} \tag{6.2.29}$$

For $\omega = 0$ we get the least squares estimator

$$\hat{\gamma} = \begin{bmatrix}1\\ -0.5\end{bmatrix}.$$

Figure 6.2.1 displays the components of $\tilde{\gamma}_\omega$ in dependence of ω for $\psi_1 = 1$.

•

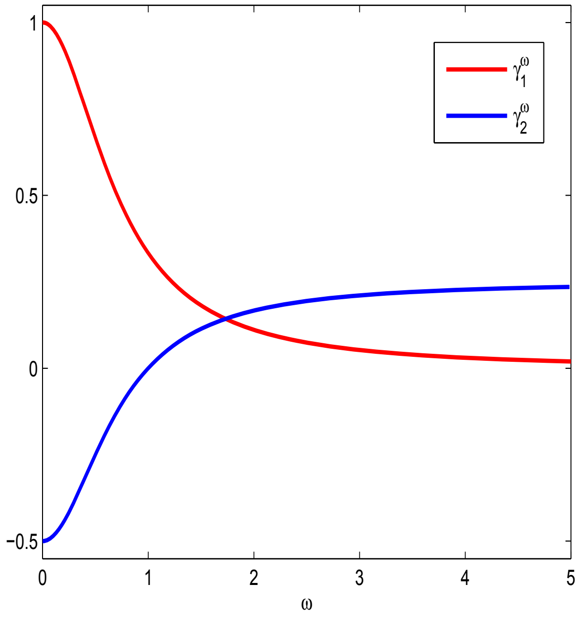


FIGURE 6.2.1. Disturbed least squares estimator in dependence of ω

6.2.1 The Covariance Matrix of the Disturbed Least Squares Estimator

With the same arguments as in (5.1.3), the covariance matrix of $\tilde{\boldsymbol{\gamma}}_\omega =: \begin{bmatrix} \tilde{\gamma}_1^\omega & \dots & \tilde{\gamma}_p^\omega \end{bmatrix}^T$ is given by

$$\begin{aligned}
\Sigma(\tilde{\boldsymbol{\gamma}}_\omega) &= (\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\Sigma(\boldsymbol{y}^*)\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} \\
&= (\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\Sigma(\boldsymbol{\varepsilon}^*)\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} \\
&= (\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\boldsymbol{P}\Sigma(\boldsymbol{\varepsilon})\boldsymbol{P}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} \\
&= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^T\right)\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} \\
&= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}, \qquad (6.2.30)
\end{aligned}$$

because $\boldsymbol{Z}^T\boldsymbol{1}_n$ is a null matrix due to the centered matrix $\boldsymbol{Z}$. It follows with Corollary 6.1.9

$$\begin{aligned}
\Sigma(\tilde{\boldsymbol{\gamma}}_\omega) &= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}-\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} \\
&= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} - \sigma^2\omega^2(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} \\
&= \sigma^2\frac{n\omega^2\tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|} \\
&\quad - \sigma^2\omega^2\frac{n\omega^2\tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\frac{n\omega^2\tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|}. \qquad (6.2.31)
\end{aligned}$$

In order to simplify equation (6.2.31), it is convenient to prove the following lemma.

LEMMA 6.2.2. *For any matrix $\boldsymbol{A} \in \mathbb{R}^{p\times(p+1)}$*

$$\left|\boldsymbol{A}_{\{r\}[l]}\right| = (-1)^{(l-r+1)}\left|\boldsymbol{A}_{\{l\}[r]}\right|, \quad 1 \le r, l \le p,\ l \ne r, \qquad (6.2.32)$$

where $\boldsymbol{A}_{\{r\}[l]}$ is formed out of $\boldsymbol{A}$ by replacing the l–th column by a column of ones and then striking out the r–th column.

PROOF. The matrices $\boldsymbol{A}_{\{r\}[l]}$ and $\boldsymbol{A}_{\{l\}[r]}$ consist of the same columns, but the order of the columns is different. Thus the determinants in (6.2.32) may only differ in sign. For $l > r$ we get $\boldsymbol{A}_{\{r\}[l]}$ out of $\boldsymbol{A}_{\{l\}[r]}$ by interchanging the r-th column (which is the column of ones) and the $(r+1)$-th column and then by interchanging the $(r+1)$-th column (which is now the column of ones) and the $(r+2)$-th column and so on, until the column of ones is in the $(l-1)$-th column of $\boldsymbol{A}_{\{l\}[r]}$. These successive interchanges of columns can be expressed by the following permutation

$$\sigma = \begin{pmatrix} l-2 & l-1 \end{pmatrix}\dots\begin{pmatrix} r+1 & r+2 \end{pmatrix}\begin{pmatrix} r & r+1 \end{pmatrix}.$$

With (A.2.3) we have

$$\operatorname{sign}(\sigma) = (-1)^{(l-1)-r} = (-1)^{l-r+1}.$$

It is well known that permuting the rows or columns of a matrix by σ, changes the sign of the determinant by $\operatorname{sign}(\sigma)$ (see also (A.2.5)). Thus

$$\left|\boldsymbol{A}_{\{r\}[l]}\right| = (-1)^{(l-r+1)}\left|\boldsymbol{A}_{\{l\}[r]}\right|, \quad 1 \le r < l \le p.$$

Interchanging the indices l and r for $l < r$ implies

$$\left|\boldsymbol{A}_{\{l\}[r]}\right| = (-1)^{(r-l+1)}\left|\boldsymbol{A}_{\{r\}[l]}\right|$$

and we get

$$\left|\boldsymbol{A}_{\{r\}[l]}\right| = (-1)^{(l-r+1)}\left|\boldsymbol{A}_{\{l\}[r]}\right|.$$

This completes the proof.

□

By making use of Lemma 6.2.2, we obtain the following Lemma.

Lemma 6.2.3. *Let $\tilde{\boldsymbol{M}}_x^{quad}$ be defined as in Lemma 6.1.7. Then*

$$\boldsymbol{\Psi}\tilde{\boldsymbol{M}}_x^{quad} = \boldsymbol{0}.$$

Proof. Consider the (u,v)-th element of the matrix $\boldsymbol{\Psi}\tilde{\boldsymbol{M}}_x^{quad} \in \mathbb{R}^{n\times p}$. It is independent of u and given by

$$\boldsymbol{\Psi}\tilde{\boldsymbol{M}}_x^{quad}(u,v) = \sum_{l=1}^{p}(-1)^{l+v}\psi_l\boldsymbol{\psi}_{\{l\}}{}^T\tilde{\boldsymbol{A}}_x^{(lv)}\boldsymbol{\psi}_{\{v\}},$$

with $\tilde{\boldsymbol{A}}_x^{(lv)} = \left[\left|\boldsymbol{X}_{\{l\}[r]}{}^T\boldsymbol{X}_{\{v\}[s]}\right|\right]_{\substack{1\le r,s\le p\\ l\ne r;v\ne s}}$ defined like in Lemma 6.1.7. It follows

$$\begin{aligned}\boldsymbol{\Psi}\tilde{\boldsymbol{M}}_x^{quad}(u,v) &= \sum_{l=1}^{p}\sum_{\substack{s=1\\ s\ne v}}^{p}\sum_{\substack{r=1\\ r\ne l}}^{p}(-1)^{l+v}\psi_r\psi_s\psi_l\left|\boldsymbol{X}_{\{l\}[r]}{}^T\boldsymbol{X}_{\{v\}[s]}\right|\\ &= \sum_{\substack{s=1\\ s\ne v}}^{p}\left(\sum_{l=2}^{p}\sum_{r=1}^{l-1}(-1)^{l+v}\psi_r\psi_l\left|\boldsymbol{X}_{\{l\}[r]}{}^T\boldsymbol{X}_{\{v\}[s]}\right|\right.\\ &\left.+\sum_{r=2}^{p}\sum_{l=1}^{r-1}(-1)^{l+v}\psi_r\psi_l\left|\boldsymbol{X}_{\{l\}[r]}{}^T\boldsymbol{X}_{\{v\}[s]}\right|\right)\psi_s\end{aligned}$$

$$= \sum_{\substack{s=1\\ s\neq v}}^{p} \left(\sum_{l=2}^{p} \sum_{r=1}^{l-1} (-1)^{l+v} \psi_r \psi_l \left| \boldsymbol{X}_{\{l\}[r]}{}^T \boldsymbol{X}_{\{v\}[s]} \right| \right.$$

$$\left. + \sum_{l=2}^{p} \sum_{r=1}^{l-1} (-1)^{r+v} \psi_r \psi_l \left| \boldsymbol{X}_{\{r\}[l]}{}^T \boldsymbol{X}_{\{v\}[s]} \right| \right) \psi_s.$$

Applying the Cauchy–Binet formula (A.2.6) and (6.2.32) in Lemma 6.2.2 implies for $\tilde{s} = \binom{n}{p-1}$

$$\begin{aligned} \left| \boldsymbol{X}_{\{r\}[l]}{}^T \boldsymbol{X}_{\{v\}[s]} \right| &= \sum_{k=1}^{\tilde{s}} \left| \boldsymbol{X}_{\{r\}[l]}{}^T \langle k \rangle \right| \left| \boldsymbol{X}_{\{v\}[s]}{}^T \langle k \rangle \right| \\ &= (-1)^{l-r+1} \sum_{k=1}^{\tilde{s}} \left| \boldsymbol{X}_{\{l\}[r]}{}^T \langle k \rangle \right| \left| \boldsymbol{X}_{\{v\}[s]}{}^T \langle k \rangle \right| \\ &= (-1)^{l-r+1} \left| \boldsymbol{X}_{\{l\}[r]}{}^T \boldsymbol{X}_{\{v\}[s]} \right|. \end{aligned}$$

Thus $\boldsymbol{\Psi} \tilde{\boldsymbol{M}}_x^{quad}$ is an $n \times p$ null matrix.

□

Lemma 6.2.3 is also valid for the special case of a standardized design matrix $\boldsymbol{Z}$ and thus we get for $\tilde{\boldsymbol{M}}^{quad}$ given in Corollary 6.1.9

$$\boldsymbol{\Psi} \tilde{\boldsymbol{M}}^{quad} = \boldsymbol{0} \in \mathbb{R}^{n \times p}$$

and (6.2.31) simplifies to

$$\begin{aligned} \Sigma(\tilde{\gamma}_\omega) = \sigma^2 & \frac{n\omega^2 \tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|} \\ & - \frac{\sigma^2 \omega^2}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^2} \tilde{\boldsymbol{M}}^{const} \boldsymbol{\Psi}^T \boldsymbol{\Psi} \tilde{\boldsymbol{M}}^{const}. \end{aligned} \tag{6.2.33}$$

We have

$$\begin{aligned} & \tilde{\boldsymbol{M}}^{const} \boldsymbol{\Psi}^T \boldsymbol{\Psi} \tilde{\boldsymbol{M}}^{const} \\ & \qquad = \left[n \sum_{s=1}^{p} \sum_{r=1}^{p} (-1)^{u+v+r+s} \psi_r \psi_s |\boldsymbol{Z}_{\{u\}}{}^T \boldsymbol{Z}_{\{r\}}| |\boldsymbol{Z}_{\{v\}}{}^T \boldsymbol{Z}_{\{s\}}| \right]_{1 \le u,v \le p} \end{aligned}$$

and the trace of $\tilde{\boldsymbol{M}}^{const} \boldsymbol{\Psi}^T \boldsymbol{\Psi} \tilde{\boldsymbol{M}}^{const}$ is given by

$$\operatorname{tr} \left(\tilde{\boldsymbol{M}}^{const} \boldsymbol{\Psi}^T \boldsymbol{\Psi} \tilde{\boldsymbol{M}}^{const} \right) = n \sum_{j=1}^{p} \left(\sum_{r=1}^{p} (-1)^r \psi_r |\boldsymbol{Z}_{\{j\}}{}^T \boldsymbol{Z}_{\{r\}}| \right)^2.$$

Thus we can deduce the following Corollary.

Corollary 6.2.4. *For fixed* $\boldsymbol{\Psi}$ *and arbitrary* ω *the elements of the covariance matrix are given by*

$$\begin{aligned}\operatorname{cov}(\tilde{\gamma}_u^\omega, \tilde{\gamma}_v^\omega) &= \sigma^2 \frac{n\omega^2 \tilde{m}_{u,v}^{quad} + \tilde{m}_{u,v}^{const}}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|} \\ &- \frac{\sigma^2 \omega^2 n \sum_{s=1}^p \sum_{r=1}^p (-1)^{u+v+r+s} \psi_r \psi_s |\boldsymbol{Z}_{\{u\}}{}^T \boldsymbol{Z}_{\{r\}}| |\boldsymbol{Z}_{\{v\}}{}^T \boldsymbol{Z}_{\{s\}}|}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^2}, \quad u, v = 1, \ldots, p,\end{aligned}$$

with $\tilde{m}_{u,v}^{quad}$ *and* $\tilde{m}_{u,v}^{const}$ *defined like in Corollary 6.1.9. Thus the total variance of* $\tilde{\boldsymbol{\gamma}}_\omega$ *can be written as*

$$\begin{aligned}\operatorname{tr}(\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)) &= \sum_{j=1}^p \operatorname{var}(\tilde{\gamma}_j^\omega) \\ &= \sigma^2 \sum_{j=1}^p \left(\frac{n\omega^2 \boldsymbol{\psi}_{\{j\}}{}^T \tilde{\boldsymbol{A}}^{(jj)} \boldsymbol{\psi}_{\{j\}} + |\boldsymbol{Z}_{\{j\}}{}^T \boldsymbol{Z}_{\{j\}}|}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|} \right. \\ &\qquad \left. - \frac{n\omega^2 \left(\sum_{r=1}^p (-1)^r \psi_r |\boldsymbol{Z}_{\{j\}}{}^T \boldsymbol{Z}_{\{r\}}| \right)^2}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^2} \right).\end{aligned}$$

Example 6.2.5. With the help of Example 6.1.10 and (6.2.33), the covariance matrix of the disturbed least squares estimator (6.2.29) of Example 6.2.1 is given

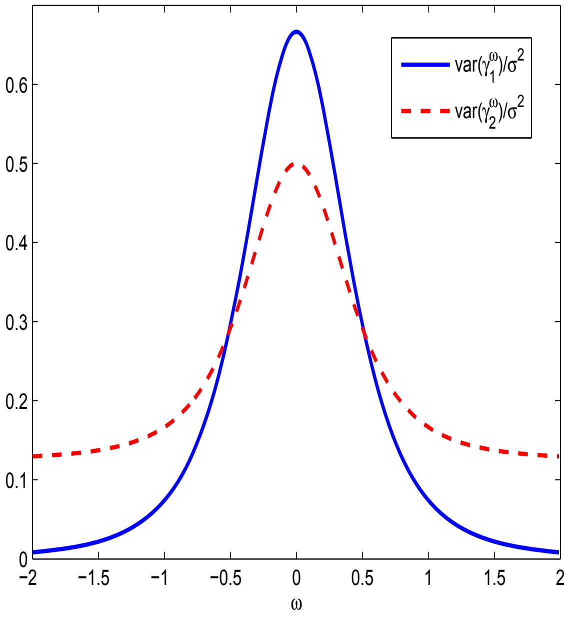


Figure 6.2.2. Variances of the components of $\tilde{\boldsymbol{\gamma}}_\omega$ in dependence of ω

by

$$\begin{aligned}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega) &= \sigma^2 \frac{n\omega^2 \tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|} \\ &\quad - \frac{\sigma^2\omega^2}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^2} \tilde{\boldsymbol{M}}^{const} \boldsymbol{\Psi}^T \boldsymbol{\Psi} \tilde{\boldsymbol{M}}^{const} \\ &= \frac{\sigma^2}{24\omega^2\psi_1^2 + 12} \begin{bmatrix} 8 & -6 \\ -6 & 6 + 3\omega^2\psi_1^2 \end{bmatrix} - \frac{\sigma^2\omega^2\psi_1^2}{(24\omega^2\psi_1^2 + 12)^2} \begin{bmatrix} 192 & -144 \\ -144 & 108 \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} \frac{\frac{2}{3}}{(2\omega^2\psi_1^2+1)^2} & \frac{-\frac{1}{2}}{(2\omega^2\psi_1^2+1)^2} \\ \frac{-\frac{1}{2}}{(2\omega^2\psi_1^2+1)^2} & \frac{\frac{1}{2}(\omega^4\psi_1^4+\omega^2\psi_1^2+1)}{(2\omega^2\psi_1^2+1)^2} \end{bmatrix}.\end{aligned}$$

The numerator of $\text{var}(\tilde{\gamma}_1^\omega)$ does not depend on ω and ψ_1 and the denominator is an open upward parabola in ω for fixed ψ_1. For $\tau := \omega\psi$ the first derivative of $\text{var}(\tilde{\gamma}_1^\omega)$ is given by

$$\frac{\partial \text{var}(\tilde{\gamma}_1^\omega)}{\partial \tau} = \frac{-16\tau}{3(2\tau^2 + 1)^3}$$

and thus $\text{var}(\tilde{\gamma}_1^\omega)$ has got a maximum in $\tau = 0$. Thus we can find a $\tau \neq 0$, such that the variance of $\tilde{\gamma}_1^\omega$ is smaller than the corresponding one of the least squares estimate $\hat{\gamma}_1$. Figure 6.2.2 displays the variances of $\tilde{\gamma}_1^\omega/\sigma^2$ and $\tilde{\gamma}_2^\omega/\sigma^2$ in dependence of ω for $\psi_1 = 1$.

•

6.2.2 The Bias of the Disturbed Least Squares Estimator From (6.2.27) the bias of $\tilde{\boldsymbol{\gamma}}_\omega$ is given by

$$\begin{aligned}\text{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) &= \text{E}(\tilde{\boldsymbol{\gamma}}_\omega) - \boldsymbol{\gamma} \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^{-1} \boldsymbol{Z}^T \text{E}(\boldsymbol{y}^*) - \boldsymbol{\gamma} \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^{-1} \boldsymbol{Z}^T \left(\boldsymbol{Z}\boldsymbol{\gamma} + \text{E}(\boldsymbol{\varepsilon}^*)\right) - \boldsymbol{\gamma} \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^{-1} \boldsymbol{Z}^T \left(\boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{P}\text{E}(\boldsymbol{\varepsilon})\right) - \boldsymbol{\gamma} \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^{-1} \boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\gamma} - \boldsymbol{\gamma} \\ &= \left(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^{-1} \left(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi} - \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)\boldsymbol{\gamma} - \boldsymbol{\gamma} \\ &= -\omega^2(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma} \\ &= -\frac{\omega^2 \tilde{\boldsymbol{M}} \boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|} \\ &= -\frac{\omega^2 \tilde{\boldsymbol{M}}^{const} \boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|} \in \mathbb{R}^{p\times 1}, \qquad (6.2.34)\end{aligned}$$

because of Lemma 6.2.3. Thus we can deduce the following corollary.

Corollary 6.2.6. *For fixed $\boldsymbol{\Psi}$ and arbitrary ω the squared bias of $\tilde{\boldsymbol{\gamma}}_\omega$ is expressible as*

$$\mathrm{Bias}^T(\tilde{\boldsymbol{\gamma}}_\omega)\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) = \frac{\omega^4\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\tilde{\boldsymbol{M}}^{const})^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^2} \in \mathbb{R}.$$

Example 6.2.7. Continuing Example 6.2.1 the bias of (6.2.29) is given by

$$\begin{aligned}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) &= -\frac{\omega^2\tilde{\boldsymbol{M}}^{const}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|} \\ &= -\frac{\omega^2\begin{bmatrix}8 & -6\\ -6 & 6\end{bmatrix}\begin{bmatrix}\psi_1 & \psi_1 & \psi_1\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\psi_1 & 0\\ \psi_1 & 0\\ \psi_1 & 0\end{bmatrix}}{24\omega^2\psi_1^2 + 12}\boldsymbol{\gamma}.\end{aligned}$$

With $\boldsymbol{\gamma} = \begin{bmatrix}1 & 1\end{bmatrix}^T$ we have

$$\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) = \begin{bmatrix}\frac{-\omega^2}{2\omega^2\psi_1^2+1}\\ \frac{-\omega^2(2+7\omega^2\psi_1^2)}{2(2\omega^2\psi_1^2+1)}.\end{bmatrix}$$

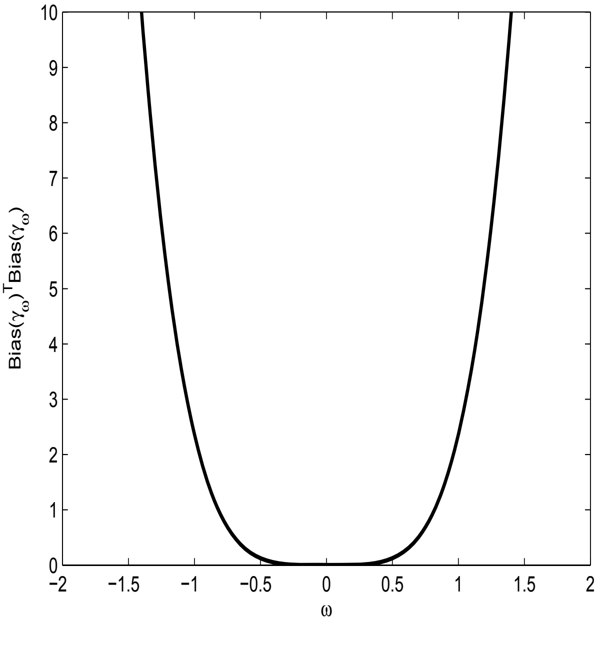


Figure 6.2.3. Squared Bias in dependence of ω

Thus the squared bias of $\tilde{\gamma}_\omega$ is given by

$$\mathrm{Bias}^T(\tilde{\gamma}_\omega)\mathrm{Bias}(\tilde{\gamma}_\omega) = \frac{\frac{1}{4}\omega^4(8 + 28\omega 2\psi_1^2 + 49\omega^4\psi_1^4)}{(2\omega^2\psi_1^2 + 1)^2}.$$

Figure 6.2.3 displays the squared bias in dependence of ω for $\psi_1 = 1$. Because of the unbiasedness of the least squares estimator there is a double root in $\omega = 0$.

•

6.3. Mean Squared Error Properties of the DLSE

The mean squared error for the standardized model can be written as

$$\mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega) = \sum_{j=1}^{p} \mathrm{var}(\tilde{\gamma}_j^\omega) + \mathrm{Bias}^T(\tilde{\boldsymbol{\gamma}}_\omega)\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega), \tag{6.3.35}$$

with

$$\sum_{j=1}^{p} \mathrm{var}(\tilde{\gamma}_j^\omega) = \sigma^2 \sum_{j=1}^{p} \left(\frac{n\omega^2 \boldsymbol{\psi}_{\{j\}}{}^T \tilde{\boldsymbol{A}}^{(jj)} \boldsymbol{\psi}_{\{j\}} + |\boldsymbol{Z}_{\{j\}}{}^T \boldsymbol{Z}_{\{j\}}|}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|} \right.$$
$$\left. - \frac{n\omega^2 \left(\sum_{r=1}^{p} (-1)^r \psi_r |\boldsymbol{Z}_{\{j\}}{}^T \boldsymbol{Z}_{\{r\}}| \right)^2}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^2} \right)$$

given in Corollary 6.2.4 and

$$\mathrm{Bias}^T(\tilde{\boldsymbol{\gamma}}_\omega)\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) = \frac{\omega^4 \boldsymbol{\gamma}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} (\tilde{\boldsymbol{M}}^{const})^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi} \boldsymbol{\gamma}}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^2},$$

given in Corollary 6.2.6. Obviously $\mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega)$ is symmetric in ω. To sketch the curve $\mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega)$ with respect to ω, we consider the first derivative of $\mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega)$ with respect to omega

$$\frac{\partial}{\partial\omega}\mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega) = \sum_{j=1}^{p} \frac{\partial}{\partial\omega} \mathrm{var}\left(\tilde{\gamma}_j^\omega\right) + \frac{\partial}{\partial\omega}\left(\mathrm{Bias}^T(\tilde{\boldsymbol{\gamma}}_\omega)\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)\right), \tag{6.3.36}$$

where

$$\begin{aligned}
\frac{\partial}{\partial\omega}\left(\mathrm{Bias}^T(\tilde{\boldsymbol{\gamma}}_\omega)\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)\right) &= \frac{\partial}{\partial\omega}\left(\frac{\omega^4 \boldsymbol{\gamma}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} (\tilde{\boldsymbol{M}}^{const})^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi} \boldsymbol{\gamma}}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^2} \right) \\
&= \frac{4\omega^3 \boldsymbol{\gamma}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} (\tilde{\boldsymbol{M}}^{const})^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi} \boldsymbol{\gamma} (n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^3} \\
&\quad - \frac{4n\omega^5 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} \boldsymbol{\gamma}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} (\tilde{\boldsymbol{M}}^{const})^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi} \boldsymbol{\gamma}}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^3} \\
&= \frac{4\omega^3 |\boldsymbol{Z}^T \boldsymbol{Z}| \boldsymbol{\gamma}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} (\tilde{\boldsymbol{M}}^{const})^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi} \boldsymbol{\gamma}}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^3},
\end{aligned} \tag{6.3.37}$$

and

$$\begin{aligned}
\frac{\partial}{\partial\omega}\sum_{j=1}^{p}\operatorname{var}(\tilde{\gamma}_j^{\omega}) &= \sigma^2\sum_{j=1}^{p}\frac{\partial}{\partial\omega}\left(\frac{n\omega^2\boldsymbol{\psi}_{\{j\}}{}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}+|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{j\}}|}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|}\right.\\
&\qquad\left.-\frac{n\omega^2\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^2}\right)\\
&= \sigma^2\sum_{j=1}^{p}\left(\frac{2n\omega\boldsymbol{\psi}_{\{j\}}{}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^2}\right.\\
&\qquad-\frac{2n\omega\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}\left(n\omega^2\boldsymbol{\psi}_{\{j\}}{}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}+|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{j\}}|\right)}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^2}\\
&\qquad-\frac{2n\omega\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}\\
&\qquad\left.+\frac{4n^2\omega^3\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}\right)\\
&= \sigma^2\sum_{j=1}^{p}\left(\frac{2n\omega|\boldsymbol{Z}^T\boldsymbol{Z}|\boldsymbol{\psi}_{\{j\}}{}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}-2n\omega\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{j\}}|}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^2}\right.\\
&\qquad\left.-\frac{\left(2n\omega|\boldsymbol{Z}^T\boldsymbol{Z}|-2n^2\omega^3\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}\right)\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}\right)\\
&= 2n\omega\sigma^2\sum_{j=1}^{p}\left(\frac{\boldsymbol{\psi}_{\{j\}}{}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}|\boldsymbol{Z}^T\boldsymbol{Z}|-\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{j\}}|}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^2}\right.\\
&\qquad\left.-\frac{\left(|\boldsymbol{Z}^T\boldsymbol{Z}|-n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}\right)\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}\right)\\
&= 2n\sigma^2\sum_{j=1}^{p}\frac{n\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}s_1(j)\omega^3+|\boldsymbol{Z}^T\boldsymbol{Z}|s_2(j)\omega}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3},
\end{aligned}\tag{6.3.38}$$

with

$$\begin{aligned}
s_1(j) &:= \boldsymbol{\psi}_{\{j\}}{}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}|\boldsymbol{Z}^T\boldsymbol{Z}|-\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{j\}}|\\
&\qquad+\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2,\\
s_2(j) &:= \boldsymbol{\psi}_{\{j\}}{}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}|\boldsymbol{Z}^T\boldsymbol{Z}|-\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{j\}}|\\
&\qquad-\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2.
\end{aligned}\tag{6.3.39}$$

Thus (6.3.36) can be written as

$$\frac{\partial \mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega)}{\partial \omega} = 2n\sigma^2 \sum_{j=1}^{p} \frac{n\boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} s_1(j)\omega^3 + |\boldsymbol{Z}^T\boldsymbol{Z}| s_2(j)\omega}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^3} + \frac{4\omega^3 |\boldsymbol{Z}^T\boldsymbol{Z}| \boldsymbol{\gamma}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} (\tilde{\boldsymbol{M}}^{const})^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi} \boldsymbol{\gamma}}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^3}. \tag{6.3.40}$$

We have

$$|\boldsymbol{Z}^T\boldsymbol{Z}| > 0$$

and from (6.1.24) we concluded that $\tilde{\boldsymbol{M}}^{const}$ is positive definite and thus

$$n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} > 0. \tag{6.3.41}$$

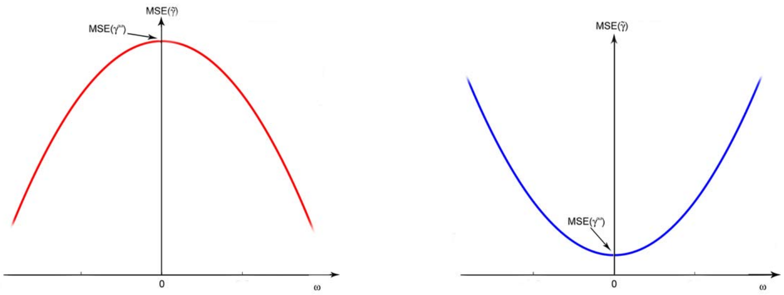
As a consequence the denominator in (6.3.40) is always bigger than zero. With (6.3.40) it is easy to see, that

$$\left.\frac{\partial \mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega)}{\partial \omega}\right|_{\omega=0} = 0,$$

i.e. the mean squared error of $\tilde{\boldsymbol{\gamma}}_\omega$ probably has got an extremum in $\omega = 0$ in case of a standardized design matrix. Let $[-\varepsilon, \varepsilon]$ be an ε-neighborhood of ω about zero. If there is a maximum in $\omega = 0$ we can choose ε small enough, such that the mean squared error of the disturbed estimator $\tilde{\boldsymbol{\gamma}}_\omega$ is smaller than the corresponding one of the least squares estimator for $\omega \in [-\varepsilon, 0)$ and $\omega \in (0, \varepsilon]$, i.e.

$$\forall\ \Psi \neq \mathbf{0} \quad \exists\ \varepsilon > 0 \quad \forall\ \omega \in [-\varepsilon, \varepsilon]\backslash\{0\} : \mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega) < \mathrm{MSE}(\hat{\boldsymbol{\gamma}}). \tag{6.3.42}$$

Otherwise if there is a minimum in $\omega = 0$, we will not find any ε, such that (6.3.42) is fulfilled. Of course it will be desirable to have a maximum in $\omega = 0$. With the help of the following Lemma it will be possible to simplify (6.3.40) and thus to determine the kind of the extremum.



(a) Maximum in $\omega = 0$

(b) Minimum in $\omega = 0$

FIGURE 6.3.4. Extremum in $\omega = 0$

Lemma 6.3.1. *For any positive definite matrix $\boldsymbol{A} = [a_{r,s}]_{1 \leq r,s \leq p}$,*

$$|\boldsymbol{A}_{\{i,j\}}||\boldsymbol{A}_{\{k,l\}}| - |\boldsymbol{A}_{\{i,l\}}||\boldsymbol{A}_{\{j,k\}}| = \begin{cases} (-1)^{i+j}|\boldsymbol{A}| & p = 2 \\ |\boldsymbol{A}_{\{ik,jl\}}||\boldsymbol{A}| & p \geq 3 \end{cases}, \quad i \neq k,\ j \neq l,$$

where $\boldsymbol{A}_{\{q,v\}} \in \mathbb{R}^{(p-1)\times(p-1)}$, $q, v = i, j, k, l$ is formed out of $\boldsymbol{A}$ by striking out the q-th row and v-th column. $\boldsymbol{A}_{\{ik,jl\}} \in \mathbb{R}^{(p-2)\times(p-2)}$ means that both, the i-th and k-th row and the j-th and l-th column are missing.

Proof. For $p = 2$ we have $i, j, k, l = 1, 2$ with $i \neq k$ and $j \neq l$. As a consequence we can only choose $i = j$ and $k = l$ or $i \neq j$ and $k \neq l$. Thus only the following four cases are possible

$$\begin{aligned} &(1)\ i = 1,\ j = 1,\ k = 2,\ l = 2 \\ &(2)\ i = 2,\ j = 1,\ k = 1,\ l = 2 \\ &(3)\ i = 1,\ j = 2,\ k = 2,\ l = 1 \\ &(4)\ i = 2,\ j = 2,\ k = 1,\ l = 1. \end{aligned}$$

For (1) and (4) we get

$$|\boldsymbol{A}_{\{1,1\}}||\boldsymbol{A}_{\{2,2\}}| - |\boldsymbol{A}_{\{1,2\}}|^2 = a_{1,1}a_{2,2} - a_{1,2}^2 = |\boldsymbol{A}| = (-1)^{i+j}|\boldsymbol{A}|$$

and for (2) and (3)

$$|\boldsymbol{A}_{\{1,2\}}|^2 - |\boldsymbol{A}_{\{1,1\}}||\boldsymbol{A}_{\{2,2\}}| = a_{1,2}^2 - a_{1,1}a_{2,2} = -|\boldsymbol{A}| = (-1)^{i+j}|\boldsymbol{A}|.$$

Let $p \geq 3$. Moving the i-th row of $\boldsymbol{A}$ into the first row changes the sign of the determinant to $(-1)^{i-1}$, because $(i-1)$ interchanges of the rows are necessary. For moving the i-th column of $\boldsymbol{A}$ into the first column of $\boldsymbol{A}$ we additionally need $(i-1)$ interchanges. Thus in total the sign of the determinant does not change, because there are as much interchanges necessary for the rows as are for the columns. Moving analogously the j–th, k–th and l–th row of $\boldsymbol{A}$ into the second, third and fourth row and respectively the j–th, k–th and l–th column into the second, third and fourth column results in

$$|\boldsymbol{A}| = \left| \left(\begin{array}{cccc|c} a_{i,i} & a_{i,j} & a_{i,k} & a_{i,l} & - \ a_i^* \ - \\ a_{i,j} & a_{j,j} & a_{j,k} & a_{j,l} & - \ a_j^* \ - \\ a_{i,k} & a_{j,k} & a_{k,k} & a_{k,l} & - \ a_k^* \ - \\ a_{i,l} & a_{j,l} & a_{k,l} & a_{l,l} & - \ a_l^* \ - \\ \hline | & | & | & | & \\ a_i^{*T} & a_j^{*T} & a_k^{*T} & a_l^{*T} & \boldsymbol{A}_{\{ijkl,ijkl\}} \\ | & | & | & | & \end{array} \right) \right|,$$

where

$$
\begin{aligned}
a_q^* = \Big[&a_{q,1}, \quad \ldots \quad , a_{q,i-1}, \quad a_{q,i+1}, \quad \ldots \quad , a_{q,j-1} \quad , a_{q,j+1} \quad , \ldots \\
&\ldots \quad , a_{q,k-1}, \quad a_{q,k+1}, \quad \ldots \quad , a_{q,l-1} \quad , a_{q,l+1}, \quad \ldots \quad , a_{q,p}\Big] \\
&\in \mathbb{R}^{1\times(p-4)}, \;\; q = i, j, k, l.
\end{aligned}
$$

$\boldsymbol{A}_{\{ijkl,ijkl\}}$ is formed out of $\boldsymbol{A}$ by striking out the i-th, j-th, k-th and l-th row and also the i-th, j-th, k-th and l-th column. To show that $\boldsymbol{A}_{\{ijkl,ijkl\}}$ is also positive definite, consider any vector $\boldsymbol{p} \in \mathbb{R}^{p\times 1}$ which i–th, j–th, k–th and l-th components are zero, i.e.

$$
\begin{aligned}
\boldsymbol{p}^T := \Big[&p_1, \quad \ldots \quad , p_{i-1}, \quad 0, \quad p_{i+1}, \quad \ldots \quad , p_{j-1}, \quad 0, \quad p_{j+1}, \quad \ldots \\
&\ldots \quad , p_{k-1}, \quad 0, \quad p_{k+1}, \quad \ldots \quad , p_{l-1}, \quad 0, \quad p_{l+1}, \quad \ldots \quad , p_p\Big] \quad i \neq k, \; j \neq l.
\end{aligned}
$$

Then we have

$$
\boldsymbol{p}^T \boldsymbol{A} \boldsymbol{p} = \boldsymbol{p}_{\{ijkl\}}{}^T \boldsymbol{A}_{\{ijkl,ijkl\}} \boldsymbol{p}_{\{ijkl\}},
$$

with

$$
\begin{aligned}
\boldsymbol{p}_{\{ijkl\}}{}^T := \Big[&p_1, \quad \ldots \quad , p_{i-1}, \quad p_{i+1}, \quad \ldots \quad , p_{j-1}, \quad p_{j+1}, \quad \ldots \\
&\ldots \quad , p_{k-1}, \quad p_{k+1}, \quad \ldots \quad , p_{l-1}, \quad p_{l+1}, \quad \ldots \quad , p_p\Big] \in \mathbb{R}^{1\times(p-4)}.
\end{aligned}
$$

Because $\boldsymbol{A}$ is positive definite by assumption, we have $\boldsymbol{p}^T\boldsymbol{A}\boldsymbol{p} > 0$ for all non–zero $\boldsymbol{p}$ and thus in particular $\boldsymbol{p}_{\{ijkl\}}{}^T \boldsymbol{A}_{\{ijkl,ijkl\}} \boldsymbol{p}_{\{ijkl\}} > 0$ for all non–zero $\boldsymbol{p}_{\{ijkl\}}$. Hence $\boldsymbol{A}_{\{ijkl,ijkl\}}$ is also positive definite and invertible. Let

$$
\boldsymbol{W} = [w_{r,s}]_{1\leq r,s\leq p-4} := (\boldsymbol{A}_{\{ijkl,ijkl\}})^{-1}.
$$

With Theorem A.5.1 it follows

$$
\begin{aligned}
|\boldsymbol{A}| &= \left|\boldsymbol{A}_{\{ijkl,ijkl\}}\right| \left| \begin{bmatrix} a_{i,i} & a_{i,j} & a_{i,k} & a_{i,l} \\ a_{i,j} & a_{j,j} & a_{j,k} & a_{j,l} \\ a_{i,k} & a_{j,k} & a_{k,k} & a_{k,l} \\ a_{i,l} & a_{j,l} & a_{k,l} & a_{l,l} \end{bmatrix} - \begin{bmatrix} - \, a_i^* - \\ - \, a_j^* - \\ - \, a_k^* - \\ - \, a_l^* - \end{bmatrix} \boldsymbol{W} \begin{bmatrix} | & | & | & | \\ a_i^{*T} & a_j^{*T} & a_k^{*T} & a_l^{*T} \\ | & | & | & | \end{bmatrix} \right| \\
&=: |\boldsymbol{W}|^{-1} \left| \begin{bmatrix} a_{i,i} & a_{i,j} & a_{i,k} & a_{i,l} \\ a_{i,j} & a_{j,j} & a_{j,k} & a_{j,l} \\ a_{i,k} & a_{j,k} & a_{k,k} & a_{k,l} \\ a_{i,l} & a_{j,l} & a_{k,l} & a_{l,l} \end{bmatrix} - \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\ m_{1,2} & m_{2,2} & m_{2,3} & m_{2,4} \\ m_{1,3} & m_{2,3} & m_{3,3} & m_{3,4} \\ m_{1,4} & m_{2,4} & m_{3,4} & m_{4,4} \end{bmatrix} \right|. \qquad (6.3.43)
\end{aligned}
$$

With an analogous argumentation we can write

$$
|\boldsymbol{A}_{\{i,j\}}| = (-1)^{i+j} |\boldsymbol{W}|^{-1} \left| \begin{bmatrix} a_{i,j} & a_{j,k} & a_{j,l} \\ a_{i,k} & a_{k,k} & a_{k,l} \\ a_{i,l} & a_{k,l} & a_{l,l} \end{bmatrix} - \begin{bmatrix} - \, a_j^* - \\ - \, a_k^* - \\ - \, a_l^* - \end{bmatrix} \boldsymbol{W} \begin{bmatrix} | & | & | \\ a_i^{*T} & a_k^{*T} & a_l^{*T} \\ | & | & | \end{bmatrix} \right|
$$

$$= (-1)^{i+j}|\boldsymbol{W}|^{-1}\left|\begin{bmatrix} a_{i,j} & a_{j,k} & a_{j,l}\\ a_{i,k} & a_{k,k} & a_{k,l}\\ a_{i,l} & a_{k,l} & a_{l,l}\end{bmatrix} - \begin{bmatrix} m_{1,2} & m_{2,3} & m_{2,4}\\ m_{1,3} & m_{3,3} & m_{3,4}\\ m_{1,4} & m_{3,4} & m_{4,4}\end{bmatrix}\right|,$$

but here we have to consider the sign of the determinant: Moving the j-th row of $\boldsymbol{A}_{\{i,j\}}$ into the first row changes the sign of the determinant to $(-1)^{j-1}$, because $(j-1)$ interchanges of the rows are necessary. For moving the i-th column of $\boldsymbol{A}_{\{i,j\}}$ into the first column of $\boldsymbol{A}_{\{i,j\}}$we additionally need $(i-1)$ interchanges. Moving the k-th and l-th row of $\boldsymbol{A}_{\{i,j\}}$ into the second and third row and respectively the k-th and l-th column into the second and third column does not change the sign. Thus in total the sign of the determinant is $(-1)^{i+j-2} = (-1)^{i+j}$. Hence we get

$$\begin{aligned}
|\boldsymbol{A}_{\{i,j\}}||\boldsymbol{A}_{\{k,l\}}| - |\boldsymbol{A}_{\{i,l\}}||\boldsymbol{A}_{\{j,k\}}| &= (-1)^{i+j+k+l}|\boldsymbol{W}|^{-2}\\
&\cdot\Bigg(\left|\begin{bmatrix} a_{i,j} & a_{j,k} & a_{j,l}\\ a_{i,k} & a_{k,k} & a_{k,l}\\ a_{i,l} & a_{k,l} & a_{l,l}\end{bmatrix} - \begin{bmatrix} m_{1,2} & m_{2,3} & m_{2,4}\\ m_{1,3} & m_{3,3} & m_{3,4}\\ m_{1,4} & m_{3,4} & m_{4,4}\end{bmatrix}\right|\\
&\cdot\left|\begin{bmatrix} a_{i,i} & a_{i,j} & a_{i,k}\\ a_{i,j} & a_{j,j} & a_{j,k}\\ a_{i,l} & a_{j,l} & a_{k,l}\end{bmatrix} - \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3}\\ m_{1,2} & m_{2,2} & m_{2,3}\\ m_{1,4} & m_{2,4} & m_{3,4}\end{bmatrix}\right|\\
&-\left|\begin{bmatrix} a_{i,j} & a_{j,j} & a_{j,k}\\ a_{i,k} & a_{j,k} & a_{k,k}\\ a_{i,l} & a_{j,l} & a_{k,l}\end{bmatrix} - \begin{bmatrix} m_{1,2} & m_{2,2} & m_{2,3}\\ m_{1,3} & m_{2,3} & m_{3,3}\\ m_{1,4} & m_{2,4} & m_{3,4}\end{bmatrix}\right|\\
&\cdot\left|\begin{bmatrix} a_{i,i} & a_{i,j} & a_{i,l}\\ a_{i,k} & a_{j,k} & a_{k,l}\\ a_{i,l} & a_{j,l} & a_{l,l}\end{bmatrix} - \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,4}\\ m_{1,3} & m_{2,3} & m_{3,4}\\ m_{1,4} & m_{2,4} & m_{4,4}\end{bmatrix}\right|\Bigg). \qquad (6.3.44)
\end{aligned}$$

Expanding the product of the determinants of the 3×3 matrices in (6.3.44) and rearranging the remaining terms results in

$$\begin{aligned}
&|\boldsymbol{A}_{\{i,j\}}||\boldsymbol{A}_{\{k,l\}}| - |\boldsymbol{A}_{\{i,l\}}||\boldsymbol{A}_{\{j,k\}}|\\
&\quad= (-1)^{i+j+k+l}|\boldsymbol{W}|^{-2}\left|\begin{bmatrix} a_{i,j} & a_{j,k}\\ a_{i,l} & a_{k,l}\end{bmatrix} - \begin{bmatrix} m_{1,2} & m_{2,3}\\ m_{1,4} & m_{3,4}\end{bmatrix}\right|\\
&\times\left|\begin{bmatrix} a_{i,i} & a_{i,j} & a_{i,k} & a_{i,l}\\ a_{i,j} & a_{j,j} & a_{j,k} & a_{j,l}\\ a_{i,k} & a_{j,k} & a_{k,k} & a_{k,l}\\ a_{i,l} & a_{j,l} & a_{k,l} & a_{l,l}\end{bmatrix} - \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4}\\ m_{1,2} & m_{2,2} & m_{2,3} & m_{2,4}\\ m_{1,3} & m_{2,3} & m_{3,3} & m_{3,4}\\ m_{1,4} & m_{2,4} & m_{3,4} & m_{4,4}\end{bmatrix}\right|\\
&\qquad= |\boldsymbol{A}_{\{ik,jl\}}||\boldsymbol{A}|,
\end{aligned}$$

because

$$
\begin{aligned}
|\boldsymbol{A}_{\{ik,jl\}}| &= (-1)^{i+j+k+l} \left| \left(\begin{array}{cc|c} a_{j,i} & a_{j,k} & - \, a_j^* \, - \\ a_{l,i} & a_{l,k} & - \, a_l^* \, - \\ \hline | & | & \\ a_i^{*T} & a_k^{*T} & \boldsymbol{A}_{\{ijkl,ijkl\}} \\ | & | & \end{array} \right) \right| \\
&= (-1)^{i+j+k+l} |\boldsymbol{W}|^{-1} \left| \begin{bmatrix} a_{i,j} & a_{j,k} \\ a_{i,l} & a_{k,l} \end{bmatrix} - \begin{bmatrix} - \, a_j^* - \\ - \, a_l^* - \end{bmatrix} \boldsymbol{W} \begin{bmatrix} | & | \\ a_i^{*T} & a_k^{*T} \\ | & | \end{bmatrix} \right| \\
&= (-1)^{i+j+k+l} |\boldsymbol{W}|^{-1} \left| \begin{bmatrix} a_{i,j} & a_{j,k} \\ a_{i,l} & a_{k,l} \end{bmatrix} - \begin{bmatrix} m_{1,2} & m_{2,3} \\ m_{1,4} & m_{3,4} \end{bmatrix} \right| .
\end{aligned}
$$

(Actually these calculations can be done with the help of the symbolic toolbox in `MATLAB`). Thus everything is proved.

□

To simplify the expressions for $s_1(j)$ and $s_2(j)$ in (6.3.39) we consider

$$
s^* := \sum_{j=1}^{p} \left({\boldsymbol{\psi}_{\{j\}}}^T \tilde{\boldsymbol{A}}^{(jj)} \boldsymbol{\psi}_{\{j\}} |\boldsymbol{Z}^T \boldsymbol{Z}| - \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{j\}}| \right) .
$$

From Corollary 6.1.9 we have for $p = 2$

$$
\begin{aligned}
s^* = \psi_2^2 |\boldsymbol{Z}^T \boldsymbol{Z}| - |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{1\}}| \sum_{s=1}^{2} \sum_{r=1}^{2} (-1)^{r+s} \psi_r \psi_s |{\boldsymbol{Z}_{\{r\}}}^T \boldsymbol{Z}_{\{s\}}| + \psi_1^2 |\boldsymbol{Z}^T \boldsymbol{Z}| \\
- |{\boldsymbol{Z}_{\{2\}}}^T \boldsymbol{Z}_{\{2\}}| \sum_{s=1}^{2} \sum_{r=1}^{2} (-1)^{r+s} \psi_r \psi_s |{\boldsymbol{Z}_{\{r\}}}^T \boldsymbol{Z}_{\{s\}}|.
\end{aligned}
$$

With

$$
\begin{aligned}
\sum_{s=1}^{2} \sum_{r=1}^{2} (-1)^{r+s} \psi_r \psi_s |{\boldsymbol{Z}_{\{r\}}}^T \boldsymbol{Z}_{\{s\}}| = \psi_1^2 |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{1\}}| - 2\psi_1 \psi_2 |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{2\}}| \\
+ \psi_2^2 |{\boldsymbol{Z}_{\{2\}}}^T \boldsymbol{Z}_{\{2\}}|
\end{aligned}
$$

it follows

$$
\begin{aligned}
s^* &= \psi_2^2 \left(|\boldsymbol{Z}^T \boldsymbol{Z}| - |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{1\}}| |{\boldsymbol{Z}_{\{2\}}}^T \boldsymbol{Z}_{\{2\}}| \right) \\
&\quad + \psi_1^2 \left(|\boldsymbol{Z}^T \boldsymbol{Z}| - |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{1\}}| |{\boldsymbol{Z}_{\{2\}}}^T \boldsymbol{Z}_{\{2\}}| \right) + 2\psi_1 \psi_2 |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{1\}}| |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{2\}}| \\
&\qquad + 2\psi_1 \psi_2 |{\boldsymbol{Z}_{\{2\}}}^T \boldsymbol{Z}_{\{2\}}| |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{2\}}| - \psi_1^2 |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{1\}}|^2 - \psi_2^2 |{\boldsymbol{Z}_{\{2\}}}^T \boldsymbol{Z}_{\{2\}}|^2.
\end{aligned}
$$

Application of Lemma 6.3.1 for p=2 implies

$$
|\boldsymbol{Z}^T \boldsymbol{Z}| - |{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{1\}}| |{\boldsymbol{Z}_{\{2\}}}^T \boldsymbol{Z}_{\{2\}}| = -|{\boldsymbol{Z}_{\{1\}}}^T \boldsymbol{Z}_{\{2\}}|^2
$$

and thus

$$\begin{aligned} s^* = &-\left(\psi_1^2|{\boldsymbol{Z}_{\{1\}}}^T\boldsymbol{Z}_{\{1\}}|^2 - 2\psi_1\psi_2|{\boldsymbol{Z}_{\{1\}}}^T\boldsymbol{Z}_{\{1\}}||{\boldsymbol{Z}_{\{1\}}}^T\boldsymbol{Z}_{\{2\}}| + \psi_2^2|{\boldsymbol{Z}_{\{1\}}}^T\boldsymbol{Z}_{\{2\}}|^2\right) \\ &-\left(\psi_1^2|{\boldsymbol{Z}_{\{1\}}}^T\boldsymbol{Z}_{\{2\}}|^2 - 2\psi_1\psi_2|{\boldsymbol{Z}_{\{2\}}}^T\boldsymbol{Z}_{\{2\}}||{\boldsymbol{Z}_{\{1\}}}^T\boldsymbol{Z}_{\{2\}}| + \psi_2^2|{\boldsymbol{Z}_{\{2\}}}^T\boldsymbol{Z}_{\{2\}}|^2\right) \\ &= -\sum_{j=1}^{2}\left(\sum_{r=1}^{2}(-1)^r\psi_r|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{r\}}|\right)^2 . \end{aligned}$$

For $p = 3$, s^* can be written as

$$\begin{aligned} s^* &= \sum_{j=1}^{p}\left(|\boldsymbol{Z}^T\boldsymbol{Z}|\sum_{\substack{s=1\\ s\neq j}}^{p}\sum_{\substack{r=1\\ r\neq j}}^{p}(-1)^{r+s}\psi_r\psi_s|{\boldsymbol{Z}_{\{jr\}}}^T\boldsymbol{Z}_{\{js\}}|\right. \\ &\qquad \left. -|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{j\}}|\sum_{s=1}^{p}\sum_{r=1}^{p}(-1)^{r+s}\psi_r\psi_s|{\boldsymbol{Z}_{\{r\}}}^T\boldsymbol{Z}_{\{s\}}|\right) \\ &= \sum_{j=1}^{p}\left(|\boldsymbol{Z}^T\boldsymbol{Z}|\sum_{\substack{s=1\\ s\neq j}}^{p}\sum_{\substack{r=1\\ r\neq j}}^{p}(-1)^{r+s}\psi_r\psi_s|{\boldsymbol{Z}_{\{jr\}}}^T\boldsymbol{Z}_{\{js\}}| - |{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{j\}}|\right. \\ &\qquad \left(\psi_j^2|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{j\}}| + 2\sum_{\substack{r=1\\ r\neq j}}^{p}(-1)^{r+j}\psi_r\psi_j|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{r\}}|\right. \\ &\qquad \left.\left. +\sum_{\substack{s=1\\ s\neq j}}^{p}\sum_{\substack{r=1\\ r\neq j}}^{p}(-1)^{r+s}\psi_r\psi_s|{\boldsymbol{Z}_{\{r\}}}^T\boldsymbol{Z}_{\{s\}}|\right)\right) \\ &= \sum_{j=1}^{p}\left(-\psi_j^2|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{j\}}|^2 - 2|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{j\}}|\sum_{\substack{r=1\\ r\neq j}}^{p}(-1)^{r+j}\psi_r\psi_j|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{r\}}|\right. \\ &\left. +\sum_{\substack{s=1\\ s\neq j}}^{p}\sum_{\substack{r=1\\ r\neq j}}^{p}(-1)^{r+s}\psi_r\psi_s\left(|\boldsymbol{Z}^T\boldsymbol{Z}||{\boldsymbol{Z}_{\{jr\}}}^T\boldsymbol{Z}_{\{js\}}| - |{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{j\}}||{\boldsymbol{Z}_{\{r\}}}^T\boldsymbol{Z}_{\{s\}}|\right)\right) . \end{aligned} \tag{6.3.45}$$

Using Lemma 6.3.1 for $i = j$, $k = r$, $l = s$ results in

$$|\boldsymbol{Z}^T\boldsymbol{Z}||{\boldsymbol{Z}_{\{jr\}}}^T\boldsymbol{Z}_{\{js\}}| - |{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{j\}}||{\boldsymbol{Z}_{\{r\}}}^T\boldsymbol{Z}_{\{s\}}| = -|{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{r\}}||{\boldsymbol{Z}_{\{j\}}}^T\boldsymbol{Z}_{\{s\}}|$$

and it follows for (6.3.45)

$$
\begin{aligned}
s^* = \sum_{j=1}^{p} \Bigg(& -\psi_j^2 |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{j\}}|^2 - 2|{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{j\}}| \sum_{\substack{r=1\\ r\neq j}}^{p} (-1)^{r+j} \psi_r \psi_j |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| \\
& - \sum_{\substack{s=1\\ s\neq j}}^{p} \sum_{\substack{r=1\\ r\neq j}}^{p} (-1)^{r+s} \psi_r \psi_s |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{s\}}| \Bigg) \\
= & - \sum_{j=1}^{p} \left(\sum_{s=1}^{p} \sum_{r=1}^{p} (-1)^{r+s} \psi_r \psi_s |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{s\}}| \right) \\
= & - \sum_{j=1}^{p} \left(\sum_{r=1}^{p} (-1)^{r} \psi_r |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| \right)^2 , \qquad (6.3.46)
\end{aligned}
$$

for $\boldsymbol{Z}$ having full rank. Hence it follows for (6.3.39) and $p \geq 2$

$$
\begin{aligned}
\sum_{j=1}^{p} s_1(j) = \sum_{j=1}^{p} \Bigg(& {\boldsymbol{\psi}_{\{j\}}}^T \tilde{\boldsymbol{A}}^{(jj)} \boldsymbol{\psi}_{\{j\}} |\boldsymbol{Z}^T \boldsymbol{Z}| - \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{j\}}| \\
& + \left(\sum_{r=1}^{p} (-1)^{r} \psi_r |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| \right)^2 \Bigg) = 0 \qquad (6.3.47)
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{j=1}^{p} s_2(j) = \sum_{j=1}^{p} \Bigg(& {\boldsymbol{\psi}_{\{j\}}}^T \tilde{\boldsymbol{A}}^{(jj)} \boldsymbol{\psi}_{\{j\}} |\boldsymbol{Z}^T \boldsymbol{Z}| - \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{j\}}| \\
& - \left(\sum_{r=1}^{p} (-1)^{r} \psi_r |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| \right)^2 \Bigg) \\
= & -2 \sum_{j=1}^{p} \left(\sum_{r=1}^{p} (-1)^{r} \psi_r |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| \right)^2 . \qquad (6.3.48)
\end{aligned}
$$

Thus with (6.3.40) the first derivative of $\mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega)$ with respect to ω can be written as

$$
\begin{aligned}
\frac{\partial \mathrm{MSE}(\tilde{\boldsymbol{\gamma}}_\omega)}{\partial \omega} = & -4n\sigma^2 \sum_{j=1}^{p} \frac{|\boldsymbol{Z}^T \boldsymbol{Z}| \left(\sum_{r=1}^{p} (-1)^r \psi_r |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| \right)^2 \omega}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^3} \\
& + \frac{4\omega^3 |\boldsymbol{Z}^T \boldsymbol{Z}| \boldsymbol{\gamma}^T \boldsymbol{\Psi}^T \boldsymbol{\Psi} (\tilde{\boldsymbol{M}}^{const})^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi} \boldsymbol{\gamma}}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^3} . \qquad (6.3.49)
\end{aligned}
$$

where

$$
\frac{\partial}{\partial \omega} \sum_{j=1}^{p} \mathrm{var}(\tilde{\gamma}_j^\omega) = -4n\sigma^2 \sum_{j=1}^{p} \frac{|\boldsymbol{Z}^T \boldsymbol{Z}| \left(\sum_{r=1}^{p} (-1)^r \psi_r |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{r\}}| \right)^2 \omega}{(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T \boldsymbol{Z}|)^3}
$$

and

$$\frac{\partial}{\partial\omega}\left(\mathrm{Bias}^T(\tilde{\gamma}_\omega)\mathrm{Bias}(\tilde{\gamma}_\omega)\right) = \frac{4\omega^3|\boldsymbol{Z}^T\boldsymbol{Z}|\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\tilde{\boldsymbol{M}}^{const})^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}. \tag{6.3.50}$$

Now we are in a position to deduce the main theorem of this Chapter.

Theorem 6.3.2 (Existence Theorem I). *For $\boldsymbol{Z}\in\mathbb{R}^{n\times p}$, $n>p$, $p\geq 2$ having full rank*

$$\forall\ \Psi\neq\mathbf{0}\quad \exists\ \varepsilon>0\quad \forall\ \omega\in[-\varepsilon,\varepsilon]\backslash\{0\}:\mathrm{MSE}(\tilde{\gamma}_\omega)<\mathrm{MSE}(\hat{\gamma}).$$

Proof. From (6.3.41) we know, that the denominator of (6.3.49) is always bigger than zero for $\boldsymbol{Z}$ having full column rank. Let $[-\varepsilon,\varepsilon]$ be an ε–neighborhood of ω about zero that can be made arbitrarily small. Thus choose ε small enough, such that we can write for (6.3.49) and $\omega\in[-\varepsilon,\varepsilon]$

$$\begin{aligned}\frac{\partial\mathrm{MSE}(\tilde{\gamma}_\omega)}{\partial\omega} = -4n\sigma^2\sum_{j=1}^{p}\frac{|\boldsymbol{Z}^T\boldsymbol{Z}|\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2\omega}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}\\ +\frac{\mathcal{O}(\omega^3)}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}.\end{aligned} \tag{6.3.51}$$

We have $|\boldsymbol{Z}^T\boldsymbol{Z}|>0$. If $\sum_{j=1}^{p}\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2>0$, $\mathrm{MSE}(\tilde{\gamma}_\omega)$ has got a maximum in $\omega=0$, because

$$\frac{\partial\mathrm{MSE}(\tilde{\gamma}_\omega)}{\partial\omega}\begin{cases}>0, & \omega\in[-\varepsilon,0)\\ \ \ 0, & \omega=0\\ <0, & \omega\in(0,\varepsilon]\end{cases}.$$

Because

$$\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2\geq 0,$$

a solution to

$$\sum_{j=1}^{p}\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2=\sum_{j=1}^{p}\left(\sum_{r=1}^{p}(-1)^{r+j}\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2=0$$

is only given for $\left(\sum_{r=1}^{p}(-1)^{r+j}\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)=0$ for all $j=1,\ldots,p$. This solution can be found by solving the homogeneous system of linear equations

$$\begin{bmatrix}|\boldsymbol{Z}_{\{1\}}{}^T\boldsymbol{Z}_{\{1\}}| & \ldots & (-1)^{1+p}|\boldsymbol{Z}_{\{1\}}{}^T\boldsymbol{Z}_{\{p\}}|\\ \vdots & \ddots & \vdots\\ (-1)^{1+p}|\boldsymbol{Z}_{\{1\}}{}^T\boldsymbol{Z}_{\{p\}}| & \ldots & |\boldsymbol{Z}_{\{p\}}{}^T\boldsymbol{Z}_{\{p\}}|\end{bmatrix}\begin{bmatrix}\psi_1\\ \vdots\\ \psi_p\end{bmatrix}=\mathbf{0}. \tag{6.3.52}$$

But the matrix on the left hand side of (6.3.52) is just the adjoint matrix of $\boldsymbol{Z}^T\boldsymbol{Z}$. From (6.1.24) we know, that the adjoint matrix of $\boldsymbol{Z}^T\boldsymbol{Z}$ is also positive definite and thus not singular. As a consequence (6.3.52) has a unique solution (see Lancaster (1985,[**30**]), p. 94), namely the zero vector, i.e. $\psi_1 = \ldots = \psi_p = 0$. Hence for $\boldsymbol{\psi} \neq \boldsymbol{0}$ it is

$$\sum_{j=1}^{p}\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2 > 0$$

and we will have a maximum in $\omega = 0$.

□

The numerator of (6.3.49) may have three roots. From Theorem 6.3.2 we know, that there is a maximum in $\omega_1 = 0$. Thus there will be two minima in

$$\omega_{2,3} = \pm\sigma\sqrt{\frac{n\sum_{j=1}^{p}\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2}{\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\tilde{\boldsymbol{M}}^{const})^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}}. \tag{6.3.53}$$

The numerator and denominator of (6.3.53) are bigger than zero for $\boldsymbol{Z}$ having full rank (see proof of Theorem 6.3.2). Thus the minima exist. From (6.3.35) it is easy to see, that

$$\lim_{\omega\to\pm\infty}\operatorname{MSE}(\tilde{\gamma}_\omega) = \sigma^2\frac{\sum_{j=1}^{p}\boldsymbol{\psi}_{\{j\}}^T\tilde{\boldsymbol{A}}^{(jj)}\boldsymbol{\psi}_{\{j\}}}{\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}} + \frac{\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\tilde{\boldsymbol{M}}^{const})^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{n^2\left(\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}\right)^2}.$$

Example 6.3.3. Choosing $\psi_1 \neq 0$, but $\psi_2 = \ldots = \psi_p = 0$ results in

$$\sum_{j=1}^{p}s_2(j) = -2\psi_1^2|\boldsymbol{Z}^T\boldsymbol{Z}|\sum_{j=1}^{p}|\boldsymbol{Z}_{\{1\}}{}^T\boldsymbol{Z}_{\{j\}}|^2$$

and

$$\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\tilde{\boldsymbol{M}}^{const})^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma} = n^2\psi_1^4\gamma_1^2\sum_{j=1}^{p}|\boldsymbol{Z}_{\{1\}}{}^T\boldsymbol{Z}_{\{j\}}|^2 > 0$$

and thus there will be a minimum in

$$\omega_2 = \sqrt{\frac{\sigma^2}{n\psi_1^2\gamma_1^2}}$$

and

$$\omega_3 = -\sqrt{\frac{\sigma^2}{n\psi_1^2\gamma_1^2}}. \tag{6.3.54}$$

The mean squared error of (6.2.33) of Example 6.1.10 is given by

$$
\begin{aligned}
\mathrm{MSE}(\tilde{\gamma}_\omega) &= \sigma^2 \left(\frac{\frac{2}{3}}{(2\omega^2\psi_1^2+1)^2} + \frac{\frac{1}{2}(\omega^4\psi_1^4+\omega^2\psi_1^2+1)}{(2\omega^2\psi_1^2+1)^2} \right) \\
&\qquad + \frac{\frac{1}{4}\omega^4(8+28\omega^2\psi_1^2+49\omega^4\psi_1^4)}{(2\omega^2\psi_1^2+1)^2} \\
&= \frac{1}{12}\frac{\left(14\sigma^2+6\sigma^2\omega^2\psi_1^2+6\sigma^2\omega^4\psi_1^4+24\omega^4\psi_1^4+84\omega^6\psi_1^6+147\omega^8\psi_1^8\right)}{(2\omega^2\psi_1^2+1)^2}.
\end{aligned}
$$

Figure 6.3.5 shows the total variance, the squared bias and the mean squared error of $\tilde{\gamma}_\omega$ for $\sigma^2 = \psi_1^2 = 1$. The minima are given by (6.3.54)

$$\omega_{2,3} = \pm\sqrt{\frac{1}{3}} = \pm 0.5774.$$

•

6.4. The Matrix Mean Squared Error of the DLSE

From Chapter 2 the matrix mean squared error of $\tilde{\gamma}$ is defined by

$$\mathrm{MSE}(\tilde{\gamma}) = \Sigma(\tilde{\gamma}) + \mathrm{Bias}(\tilde{\gamma})\mathrm{Bias}^T(\tilde{\gamma})$$

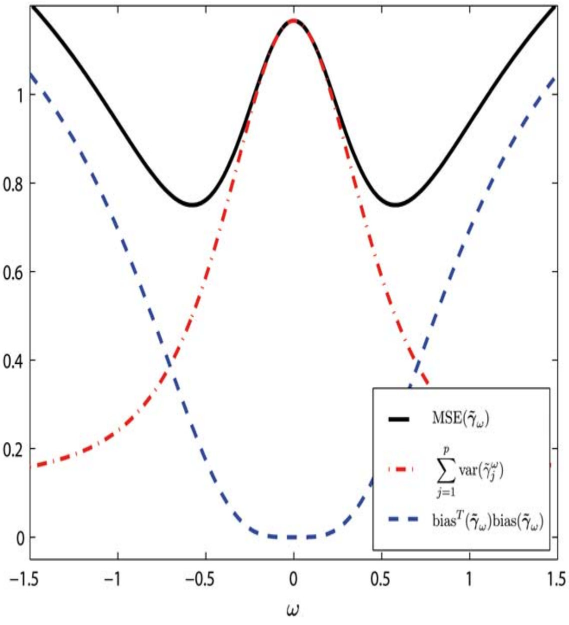


FIGURE 6.3.5. Mean squared error in dependence of ω

and with (6.2.30) and (6.2.34) it follows

$$\begin{aligned}\mathrm{MtxMSE}(\tilde{\gamma}) &= \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\\ &\quad+\omega^4(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\\ &= (\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\left(\sigma^2\boldsymbol{Z}^T\boldsymbol{Z}+\omega^4\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}.\end{aligned}$$

To prefer our disturbed least squares estimator $\tilde{\boldsymbol{\gamma}}$ to the least squares estimator $\hat{\boldsymbol{\gamma}}$, the matrix

$$\Delta := \mathrm{MtxMSE}(\hat{\gamma}) - \mathrm{MtxMSE}(\tilde{\gamma}) \geq 0 \tag{6.4.55}$$

has to be positive semidefinit. Because

$$\mathrm{MtxMSE}(\hat{\gamma}) = \sigma^2(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$$

(6.4.55) is equivalent to

$$\begin{aligned}\sigma^2\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}&\left(\sigma^2\boldsymbol{Z}^T\boldsymbol{Z}\right.\\ &\left.+\omega^4\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1} \geq 0. \end{aligned}\tag{6.4.56}$$

With the same arguments as in the proof of Theorem 5.4.2, (6.4.56) can be written as

$$\begin{aligned}&\sigma^2\left(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})\\ &\qquad -\sigma^2\boldsymbol{Z}^T\boldsymbol{Z}-\omega^4\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\geq 0\\ &\Leftrightarrow \sigma^2\left(\boldsymbol{Z}^T\boldsymbol{Z}+2\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}+\omega^4\boldsymbol{\Psi}^T\boldsymbol{\Psi}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)\\ &\qquad -\sigma^2\boldsymbol{Z}^T\boldsymbol{Z}-\omega^4\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\geq 0\\ &\Leftrightarrow \omega^2\left(2\sigma^2\boldsymbol{I}_p-\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\right)\boldsymbol{\Psi}^T\boldsymbol{\Psi}+\sigma^2\omega^4\boldsymbol{\Psi}^T\boldsymbol{\Psi}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\geq 0. \end{aligned}\tag{6.4.57}$$

From (6.4.57) we can deduce the following Theorem.

THEOREM 6.4.1. *The disturbed least squares estimator $\tilde{\boldsymbol{\gamma}}$ has to be preferred to the least squares estimator $\hat{\boldsymbol{\gamma}}$, i.e. $\Delta \geq 0$, if*

$$\left(2\sigma^2 - n\omega^2\sum_{s=1}^{p}\sum_{r=1}^{p}\psi_r\psi_s\gamma_r\gamma_s\right)\geq 0.$$

PROOF. It is

$$\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T = n\begin{bmatrix}\sum_{r=1}^p\psi_1\psi_r\gamma_1\gamma_r & \cdots & \sum_{r=1}^p\psi_1\psi_r\gamma_p\gamma_r\\ \vdots & & \vdots\\ \sum_{r=1}^p\psi_p\psi_r\gamma_1\gamma_r & \cdots & \sum_{r=1}^p\psi_p\psi_r\gamma_p\gamma_r\end{bmatrix}$$

and thus

$$\begin{aligned}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi} &= n^2\begin{bmatrix} \sum_{r=1}^{p}\sum_{s=1}^{p}\psi_1^2\psi_r\psi_s\gamma_r\gamma_s & \dots \sum_{r=1}^{p}\sum_{s=1}^{p}\psi_1\psi_p\psi_r\psi_s\gamma_r\gamma_s \\ \vdots & \vdots \\ \sum_{r=1}^{p}\sum_{s=1}^{p}\psi_1\psi_p\psi_r\psi_s\gamma_r\gamma_s \dots & \sum_{r=1}^{p}\sum_{s=1}^{p}\psi_p^2\psi_r\psi_s\gamma_r\gamma_s \end{bmatrix} \\ &= n\sum_{r=1}^{p}\sum_{s=1}^{p}\psi_r\psi_s\gamma_r\gamma_s\boldsymbol{\Psi}^T\boldsymbol{\Psi}.\end{aligned}$$

Hence it follows with (6.4.57)

$$2\sigma^2\boldsymbol{\Psi}^T\boldsymbol{\Psi} - \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi} = \left(2\sigma^2 - n\omega^2\sum_{r=1}^{p}\sum_{s=1}^{p}\psi_r\psi_s\gamma_r\gamma_s\right)\boldsymbol{\Psi}^T\boldsymbol{\Psi}. \tag{6.4.58}$$

For $c := \left(2\sigma^2 - n\omega^2\sum_{s=1}^{p}\sum_{r=1}^{p}\psi_r\psi_s\gamma_r\gamma_s\right) \geq 0$, (6.4.58) will be a positive semidefinite matrix, because with

$$\boldsymbol{p}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{p} \geq 0,$$

it follows

$$\boldsymbol{p}^T c\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{p} = c\boldsymbol{p}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{p} \geq 0$$

for any $\boldsymbol{p} \in \mathbb{R}^{p\times 1}$. $\boldsymbol{Z}^T\boldsymbol{Z}$ is positive definite by assumption and with $\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}^{-\frac{1}{2}}$ and Appendix A.8 the inverse of $\boldsymbol{Z}^T\boldsymbol{Z}$ can be written as

$$\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} = \left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\right)^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^T = \left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^T\right)^T\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^T\right)$$

and thus

$$\begin{aligned}&\boldsymbol{p}^T\left(c\boldsymbol{\Psi}^T\boldsymbol{\Psi} + \sigma^2\omega^4\boldsymbol{\Psi}^T\boldsymbol{\Psi}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)\boldsymbol{p} \\ &\qquad = c\boldsymbol{p}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{p} + \sigma^2\omega^4\boldsymbol{p}^T\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)^T\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\right)\boldsymbol{p} \geq 0,\end{aligned}$$

for $c \geq 0$. Hence (6.4.57) is positive semidefinite for $c \geq 0$. Otherwise for $c < 0$ no conclusion can be drawn.

□

6.5. Mean Squared Error Properties of $\tilde{\beta}_\omega$

We showed in (3.2.14), that the relationship between the original and standardized least squares estimates of the regression coefficients is given by

$$\hat{\beta}_j = \hat{\gamma}_j \left(\frac{1}{S_{jj}}\right)^{\frac{1}{2}}, \quad j = 1, \dots, p$$

and

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{X}_j. \tag{6.5.59}$$

Thus we can define the disturbed least squares estimator of β_0 by

$$\tilde{\beta}_0^\omega = \bar{y} - \tilde{\gamma}_1^\omega \bar{X}_1 - \dots - \tilde{\gamma}_p^\omega \bar{X}_p$$

and of the remaining components by

$$\tilde{\beta}_j^\omega = \tilde{\gamma}_j^\omega \left(\frac{1}{S_{jj}}\right)^{\frac{1}{2}}, \quad j = 1, \dots, p. \tag{6.5.60}$$

The disturbed least squares estimator $\tilde{\boldsymbol{\beta}}_\omega$ of $\boldsymbol{\beta}$ can then be written as

$$\tilde{\boldsymbol{\beta}}_\omega = \tau + \boldsymbol{Q}\boldsymbol{D}^{-1}\tilde{\boldsymbol{\gamma}}_\omega \in \mathbb{R}^{(p+1)\times 1},$$

with

$$\tau = \begin{bmatrix} \bar{y}, & 0, & \dots & ,0 \end{bmatrix}^T \in \mathbb{R}^{(p+1)\times 1},$$

$$\boldsymbol{Q} = \begin{bmatrix} -\bar{X}_1 & -\bar{X}_2 & \dots & -\bar{X}_p \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{(p+1)\times p}$$

and

$$\boldsymbol{D} = \begin{bmatrix} \sqrt{S_{11}} & & \\ & \ddots & \\ & & \sqrt{S_{pp}} \end{bmatrix} \in \mathbb{R}^{p\times p}. \tag{6.5.61}$$

It follows

$$\text{Bias}(\tilde{\boldsymbol{\beta}}_\omega) = \text{E}(\boldsymbol{Q}\boldsymbol{D}^{-1}\tilde{\boldsymbol{\gamma}}_\omega + \tau) - \boldsymbol{\beta} = \boldsymbol{Q}\boldsymbol{D}^{-1}\text{E}(\tilde{\boldsymbol{\gamma}}_\omega) + \text{E}(\tau) - \boldsymbol{\beta}. \tag{6.5.62}$$

From (3.1.2) we have

$$\begin{aligned}
\mathrm{E}(\tau) - \boldsymbol{\beta} &= \begin{bmatrix} \beta_0 + \beta_1\bar{X}_1 + \ldots + \beta_p\bar{X}_p \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \\
&= \begin{bmatrix} \beta_1\bar{X}_1 + \ldots + \beta_p\bar{X}_p \\ -\beta_1 \\ \vdots \\ -\beta_p \end{bmatrix} = -\boldsymbol{Q}\boldsymbol{\beta}_{\{\beta_0\}}
\end{aligned}$$

with $\boldsymbol{\beta}^T_{\{\beta_0\}} = \begin{bmatrix}\beta_1, & \ldots & ,\beta_p\end{bmatrix}$. From $\boldsymbol{\beta}_{\{\beta_0\}} = \boldsymbol{D}^{-1}\boldsymbol{\gamma}$ (see (3.2.10)) it follows for (6.5.62)

$$\begin{aligned}
\mathrm{Bias}(\tilde{\boldsymbol{\beta}}_\omega) &= \boldsymbol{Q}\boldsymbol{D}^{-1}\mathrm{E}(\tilde{\gamma}_\omega) - \boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{\gamma} \\
&= \boldsymbol{Q}\boldsymbol{D}^{-1}\mathrm{Bias}(\tilde{\gamma}_\omega). \qquad (6.5.63)
\end{aligned}$$

Furthermore we have

$$\Sigma(\tilde{\boldsymbol{\beta}}_\omega) = \Sigma(\boldsymbol{Q}\boldsymbol{D}^{-1}\tilde{\gamma}_\omega + \tau) = \boldsymbol{Q}\boldsymbol{D}^{-1}\Sigma(\tilde{\gamma}_\omega)\boldsymbol{D}^{-1}\boldsymbol{Q}^T$$

and thus with (A.1.2)

$$\begin{aligned}
\mathrm{MSE}(\tilde{\boldsymbol{\beta}}_\omega) &= \mathrm{tr}\left(\boldsymbol{Q}\boldsymbol{D}^{-1}\Sigma(\tilde{\gamma}_\omega)\boldsymbol{D}^{-1}\boldsymbol{Q}^T\right) + \mathrm{Bias}^T(\tilde{\gamma}_\omega)\boldsymbol{D}^{-1}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{D}^{-1}\mathrm{Bias}(\tilde{\gamma}_\omega) \\
&= \mathrm{tr}\left(\boldsymbol{D}^{-1}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{D}^{-1}\Sigma(\tilde{\gamma}_\omega)\right) + \mathrm{Bias}^T(\tilde{\gamma}_\omega)\boldsymbol{D}^{-1}\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{D}^{-1}\mathrm{Bias}(\tilde{\gamma}_\omega).
\end{aligned}$$

Because

$$\begin{aligned}
\boldsymbol{Q}^T\boldsymbol{Q} &= \begin{bmatrix} \bar{X}_1^2+1 & \bar{X}_1\bar{X}_2 & \ldots & \bar{X}_1\bar{X}_p \\ \bar{X}_1\bar{X}_2 & \bar{X}_2^2+1 & \ldots & \bar{X}_2\bar{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{X}_1\bar{X}_p & \bar{X}_2\bar{X}_p & \ldots & \bar{X}_p^2+1 \end{bmatrix} \\
&= \begin{bmatrix} \bar{X}_1^2 & \bar{X}_1\bar{X}_2 & \ldots & \bar{X}_1\bar{X}_p \\ \bar{X}_1\bar{X}_2 & \bar{X}_2^2 & \ldots & \bar{X}_2\bar{X}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{X}_1\bar{X}_p & \bar{X}_2\bar{X}_p & \ldots & \bar{X}_p^2 \end{bmatrix} + \boldsymbol{I}_p \\
&= \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix} \begin{bmatrix}\bar{X}_1, & \ldots & ,\bar{X}_p\end{bmatrix} + \boldsymbol{I}_p =: \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T + \boldsymbol{I}_p
\end{aligned}$$

and

$$\boldsymbol{D}^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{D}^{-1} = \begin{bmatrix} \frac{\bar{X}_1^2}{S_{11}} & \frac{\bar{X}_1\bar{X}_2}{\sqrt{S_{11}}\sqrt{S_{22}}} & \cdots & \frac{\bar{X}_1\bar{X}_p}{\sqrt{S_{11}}\sqrt{S_{pp}}} \\ \frac{\bar{X}_1\bar{X}_2}{\sqrt{S_{11}}\sqrt{S_{22}}} & \frac{\bar{X}_2^2}{S_{22}} & \cdots & \frac{\bar{X}_2\bar{X}_p}{\sqrt{S_{22}}\sqrt{S_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\bar{X}_1\bar{X}_p}{\sqrt{S_{11}}\sqrt{S_{pp}}} & \frac{\bar{X}_2\bar{X}_p}{\sqrt{S_{22}}\sqrt{S_{pp}}} & \cdots & \frac{\bar{X}_p^2}{S_{pp}} \end{bmatrix}, \tag{6.5.64}$$

we get with (A.1.1)

$$\begin{aligned} \mathrm{MSE}(\tilde{\boldsymbol{\beta}}_\omega) &= \mathrm{tr}\left(\boldsymbol{D}^{-1}(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T + \boldsymbol{I}_p)\boldsymbol{D}^{-1}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)\right) \\ &\quad + \mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)^T\boldsymbol{D}^{-1}(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T + \boldsymbol{I}_p)\boldsymbol{D}^{-1}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) \\ &= \mathrm{tr}\left(\boldsymbol{D}^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{D}^{-1}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)\right) + \mathrm{tr}\left(\boldsymbol{D}^{-2}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)\right) \\ &+ \mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)^T\boldsymbol{D}^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{D}^{-1}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) + \mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)^T\boldsymbol{D}^{-2}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega). \end{aligned} \tag{6.5.65}$$

Consider $\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}} := \left[\tilde{\beta}^\omega_1, \quad \dots \quad , \tilde{\beta}^\omega_p\right]$. With (6.5.60) and (A.1.2) the total variance of $\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}}$ is given by

$$\sum_{j=1}^p \mathrm{var}(\tilde{\beta}^\omega_j) = \mathrm{tr}\left(\Sigma(\boldsymbol{D}^{-1}\tilde{\boldsymbol{\gamma}}_\omega)\right) = \mathrm{tr}\left(\boldsymbol{D}^{-1}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)\boldsymbol{D}^{-1}\right) = \mathrm{tr}\left(\boldsymbol{D}^{-2}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)\right).$$

From (6.5.63) and the definition of $\boldsymbol{Q}$ the bias of $\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}}$ is given by

$$\mathrm{Bias}(\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}}) = \boldsymbol{D}^{-1}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega).$$

Thus we have

$$\begin{aligned} \mathrm{MSE}(\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}}) &= \mathrm{tr}\left(\boldsymbol{D}^{-2}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)\right) + \mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)^T\boldsymbol{D}^{-2}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) \\ &= \sum_{j=1}^p \frac{\mathrm{var}(\tilde{\gamma}^\omega_j)}{\mathrm{S}_{jj}} + \mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)^T\boldsymbol{D}^{-2}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) \end{aligned} \tag{6.5.66}$$

and with (6.5.65), (6.5.64) and (6.2.34) we can write

$$\begin{aligned} \mathrm{MSE}(\tilde{\beta}^\omega_0) = \mathrm{tr}\left(\boldsymbol{D}^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{D}^{-1}\Sigma(\tilde{\boldsymbol{\gamma}}_\omega)\right) &+ \mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)^T\boldsymbol{D}^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{D}^{-1}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega) \\ = \sum_{j=1}^p\sum_{i=1}^p &\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\mathrm{cov}(\tilde{\gamma}^\omega_i, \tilde{\gamma}^\omega_j) \\ &+ \frac{\omega^4\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\tilde{\boldsymbol{M}}^{const}\boldsymbol{D}^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{D}^{-1}\tilde{\boldsymbol{M}}^{const}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^2}. \end{aligned} \tag{6.5.67}$$

Analogously to the last section, we can proof an Existence Theorem for the disturbed least squares estimator of $\tilde{\boldsymbol{\beta}}_\omega$.

THEOREM 6.5.1 (Existence Theorem II). *For $\boldsymbol{X}$ having full rank,*

$$\forall\ \Psi \neq \mathbf{0} \quad \exists\ \varepsilon > 0 \quad \forall\ \omega \in [-\varepsilon, \varepsilon]\backslash\{0\} : \mathrm{MSE}(\tilde{\boldsymbol{\beta}}_\omega) < \mathrm{MSE}(\hat{\boldsymbol{\beta}}).$$

PROOF. With (6.5.66), the first derivative of the mean squared error of $\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}}$ is given by

$$\frac{\partial}{\partial\omega}\mathrm{MSE}(\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}}) = \frac{\partial}{\partial\omega}\sum_{j=1}^{p}\frac{\mathrm{var}(\tilde{\gamma}^\omega_j)}{\mathrm{S}_{jj}} + \frac{\partial}{\partial\omega}\left(\mathrm{Bias}^T(\tilde{\boldsymbol{\gamma}}_\omega)\boldsymbol{D}^{-2}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)\right).$$

With the help of (6.3.50) we get

$$\frac{\partial}{\partial\omega}\sum_{j=1}^{p}\frac{\mathrm{var}(\tilde{\gamma}^\omega_j)}{\mathrm{S}_{jj}} = -4n\sigma^2\sum_{j=1}^{p}\frac{|\boldsymbol{Z}^T\boldsymbol{Z}|\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2\omega}{\mathrm{S}_{jj}(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^3} \qquad (6.5.68)$$

and

$$\frac{\partial}{\partial\omega}\left(\mathrm{Bias}^T(\tilde{\boldsymbol{\gamma}}_\omega)\boldsymbol{D}^{-2}\mathrm{Bias}(\tilde{\boldsymbol{\gamma}}_\omega)\right) = \frac{4\omega^3|\boldsymbol{Z}^T\boldsymbol{Z}|\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\tilde{\boldsymbol{M}}^{const}\boldsymbol{D}^{-2}\tilde{\boldsymbol{M}}^{const}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^3}. \qquad (6.5.69)$$

Thus it is easy to see that

$$\left.\frac{\partial}{\partial\omega}\mathrm{MSE}(\tilde{\boldsymbol{\beta}}^\omega_{\{\tilde{\beta}_0\}})\right|_{\omega=0} = 0.$$

For $\omega \in [-\varepsilon_1, \varepsilon_1]\backslash\{0\}$ and ε_1 small enough it follows

$$\frac{\partial}{\partial\omega}\mathrm{MSE}(\tilde{\boldsymbol{\beta}}_{\{\tilde{\beta}_0\}}) = -4n\omega\sigma^2|\boldsymbol{Z}^T\boldsymbol{Z}|\sum_{j=1}^{p}\frac{\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2}{\mathrm{S}_{jj}(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^3} + \frac{\mathcal{O}(\omega^3)}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^3}. \qquad (6.5.70)$$

In the proof of Theorem 6.3.2 we showed that

$$\left(\sum_{r=1}^{p}(-1)^r\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)^2 > 0$$

and

$$(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|)^3 > 0$$

for $\boldsymbol{Z}$ having full rank. From the definition of S_{jj}, given in (3.2.9), we know that $S_{jj} > 0$.

Hence it follows with (6.5.70)

$$\frac{\partial}{\partial\omega}\mathrm{MSE}(\tilde{\boldsymbol{\beta}}_{\{\tilde{\beta}_0\}})\begin{cases} > 0 & ,\forall\ \omega\in[-\varepsilon_1,0) \\ \ \ 0 & ,\omega=0 \\ < 0 & ,\omega\in(0,\varepsilon_1] \end{cases} . \tag{6.5.71}$$

Now we have to show, that there exists any ε_2, such that this case differentiation is also true for $\mathrm{MSE}(\tilde{\beta}_0^{\omega})$. We know from (6.5.67) that

$$\mathrm{var}(\tilde{\beta}_0^{\omega})=\sum_{j=1}^{p}\sum_{i=1}^{p}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\mathrm{cov}(\tilde{\gamma}_i^{\omega},\tilde{\gamma}_j^{\omega}) \tag{6.5.72}$$

and it follows from Corollary 6.2.4

$$\begin{aligned}\mathrm{cov}(\tilde{\gamma}_i^{\omega},\tilde{\gamma}_j^{\omega})=\sigma^2&\frac{n\omega^2\tilde{m}_{i,j}^{quad}+\tilde{m}_{i,j}^{const}}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|}\\ &-\frac{\sigma^2\omega^2 n\sum_{s=1}^{p}\sum_{r=1}^{p}(-1)^{i+j+r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{r\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{s\}}|}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^2},\end{aligned}$$

with

$$\begin{aligned}\tilde{\boldsymbol{M}}^{quad}&=\left[\tilde{m}_{i,j}^{quad}\right]_{1\le i,j\le p}=\left[(-1)^{i+j}\boldsymbol{\psi}_{\{i\}}{}^T\tilde{\boldsymbol{A}}^{(ij)}\boldsymbol{\psi}_{\{j\}}\right]_{1\le i,j\le p},\\ \tilde{\boldsymbol{M}}^{const}&=\left[\tilde{m}_{i,j}^{const}\right]_{1\le i,j\le p}=\left[(-1)^{i+j}\left|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}\right|\right]_{1\le i,j\le p}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\psi}_{\{i\}}&=[\psi_r]_{\substack{1\le r\le p\\ r\neq i}}\in\mathbb{R}^{(p-1)\times 1},\\ \tilde{\boldsymbol{A}}^{(ij)}&=\begin{cases}1 & ,p=2\\ \left[(-1)^{r+s}|\boldsymbol{Z}_{\{ir\}}{}^T\boldsymbol{Z}_{\{js\}}|\right]_{\substack{1\le r,s\le p\\ i\neq r;j\neq s}} & ,p\ge 3\end{cases}\in\mathbb{R}^{(p-1)\times(p-1)},\end{aligned}$$

defined like in Corollary 6.1.9. Thus we have

$$\begin{aligned}\mathrm{var}(\tilde{\beta}_0^{\omega})=\sigma^2\sum_{j=1}^{p}\sum_{i=1}^{p}(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}&\left(\frac{n\omega^2\boldsymbol{\psi}_{\{i\}}{}^T\tilde{\boldsymbol{A}}^{(ij)}\boldsymbol{\psi}_{\{j\}}+|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|}{n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|}\right.\\ &\left.-\frac{n\omega^2\sum_{s=1}^{p}\sum_{r=1}^{p}(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{r\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{s\}}|}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^2}\right).\end{aligned} \tag{6.5.73}$$

Therefore we can write with (6.5.73)

$$
\begin{aligned}
\frac{\partial}{\partial\omega}\mathrm{var}(\tilde{\beta}_0^\omega) = \sigma^2 \sum_{j=1}^p \sum_{i=1}^p (-1)^{i+j} \frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}} \frac{\partial}{\partial\omega} \Bigg(& \frac{n\omega^2 {\boldsymbol{\psi}_{\{i\}}}^T \tilde{\boldsymbol{A}}^{(ij)} \boldsymbol{\psi}_{\{j\}} + |{\boldsymbol{Z}_{\{i\}}}^T \boldsymbol{Z}_{\{j\}}|}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|} \\
& - \frac{n\omega^2 \sum_{s=1}^p \sum_{r=1}^p (-1)^{r+s} \psi_r\psi_s |{\boldsymbol{Z}_{\{i\}}}^T \boldsymbol{Z}_{\{r\}}| |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{s\}}|}{\left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)^2} \Bigg) \\
= \sigma^2 \sum_{j=1}^p \sum_{i=1}^p (-1)^{i+j} \frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}} \Bigg(& \frac{2n\omega {\boldsymbol{\psi}_{\{i\}}}^T \tilde{\boldsymbol{A}}^{(ij)} \boldsymbol{\psi}_{\{j\}} \left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)}{\left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)^2} \\
& - \frac{2n\omega \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} \left(n\omega^2 {\boldsymbol{\psi}_{\{i\}}}^T \tilde{\boldsymbol{A}}^{(ij)} \boldsymbol{\psi}_{\{j\}} + |{\boldsymbol{Z}_{\{i\}}}^T \boldsymbol{Z}_{\{j\}}|\right)}{\left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)^2} \\
& - \frac{2n\omega \sum_{s=1}^p \sum_{r=1}^p (-1)^{r+s} \psi_r\psi_s |{\boldsymbol{Z}_{\{i\}}}^T \boldsymbol{Z}_{\{r\}}| |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{s\}}|}{\left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)^2} \\
& + \frac{4n^2\omega^3 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} \sum_{s=1}^p \sum_{r=1}^p (-1)^{r+s} \psi_r\psi_s |{\boldsymbol{Z}_{\{i\}}}^T \boldsymbol{Z}_{\{r\}}| |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{s\}}|}{\left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3} \Bigg) \\
= 2n\omega\sigma^2 \sum_{j=1}^p \sum_{i=1}^p (-1)^{i+j} & \frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}} \\
\times \Bigg(& \frac{{\boldsymbol{\psi}_{\{i\}}}^T \tilde{\boldsymbol{A}}^{(ij)} \boldsymbol{\psi}_{\{j\}} |\boldsymbol{Z}^T\boldsymbol{Z}| - |{\boldsymbol{Z}_{\{i\}}}^T \boldsymbol{Z}_{\{j\}}| \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi}}{\left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)^2} \\
& - \frac{\left(|\boldsymbol{Z}^T\boldsymbol{Z}| - n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi}\right)}{\left(n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3} \\
& \times \sum_{s=1}^p \sum_{r=1}^p (-1)^{r+s} \psi_r\psi_s |{\boldsymbol{Z}_{\{i\}}}^T \boldsymbol{Z}_{\{r\}}| |{\boldsymbol{Z}_{\{j\}}}^T \boldsymbol{Z}_{\{s\}}| \Bigg).
\end{aligned}
$$

For $\omega \in [-\varepsilon_2, \varepsilon_2]\backslash\{0\}$ and ε_2 small enough, it follows

$$\frac{\partial}{\partial\omega}\mathrm{var}(\tilde{\beta}_0^\omega) = 2n\omega\sigma^2|\boldsymbol{Z}^T\boldsymbol{Z}|\sum_{j=1}^p\sum_{i=1}^p(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}$$
$$\left(\frac{\boldsymbol{\psi}_{\{i\}}{}^T\tilde{\boldsymbol{A}}^{(ij)}\boldsymbol{\psi}_{\{j\}}|\boldsymbol{Z}^T\boldsymbol{Z}| - |\boldsymbol{Z}_{\{i\}}^T\boldsymbol{Z}_{\{j\}}|\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3}\right.$$
$$\left.-\frac{\sum_{s=1}^p\sum_{r=1}^p(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{r\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{s\}}|}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3}\right)$$
$$+\frac{\mathcal{O}(\omega^3)}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3} \qquad (6.5.74)$$

and from (6.5.67)

$$\frac{\partial}{\partial\omega}\mathrm{Bias}(\tilde{\beta}_0^\omega)^T\mathrm{Bias}(\tilde{\beta}_0^\omega) = \frac{4\omega^3|\boldsymbol{Z}^T\boldsymbol{Z}|\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}\tilde{\boldsymbol{M}}^{const}\boldsymbol{D}^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{D}^{-1}\tilde{\boldsymbol{M}}^{const}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma}}{(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|)^3}$$
$$= \frac{\mathcal{O}(\omega^3)}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3}. \qquad (6.5.75)$$

Again it is easy to see, that

$$\mathrm{MSE}(\tilde{\beta}_0^\omega)\Big|_{\omega=0} = 0.$$

Furthermore we will show, that (6.5.74) is positive for $\omega\in[-\varepsilon_2,0)$ and negative for $\omega\in(0,\varepsilon_2]$. We already proved this for $\mathrm{MSE}(\tilde{\boldsymbol{\beta}}_{\{\tilde{\beta}_0\}})$ in (6.5.70) and as a consequence $\mathrm{MSE}(\tilde{\boldsymbol{\beta}}_\omega)$ will have a maximum in $\omega=0$. Define for the first numerator of (6.5.74)

$$s^+ := \sum_{j=1}^p\sum_{i=1}^p(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\Big(|\boldsymbol{Z}^T\boldsymbol{Z}|\boldsymbol{\psi}_{\{i\}}{}^T\tilde{\boldsymbol{A}}^{(ij)}\boldsymbol{\psi}_{\{j\}}$$
$$-|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}\Big). \qquad (6.5.76)$$

Let $p\geq 3$. It follows with Corollary 6.1.9

$$s^+ = \sum_{j=1}^p\sum_{i=1}^p(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\left(|\boldsymbol{Z}^T\boldsymbol{Z}|\sum_{\substack{s=1\\s\neq j}}^p\sum_{\substack{r=1\\r\neq i}}^p(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{ir\}}{}^T\boldsymbol{Z}_{\{js\}}|\right.$$
$$\left.-|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|\sum_{s=1}^p\sum_{r=1}^p(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{s\}}|\right). \qquad (6.5.77)$$

We can write

$$\sum_{s=1}^{p}\sum_{r=1}^{p}(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{s\}}| = (-1)^{i+j}\psi_i\psi_j|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}| + \sum_{\substack{s=1\\s\neq j}}^{p}(-1)^{i+s}\psi_i\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}| + \sum_{\substack{r=1\\r\neq i}}^{p}(-1)^{r+j}\psi_r\psi_j|\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{j\}}| + \sum_{\substack{s=1\\s\neq j}}^{p}\sum_{\substack{r=1\\r\neq i}}^{p}(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{s\}}|$$

and it follows for (6.5.76)

$$s^+ = \sum_{j=1}^{p}\sum_{i=1}^{p}(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\Bigg(- (-1)^{i+j}\psi_i\psi_j|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|^2 - |\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|\sum_{\substack{s=1\\s\neq j}}^{p}(-1)^{i+s}\psi_i\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}| - |\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|\sum_{\substack{r=1\\r\neq i}}^{p}(-1)^{r+j}\psi_r\psi_j|\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{j\}}| + \sum_{\substack{s=1\\s\neq j}}^{p}\sum_{\substack{r=1\\r\neq i}}^{p}(-1)^{r+s}\psi_r\psi_s\left(|\boldsymbol{Z}^T\boldsymbol{Z}||\boldsymbol{Z}_{\{ir\}}{}^T\boldsymbol{Z}_{\{js\}}| - |\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}||\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)\Bigg). \tag{6.5.78}$$

Using Lemma 6.3.1 implies

$$s^+ = \sum_{j=1}^{p}\sum_{i=1}^{p}(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\Bigg(- (-1)^{i+j}\psi_i\psi_j|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|^2 - |\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|\sum_{\substack{s=1\\s\neq j}}^{p}(-1)^{i+s}\psi_i\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}| - |\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|\sum_{\substack{r=1\\r\neq i}}^{p}(-1)^{r+j}\psi_r\psi_j|\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{j\}}| - \sum_{\substack{s=1\\s\neq j}}^{p}\sum_{\substack{r=1\\r\neq i}}^{p}(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\Bigg)$$

$$= -\sum_{j=1}^{p}\sum_{i=1}^{p}(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\left(\sum_{s=1}^{p}\sum_{r=1}^{p}(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|\right)$$

$$= -\sum_{i=1}^{p}\sum_{s=1}^{p}(-1)^{i+s}\frac{\bar{X}_i}{\sqrt{\mathrm{S}_{ii}}}\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}|\sum_{j=1}^{p}\sum_{r=1}^{p}(-1)^{j+r}\frac{\bar{X}_j}{\sqrt{\mathrm{S}_{jj}}}\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|$$

$$= -\left(\sum_{s=1}^{p}\sum_{i=1}^{p}(-1)^{i+s}\frac{\bar{X}_i}{\sqrt{\mathrm{S}_{ii}}}\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)^2 . \quad (6.5.79)$$

For $p = 2$ we have from Corollary 6.1.9

$$s^+ = \sum_{j=1}^{2}\sum_{i=1}^{2}(-1)^{i+j}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\left(|\boldsymbol{Z}^T\boldsymbol{Z}|\sum_{\substack{s=1\\ s\neq j}}^{2}\sum_{\substack{r=1\\ r\neq i}}^{2}\psi_r\psi_s\right.$$
$$\left. -|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}|\sum_{s=1}^{2}\sum_{r=1}^{2}(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{s\}}|\right) . \quad (6.5.80)$$

With the help of Lemma 6.3.1 we get for $p = 2$

$$\sum_{\substack{s=1\\ s\neq j}}^{2}\sum_{\substack{r=1\\ r\neq i}}^{2}\psi_r\psi_s\left(|\boldsymbol{Z}^T\boldsymbol{Z}| - (-1)^{r+s}|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}||\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)$$
$$= \sum_{\substack{s=1\\ s\neq j}}^{2}\sum_{\substack{r=1\\ r\neq i}}^{2}\psi_r\psi_s\left((-1)^{i+j+r+s}|\boldsymbol{Z}^T\boldsymbol{Z}| - (-1)^{r+s}|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{j\}}||\boldsymbol{Z}_{\{r\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)$$
$$= -\sum_{\substack{s=1\\ s\neq j}}^{2}\sum_{\substack{r=1\\ r\neq i}}^{2}(-1)^{r+s}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|.$$

Hence the presentation of s^+ in (6.5.79) is also valid for $p = 2$. On the other hand we have for the second numerator in (6.5.74)

$$\sum_{j=1}^{p}\sum_{i=1}^{p}\sum_{s=1}^{p}\sum_{r=1}^{p}(-1)^{i+j+r+s}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{r\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{s\}}|$$
$$= \sum_{j=1}^{p}\sum_{i=1}^{p}\sum_{s=1}^{p}\sum_{r=1}^{p}(-1)^{i+j+r+s}\frac{\bar{X}_i\bar{X}_j}{\sqrt{\mathrm{S}_{ii}}\sqrt{\mathrm{S}_{jj}}}\psi_r\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}||\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|$$
$$= \sum_{i=1}^{p}\sum_{s=1}^{p}(-1)^{s+i}\frac{\bar{X}_i}{\sqrt{\mathrm{S}_{ii}}}\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}|\sum_{j=1}^{p}\sum_{r=1}^{p}(-1)^{r+j}\frac{\bar{X}_j}{\sqrt{\mathrm{S}_{jj}}}\psi_r|\boldsymbol{Z}_{\{j\}}{}^T\boldsymbol{Z}_{\{r\}}|$$
$$= \left(\sum_{i=1}^{p}\sum_{s=1}^{p}(-1)^{i+s}\frac{\bar{X}_i}{\sqrt{\mathrm{S}_{ii}}}\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)^2$$

and thus it follows for (6.5.74)

$$\frac{\partial}{\partial\omega}\mathrm{var}(\tilde{\beta}_0^\omega) = -4n\omega\sigma^2|\boldsymbol{Z}^T\boldsymbol{Z}|\frac{\left(\sum_{s=1}^p\sum_{i=1}^p(-1)^{i+s}\frac{\bar{X}_i}{\sqrt{\mathrm{S}_{ii}}}\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)^2}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3} + \frac{\mathcal{O}(\omega^3)}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3}$$

for $\boldsymbol{Z}$ having full rank. With the squared bias in (6.5.75) it follows

$$\frac{\partial}{\partial\omega}\mathrm{MSE}(\tilde{\beta}_0^\omega) = -4n\omega\sigma^2|\boldsymbol{Z}^T\boldsymbol{Z}|\frac{\left(\sum_{s=1}^p\sum_{i=1}^p(-1)^{i+s}\frac{\bar{X}_i}{\sqrt{\mathrm{S}_{ii}}}\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)^2}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3} + \frac{\mathcal{O}(\omega^3)}{\left(n\omega^2\boldsymbol{\psi}^T\tilde{\boldsymbol{M}}^{const}\boldsymbol{\psi}+|\boldsymbol{Z}^T\boldsymbol{Z}|\right)^3}.$$

We have

$$t := \left(\sum_{s=1}^p\sum_{i=1}^p(-1)^{i+s}\frac{\bar{X}_i}{\sqrt{\mathrm{S}_{ii}}}\psi_s|\boldsymbol{Z}_{\{i\}}{}^T\boldsymbol{Z}_{\{s\}}|\right)^2 \geq 0$$

and for $t = 0$ everything is proved with (6.5.71).
But also for $t > 0$ we can find an ε_2, such that $\frac{\partial}{\partial\omega}\mathrm{MSE}(\tilde{\beta}_0^\omega)$ is positive for $\omega \in [-\varepsilon_2, 0)$ and negative for $\omega \in (0, \varepsilon_2]$. With

$$\varepsilon := \begin{cases}\min(\varepsilon_1, \varepsilon_2) & , t > 0\\ \varepsilon_1 & , t = 0\end{cases}$$

everything is proved.

□

Note 6.5.2. Obviously $\tilde{\gamma}_\omega$ and $\tilde{\boldsymbol{\beta}}_\omega$ do not only depend on ω, but also on $\boldsymbol{\psi}$. Unless there is a known systematic error in the data, $\boldsymbol{\psi}$ will usually be unknown in applied work. The Existence Theorems 6.3.2 and 6.5.1 are valid for arbitrary $\boldsymbol{\psi}$, i.e. we do not have any restrictions on choosing $\boldsymbol{\psi}$. Thus it would be suggestive to choose $\boldsymbol{\psi}$ optimal in some kind of way.

6.6. The DLSE for Unstandardized Data

All calculations made in Chapter 6 were based on a standardized design matrix $\boldsymbol{Z}$. Because of the controversy of the usefulness of standardization in literature (see Section 5.3 in Chapter 5) it is eligible to ask what happens with the disturbed least squares estimator in case of unstandardized data.

Therefore consider the regression model (1.0.1) of Chapter 1 with the unstandardized design matrix $\boldsymbol{X}$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{bmatrix} \in \mathbb{R}^{n\times(p+1)}. \tag{6.6.81}$$

The disturbed least squares estimator is then given by

$$\tilde{\boldsymbol{\beta}}_\omega = \Big((\boldsymbol{X} + \omega\boldsymbol{\Psi})^T (\boldsymbol{X} + \omega\boldsymbol{\Psi}) \Big)^{-1} (\boldsymbol{X} + \omega\boldsymbol{\Psi})^T \boldsymbol{y}. \tag{6.6.82}$$

With Lemma 6.1.7 and Lemma 6.2.3 it follows

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_\omega &= \frac{\omega^2 \tilde{\boldsymbol{M}}_x^{quad} + \omega \tilde{\boldsymbol{M}}_x^{lin} + \tilde{\boldsymbol{M}}_x^{const}}{\omega^2 \boldsymbol{\psi}^T \boldsymbol{A}_x \boldsymbol{\psi} + 2\omega \boldsymbol{b}_x^T \boldsymbol{\psi} + |\boldsymbol{X}^T\boldsymbol{X}|} (\boldsymbol{X} + \omega\boldsymbol{\Psi})^T \boldsymbol{y} \\ &= \frac{\omega^2 (\tilde{\boldsymbol{M}}_x^{quad} \boldsymbol{X}^T\boldsymbol{y} + \tilde{\boldsymbol{M}}_x^{lin} \boldsymbol{\psi}^T \boldsymbol{y}) + \omega (\tilde{\boldsymbol{M}}_x^{lin} \boldsymbol{X}^T \boldsymbol{y} + \tilde{\boldsymbol{M}}_x^{const} \boldsymbol{\psi}^T \boldsymbol{y}) + \tilde{\boldsymbol{M}}_x^{const} \boldsymbol{X}^T \boldsymbol{y}}{\omega^2 \boldsymbol{\psi}^T \boldsymbol{A}_x \boldsymbol{\psi} + 2\omega \boldsymbol{b}_x^T \boldsymbol{\psi} + |\boldsymbol{X}^T\boldsymbol{X}|}. \end{aligned} \tag{6.6.83}$$

The first column of the design matrix $\boldsymbol{X}$ consists only of ones and thus it follows

$$|\boldsymbol{X}_{[j]}{}^T \boldsymbol{X}_{[j]}| = |\boldsymbol{X}^T \boldsymbol{X}_{[j]}| = 0, \quad j = 2, \dots, p$$

and

$$|\boldsymbol{X}_{[1]}{}^T \boldsymbol{X}_{[1]}| = |\boldsymbol{X}^T \boldsymbol{X}_{[1]}| = |\boldsymbol{X}^T\boldsymbol{X}|.$$

With the definitions of Lemma 6.1.1 it is

$$\boldsymbol{b}_x^T = \begin{bmatrix} |\boldsymbol{X}^T\boldsymbol{X}| & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{1\times p},$$

and

$$\boldsymbol{A}_x = \begin{bmatrix} |\boldsymbol{X}^T\boldsymbol{X}| & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{p\times p}.$$

Hence

$$\begin{aligned} |\boldsymbol{M}_x| = \omega^2 |\boldsymbol{X}^T\boldsymbol{X}| \psi_1^2 + 2\omega |\boldsymbol{X}^T\boldsymbol{X}| \psi_1 + |\boldsymbol{X}^T\boldsymbol{X}| \\ = |\boldsymbol{X}^T\boldsymbol{X}| (\omega^2 \psi_1^2 + 2\omega\psi_1 + 1). \end{aligned} \tag{6.6.84}$$

In the same way we get

$$\begin{aligned} |\boldsymbol{X}_{\{u\}}{}^T \boldsymbol{X}_{\{v\}[r]}| &= 0, \quad v \neq r \neq 1 \\ |\boldsymbol{X}_{\{u\}[r]}{}^T \boldsymbol{X}_{\{v\}[s]}| &= 0, \quad v \neq s \neq 1 \vee u \neq r \neq 1. \end{aligned} \tag{6.6.85}$$

With the help of (6.6.85) we get

$$\begin{aligned}
\tilde{\boldsymbol{b}}_x^{(uv)T}\boldsymbol{\psi}_{\{v\}} &= \begin{bmatrix} |\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{v\}[1]}| & 0 & \dots & 0 \end{bmatrix}\boldsymbol{\psi}_{\{v\}} \\
&= \psi_1|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{v\}[1]}|, \quad v \neq 1, \\
\tilde{\boldsymbol{b}}_x^{(u1)T}\boldsymbol{\psi}_{\{1\}} &= \begin{bmatrix} |\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{1\}[2]}| & \dots & |\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{1\}[p]}| \end{bmatrix}\boldsymbol{\psi}_{\{1\}} \\
&= \sum_{r=2}^{p}\psi_r|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{1\}[r]}|,
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\psi}_{\{1\}}^T\tilde{\boldsymbol{A}}_x^{(11)}\boldsymbol{\psi}_{\{1\}} &= \boldsymbol{\psi}_{\{1\}}{}^T\begin{bmatrix} |\boldsymbol{X}_{\{1\}[2]}{}^T\boldsymbol{X}_{\{1\}[2]}| & \dots & |\boldsymbol{X}_{\{1\}[2]}{}^T\boldsymbol{X}_{\{1\}[p]}| \\ \vdots & \ddots & \vdots \\ |\boldsymbol{X}_{\{1\}[p]}{}^T\boldsymbol{X}_{\{1\}[2]}| & \dots & |\boldsymbol{X}_{\{1\}[p]}{}^T\boldsymbol{X}_{\{1\}[p]}| \end{bmatrix}\boldsymbol{\psi}_{\{1\}} \\
&= \sum_{r=2}^{p}\psi_r^2|\boldsymbol{X}_{\{1\}[r]}{}^T\boldsymbol{X}_{\{1\}[r]}| + 2\sum_{s=3}^{p}\sum_{r=2}^{s-1}\psi_r\psi_s|\boldsymbol{X}_{\{1\}[r]}{}^T\boldsymbol{X}_{\{1\}[s]}|, \\
\boldsymbol{\psi}_{\{u\}}{}^T\tilde{\boldsymbol{A}}_x^{(uv)}\boldsymbol{\psi}_{\{v\}} &= \boldsymbol{\psi}_{\{u\}}{}^T\begin{bmatrix} |\boldsymbol{X}_{\{u\}[1]}{}^T\boldsymbol{X}_{\{v\}[1]}| & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}\boldsymbol{\psi}_{\{v\}} \\
&= \psi_1^2|\boldsymbol{X}_{\{u\}[1]}{}^T\boldsymbol{X}_{\{v\}[1]}|, \quad u, v \neq 1, \\
\boldsymbol{\psi}_{\{1\}}{}^T\tilde{\boldsymbol{A}}_x^{(1v)}\boldsymbol{\psi}_{\{v\}} &= \boldsymbol{\psi}_{\{1\}}{}^T\begin{bmatrix} |\boldsymbol{X}_{\{1\}[2]}{}^T\boldsymbol{X}_{\{v\}[1]}| & 0 & \dots & 0 \\ |\boldsymbol{X}_{\{1\}[3]}{}^T\boldsymbol{X}_{\{v\}[1]}| & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ |\boldsymbol{X}_{\{1\}[p]}{}^T\boldsymbol{X}_{\{v\}[1]}| & 0 & \dots & 0 \end{bmatrix}\boldsymbol{\psi}_{\{v\}} \\
&= \psi_1\sum_{r=2}^{p}\psi_r|\boldsymbol{X}_{\{1\}[r]}{}^T\boldsymbol{X}_{\{v\}[1]}|, \quad v \neq 1, \\
\boldsymbol{\psi}_{\{u\}}{}^T\tilde{\boldsymbol{A}}_x^{(u1)}\boldsymbol{\psi}_{\{1\}} &= \boldsymbol{\psi}_{\{u\}}{}^T\begin{bmatrix} |\boldsymbol{X}_{\{u\}[1]}{}^T\boldsymbol{X}_{\{1\}[2]}| & \dots & |\boldsymbol{X}_{\{u\}[1]}{}^T\boldsymbol{X}_{\{1\}[p]}| \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{bmatrix}\boldsymbol{\psi}_{\{1\}} \\
&= \psi_1\sum_{r=2}^{p}\psi_r|\boldsymbol{X}_{\{u\}[1]}{}^T\boldsymbol{X}_{\{1\}[r]}|, \quad u \neq 1. \qquad (6.6.86)
\end{aligned}$$

The following lemma helps to simplify expression (6.6.83) for the disturbed least squares estimator in the case of a regression model with intercept.

Lemma 6.6.1. *With the notation of Lemma 6.1.7 and $\boldsymbol{X}$ given as in (6.6.81) the following equations hold for $u \neq 1$*

$$\left[\tilde{\boldsymbol{M}}_x^{quad}\boldsymbol{X}^T + \tilde{\boldsymbol{M}}_x^{lin}\boldsymbol{\Psi}^T\right]_{\substack{1\le u\le p\\1\le v\le n}} = \psi_1^2\left[\tilde{\boldsymbol{M}}_x^{const}\boldsymbol{X}^T\right]_{\substack{1\le u\le p\\1\le v\le n}}, \tag{6.6.87}$$

$$\left[\tilde{\boldsymbol{M}}_x^{lin}\boldsymbol{X}^T + \tilde{\boldsymbol{M}}_x^{const}\boldsymbol{\Psi}^T\right]_{\substack{1\le u\le p\\1\le v\le n}} = 2\psi_1\left[\tilde{\boldsymbol{M}}_x^{const}\boldsymbol{X}^T\right]_{\substack{1\le u\le p\\1\le v\le n}}. \tag{6.6.88}$$

Proof. Proof of (6.6.87): Consider the (u,v)–th element of $\tilde{\boldsymbol{M}}_x^{quad}\boldsymbol{X}^T \in \mathbb{R}^{p\times n}$. With (6.6.86) it is for $u \neq 1$

$$\begin{aligned}
\tilde{\boldsymbol{M}}_x^{quad}\boldsymbol{X}^T(u,v) &= \sum_{r=1}^{p}(-1)^{u+r}x_{v,r}\boldsymbol{\psi}_{\{u\}}{}^T\tilde{\boldsymbol{A}}_x^{(ur)}\boldsymbol{\psi}_{\{r\}}\\
&= (-1)^{u+1}x_{v,1}\boldsymbol{\psi}_{\{u\}}{}^T\tilde{\boldsymbol{A}}_x^{(u1)}\boldsymbol{\psi}_{\{1\}} + \sum_{r=2}^{p}(-1)^{u+r}x_{v,r}\boldsymbol{\psi}_{\{u\}}{}^T\tilde{\boldsymbol{A}}_x^{(ur)}\boldsymbol{\psi}_{\{r\}}\\
&= (-1)^{u+1}\psi_1 x_{v,1}\sum_{r=2}^{p}\psi_r|\boldsymbol{X}_{\{u\}[1]}{}^T\boldsymbol{X}_{\{1\}[r]}|\\
&\quad + \psi_1^2\sum_{r=2}^{p}(-1)^{u+r}x_{v,r}|\boldsymbol{X}_{\{u\}[1]}{}^T\boldsymbol{X}_{\{r\}[1]}|.
\end{aligned}$$

With (6.6.85) and Lemma 6.2.2 it follows

$$\begin{aligned}
\tilde{\boldsymbol{M}}_x^{quad}\boldsymbol{X}^T(u,v) &= (-1)^{u+1}\psi_1\sum_{r=2}^{p}\psi_r|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{1\}[r]}|\\
&\quad + \psi_1^2\sum_{r=2}^{p}(-1)^{u+r}x_{v,r}|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{r\}}|\\
&= \psi_1\sum_{r=2}^{p}(-1)^{u+r+1}\psi_r|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{r\}}| + \psi_1^2\sum_{r=2}^{p}(-1)^{u+r}x_{v,r}|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{r\}}|,
\end{aligned}$$

because $x_{v,1} = 1,\ v = 1,\ldots,n$. With a similar argumentation we get for the (u,v)–th element of $\tilde{\boldsymbol{M}}_x^{lin}\boldsymbol{\Psi}^T$

$$\begin{aligned}
\tilde{\boldsymbol{M}}_x^{lin}\boldsymbol{\Psi}^T(u,v) &= \sum_{r=1}^{p}(-1)^{u+r}\psi_r\left(\tilde{\boldsymbol{b}}_x^{(ur)T}\boldsymbol{\psi}_{\{r\}} + \tilde{\boldsymbol{b}}_x^{(ru)T}\boldsymbol{\psi}_{\{u\}}\right)\\
&= (-1)^{u+1}\psi_1\left(\tilde{\boldsymbol{b}}_x^{(u1)T}\boldsymbol{\psi}_{\{1\}} + \tilde{\boldsymbol{b}}_x^{(1u)T}\boldsymbol{\psi}_{\{u\}}\right)\\
&\quad + \sum_{r=2}^{p}(-1)^{u+r}\psi_r\left(\tilde{\boldsymbol{b}}_x^{(ur)T}\boldsymbol{\psi}_{\{r\}} + \tilde{\boldsymbol{b}}_x^{(ru)T}\boldsymbol{\psi}_{\{u\}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= (-1)^{u+1}\psi_1 \left(\sum_{r=2}^{p} \psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{1\}[r]}| + \psi_1 |\boldsymbol{X}_{\{1\}}{}^{T}\boldsymbol{X}_{\{u\}[1]}| \right) \\
&\quad + \psi_1 \sum_{r=2}^{p} (-1)^{u+r}\psi_r \left(|\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}[1]}| + |\boldsymbol{X}_{\{r\}}{}^{T}\boldsymbol{X}_{\{u\}[1]}| \right) \\
&= (-1)^{u+1}\psi_1^2 x_{v,1} |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{1\}}| + (-1)^{u+1}\psi_1 \sum_{r=2}^{p} \psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{1\}[r]}| \\
&\quad + 2\psi_1 \sum_{r=2}^{p} (-1)^{u+r}\psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}| \\
&= (-1)^{u+1}\psi_1^2 x_{v,1} |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{1\}}| + (-1)^{u+1}\psi_1 \sum_{r=2}^{p} (-1)^{r}\psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}| \\
&\quad + 2\psi_1 \sum_{r=2}^{p} (-1)^{u+r}\psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}| \\
&= (-1)^{u+1}\psi_1^2 x_{v,1} |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{1\}}| + \psi_1 \sum_{r=2}^{p} (-1)^{u+r}\psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}|.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\left(\tilde{\boldsymbol{M}}_x^{quad}\boldsymbol{X}^T + \tilde{\boldsymbol{M}}_x^{lin}\boldsymbol{\Psi}^T \right)(u,v) &= \psi_1^2 \sum_{r=1}^{p} (-1)^{u+r} x_{v,r} |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}| \\
&= \psi_1^2 \tilde{\boldsymbol{M}}_x^{cons}\boldsymbol{X}^T(u,v)
\end{aligned}
$$

completes the proof.

Proof of (6.6.88): Analogous to the previous proof we have

$$
\begin{aligned}
\tilde{\boldsymbol{M}}_x^{lin}\boldsymbol{X}^T(u,v) &= \sum_{r=1}^{p} (-1)^{u+r} x_{v,r} \left(\tilde{\boldsymbol{b}}_x^{(ur)T}\boldsymbol{\psi}_{\{r\}} + \tilde{\boldsymbol{b}}_x^{(ru)T}\boldsymbol{\psi}_{\{u\}} \right) \\
&= (-1)^{u+1} x_{v,1} \left(\tilde{\boldsymbol{b}}_x^{(u1)T}\boldsymbol{\psi}_{\{1\}} + \tilde{\boldsymbol{b}}_x^{(1u)T}\boldsymbol{\psi}_{\{u\}} \right) \\
&\quad + \sum_{r=2}^{p} (-1)^{u+r} x_{v,r} \left(\tilde{\boldsymbol{b}}_x^{(ur)T}\boldsymbol{\psi}_{\{r\}} + \tilde{\boldsymbol{b}}_x^{(ru)T}\boldsymbol{\psi}_{\{u\}} \right) \\
&= (-1)^{u+1}\psi_1 |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{1\}}| + \sum_{r=2}^{p} (-1)^{u+r+1}\psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}| \\
&\quad + 2\psi_1 \sum_{r=2}^{p} (-1)^{u+r} x_{v,r} |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}|, \\
\tilde{\boldsymbol{M}}_x^{const}\boldsymbol{\Psi}^T(u,v) &= (-1)^{u+1}\psi_1 |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{1\}}| + \sum_{r=2}^{p} (-1)^{u+r}\psi_r |\boldsymbol{X}_{\{u\}}{}^{T}\boldsymbol{X}_{\{r\}}|
\end{aligned}
$$

and thus

$$
\begin{aligned}
\left(\tilde{\boldsymbol{M}}_x^{lin}\boldsymbol{X}^T + \tilde{\boldsymbol{M}}_x^{const}\boldsymbol{\Psi}^T\right)(u,v) &= 2(-1)^{u+1}\psi_1 x_{v,1}|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{1\}}| \\
&\quad + 2\psi_1\sum_{r=2}^{p}(-1)^{u+r}x_{v,r}|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{r\}}| \\
&= 2\psi_1\sum_{r=1}^{p}(-1)^{u+r}x_{v,r}|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{r\}}| \\
&\qquad\qquad = 2\psi_1\tilde{\boldsymbol{M}}_x^{const}\boldsymbol{X}^T(u,v).
\end{aligned}
$$

□

With the help of Lemma 6.6.1 we get the following corollary.

Corollary 6.6.2. *For a regression model with intercept (1.0.1) we get*

$$\tilde{\beta}_j = \hat{\beta}_j, \quad j = 2,\ldots,p,$$

i.e. excluding the intercept, the coefficients of the disturbed least squares estimator are equal to the corresponding ones of the least squares estimator.

Proof. Let $\tilde{\boldsymbol{\beta}}^{\omega}_{\{\tilde{\beta}_0\}}$ denote the vector, which contains all coefficients of the disturbed least squares estimator $\tilde{\boldsymbol{\beta}}_{\omega}$, except the coefficient for the intercept. With (6.6.83) and Lemma 6.6.1 we get

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}}^{\omega}_{\{\tilde{\beta}_0\}} &= \frac{1}{|\boldsymbol{X}^T\boldsymbol{X}|(\omega^2\psi_1^2+2\omega\psi_1+1)}\Big(\omega^2(\tilde{\boldsymbol{M}}_x^{quad}{}_{\{1\}}\boldsymbol{X}^T + \tilde{\boldsymbol{M}}_x^{lin}{}_{\{1\}}\boldsymbol{\Psi}^T)\boldsymbol{y} \\
&\qquad + \omega(\tilde{\boldsymbol{M}}_x^{lin}{}_{\{1\}}\boldsymbol{X}^T + \tilde{\boldsymbol{M}}_x^{const}{}_{\{1\}}\boldsymbol{\Psi}^T)\boldsymbol{y} + \tilde{\boldsymbol{M}}_x^{const}{}_{\{1\}}\boldsymbol{X}^T\boldsymbol{y}\Big) \\
&= \frac{\omega^2\psi_1^2\tilde{\boldsymbol{M}}_1^{const}{}_{\{1\}}\boldsymbol{X}^T\boldsymbol{y} + 2\omega\psi_1\tilde{\boldsymbol{M}}_x^{const}{}_{\{1\}}\boldsymbol{X}^T\boldsymbol{y} + \tilde{\boldsymbol{M}}_x^{const}{}_{\{1\}}\boldsymbol{X}^T\boldsymbol{y}}{|\boldsymbol{X}^T\boldsymbol{X}|(\omega^2\psi_1^2+2\omega\psi_1+1)} \\
&\qquad\qquad = \frac{\tilde{\boldsymbol{M}}_x^{const}{}_{\{1\}}\boldsymbol{X}^T\boldsymbol{y}}{|\boldsymbol{X}^T\boldsymbol{X}|} = \hat{\boldsymbol{\beta}}_{\{\beta_0\}},
\end{aligned}
$$

where $\tilde{\boldsymbol{M}}_x^{quad}{}_{\{1\}}$, $\tilde{\boldsymbol{M}}_x^{lin}{}_{\{1\}}$ and $\tilde{\boldsymbol{M}}_x^{const}{}_{\{1\}}$ denote $(p-1)\times p$ submatrices of the original matrices, obtained by striking out the first row.

□

The covariance matrix of (6.6.82) is given by

$$\Sigma(\tilde{\boldsymbol{\beta}}_{\omega}) = \sigma^2\left((\boldsymbol{X}+\omega\boldsymbol{\Psi})^T(\boldsymbol{X}+\omega\boldsymbol{\Psi})\right)^{-1}$$

and it follows with Lemma 6.1.7

$$\Sigma(\tilde{\boldsymbol{\beta}}_{\omega}) = \sigma^2\frac{\omega^2\tilde{\boldsymbol{M}}_x^{quad} + \omega\tilde{\boldsymbol{M}}_x^{lin} + \tilde{\boldsymbol{M}}_x^{const}}{\omega^2\boldsymbol{\psi}^T\boldsymbol{A}_x\boldsymbol{\psi} + 2\omega\boldsymbol{b}_x^T\boldsymbol{\psi} + |\boldsymbol{X}^T\boldsymbol{X}|}. \tag{6.6.89}$$

With the equations given in (6.6.86) we get for the j–th, $j = 2, \dots, p$ diagonal element of (6.6.89)

$$\begin{aligned}\operatorname{var}(\tilde{\beta}_j^{\omega}) &= \sigma^2 \frac{\omega^2 {\tilde{m}_x^{quad}}_{j,j} + \omega {\tilde{m}_x^{lin}}_{j,j} + {\tilde{m}_x^{const}}_{j,j}}{\omega^2 \boldsymbol{\psi}^T \boldsymbol{A}_x \boldsymbol{\psi} + 2\omega \boldsymbol{b}_x^T \boldsymbol{\psi} + |\boldsymbol{X}^T \boldsymbol{X}|} \\ &= \sigma^2 \frac{|\boldsymbol{X}_{\{j\}}^T \boldsymbol{X}_{\{j\}}|(\omega^2 \psi_1^2 + 2\omega\psi_1 + 1)}{|\boldsymbol{X}^T \boldsymbol{X}|(\omega^2 \psi_1^2 + 2\omega\psi_1 + 1)} = \operatorname{var}(\hat{\beta}_j).\end{aligned}$$

Thus we cannot get any improvement by using the disturbed model. Hence in case of a regression model including an intercept, we have to standardize or at least to center the design matrix $\boldsymbol{X}$.

6.7. The DLSE as Ridge Regression Estimator

In (5.4.17) we introduced the generalized ridge estimator of C.R. Rao. He proposed adding a positive semidefinite matrix $\boldsymbol{H}$ on $\boldsymbol{X}^T \boldsymbol{X}$. Hence the disturbed least squares estimator (6.2.27) is a special case of the generalized ridge estimator, with

$$\boldsymbol{H} = \boldsymbol{\Psi}^T \boldsymbol{\Psi}.$$

But only due to the special structure of $\boldsymbol{\Psi}$ the calculation of the inverse of $\boldsymbol{M} = (\boldsymbol{Z}^T \boldsymbol{Z} + \omega^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi})$ could be simplified. In the proof of Lemma 6.1.1 or Lemma 6.1.7, the application of Theorem A.4.1 about the determinant of the sum of two matrices only simplified, because of $\boldsymbol{\Psi}$ having rank one. As a consequence the disturbed least squares estimator and its properties (e.g. its mean squared error) could be described in dependence of ω and/or $\boldsymbol{\psi}$. For another positive semidefinite matrix these calculations would be more complicated or even impossible. This fact is the main advantage of the disturbed least squares estimator.

As shown in Section 6.6, the disturbed least squares estimator is not applicable to unstandardized (or uncentered) data including an intercept. But then it is possible to apply our results on the generalized ridge estimator of C.R. Rao

$$\hat{\boldsymbol{\beta}}_{rao} := \left(\boldsymbol{X}^T \boldsymbol{X} + \omega^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{6.7.90}$$

and to describe it in dependence of ω^2. For (6.7.90), Corollary 6.1.9 is also valid for unstandardized data.

Corollary 6.7.1. *For an arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $n \geq p$, $p \geq 2$ and $\boldsymbol{\Psi}$ defined like in (6.1.2) we have*

$$\left(\boldsymbol{X}^T \boldsymbol{X} + \omega^2 \boldsymbol{\Psi}^T \boldsymbol{\Psi}\right)^{-1} = \frac{n\omega^2 \tilde{\boldsymbol{M}}^{quad} + \tilde{\boldsymbol{M}}^{const}}{n\omega^2 \boldsymbol{\psi}^T \tilde{\boldsymbol{M}}^{const} \boldsymbol{\psi} + |\boldsymbol{X}^T \boldsymbol{X}|}, \quad \omega \in \mathbb{R},$$

with

$$\tilde{\boldsymbol{M}}^{quad} = \left[\tilde{m}_{u,v}^{quad}\right]_{1\leq u,v\leq p} = \left[(-1)^{u+v}\boldsymbol{\psi}_{\{u\}}{}^T\tilde{\boldsymbol{A}}^{(uv)}\boldsymbol{\psi}_{\{v\}}\right]_{1\leq u,v\leq p},$$

$$\tilde{\boldsymbol{M}}^{const} = \left[\tilde{m}_{u,v}^{const}\right]_{1\leq u,v\leq p} = \left[(-1)^{u+v}\left|\boldsymbol{X}_{\{u\}}{}^T\boldsymbol{X}_{\{v\}}\right|\right]_{1\leq u,v\leq p}$$

and

$$\boldsymbol{\psi}_{\{u\}} = \left[\psi_r\right]_{\substack{1\leq r\leq p\\ r\neq u}} \in \mathbb{R}^{(p-1)\times 1},$$

$$\tilde{\boldsymbol{A}}^{(uv)} = \begin{cases} 1 & p=2 \\ \left[(-1)^{r+s}|\boldsymbol{X}_{\{ur\}}{}^T\boldsymbol{X}_{\{vs\}}|\right]_{\substack{1\leq r,s\leq p\\ u\neq r;v\neq s}} & p\geq 3\end{cases} \in \mathbb{R}^{(p-1)\times(p-1)}.$$

analogous to Corollary 6.1.9 in Chapter 6.

As a consequence all results of Section 6.1 until 6.4 can be carried to the generalized ridge estimator of C.R. Rao for $\boldsymbol{H} = \boldsymbol{\Psi}^T\boldsymbol{\Psi}$.

6.8. Example: The DLSE of the Economic Data

With (6.2.27) the disturbed least squares estimator of the standardized Economic Data of Example 4.4 is given by

$$\tilde{\gamma}_\omega = (\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\boldsymbol{y}^*,$$

with

$$\boldsymbol{\Psi} = \begin{bmatrix}\psi_1 & \psi_2 & \psi_3\\ \vdots & \vdots & \vdots \\ \psi_1 & \psi_2 & \psi_3\end{bmatrix} \in \mathbb{R}^{17\times 3}$$

and $\boldsymbol{\psi}^T = \begin{bmatrix}\psi_1, & \psi_2, & \psi_3\end{bmatrix}$. In contrast to the ridge estimator of Chapter 5, the disturbed least squares estimator depends on the three unknown parameters ψ_1, ψ_2, ψ_3, which have to be estimated. Therefore we will also use some of the methods for choosing the biasing factor k, introduced in Section 5.2.

- The first possibility is to determine the matrix $\boldsymbol{\Psi}_{\min}$ in a way, such that the estimated mean squared error is minimized. From (6.2.30) and (6.2.34) we know, that the mean squared error of $\tilde{\gamma}_\omega$ can be written as

$$\begin{aligned}\mathrm{MSE}(\tilde{\gamma}_\omega) &= \mathrm{tr}\,(\Sigma(\tilde{\gamma}_\omega)) + \mathrm{Bias}^T(\tilde{\gamma}_\omega)\mathrm{Bias}(\tilde{\gamma}_\omega)\\ &= \sigma^2\mathrm{tr}\left((\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\right)\\ &\qquad + \omega^4\boldsymbol{\gamma}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\boldsymbol{Z}^T\boldsymbol{Z}+\omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-2}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\boldsymbol{\gamma} \quad .\end{aligned}$$

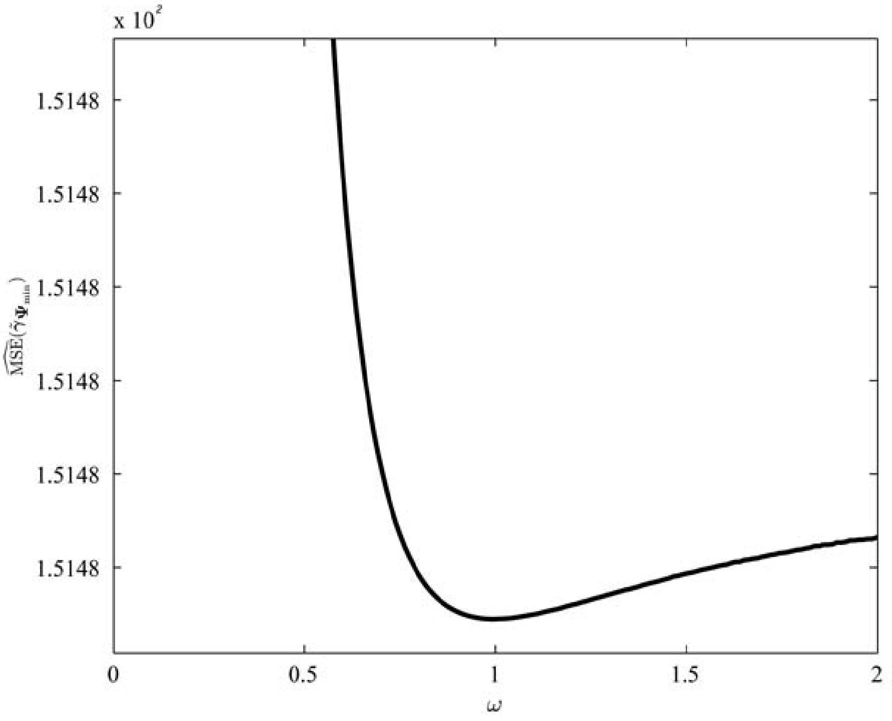


FIGURE 6.8.6. Estimated mean squared error in dependence of ω

Of course σ^2 and $\boldsymbol{\gamma}$ are unknown parameters, which have to be estimated. From Table 5.5.4 we know

$$\hat{\boldsymbol{\gamma}}^T = \begin{bmatrix}-19.106, & 24.356, & 6.421\end{bmatrix}$$

and with the help of Table 4.4.3 we get

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}})}{n-p-1} = \frac{11.374}{17-3-1} = 0.8745.$$

Thus an estimator for the mean squared error is given by

$$\widehat{\mathrm{MSE}}(\tilde{\boldsymbol{\gamma}}_\omega) = \hat{\sigma}^2 \mathrm{tr}\left((\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1}\right) \\ + \omega^4\hat{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}^T\boldsymbol{\Psi}(\boldsymbol{Z}^T\boldsymbol{Z} + \omega^2\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-2}\boldsymbol{\Psi}^T\boldsymbol{\Psi}\hat{\boldsymbol{\gamma}}. \qquad (6.8.91)$$

For $\omega = 0$ we get the estimated mean squared error of the least squares estimator

$$\widehat{\mathrm{MSE}}(\hat{\boldsymbol{\gamma}}) = \hat{\sigma}^2 \mathrm{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} = 927.18.$$

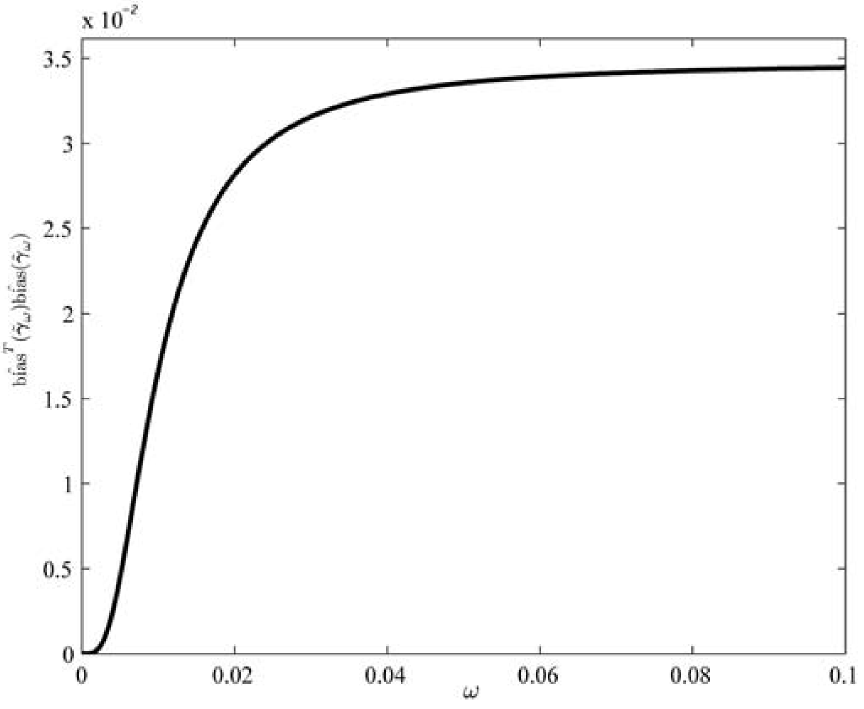


FIGURE 6.8.7. Estimated squared bias in dependence of ω

With the help of `MATLAB` we can find the matrix

$$\mathbf{\Psi}_{\text{min}} = \begin{bmatrix} 4.0149 & 2.4507 & 2.6137 \\ \vdots & \vdots & \vdots \\ 4.0149 & 2.4507 & 2.6137 \end{bmatrix},$$

which minimizes the estimated mean squared error $\widehat{\text{MSE}}(\tilde{\gamma}_\omega)$ for $\omega = 1$. The function value of the estimated mean squared error for $\mathbf{\Psi}_{\text{min}}$ is given by

$$\widehat{\text{MSE}}(\tilde{\gamma}_{\mathbf{\Psi}_{\text{min}}}) = 151.48,$$

whereas in case of the ridge estimator we got in (5.5.34)

$$\widehat{\text{MSE}}(\hat{\gamma}_{k_{\text{min}}}) = 517.06. \tag{6.8.92}$$

Figure 6.8.6 displays the estimated mean squared error in dependence of ω and Figure 6.8.7 shows the squared bias in dependence of ω. In contrast to the total variance, the squared bias of $\tilde{\gamma}_{\mathbf{\Psi}_{\text{min}}}$ is negligible small and the improvement of the estimated mean squared error is mainly due to the improvement of the total variance. In case of the ridge estimator in Section 5.5 the squared bias has an essential influence on the estimated mean squared error for non–zero k (see Figure 6.8.91). Taking $\boldsymbol{\Psi}_{\text{min}}$, the disturbed least squares estimator of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix}5.5256, & -4.2966, & 3.1546, & 0.002855\end{bmatrix}^T .$$

For the ridge estimator there is only the biasing factor k, which has to be estimated. Therefore suppose

$$\boldsymbol{\Psi}^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{17\times 4}.$$

Now ω is also the only unknown parameter. From (6.3.54) we know that

$$\omega_{\text{min}} = \sqrt{\frac{\sigma^2}{n\psi_1^2\hat{\gamma}_1^2}} = 0.012$$

and we get

$$\widehat{\text{MSE}}(\tilde{\gamma}_{\omega_{\text{min}}}) = 470.72.$$

Thus the estimated mean squared error is still smaller than the corresponding one of the ridge estimator given in (6.8.92).

- We can also apply the method of McDonald and Galarneau (see Section 5.2.2, (3)) to get an unbiased estimator of $\boldsymbol{\gamma}^T\boldsymbol{\gamma}$. From Section 5.5 we have

$$\boldsymbol{Q}_Z = \hat{\boldsymbol{\gamma}}^T\hat{\boldsymbol{\gamma}} - \hat{\sigma}_Z^2 \text{tr}\left((\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\right) = 138,$$

where $\hat{\sigma}_Z^2$ is given in (5.5.31). With the help of `MATLAB` we find

$$\hat{\boldsymbol{\Psi}} = \begin{bmatrix} 0.0084273 & -0.010635 & -0.0021998 \\ \vdots & \vdots & \vdots \\ 0.0084273 & -0.010635 & -0.0021998 \end{bmatrix},$$

such that

$$\tilde{\gamma}_\omega(\hat{\boldsymbol{\Psi}})^T\tilde{\gamma}_\omega(\hat{\boldsymbol{\Psi}}) \approx \boldsymbol{Q}$$

for $\omega = 1$. Then we have

$$\widehat{\text{MSE}}(\tilde{\gamma}_\omega) = 794.64,$$

for $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$. Transforming back implies

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix}-4.707, & 0.019415, & 1.5279, & -8.5845\cdot 10^{-5}\end{bmatrix}^T .$$

CHAPTER 7

Simulation Study

In the following simulation study we try to evaluate the performance of the proposed disturbed least squares estimator (6.2.27) compared to the ridge estimator and the least squares estimator. The simulation design will essentially follow those of D. Trenkler and G. Trenkler (1984,[**60**]), which is geared to the approach and simulation study of McDonald and Galarneau (1975,[**37**]), mentioned in (3) of Section 5.2.2.
Consider the following linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_5 x_{i,5} + \varepsilon_i, \quad i = 1, \ldots, 30$$

or in vector notation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{7.0.1}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{30})$. Our aim will be to compute regression models of the form (7.0.1) with the help of random numbers for the design matrix
$\boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{30} & X_1 & \ldots & X_5 \end{bmatrix}$ and the error vector $\boldsymbol{\varepsilon}$. These should be generated for different values of σ^2 and $\boldsymbol{\beta}$. To examine the performance of the estimators also for different degrees of multicollinearity, the design matrix $\boldsymbol{X}$ should be computed in dependence of the pairwise correlation of the regressors.
Therefore consider the random variables

$$\begin{aligned} R_1 &:= \sqrt{(1-\rho^2)} U_1 + \rho U_3, \\ R_2 &:= \sqrt{(1-\rho_*^2)} U_2 + \rho_* U_3, \quad \rho, \rho_* \in \mathbb{R}, \end{aligned}$$

where U_1, U_2 and U_3 are independent, standard normal distributed random variables. It is

$$\begin{aligned} \mathrm{E}(R_i) &= 0, \\ \mathrm{var}(R_i) &= 1, \quad i = 1, 2. \end{aligned}$$

Because of the independence of U_1 and U_2 it follows

$$\begin{aligned} \mathrm{cov}(R_1, R_2) &= \mathrm{E}\left((\sqrt{(1-\rho^2)} U_1 + \rho U_3)(\sqrt{(1-\rho_*^2)} U_2 + \rho_* U_3) \right) \\ &= \mathrm{E}(\sqrt{(1-\rho^2)}\sqrt{(1-\rho_*^2)} U_1 U_2) + \mathrm{E}(\sqrt{(1-\rho^2)} \rho_* U_1 U_3) + \mathrm{E}(\sqrt{(1-\rho_*^2)} \rho U_2 U_3) \\ &\quad + \mathrm{E}(\rho_* \rho U_3^2) = \mathrm{var}(\sqrt{\rho_* \rho} U_3) = \rho_* \rho. \end{aligned} \tag{7.0.2}$$

Hence we get for the correlation of R_1 and R_2

$$\operatorname{corr}(R_1, R_2) = \rho_* \rho.$$

This structure of random variables will be helpful to compute correlated random numbers for the regressors in the simulation study.
To test the performance of the ridge and disturbed least squares estimator for different regression models, we have to find suitable estimates of k and $\boldsymbol{\psi}$. Therefore we will use the approach of G. Trenkler (1981,[**58**]) (see also Section 5.2.2, (3)), i.e. we will determine k and $\boldsymbol{\Psi}$ in a way, such that

$$\boldsymbol{g}^T \boldsymbol{g} \approx \operatorname{abs}\left(\hat{\gamma}^T \hat{\gamma} - \hat{\sigma}^2 \operatorname{tr}\left(\boldsymbol{Z}^T \boldsymbol{Z}\right)^{-1}\right),$$

where $\boldsymbol{g}$ denotes the ridge or disturbed least squares estimator of the standardized model

$$\boldsymbol{y}^* = \boldsymbol{Z}\gamma + \varepsilon^*$$

of (7.0.1).

7.1. The Algorithm

Consider the following steps for computing the desired regression models:

Step 0: A set of standard normal distributed random vectors $N_j \in \mathbb{R}^{30\times 1}$, $j = 1, \ldots, 6$ is generated.

Step 1: For fixed ρ and $\rho_* \in \mathbb{R}$, the regressors are computed in the following way:

$$X_j \quad := \begin{cases} \mathbf{1}_{30} & , j = 1 \\ \sqrt{(1-\rho^2)} N_j + \rho N_6 & , 1 \leq j \leq 3 \\ \sqrt{(1-\rho_*^2)} N_j + \rho_* N_6 & , 4 \leq j \leq 5 \end{cases} . \tag{7.1.3}$$

$\rho_* \backslash \rho$	0	0.3	0.5	0.7	0.8	0.9	0.95	0.99	0.995
0	1.00	1.30	2.00	3.88	6.33	13.79	28.77	148.75	298.75
0.3	1.20	1.49	2.29	4.34	7.00	15.09	31.32	161.30	323.80
0.5	1.67	2.11	2.67	4.99	7.98	17.01	35.15	180.34	361.85
0.7	2.92	3.67	4.61	5.80	9.20	19.47	40.04	204.77	410.69
0.8	4.56	5.67	7.09	8.87	9.89	20.84	42.80	218.56	438.27
0.9	9.53	11.72	14.57	18.11	20.13	22.32	45.74	233.27	467.69
0.95	19.51	23.87	29.56	36.62	40.64	44.99	47.28	240.96	483.06
0.99	99.50	121.12	149.55	184.80	204.82	226.43	237.84	247.26	495.66
0.995	199.50	242.69	299.55	370.03	410.04	453.24	476.04	494.86	497.25

TABLE 7.0.1. Condition numbers of $\boldsymbol{C}(\boldsymbol{X}_{\{1\}})$

With (7.0.2) it is easy to see that the theoretical correlation matrix of $\boldsymbol{X}_{\{1\}} := \begin{bmatrix} X_1 & \dots & X_5 \end{bmatrix}$ is given by

$$\boldsymbol{C}(\boldsymbol{X}_{\{1\}}) := \begin{bmatrix} 1 & \rho^2 & \rho^2 & \rho\rho_* & \rho\rho_* \\ \rho^2 & 1 & \rho^2 & \rho\rho_* & \rho\rho_* \\ \rho^2 & \rho^2 & 1 & \rho\rho_* & \rho\rho_* \\ \rho\rho_* & \rho\rho_* & \rho\rho_* & 1 & \rho_*^2 \\ \rho\rho_* & \rho\rho_* & \rho\rho_* & \rho_*^2 & 1 \end{bmatrix}. \tag{7.1.4}$$

The correlations ρ and ρ_* take the following nine different values, reflecting the range from none to extreme multicollinearity

$$\rho,\ \rho_* = 0, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95, 0.99, 0.995.$$

Hence in total we consider 81 different combinations of ρ and ρ_*. Table 7.0.1 displays the condition numbers of $\boldsymbol{C}(\boldsymbol{X}_{\{1\}})$ in dependence of ρ and ρ_*.

Step 2: For each set of regressors constructed that way, two choices for the true coefficient $\boldsymbol{\beta}$ in (7.0.1) are considered.

$\phi = 0$: $\beta_0 = 0$ and $\boldsymbol{\beta}_{\{\beta_0\}} = V_{\min}$, where $V_{\min}$ is the normalized eigenvector belonging to smallest eigenvalue $\lambda_{\min}$ of the correlation matrix of $\boldsymbol{X}_{\{1\}}$ given in (7.1.4).

$\phi = 1$: $\beta_0 = 0$ and $\boldsymbol{\beta}_{\{\beta_0\}} = V_{\max}$, where $V_{\max}$ is the normalized eigenvector belonging to largest eigenvalue $\lambda_{\max}$ of the correlation matrix of $\boldsymbol{X}_{\{1\}}$ given in (7.1.4).

As mentioned in Note 5.2.3 these choices minimize or maximize the improvement of the ridge estimator compared to the least squares estimator.

Step 3: Observations on the dependent variable are determined by (7.0.1) for $\boldsymbol{\beta}^T := \begin{bmatrix} 0, & V_{\min} \end{bmatrix}$ and $\boldsymbol{\beta}^T := \begin{bmatrix} 0, & V_{\max} \end{bmatrix}$ and for the following seven different values of σ^2

$$\sigma^2 = 0.01, 0.1, 0.3, 0.5, 1, 3, 5.$$

Normally distributed random numbers with mean 0 and variance σ^2 are used as realizations for ε_i in (7.0.1).

To illustrate the proceeding until Step 3, define by $\boldsymbol{\tau} \in \mathbb{R}^{1134\times 4}$ the matrix, which contains all possible realizations of the tupel $(\rho, \rho_*, \sigma^2, \phi)$, i.e.

$$\boldsymbol{\tau} := \begin{bmatrix} 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0.01 & 1 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0.1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0.995 & 0.995 & 5 & 1 \end{bmatrix}. \tag{7.1.5}$$

For each row $\tau(u)$ of $\boldsymbol{\tau}$, a design matrix $\boldsymbol{X}(u)$ and a normal distributed random vector $\boldsymbol{\varepsilon}(u)$ is generated with the help of normally distributed random numbers.

Afterwards 30 samples of the dependent variable $\boldsymbol{y}(u)$ can be computed for each row $\tau(u)$ of $\boldsymbol{\tau}$.

Step 4: Once the set of dependent variables is constructed in dependence of $\tau(u)$ we can write

$$\boldsymbol{y}(u) = \boldsymbol{X}(u)\boldsymbol{\beta}(u) + \boldsymbol{\varepsilon}(u). \tag{7.1.6}$$

As mentioned above, the index u should emphasize the dependence of the standardized model on the u–th row of $\boldsymbol{\tau}$ in (7.1.5). From (7.1.6) we can calculate the least squares estimator of $\boldsymbol{\beta}$ and σ^2 for the u–th tupel by

$$\begin{aligned} \hat{\boldsymbol{\beta}}(u) &= \left(\boldsymbol{X}(u)^T \boldsymbol{X}(u)\right)^{-1} \boldsymbol{X}(u)^T \boldsymbol{y}(u) \\ \hat{\boldsymbol{\sigma}}(u)^2 &= \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}}(u))}{30-5-1}. \end{aligned} \tag{7.1.7}$$

Step 5: Then (7.1.6) is standardized according to Section 3.2

$$\boldsymbol{y}^*(u) = \boldsymbol{Z}(u)\boldsymbol{\gamma}(u) + \boldsymbol{\varepsilon}^*(u), \quad u = 1, \ldots, 1134$$

and we get the least squares estimator of the standardized model by

$$\hat{\boldsymbol{\gamma}}(u) = \left(\boldsymbol{Z}(u)^T \boldsymbol{Z}(u)\right)^{-1} \boldsymbol{Z}(u)^T \boldsymbol{y}^*(u).$$

Step 6: The ridge and disturbed least squares estimates are calculated using the standardized models.

The biasing factor $k(u)$ is determined, such that

$$\hat{\boldsymbol{\gamma}}_r(u)^T \hat{\boldsymbol{\gamma}}_r(u) \approx \mathrm{abs}\left(\hat{\boldsymbol{\gamma}}(u)^T \hat{\boldsymbol{\gamma}}(u) - \hat{\sigma}(u)^2 \sum_{j=1}^{p} \frac{1}{\lambda_j(u)}\right), \tag{7.1.8}$$

with the ridge estimator

$$\hat{\boldsymbol{\gamma}}_r(u) = \left(\boldsymbol{Z}(u)^T \boldsymbol{Z}(u) + k(u)\boldsymbol{I}_5\right)^{-1} \boldsymbol{Z}(u)^T \boldsymbol{y}^*(u).$$

abs($\cdot$) denotes the absolute value and $\lambda_j(u),\ j = 1, \ldots, p$ are the eigenvalues of the matrix $\boldsymbol{Z}(u)^T\boldsymbol{Z}(u)$ (which is equal to the correlation matrix $\boldsymbol{C}(\boldsymbol{X}_{\{1\}}(u))$). If we cannot find any $k(u)$, such that (7.1.8) is fulfilled, we choose $k(u) = 0$.

Step 7: In the same way we determine a suitable matrix $\boldsymbol{\Psi}(u)$. If there exists any $\boldsymbol{\Psi}(u)$ such that

$$\tilde{\boldsymbol{\gamma}}(u)^T\tilde{\boldsymbol{\gamma}}(u) \approx \text{abs}\left(\hat{\boldsymbol{\gamma}}(u)^T\hat{\boldsymbol{\gamma}}(u) - \hat{\sigma}(u)^2\sum_{j=1}^{p}\frac{1}{\lambda_j(u)}\right)$$

is fulfilled, we choose the disturbed least squares estimator

$$\tilde{\boldsymbol{\gamma}}(u) = \left(\boldsymbol{Z}(u)^T\boldsymbol{Z}(u) + \boldsymbol{\Psi}(u)^T\boldsymbol{\Psi}(u)\right)^{-1}\boldsymbol{Z}(u)^T\boldsymbol{y}(u)^*.$$

Otherwise we choose the least squares estimator by setting $\boldsymbol{\Psi}(u) = \boldsymbol{0}$. In both cases we take (7.1.7) as an estimator for $\sigma(u)^2$. Hence we do not follow McDonald and Galarneau. As already mentioned in Section 5.2.2, they misleadingly used the residual sum of squares of the standardized model to calculate the least squares estimator of σ^2 instead of RSS($\hat{\boldsymbol{\beta}}(u)$).

Step 8: The estimated coefficients of $\hat{\boldsymbol{\gamma}}_r(u)$ and $\tilde{\boldsymbol{\gamma}}(u)$ are then transformed back into the original model (7.0.1) along the lines of formulae (3.2.14) :

$$\begin{aligned}\hat{\boldsymbol{\beta}}^r_{\{\beta_0\}}(u) &= \left[\hat{\beta}^r_j(u)\right]_{1\leq j\leq 5} := \boldsymbol{D}^{-1}(u)\hat{\boldsymbol{\gamma}}_r(u),\\ \tilde{\boldsymbol{\beta}}_{\{\beta_0\}}(u) &= \left[\tilde{\beta}_j(u)\right]_{1\leq j\leq 5} := \boldsymbol{D}^{-1}(u)\tilde{\boldsymbol{\gamma}}(u) \quad \in \mathbb{R}^{5\times 1},\end{aligned} \tag{7.1.9}$$

where $\boldsymbol{D}^{-1}(u)$ is given by (3.2.12) for the design matrix $\boldsymbol{X}(u)$. The estimates for the intercept are given by

$$\begin{aligned}\hat{\beta}^r_0(u) &:= \bar{y}(u) - \sum_{j=1}^{5}\hat{\beta}^r_j(u)\bar{X}_j(u),\\ \tilde{\beta}_0(u) &:= \bar{y}(u) - \sum_{j=1}^{5}\tilde{\beta}_j(u)\bar{X}_j(u),\end{aligned} \tag{7.1.10}$$

where $\bar{y}(u)$ denotes the mean of $\boldsymbol{y}(u)$ and $\bar{X}_j(u)$ the mean of the j–th column of $\boldsymbol{X}(u)$. Then the ridge estimator of $\boldsymbol{\beta}(u)$ is given by

$$\hat{\boldsymbol{\beta}}^r(u)^T := \left[\hat{\beta}^r_0(u),\quad \hat{\boldsymbol{\beta}}^r_{\{\beta_0\}}(u)^T\right]$$

and the disturbed least squares estimator by

$$\tilde{\boldsymbol{\beta}}(u)^T := \left[\tilde{\beta}_0(u),\quad \tilde{\boldsymbol{\beta}}_{\{\beta_0\}}(u)^T\right].$$

Now all the steps are repeated 1000 times until we have 1000 estimates of $\hat{\boldsymbol{\beta}}(u)$, $\hat{\boldsymbol{\beta}}^r(u)$ and $\tilde{\boldsymbol{\beta}}(u)$. Denote by $\hat{\boldsymbol{\beta}}(u,m)$, $\hat{\boldsymbol{\beta}}^r(u,m)$ and $\tilde{\boldsymbol{\beta}}(u,m)$, $1 \leq m \leq 1000$, the least squares, ridge and disturbed least squares estimators of the m–th run, calculated by (7.1.9) and (7.1.10) for fixed $u, 1 \leq u \leq 1134$.

Biased estimators are constructed with the aim of achieving a smaller mean squared error than the corresponding one of the least squares estimator. A measure of the obtained improvement of an arbitrary estimator $\boldsymbol{b}$ to the least squares estimator is given by

$$r := \frac{\mathrm{E}\left((\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})\right)}{\mathrm{E}\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)}.$$

If $r < 1$ the estimator $\boldsymbol{b}$ has a smaller mean squared error than the least squares estimator. Thus we use the following ratios as a measure of goodness of $\hat{\boldsymbol{\beta}}^r(u,m)$ and $\tilde{\boldsymbol{\beta}}(u,m)$ compared to the least squares estimator in dependence of ρ, ρ_*, σ^2 and ϕ.

$$\hat{\boldsymbol{r}}^r(u) := \frac{\sum_{m=1}^{1000} \left\|\hat{\boldsymbol{\beta}}^r(u,m)-\boldsymbol{\beta}\right\|_2}{\sum_{m=1}^{1000} \left\|\hat{\boldsymbol{\beta}}(u,m)-\boldsymbol{\beta}\right\|_2} \tag{7.1.11}$$

for the ridge estimators and

$$\tilde{\boldsymbol{r}}(u) := \frac{\sum_{m=1}^{1000} \left\|\tilde{\boldsymbol{\beta}}(u,m)-\boldsymbol{\beta}\right\|_2}{\sum_{m=1}^{1000} \left\|\hat{\boldsymbol{\beta}}(u,m)-\boldsymbol{\beta}\right\|_2} \tag{7.1.12}$$

for the disturbed least squares estimator, where $\|\cdot\|_2$ denotes the Euclidean norm. If $\hat{\boldsymbol{r}}^r(u)$ and $\tilde{\boldsymbol{r}}(u)$ are smaller than one, the ridge estimator and disturbed least squares estimator perform better than the least squares estimator for the u–th combination of ρ, ρ_*, σ^2 and ϕ in (7.1.5). Additionally we consider

$$\frac{\tilde{\boldsymbol{r}}(u)}{\hat{\boldsymbol{r}}^r(u)} = \frac{\sum_{m=1}^{1000} \left\|\tilde{\boldsymbol{\beta}}(u,m)-\boldsymbol{\beta}\right\|_2}{\sum_{m=1}^{1000} \left\|\hat{\boldsymbol{\beta}}^r(u,m)-\boldsymbol{\beta}\right\|_2}$$

to examine the performance of the disturbed least squares estimator compared to the ridge estimator.

7.1.1. Implementation

The algorithm is implemented in `MATLAB` and can be found on the attached CD. The steps described in Section 7.1 are programmed in the M–file `algorithm.m`, which can be started with the M-file `start.m` . There the number of simulations (here 1000), the number of observations (here 30) and the chosen values for ρ, ρ_* and σ^2 can be changed.

As output we obtain the two arrays `ridge_r` and `tilde_r` $\in \mathbb{R}^{9\times 9\times 7\times 2}$. They contain the ratios $\hat{\boldsymbol{r}}^r(u)$ and $\tilde{\boldsymbol{r}}(u)$, given in (7.1.11) and (7.1.12), for all combinations of ρ, ρ_*, σ^2 and ϕ.

We consider in more detail the implementation of Step 6 and Step 7, where we try to find an optimal k and $\boldsymbol{\Psi}$. We have to find the root of the function

$$f = \boldsymbol{g}^T\boldsymbol{g} - \text{abs}\left(\hat{\gamma}^T\hat{\gamma} - \hat{\sigma}^2\text{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\right), \tag{7.1.13}$$

where $\boldsymbol{g}$ denotes the ridge or disturbed least squares estimator of the standardized model. Since we cannot ensure the existence of a root of the function f, we only try to minimize the function

$$f^+ = \text{abs}\left(\boldsymbol{g}^T\boldsymbol{g} - \text{abs}\left(\hat{\gamma}^T\hat{\gamma} - \hat{\sigma}^2\text{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\right)\right). \tag{7.1.14}$$

Of course we have to observe that this minimum is close to zero.

For $\boldsymbol{\psi}$, this can be done with the help of the powerful procedure `fminsearch` in `MATLAB` . We write

```
[psi,fval,exitflag]=fminsearch(@function,[0;0;0;0;0],...)
```

where `@function` refers to the m-file `function.m` , which contains the implementation of (7.1.14) for the disturbed least squares estimator. Obviously we choose zero as starting point for the iteration for all components of $\boldsymbol{\psi}$. The dots are only placeholder for further required input arguments.

- The output `fval` returns us the value of the function f^+ at the solution `psi`. Here we have to ensure, that `fval` is small enough, say `fval` $<$ $0.01{\cdot}\boldsymbol{Q}$, with $\boldsymbol{Q} = \text{abs}\left(\hat{\gamma}^T\hat{\gamma} - \hat{\sigma}^2\text{tr}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\right)$. Of course this condition has been chosen arbitrarily.
- `exitflag` describes the exit condition of `fminsearch`. Thus it is possible to check, whether the algorithm has converged to a local minimum. If not we choose $\boldsymbol{\Psi} = \boldsymbol{0}$.

In case of the ridge estimator we have to solve a constrained optimization problem, because we assume $k > 0$. The command

```
[k,fval,exitflag]=fminbnd(@findk,eps,inf,...)
```

returns the minimum of the function `@findk` on the interval $(0, \infty)$. `findk` is equal to (7.1.14) in case of the ridge estimator. The output functions `fval` and `exitflag` are handled in the same way as shown above.

Note 7.1.1. For further information on the procedures `fminsearch` and `fminbnd`, see `http://www.mathworks.com/access/helpdesk/help/toolbox/optim/`.

7.2. The Simulation Results

The main results of the simulation study are presented in two–way tables, which can be found in the folder `results` on the attached CD. They show the ratios $\tilde{\boldsymbol{r}}(u)$ and $\hat{\boldsymbol{r}}^r(u)$ versus ρ and ρ_* for the different values of σ^2 and ϕ. We use these tables to compare the performance of the proposed disturbed least squares estimator with the least squares and ridge estimator with respect to the degree of multicollinearity of the design matrices. Therefore consider Table 7.2.2, which gives all calculated ratios $\tilde{\boldsymbol{r}}(\rho, \rho_*, 3, 0)$ in dependence of ρ and ρ_*, but for fixed $\sigma^2 = 3$ and $\phi = 0$. All values in this table are smaller than one, i.e. for all combinations of ρ and ρ_* the disturbed least squares estimator has a smaller estimated mean squared error than the least squares estimator. We say that for all calculated combinations $(\rho, \rho_*, 3, 0)$, the disturbed least squares estimator performs better or dominates the least squares estimator.

One way to illustrate the results of Table 7.2.2 is to plot $\tilde{\boldsymbol{r}}(\rho, \rho_*, 3, 0)$ in dependence of ρ and ρ_* and to connect all points with each other to get a surface plot like given in Figure 7.2.1, (a). But for a better interpretation of the results, it is convenient to consider a contour plot instead of the surface plot. Figure 7.2.1, (b) shows the filled contour plot, which displays the isolines calculated from Table 7.2.2. The areas between the isolines are filled using constant colors. The colorbar shows the scale for the used colors.

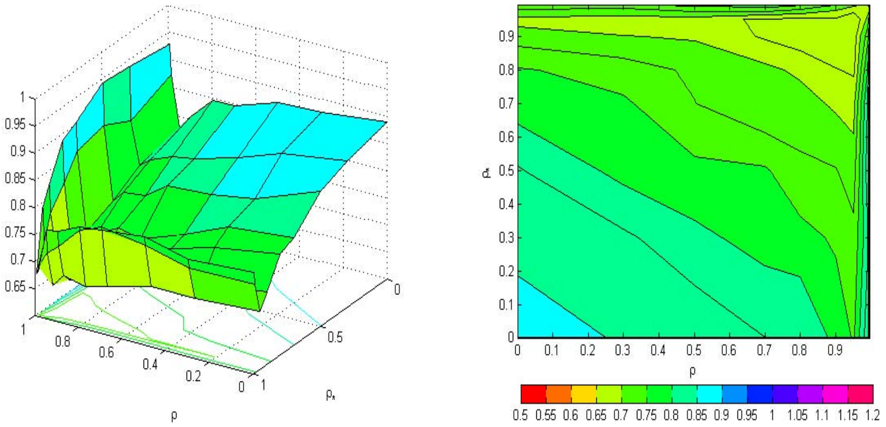
With the help of the contour plot it is easy to see, for which combinations of ρ and ρ_* the disturbed least squares estimator performs better than the least squares estimator

- red, yellow, green: $\tilde{\boldsymbol{r}}(\rho, \rho^*, \sigma^2, \phi) < 1$,
- blue: $\tilde{\boldsymbol{r}}(\rho, \rho^*, \sigma^2, \phi) \approx 1$,
- violet, pink: $\tilde{\boldsymbol{r}}(\rho, \rho^*, \sigma^2, \phi) > 1$.

Of course we have to be cautious, because the areas between the calculated values of Table 7.2.2 are only interpolated.

$\rho_*\backslash\rho$	0	0.3	0.5	0.7	0.8	0.9	0.95	0.99	0.995
0	0.89	0.88	0.87	0.85	0.85	0.80	0.78	0.87	0.93
0.3	0.87	0.85	0.82	0.80	0.79	0.78	0.76	0.84	0.94
0.5	0.85	0.80	0.79	0.79	0.76	0.75	0.73	0.87	0.94
0.7	0.80	0.79	0.75	0.72	0.73	0.71	0.71	0.85	0.90
0.8	0.79	0.77	0.74	0.72	0.71	0.69	0.68	0.81	0.89
0.9	0.73	0.72	0.71	0.68	0.67	0.66	0.66	0.77	0.82
0.95	0.71	0.70	0.70	0.68	0.67	0.68	0.65	0.73	0.79
0.99	0.76	0.76	0.78	0.79	0.78	0.74	0.72	0.69	0.68
0.995	0.79	0.78	0.80	0.79	0.79	0.77	0.76	0.68	0.68

Table 7.2.2. Results for $\tilde{\boldsymbol{r}}(u)$ for $\sigma^2 = 3$ and $\phi = 0$



(a) Surface plot

(b) Contour plot

FIGURE 7.2.1. Illustration of the result

The contour plots for all tables of the folder `results` are given in Figure 7.2.2 for $\sigma^2 = 0.01$ up to Figure 7.2.8 for $\sigma^2 = 5$. Therewith it is possible to evaluate the performance of the disturbed least squares estimator in dependence of ϕ.

$\phi = 0$: From Figure 7.2.2(b) until Figure 7.2.8(b) it is easy to see that the disturbed least squares estimator always behaves better than the least squares estimator for $\phi = 0$. For small σ^2 and weak multicollinearity (i.e. ρ and/ or ρ_* small) the blue areas dominate, i.e. there is only a slight improvement compared to the least squares estimator. For increasing σ^2 the blue areas diminish and we also have green and yellow areas for weak multicollinearity.

Thus either for large σ^2 or strong multicollinearity (i.e. $\rho, \rho_* \approx 1$) the disturbed least squares estimator performs best.

From Figure 7.2.2(d) to Figure 7.2.8(d) we can see, that the ridge estimator always dominates the least squares estimator and performs best for high variances and strong multicollinearity.

As a consequence the disturbed least squares estimator performs at least as good as the ridge estimator. Only in case of strong multicollinearity the ridge estimator performs better than the disturbed least squares estimator (violet and pink areas in Figure 7.2.2(f) to Figure 7.2.8(f)).

$\phi = 1$: Another situation is given for $\phi = 1$. Neither the disturbed least squares nor the ridge estimator performs much better than the least squares estimator for small variances until $\sigma^2 = 0.3$ (see Figure 7.2.2(c),(e) to Figure 7.2.8(c),(e)). For $\sigma^2 = 0.5$ and $\sigma^2 = 1$ the ridge and the disturbed least squares estimator only perform better than the least squares estimator for strong multicollinearity (green areas for ρ and/ or $\rho_* \approx 1$ in Figure

7.2.5(c),(e) and Figure 7.2.6(c),(e)).

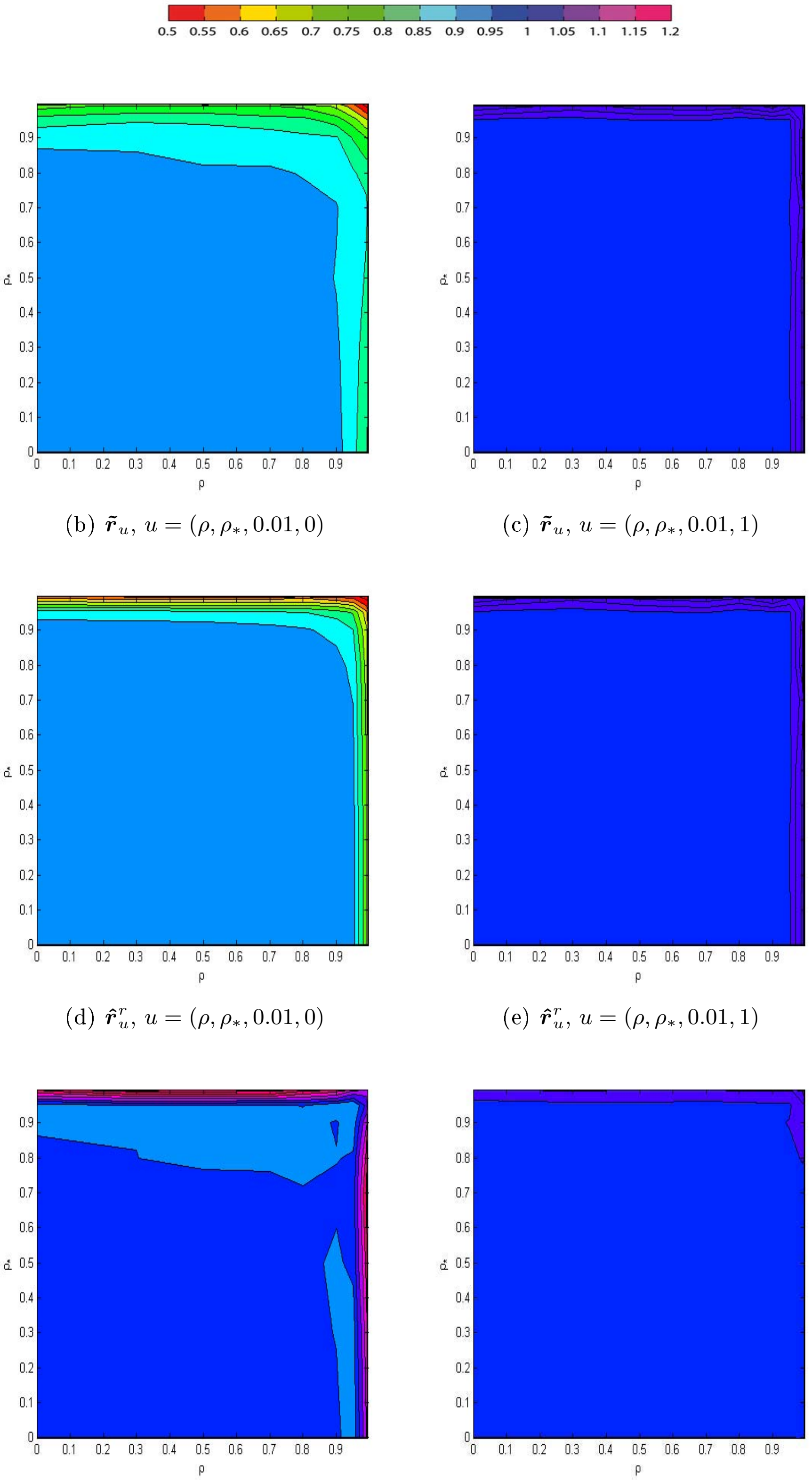
But for large σ^2, i.e. $\sigma^2 = 3$ and $\sigma^2 = 5$ they can dominate the least squares estimator (green areas in Figure 7.2.7(c),(e) and Figure 7.2.8(c),(e)), whereas there is only a slight improvement for weak multicollinearity (blue areas). Unfortunately the disturbed least squares estimator can hardly dominate the ridge estimator $\phi = 1$ (see Figure 7.2.2(g) up to Figure 7.2.8(g)), i.e. for large σ^2 the ridge estimator performs best.

Based on the performed simulation study we can conclude, that the disturbed least squares estimator performs at least as good as the ridge estimator. Besides the degree of multicollinearity, the performance heavily depends on σ^2. The bigger the variance, the better the performance of the disturbed least squares estimator for fixed ρ and ρ_*.

Note 7.2.1. Of course this simulation study can only throw a sidelight on the performance of the disturbed least squares estimator. For a more detailed examination extended studies would be necessary, for example with

- different methods for finding an optimal $\boldsymbol{\psi}$,
- other methods for estimating $\boldsymbol{\beta}$ and σ^2,
- more regressors and different $\boldsymbol{\beta}$,
- alternative loss functions as described in Note 2.2.5.

Furthermore not only the simulation study, but also the theoretical investigations could be extended to the consideration of a singular design matrix and non–normal error variables.



(b) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.01, 0)$

(c) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.01, 1)$

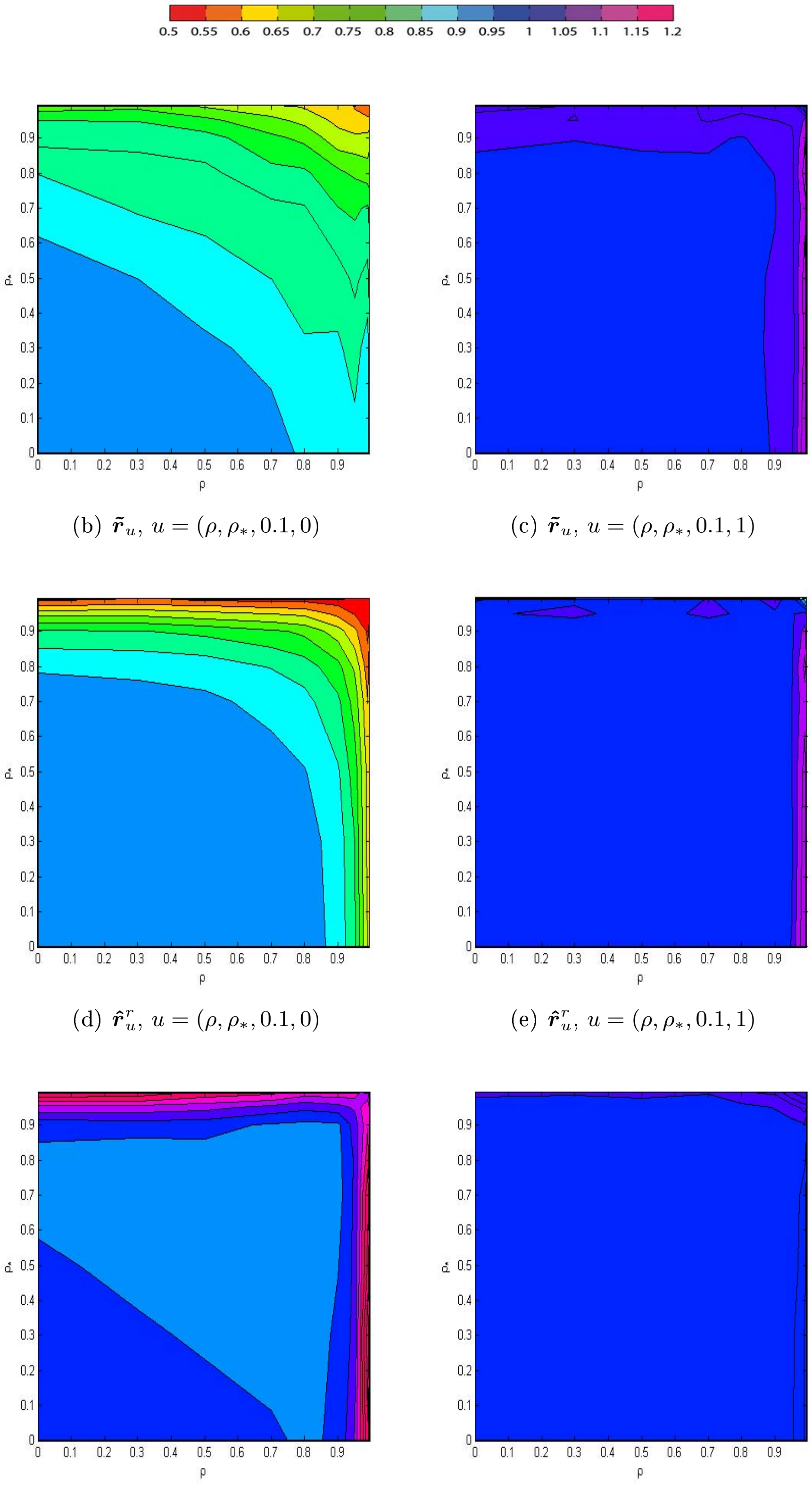
(d) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.01, 0)$

(e) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.01, 1)$

(f) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.01, 0)$

(g) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.01, 1)$

FIGURE 7.2.2. $\sigma^2 = 0.01$



(b) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.1, 0)$

(c) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.1, 1)$

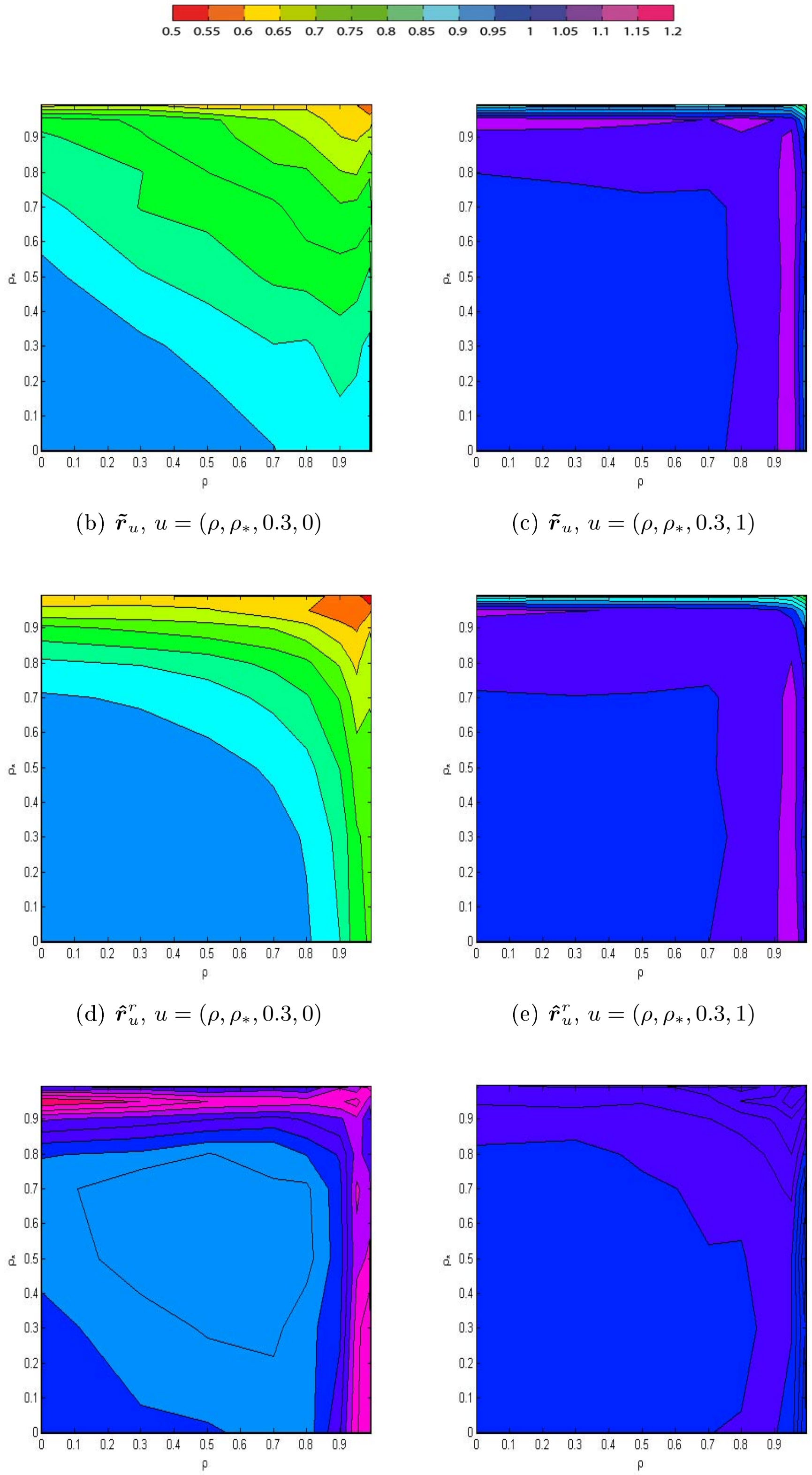
(d) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.1, 0)$

(e) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.1, 1)$

(f) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.1, 0)$

(g) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.1, 1)$

FIGURE 7.2.3. $\sigma^2 = 0.1$



(b) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.3, 0)$

(c) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.3, 1)$

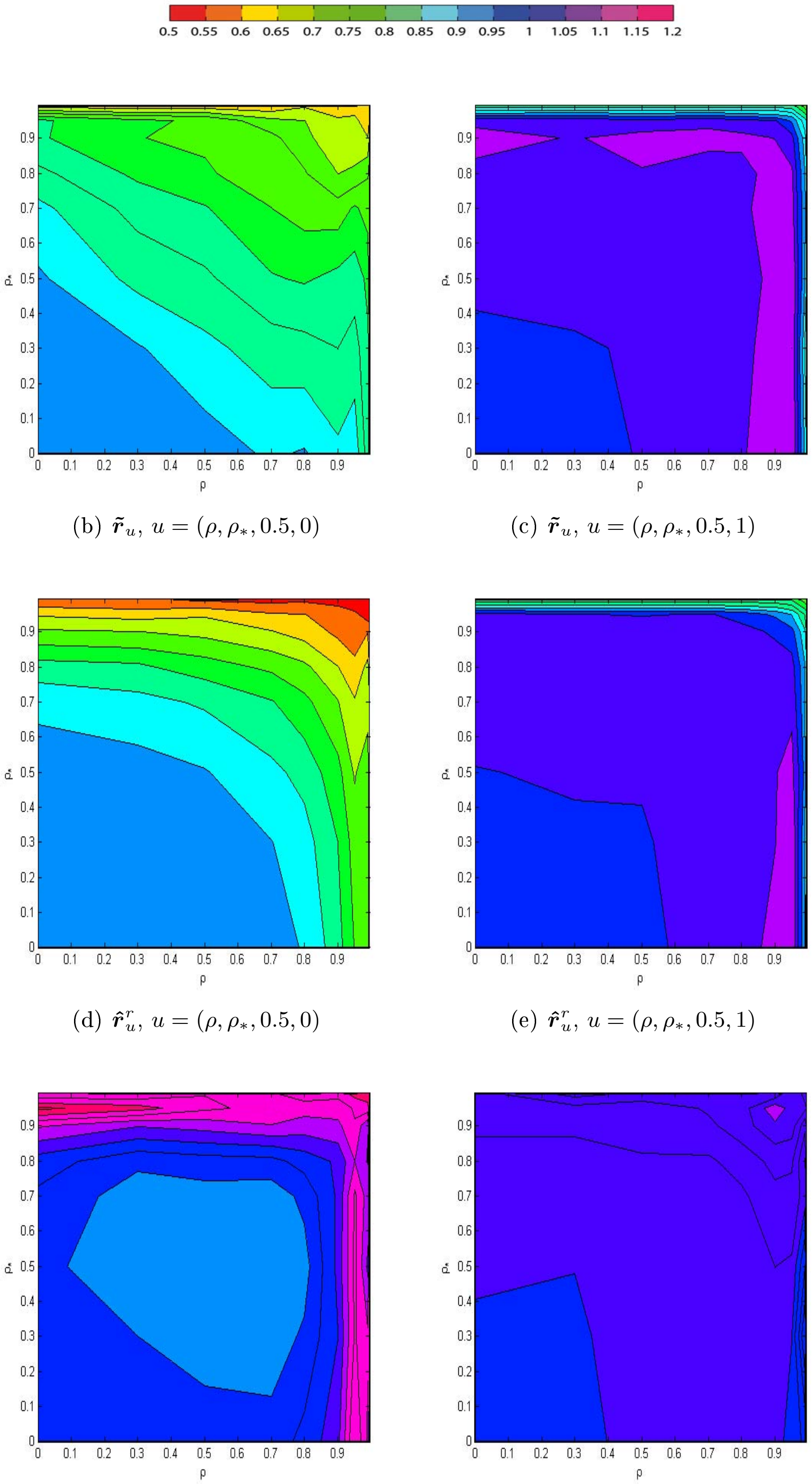
(d) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.3, 0)$

(e) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.3, 1)$

(f) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.3, 0)$

(g) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.3, 1)$

FIGURE 7.2.4. $\sigma^2 = 0.3$



(b) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.5, 0)$

(c) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 0.5, 1)$

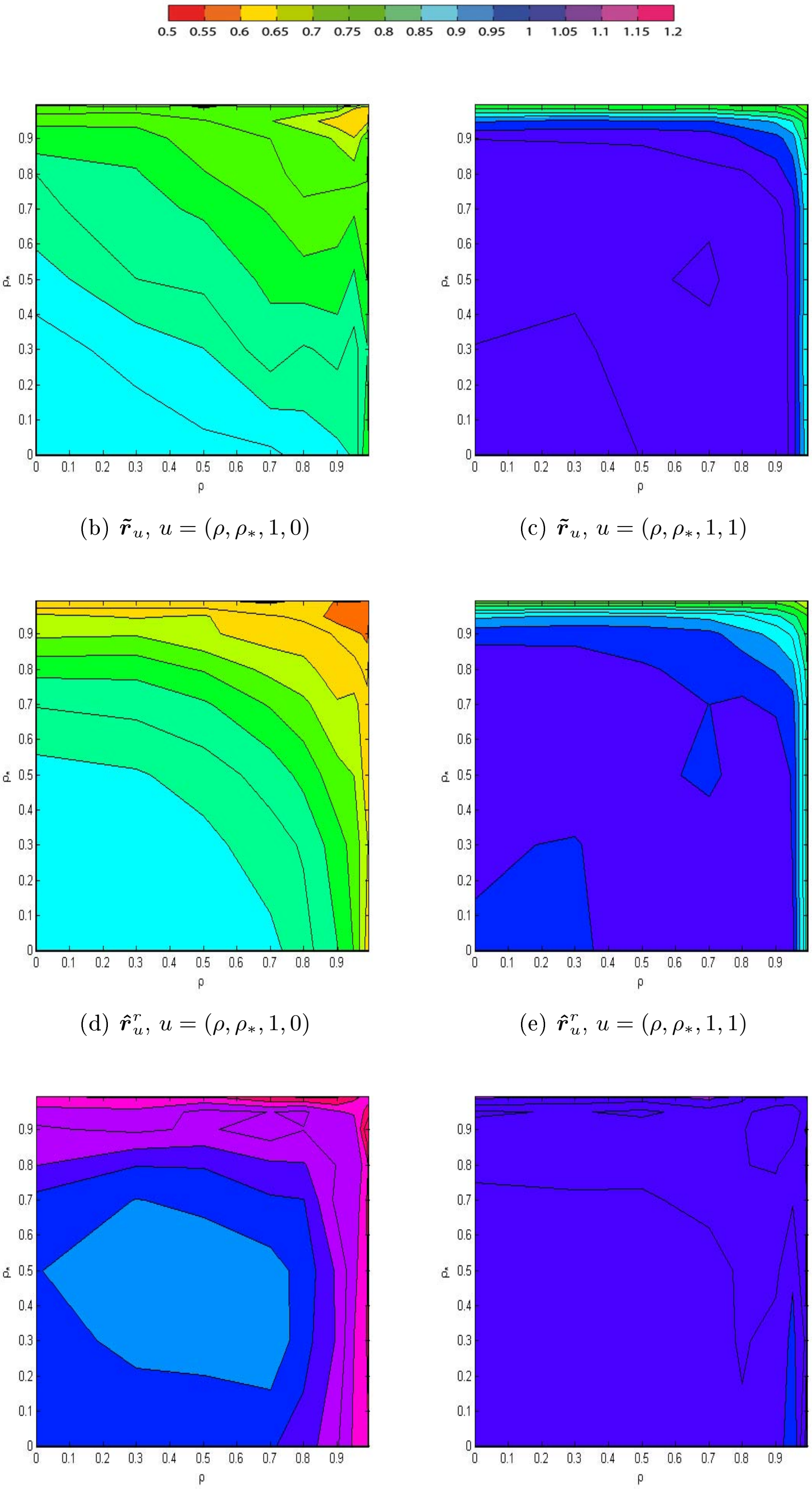
(d) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.5, 0)$

(e) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 0.5, 1)$

(f) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.5, 0)$

(g) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 0.5, 1)$

FIGURE 7.2.5. $\sigma^2 = 0.5$



(b) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 1, 0)$

(c) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 1, 1)$

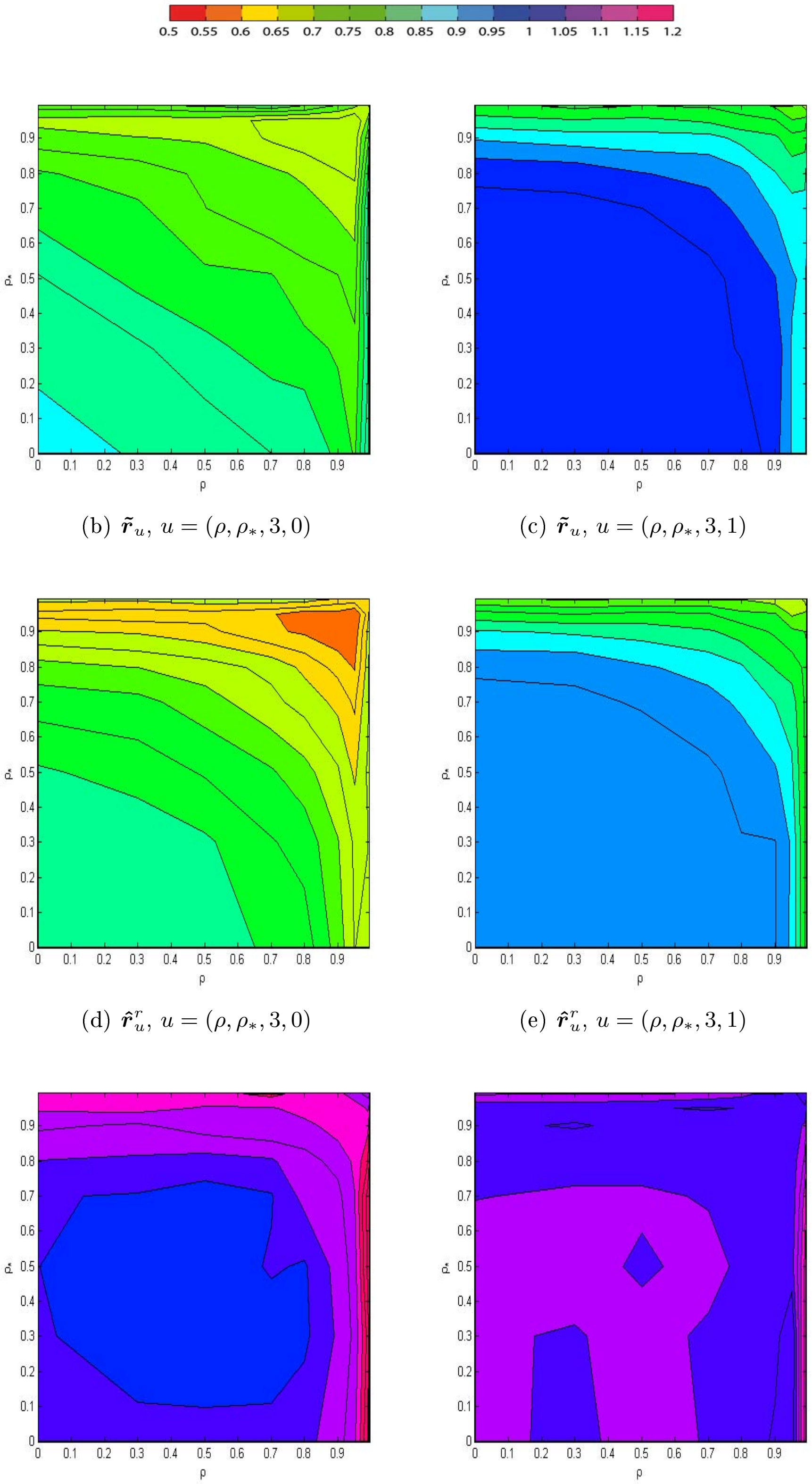
(d) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 1, 0)$

(e) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 1, 1)$

(f) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 1, 0)$

(g) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 1, 1)$

FIGURE 7.2.6. $\sigma^2 = 1$



(b) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 3, 0)$

(c) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 3, 1)$

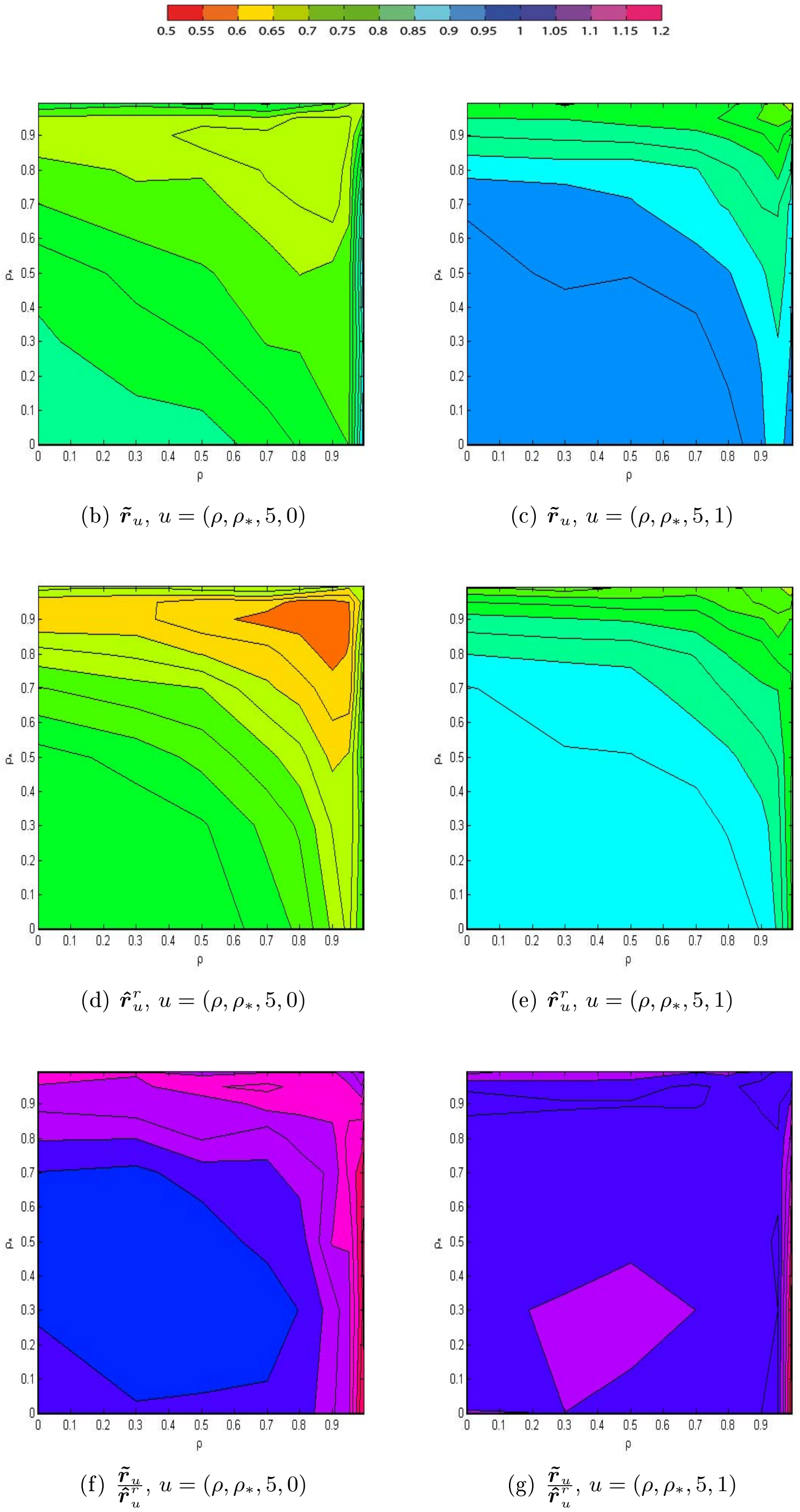
(d) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 3, 0)$

(e) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 3, 1)$

(f) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 3, 0)$

(g) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 3, 1)$

Figure 7.2.7. $\sigma^2 = 3$



(b) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 5, 0)$

(c) $\tilde{\boldsymbol{r}}_u$, $u = (\rho, \rho_*, 5, 1)$

(d) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 5, 0)$

(e) $\hat{\boldsymbol{r}}^r_u$, $u = (\rho, \rho_*, 5, 1)$

(f) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 5, 0)$

(g) $\frac{\tilde{\boldsymbol{r}}_u}{\hat{\boldsymbol{r}}^r_u}$, $u = (\rho, \rho_*, 5, 1)$

FIGURE 7.2.8. $\sigma^2 = 5$

Conclusion

Finally we summarize the most important results of this thesis.

After having introduced the least squares method and its necessary assumptions, we gave a small summary of the generally used criteria for comparing estimators. Thereby we pointed out that the mean squared error will mainly be used within this thesis. The following two chapters were dedicated to the problem of multicollinearity and the possible solution of using biased estimators, concretely ridge estimators. Besides the presentation of the harmful effects of multicollinearity, like variance inflation and possible diagnostic procedures, we also applied them on a real data set using the Economic Report of the President. For ridge estimators we presented not only the approch of Hoerl and Kennard, but also a general one, which results in a general form of ridge estimators, including the generalized ridge estimator of C.R. Rao. Furthermore we focused our attention on several procedures for estimating the biasing factor k and discussed the controversy in literature about standardization in regression and ridge regression theory. The use of the ridge estimator was also illustrated by the Economic Data set.

Next we investigated the disturbed least squares estimator, which is based on adding a small quantity $\omega\psi_j,\ j = 1, \dots, p$ on each regressor. We gave a presentation of the estimator, its total variance, squared bias and finally its mean squared error and matrix mean squared error in dependence of ω. We found out, that it is always possible to find an ω, such that for an arbitrary $\boldsymbol{\psi}$ the mean squared error of the disturbed least squares estimator is smaller than the corresponding one of the least squares estimator. But only due to the special choice of the biasing matrix $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$, it was possible to describe the estimator in dependence of $\boldsymbol{\psi}$ and thus discuss the optimality properties.

Unfortunately our approach can only be applied on standardized data. Therefore we gave a presentation of the mean squared error of the original coefficients after having transformed back the standardized coefficients. We saw, that it is also possible to find an ω, such that for arbitrary $\boldsymbol{\psi}$ the mean squared error of the disturbed least squares estimator of the original (unstandardized) model is smaller than the corresponding one of the least squares estimator.

But if the analyst does not feel up to standardize the data, it is also possible to apply all our results on the general ridge estimator of C.R. Rao for unstandardized data. This is due to the fact that the disturbed least squares estimator can be embedded into the group of the generalized ridge estimators.

In the following simulation study it was shown, that the disturbed least squares estimator mainly performs as good as the ridge estimator. Thus we found another alternative to the ridge estimator (and of course to the least squares estimator), which may be more appropriate in some situations. A disadvantage of the disturbed least squares estimator is, that it depends not only on ω, but also on $\boldsymbol{\psi}$, because in applied work the vector $\boldsymbol{\psi}$, or equivalently the matrix $\boldsymbol{\Psi}$, will be unknown.

Maybe an extended theoretical investigation and more simulation studies are necessary to develop procedures for choosing $\boldsymbol{\psi}$, in order to get an estimator with optimal statistical properties. Furthermore an examination of the performance of the disturbed least squares estimator in case of other loss functions or not normal distributed error variables requires more research.

APPENDIX A

Matrix Algebra

There are numerous books on Linear or Matrix Algebra containing helpful results used within this manuscript. In this appendix we collect some of the important results for ready reference. Proofs are referenced, wherever necessary.
We denote the row vectors of a matrix $\boldsymbol{A} := [a_{i,j}]_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$ by $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \in \mathbb{R}^n$ and define

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,n} \end{bmatrix} =: \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}.$$

Note A.0.2. We confine ourselves to real matrices of order n, although analogous results will obviously hold for complex matrices.

A.1. Trace of a Square Matrix

The trace of a square matrix $\boldsymbol{A} = [a_{i,j}]_{1\leq i,j\leq n}$ is defined to be the sum of the n diagonal elements of $\boldsymbol{A}$ and is denoted by the symbol $\operatorname{tr}(\boldsymbol{A})$. Clearly it is

$$\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}). \tag{A.1.1}$$

A further, very basic result is expressed in the following Lemma.

Lemma A.1.1. *For any $m \times n$ matrix $\boldsymbol{A}$ and $n \times m$ matrix $\boldsymbol{B}$*

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA}).$$

Proof. See Harville (1997,[**23**]), p. 51.

□

Consider now the trace of the product $\boldsymbol{ABC}$ of an $m \times n$ matrix $\boldsymbol{A}$, an $n \times p$ matrix $\boldsymbol{B}$ and a $p \times m$ matrix $\boldsymbol{C}$. Since $\boldsymbol{ABC}$ can be regarded as the product of the two matrices $\boldsymbol{AB}$ and $\boldsymbol{C}$ or, alternatively, of $\boldsymbol{A}$ and $\boldsymbol{BC}$, it follows from Lemma A.1.1

$$\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{CAB}) = \operatorname{tr}(\boldsymbol{BCA}). \tag{A.1.2}$$

A.2. Determinants

To define the determinant of an $n \times n$ matrix we require some elementary facts about permutations, which will be collected in this section.

A.2.1. Permutations

For $n \in \mathbb{N}$ the symmetric group $\boldsymbol{S}_n$ of $\{1, \ldots, n\}$ denotes the group of all bijective maps

$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}.$$

The elements of $\boldsymbol{S}_n$ are called *permutations*. The identity element of $\boldsymbol{S}_n$ is the identical map, denoted by *id*. We can write $\sigma \in \boldsymbol{S}_n$ as

$$\sigma = \begin{bmatrix} 1 & 2 & \ldots & n \\ \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{bmatrix}.$$

A permutation $\tau \in \boldsymbol{S}_n$ is called a *transposition*, if τ exchanges two elements of $\{1, \ldots, n\}$ and keeps all others fixed, i.e. there are $k, l \in \{1, \ldots, n\},\ k \neq l$, with

$$\begin{aligned} \tau(k) &= l, \\ \tau(l) &= k \text{ and} \\ \tau(i) &= i, \text{ for } i \in \{1, \ldots, n\} \backslash \{k, l\}. \end{aligned}$$

For $n \geq 2$, any permutation can (not uniquely) be decomposed into

$$\sigma = \tau_1 \circ \ldots \circ \tau_k,$$

where $\tau_1, \ldots, \tau_k \in \boldsymbol{S}_n$. The representation of a permutation as a product of transpositions is not unique, but the number k of required transposition is always either even or odd. This justifies the definition of the sign of σ by

(1) $\text{sign}(\tau) = -1$ for any transposition $\tau \in \boldsymbol{S}_n$.

(2) For $\sigma \in \boldsymbol{S}_n$ and $\sigma = \tau_1 \circ \ldots \circ \tau_k$ with transpositions $\tau_1, \ldots, \tau_k \in \boldsymbol{S}_n$, it is

$$\text{sign}(\sigma) = (-1)^k. \tag{A.2.3}$$

Now we are in a position to give the definition of a determinant of an $n \times n$ matrix.

Definition A.2.1. *For $n \geq 1$, the determinant of an $n \times n$ matrix is defined by*

$$\det : \mathbb{R}^{n \times n} \to \mathbb{R},$$

namely for $\boldsymbol{A} = [a_{i,j}]_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$

$$\det(\boldsymbol{A}) := \sum_{\sigma \in \boldsymbol{S}_n} \text{sign}(\sigma)\ a_{1,\sigma(1)} \cdot \ldots \cdot a_{n,\sigma(n)}. \tag{A.2.4}$$

Note A.2.2. Another notation for the determinant of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is

$$|\boldsymbol{A}| := \det(\boldsymbol{A}),$$

which is used within this manuscript.

A.2.2. Properties of the Determinant

The following well known theorem provides further properties of the determinant and gives some useful methods to calculate the determiant of a matrix.

THEOREM A.2.3. *A determinant* $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ *has got the following properties:*

(1) *For any* $\lambda \in \mathbb{R}$ *it is* $|\lambda \boldsymbol{A}| = \lambda^n |\boldsymbol{A}|$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

(2) *If* $\boldsymbol{A}$ *has a row (or column) consisting only of zeros, then* $|\boldsymbol{A}| = 0$.

(3) *If* $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ *is formed from* $\boldsymbol{A}$ *by interchanging two rows (or columns) of* $\boldsymbol{A}$, *then*

$$|\boldsymbol{B}| = -|\boldsymbol{A}|.$$

(4) *Let* $\boldsymbol{B}$ *represent a matrix formed from* $\boldsymbol{A}$ *by adding, to any one row of* $\boldsymbol{A}$, *scalar multiples of one or more other rows (or columns). Then*

$$|\boldsymbol{B}| = |\boldsymbol{A}|.$$

(5) $|\boldsymbol{A}| = 0 \Leftrightarrow \operatorname{rank}(\boldsymbol{A}) < n$.

(6) $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$.

(7) *If* $\boldsymbol{A}$ *is invertible, then*

$$|\boldsymbol{A}^{-1}| = \frac{1}{|\boldsymbol{A}|}.$$

(8) $|\boldsymbol{A}^T| = |\boldsymbol{A}|$

(9) *If* $\boldsymbol{A}^{-1}$ *exists, then*

$$(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1}.$$

PROOF. See Smith (1984,[**47**]), p. 232.

□

If $\boldsymbol{B}$ is formed out of $\boldsymbol{A}$ by interchanging rows (or columns) according to a permutation σ it follows from Theorem A.2.3, (3)

$$|\boldsymbol{B}| = \operatorname{sign}(\sigma)|\boldsymbol{A}| = (-1)^k |\boldsymbol{A}|. \tag{A.2.5}$$

A.2.3. Cofactor Expansion, Laplace´s Theorem and Cauchy–Binet formula

Another method for the computation of determinants is based on their reduction to determinants of matrices of smaller sizes.

Therefore we use the following notation of Lancaster (1985,[**30**]) for a matrix, which is composed of special elements of a given matrix $\boldsymbol{A} = [a_{i,j}]_{\substack{1\leq i\leq n \\ 1\leq j\leq p}} \in \mathbb{R}^{n\times p}$.

$$\boldsymbol{A}\begin{bmatrix} i_1 & \dots & i_m \\ j_1 & \dots & j_m \end{bmatrix} := \begin{bmatrix} a_{i_1,j_1} & a_{i_1,j_2} & \dots & a_{i_1,j_m} \\ a_{i_2,j_1} & a_{i_2,j_2} & \dots & a_{i_2,j_m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_m,j_1} & a_{i_m,j_2} & \dots & a_{i_m,j_m} \end{bmatrix},$$

i.e. the matrix $\boldsymbol{A}\left[\begin{smallmatrix} i_1 & \dots & i_m \\ j_1 & \dots & j_m \end{smallmatrix}\right]$ consists only of the $i_1 \leq \ldots \leq i_m$–th rows and the $j_1 \leq \ldots \leq j_m$–th columns of $\boldsymbol{A}$. The remaining rows and columns are deleted.

Definition A.2.4. *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times p}$. *For*

$$1 \leq i_1 < i_2 < \ldots < i_m \leq n \quad \textit{and} \quad 1 \leq j_1 < j_2 < \ldots < j_m \leq p, \tag{A.2.6}$$

the determinant of the $m \times m$ *submatrix* $\boldsymbol{A}\begin{bmatrix} i_1 & \dots & i_m \\ j_1 & \dots & j_m \end{bmatrix}$ *is called a minor of order* m *of* $\boldsymbol{A}$, *denoted by*

$$\left|\boldsymbol{A}\begin{bmatrix} i_1 & \dots & i_m \\ j_1 & \dots & j_m \end{bmatrix}\right|. \tag{A.2.7}$$

The minors for which $i_k = j_k$, $(k = 1, \ldots, m)$ *are called the* principal minors *of* $\boldsymbol{A}$ *of order* m *and for* $i_k = j_k = k$, $(k = 1, \ldots, m)$ *the* leading principal minors *of* $\boldsymbol{A}$.

Let $\boldsymbol{A}$ be an $n \times n$ matrix. The minor of order $(n-1)$ of $\boldsymbol{A}$, obtained by striking out the i-th row and j-th column is denoted by $|\boldsymbol{A}_{\{i,j\}}|, 1 \leq i,j \leq n$, and the signed minor $\tilde{a}_{i,j} = (-1)^{i+j}|\boldsymbol{A}_{\{i,j\}}|$ is called the *cofactor* of $a_{i,j}$. Cofactors can play an important role in computing the determinant in view of the following result.

Theorem A.2.5 (Cofactor Expansion). *Let* $\boldsymbol{A}$ *be an arbitrary* $n \times n$ *matrix. Then for any* i, j, $(1 \leq i, j \leq n)$

$$|\boldsymbol{A}| = a_{i,1}\tilde{a}_{i,1} + \ldots + a_{i,n}\tilde{a}_{i,n}$$

or similarly

$$|\boldsymbol{A}| = a_{1,j}\tilde{a}_{1,j} + \ldots + a_{n,j}\tilde{a}_{n,j},$$

where $\tilde{a}_{p,q} = (-1)^{p+q}|\boldsymbol{A}_{\{p,q\}}|$.

Proof. See Lancaster (1985,[**30**]), p. 33.

□

The following theorem is very useful for calculating the determinant of a product of two matrices.

Theorem A.2.6 (Cauchy–Binet formula). *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $m \times n$ and $n \times m$ matrices, respectively. If $n \geq m$ and $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$, then*

$$|\boldsymbol{C}| = \sum_{1 \leq j_1 \leq \ldots \leq j_m \leq n} \left| \boldsymbol{A} \begin{bmatrix} 1 & \ldots & m \\ j_1 & \ldots & j_m \end{bmatrix} \right| \left| \boldsymbol{B} \begin{bmatrix} j_1 & \ldots & j_m \\ 1 & \ldots & m \end{bmatrix} \right| \tag{A.2.8}$$

$$= \sum_{1 \leq j_1 \leq \ldots \leq j_m \leq n} \left| \boldsymbol{A} \begin{bmatrix} 1 & \ldots & m \\ j_1 & \ldots & j_m \end{bmatrix} \right| \left| \boldsymbol{B}^T \begin{bmatrix} 1 & \ldots & m \\ j_1 & \ldots & j_m \end{bmatrix} \right| . \tag{A.2.9}$$

That is, the determinant of the product $\boldsymbol{AB}$ equals the sum of the products of all possible minors of (the maximal) order m of $\boldsymbol{A}$ with the corresponding minors of $\boldsymbol{B}$ of the same order.

Proof. See Lancaster (1985,[**30**]), p. 40.

□

With

$$\left| \boldsymbol{A} \begin{bmatrix} j_1 & \ldots & j_m \\ 1 & \ldots & m \end{bmatrix} \right| = \left| \boldsymbol{A}^T \begin{bmatrix} 1 & \ldots & m \\ j_1 & \ldots & j_m \end{bmatrix} \right| ,$$

and Theorem A.2.6, it is easy to prove the following corollary (see also Lancaster (1985,[**30**]), p. 41).

Corollary A.2.7. *For any $m \times n$ matrix $\boldsymbol{A}$ and $m \leq n$ the Gram determinant*

$$|\boldsymbol{A}^T \boldsymbol{A}| = \sum_{1 \leq j_1 \leq \ldots \leq j_m \leq n} \left| \boldsymbol{A}^T \begin{bmatrix} 1 & \ldots & m \\ j_1 & \ldots & j_m \end{bmatrix} \right|^2$$

$$= \sum_{1 \leq j_1 \leq \ldots \leq j_m \leq n} \left| \boldsymbol{A} \begin{bmatrix} j_1 & \ldots & j_m \\ 1 & \ldots & m \end{bmatrix} \right|^2 \geq 0$$

with equality $(= 0)$ *holding, iff* $\operatorname{rank}(\boldsymbol{A}) < m$.

A.3. Adjoint and Inverse Matrices

Let $\boldsymbol{A} = [a_{i,j}]_{1 \leq i,j \leq n}$ be an arbitrary $n \times n$ matrix and $\tilde{a}_{i,j} = (-1)^{i+j} |\boldsymbol{A}_{\{i,j\}}|$ the cofactor of $a_{i,j}$ $(1 \leq i, j \leq n)$. The *adjoint matrix* of $\boldsymbol{A}$, written $\operatorname{adj}(\boldsymbol{A})$, is defined to be the transposed matrix of the cofactors of $\boldsymbol{A}$. Thus

$$\operatorname{adj}(\boldsymbol{A}) := [\tilde{a}_{i,j}]^T_{1 \leq i,j \leq n} .$$

Some properties of adjoint matrices follow immediately from the definition (see also Lancaster (1985,[**30**]), p. 43).

Corollary A.3.1. *For any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and any $\lambda \in \mathbb{R}$,*

(1) $\operatorname{adj}(\boldsymbol{A}^T) = (\operatorname{adj}(\boldsymbol{A}))^T$,
(2) $\operatorname{adj}(\boldsymbol{I}_n) = \boldsymbol{I}_n$,
(3) $\operatorname{adj}(\lambda \boldsymbol{A}) = \lambda^{n-1} \operatorname{adj}(\boldsymbol{A})$.

The following theorem is the main reason for the interest in the adjoint matrices.

THEOREM A.3.2. *For any $\boldsymbol{A} \in \mathbb{R}^{n\times n}$,*

$$\boldsymbol{A} \cdot \operatorname{adj}(\boldsymbol{A}) = |\boldsymbol{A}| \cdot \boldsymbol{I}_n.$$

PROOF. See Lancaster (1985,[**30**]), p. 43.

□

From Theorem A.3.2 it follows

COROLLARY A.3.3. *If $\boldsymbol{A}$ is an $n \times n$ nonsingular matrix, then*

$$\operatorname{adj}(\boldsymbol{A}) = |\boldsymbol{A}|\boldsymbol{A}^{-1},$$

or equivalently

$$\boldsymbol{A}^{-1} = \frac{\operatorname{adj}(\boldsymbol{A})}{|\boldsymbol{A}|}.$$

The following lemma gives some very basic (but useful) results on the inverse of a nonsingular sum of two square matrices.

LEMMA A.3.4. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be arbitrary $n\times n$ matrices. If $\boldsymbol{A}+\boldsymbol{B}$ is nonsingular, then*

$$\begin{aligned}
(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{A} &= \boldsymbol{I}_n - (\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}\\
\boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1} &= \boldsymbol{I}_n - \boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\\
\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{A} &= \boldsymbol{A}(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}
\end{aligned}$$

PROOF. See Harville (1997,[**23**]), p. 419.

□

A.4. Determinant of the Sum of Two Matrices

Denote by $\boldsymbol{a}_i, \boldsymbol{b}_i$ and $\boldsymbol{c}_i$ the i-th row, $i = 1, \ldots, n$ of $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ respectively. If for some $k = 1, \ldots, n$

$$\boldsymbol{c}_k = \boldsymbol{a}_k + \boldsymbol{b}_k$$

and

$$\boldsymbol{c}_i = \boldsymbol{a}_i = \boldsymbol{b}_i, \quad i = 1, \ldots, k-1, k+1, \ldots, n,$$

then

$$|\boldsymbol{C}| = |\boldsymbol{A}| + |\boldsymbol{B}|,$$

because of the linearity of the determinant.
But usually for arbitrary $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

$$|\boldsymbol{A}+\boldsymbol{B}| \neq |\boldsymbol{A}| + |\boldsymbol{B}|.$$

However, as described in the following theorem, $|\boldsymbol{A}+\boldsymbol{B}|$ can, for any particular integer $u, 1 \leq u \leq n$, be expressed as the sum of the determinants of 2^u $n \times n$ matrices. The i-th row of each of these 2^u matrices is identical to the i-th row of $\boldsymbol{A}, \boldsymbol{B}$, or $\boldsymbol{A}+\boldsymbol{B}$ $(i = 1, \ldots, n)$.

Theorem A.4.1. *For any two $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and any $u \in \{1, \ldots, n\}$*

$$|\boldsymbol{A}+\boldsymbol{B}| = \sum_{\{i_1,\ldots,i_r\}\subseteq\{1,\ldots,u\}} \left|\boldsymbol{C}_u^{\{i_1,\ldots,i_r\}}\right|,$$

where the summation runs over all 2^u subsets $\{i_1, \ldots, i_r\}$ of $\{1, \ldots, u\}$ and where $\boldsymbol{C}_u^{\{i_1,\ldots,i_r\}}$ is an $n \times n$ matrix, whose last $n-u$ rows are identical to the last $n-u$ rows of $\boldsymbol{A}+\boldsymbol{B}$, whose $i_1, \ldots, i_r$-th rows are identical to the $i_1, \ldots, i_r$-th rows of $\boldsymbol{A}$ and whose remaining $n-(n-u+r) = u-r$ rows are identical to the corresponding ones of $\boldsymbol{B}$.

Proof. See Harville (1997,[**23**]), p. 196.

□

The following special case of Theorem A.4.1 is often used within this manuscript.

Corollary A.4.2. *For any two $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,*

$$|\boldsymbol{A}+\boldsymbol{B}| = \sum_{\{i_1,\ldots,i_r\}\subseteq\{1,\ldots,n\}} \left|\boldsymbol{C}_n^{\{i_1,\ldots,i_r\}}\right|,$$

where the summation runs over all 2^n subsets $\{i_1, \ldots, i_r\}$ of $\{1, \ldots, n\}$ and where $\boldsymbol{C}_n^{\{i_1,\ldots,i_r\}}$ is an $n \times n$ matrix, whose $i_1, \ldots, i_r$-th rows are identical to the $i_1, \ldots, i_r$-th rows of $\boldsymbol{B}$ and whose remaining rows are identical to the corresponding ones of $\boldsymbol{A}$.

A.5. Determinant and Inverse of a Partitioned Matrix

The following theorem gives a helpful formula for calculating the determinant of a partitioned matrix.

Theorem A.5.1. *Let $\boldsymbol{T}$ be a $p \times p$ matrix, $\boldsymbol{U}$ a $p \times n$ matrix, $\boldsymbol{V}$ an $n \times p$ matrix and $\boldsymbol{W}$ an $n \times n$ matrix. If $\boldsymbol{T}$ is nonsingular, then*

$$\left|\begin{bmatrix} \boldsymbol{T} & \boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{W} \end{bmatrix}\right| = \left|\begin{bmatrix} \boldsymbol{W} & \boldsymbol{V} \\ \boldsymbol{U} & \boldsymbol{T} \end{bmatrix}\right| = |\boldsymbol{T}||\boldsymbol{W} - \boldsymbol{V}\boldsymbol{T}^{-1}\boldsymbol{U}|.$$

Proof. See Harville (1997,[**23**]), page 189.

□

To calculate the inverse of a partitioned matrix we can use the following theorem.

Theorem A.5.2. *Let $\boldsymbol{T}$ be a $p \times p$ matrix, $\boldsymbol{U}$ a $p \times n$ matrix, $\boldsymbol{V}$ an $n \times p$ matrix and $\boldsymbol{W}$ an $n \times n$ matrix. Suppose $\boldsymbol{T}$ is nonsingular. Then $\begin{bmatrix} \boldsymbol{T} & \boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{W} \end{bmatrix}$, or equivalently, $\begin{bmatrix} \boldsymbol{W} & \boldsymbol{V} \\ \boldsymbol{U} & \boldsymbol{T} \end{bmatrix}$, is nonsingular if and only if the $n \times n$ matrix*

$$\boldsymbol{Q} = \boldsymbol{W} - \boldsymbol{V}\boldsymbol{T}^{-1}\boldsymbol{U}$$

is nonsingular, in which case

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{W} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{T}^{-1} + \boldsymbol{T}^{-1}\boldsymbol{U}\boldsymbol{Q}^{-1}\boldsymbol{V}\boldsymbol{T}^{-1} & -\boldsymbol{T}^{-1}\boldsymbol{U}\boldsymbol{Q}^{-1} \\ -\boldsymbol{Q}^{-1}\boldsymbol{V}\boldsymbol{T}^{-1} & \boldsymbol{Q}^{-1} \end{bmatrix},$$

$$\begin{bmatrix} \boldsymbol{W} & \boldsymbol{V} \\ \boldsymbol{U} & \boldsymbol{T} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{Q}^{-1} & -\boldsymbol{Q}^{-1}\boldsymbol{V}\boldsymbol{T}^{-1} \\ -\boldsymbol{T}^{-1}\boldsymbol{U}\boldsymbol{Q}^{-1} & \boldsymbol{T}^{-1} + \boldsymbol{T}^{-1}\boldsymbol{U}\boldsymbol{Q}^{-1}\boldsymbol{V}\boldsymbol{T}^{-1} \end{bmatrix}.$$

Proof. See Harville (1997,[**23**]), page 99.

□

A.6. Projection Matrices and Expectation of a Random Quadratic Form

Projection matrices are a family of matrices with special properties and are often applied in regression theory.

Theorem A.6.1. *The projection matrix $\boldsymbol{P} := \boldsymbol{A}\left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T \in \mathbb{R}^{p\times p}$ with $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, $p \leq n$ has the following two basic properties:*

(1) *$\boldsymbol{P}$ is idempotent, i.e. $\boldsymbol{P}^2 = \boldsymbol{P}$,*
(2) *$\boldsymbol{P}$ is symmetric, i.e. $\boldsymbol{P}^T = \boldsymbol{P}$.*

Conversly, any matrix with these two properties represents a projection matrix.

Proof. See Strang (1976,[**53**]), p. 110.

□

With the help of the next theorem it will be easier to compute the expectation of a random quadratic form $\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}$, where $\boldsymbol{u} = \begin{bmatrix} u_1, & \dots & , u_n \end{bmatrix}^T$ is an n dimensional random vector.

Theorem A.6.2. *Let $\boldsymbol{u} = \begin{bmatrix} u_1, & \dots & , u_n \end{bmatrix}^T \in \mathbb{R}^{n\times 1}$ be a random vector with mean vector $\boldsymbol{\mu}$ and $n \times n$ covariance matrix $\boldsymbol{\Sigma}$. Then we have for any arbitrary $n \times n$ matrix $\boldsymbol{A}$*

$$\mathrm{E}(\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}.$$

Proof. See Falk (2002,[**11**]), p. 116.

□

A.7. Eigenvalues and Eigenvectors

DEFINITION A.7.1. *If $\boldsymbol{A}$ is an $n \times n$ matrix, then*

$$c(\lambda) := |\boldsymbol{A} - \lambda \boldsymbol{I}_n|$$

is a polynomial in λ of degree n. The n roots $\lambda_1, \dots, \lambda_n$ of the characteristic equation *$c(\lambda)$ are called* eigenvalues *of $\boldsymbol{A}$.*

The eigenvalues possibly may be complex numbers. It is easy to see (see e.g Strang (1976,[**53**]), p. 177) that each of the following conditions is necessary and sufficient for the number $\lambda_i,\ i = 1, \dots, n$ to be a *real* eigenvalue of $\boldsymbol{A}$:

(1) There is a nonzero vector $V_i \in \mathbb{R}^n$, such that $\boldsymbol{A}V_i = \lambda_i V_i$.
(2) The matrix $\boldsymbol{A} - \lambda_i \boldsymbol{I}_n$ is singular.
(3) $|\boldsymbol{A} - \lambda_i \boldsymbol{I}_n| = 0$.

V_i is called the (right) eigenvector of $\boldsymbol{A}$ for the eigenvalue λ_i. An eigenvector V_i with real components is called standardized, if $V_i^T V_i = 1,\ i = 1, \dots, n$.

A.8. Decomposition of Matrices

THEOREM A.8.1 (Spectral decomposition Theorem). *Any symmetric $n \times n$ matrix $\boldsymbol{A}$ can be written as*

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T,$$

where

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} =: diag(\lambda_1, \dots, \lambda_n)$$

is the diagonal matrix of the eigenvalues of $\boldsymbol{A}$ and

$$\boldsymbol{V} = \begin{bmatrix} V_1 & \dots & V_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$

is the orthogonal *matrix of the standardized eigenvectors V_i.*

PROOF. See Stewart (1973,[**52**]), p. 277.

□

From $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ we get $\boldsymbol{\Lambda} = \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}$. With the spectral decomposition we can define the symmetric square root decomposition of $\boldsymbol{A}$ (if $\lambda_i \geq 0,\ i = 1, \dots, n$)

$$\boldsymbol{A}^{\frac{1}{2}} := \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T, \quad \text{with} \quad \boldsymbol{\Lambda}^{\frac{1}{2}} := diag(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$$

and if $\lambda_i > 0$,

$$\boldsymbol{A}^{-\frac{1}{2}} := \boldsymbol{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^T, \quad \text{with} \quad \boldsymbol{\Lambda}^{-\frac{1}{2}} := diag(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_n}}).$$

Theorem A.8.2 (Singular value decomposition of a matrix (SVD)). *Let $\boldsymbol{A}$ be an arbitrary $m \times n$ matrix of rank r. Then there are orthogonal matrices $\boldsymbol{U} =: \begin{bmatrix} U_1 & \ldots & U_m \end{bmatrix} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{V} =: \begin{bmatrix} V_1 & \ldots & V_n \end{bmatrix} \in \mathbb{R}^{n \times n}$ such that*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Theta}\boldsymbol{V}^T,$$

with

$$\boldsymbol{\Theta} := \begin{bmatrix} \boldsymbol{\theta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

and the diagonal matrix

$$\boldsymbol{\theta} = \text{diag}(\theta_1, \ldots, \theta_r)$$

and $\theta_1 \geq \ldots \geq \theta_r > 0$.

Proof. See Stewart (1973,[**52**]), p. 319.

□

The diagonal elements $\theta_1, \ldots, \theta_n$, where $\theta_{r+1} = \ldots = \theta_n = 0$, are called the *singular values* of $\boldsymbol{A}$. Since the singular value decomposition is unique (see Stewart (1973,[**52**]), p. 319), we have

$$\boldsymbol{V}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{\Theta}^2,$$

with

$$\boldsymbol{\Theta}^2 = \begin{bmatrix} \boldsymbol{\theta}^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus $\theta_1^2, \ldots, \theta_r^2$ are the nonzero eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$, arranged in descending order and with $\theta_i \geq 0, \ i = 1, \ldots, n$. It holds

$$\theta_i = \sqrt{\lambda_i}, \quad i = 1, \ldots, n, \tag{A.8.10}$$

where λ_i denote the eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$.

If the singular value decomposition is given by Theorem A.8.2 we have (see Golub (1996,[**17**]), p. 71)

$$\text{rank}(\boldsymbol{A}) = r$$

and thus

$$\boldsymbol{A} = \sum_{i=1}^{r} \theta_i U_i V_i^T.$$

Furthermore it follows from the definition of the Euclidean norm of a matrix

$$\|\boldsymbol{A}\|_2 := \sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})} = \sigma_{\max} = \sigma_1. \tag{A.8.11}$$

The singular value decomposition enables us to adress the numerical difficulties, frequently encountered in situations where near rank deficiency prevails. For some small ε we may be interested in the ε–rank of a matrix which we define by

$$\operatorname{rank}(\boldsymbol{A}, \varepsilon) := \min_{\|\boldsymbol{A}-\boldsymbol{B}\|_2 \leq \varepsilon} \operatorname{rank}(\boldsymbol{B}).$$

The following theorem shows, that the singular values indicate how near a given matrix is to a matrix of lower rank.

Theorem A.8.3. *Let the singular value decomposition of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be given by Theorem A.8.2. If $k < r = \operatorname{rank}(\boldsymbol{A})$ and*

$$\boldsymbol{A}_k = \sum_{i=1}^{k} \theta_i U_i V_i^\top,$$

then for any matrix $\boldsymbol{B}$ of rank k it is

$$\min\{\| \boldsymbol{A} - \boldsymbol{B} \|_2 : \boldsymbol{B} \in \mathbb{R}^{m\times n}, \operatorname{rank}(\boldsymbol{B}) = k\} = \| \boldsymbol{A} - \boldsymbol{A}_k \|_2 = \theta_{k+1}.$$

Proof. See Golub (1996,[**17**]), p. 73.

□

Theorem A.8.3 states that the smallest singular value of $\boldsymbol{A}$ is the 2–norm distance of $\boldsymbol{A}$ to the set of all rank deficient matrices.

A.8.1. The Condition Number of a Matrix

The singular value decomposition provides a measure, called *condition number*, which is related to the measure of linear independence between column vectors of the matrix.

Definition A.8.4. *The condition number of a matrix $A \in \mathbb{R}^{n\times n}$ with respect to the Euclidean norm $\|\cdot\|_2$ is defined by*

$$\operatorname{cond}(\boldsymbol{A}) := \|\boldsymbol{A}\|_2 \left\|\boldsymbol{A}^{-1}\right\|_2. \tag{A.8.12}$$

Note, that the condition number can be defined with respect to any arbitrary norm. Thus cond$(\cdot)$ depends on the underlying norm. Let $\boldsymbol{A}$ be a square matrix of full rank with the singular values $\theta_i,\ i = 1, \ldots, n$. Using (A.8.11) and (A.8.10) we obtain

$$\operatorname{cond}(\boldsymbol{A}) = \sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})}\sqrt{\lambda_{\max}\left((\boldsymbol{A}^T\boldsymbol{A})^{-1}\right)} = \frac{\theta_{\max}(\boldsymbol{A})}{\theta_{\min}(\boldsymbol{A})},$$

because the singular values of $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ are given by $\frac{1}{\theta_i^2}$, $i = 1, \ldots, n$. Hence if $\boldsymbol{A}$ is rank deficient, then $\theta_{\text{min}}(\boldsymbol{A}) = 0$ and we set $\text{cond}(\boldsymbol{A}) = \infty$. It is readily shown, that the condition number of any matrix with orthonormal columns is unity and hence $\text{cond}(\boldsymbol{A})$ reaches its lower bound in this cleanest of all possible cases. It is easy to prove, that

$$\text{cond}(\boldsymbol{A}^\top\boldsymbol{A}) = (\text{cond}(\boldsymbol{A}))^2 = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}. \tag{A.8.13}$$

From Theorem A.8.2 we know, that a matrix $\boldsymbol{A}$ is of full rank, if its singular values are all non zero. If $\boldsymbol{A}$ is nearly rank deficient, then θ_{min} will be very small and as a consequence the condition number will be large. Thus the condition number can be used as a measure for the rank deficiency of a matrix.

A.9. Definite Matrices

DEFINITION A.9.1. *Let $\boldsymbol{A}$ be an arbitrary $n \times n$ matrix. $\boldsymbol{A}$ is called*

(1) *positive definite, if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$, $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ for all $\boldsymbol{x} \neq 0$,*

(2) *positive semidefinite, if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$, $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ for all $\boldsymbol{x} \neq 0$.*

We write $\boldsymbol{A} > 0$ for the first case and $\boldsymbol{A} \geq 0$ for the second.

Using this definition we can state the following well known theorem.

THEOREM A.9.2. *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix with the eigenvalues λ_i, $i = 1, \ldots, n$. Then*

(1) $0 \leq \lambda_i \in \mathbb{R}$,

(2) $\text{tr}(\boldsymbol{A}) \geq 0$,

(3) $\boldsymbol{A} = \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{A}^{\frac{1}{2}}$ *with* $\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T$,

(4) $\boldsymbol{A} + \boldsymbol{B} > 0$, *for* $0 \leq \boldsymbol{B} \in \mathbb{R}^{n\times n}$,

(5) *we have* $\boldsymbol{C}^T\boldsymbol{C} \geq 0$ *and* $\boldsymbol{C}\boldsymbol{C}^T \geq 0$ *for any matrix* $\boldsymbol{C}$,

(6) *for* $\boldsymbol{A} > 0$, *the inverse* $\boldsymbol{A}^{-1}$ *is also positive definite.*

PROOF. See Harville (1997,[**23**]), p. 543, p. 238, p. 212 and p. 214.

□

For $\boldsymbol{A} > 0$ we can replace $\leq$ by $<$ in (1) and $\geq$ by $>$ in (2) of Theorem A.9.2.

THEOREM A.9.3. *A symmetric $n \times n$ matrix $\boldsymbol{A}$ is positive semidefinite, iff there exists a matrix $\boldsymbol{P}$ (having n columns) such that $\boldsymbol{A} = \boldsymbol{P}^T\boldsymbol{P}$. $\boldsymbol{A}$ is positive definite, iff $\boldsymbol{P}$ is nonsingular.*

PROOF. See Harville (1997,[**23**]), p. 218 and p. 219.

□

A sufficient condition for the positive definiteness or positive semidefiniteness of the *Schur complement* of a partitioned matrix is given by the following theorem.

THEOREM A.9.4. *Let*

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{U} \\ \boldsymbol{U}^T & \boldsymbol{W} \end{bmatrix},$$

where $\boldsymbol{T} \in \mathbb{R}^{q \times q}, \boldsymbol{U} \in \mathbb{R}^{q \times n}$ *and* $\boldsymbol{W} \in \mathbb{R}^{n \times n}$. *If* $\boldsymbol{G}$ *is positive (semi)definite, then the Schur complement* $\boldsymbol{W} - \boldsymbol{U}^T\boldsymbol{T}^{-1}\boldsymbol{U}$ *of* $\boldsymbol{T}$ *and the Schur complement* $\boldsymbol{T} - \boldsymbol{U}\boldsymbol{W}^{-1}\boldsymbol{U}^T$ *of* $\boldsymbol{W}$ *are positive (semi)definite.*

PROOF. See Harville (1997,[**23**]), p. 242.

□

Bibliography

[1] ALLEN, D. M. (1971). The Prediction Sum of Squares as a Criterion for Selecting Predictor Variables *Technical Report* **23** Department of Statistics, University of Kentucky.

[2] BELSLEY D. A., KUH E., WELSCH R. E. (1980). *Regression Diagnostics- Identifying Influential Data and Sources of Multicollinearity*, John Wiley & Sons, Inc., New York.

[3] BELSLEY D. A., KUH E., WELSCH R. E. (1984). Demeaning Conditioning Diagnostics Through Centering *The American Statistician* **38(2)** 73–77.

[4] BELSLEY D. A. (1991). A Guide to Using the Collinearity Diagnostics *Computer Science in Economics and Management* **4** 33-50.

[5] BROOK R.J. AND MOORE, T. (1980). On the Expected Length of the Least Squares Coefficient Vector *Journal of Econometrics* **12** 245–246.

[6] CHATTERJEE, S. AND HADI, A. S. (2006). *Applied Regression Analysis*, 4nd ed. Wiley & Sons Inc., Hoboken, New Jersey.

[7] CLARK, A. E. AND TROSKIE, C. G. (2006). Ridge Regression–A Simulation Study *Regression Analysis by Example* **35** 605–619.

[8] DELANEY, N. J. AND CHATTERJEE, S. (1986). Use of the Bootstrap and Cross–Validation in Ridge Regression *Journal of Business & Economic Statistics* **4(2)** 255–262.

[9] DRAPER, N.R. AND SMITH, H. (1981). *Applied Regression Analysis*, 2nd ed. Wiley & Sons Inc., Hoboken, New Jersey.

[10] *The Economic Report of the President (2008)*
`http://www.gpoaccess.gov/eop/tables08.html`

[11] FALK M., MAROHN F., TEWES B. (2002). *Foundations of Statistical Analyses and Applications with SAS*, Birkhäuser, Basel, Bosten, Berlin.

[12] FAREBROTHER, R.W. (1976). Further Results on the Mean Squared Error of Ridge Regression *Journal of the Royal Statistical Society* **B,38** 248–250.

[13] FAREBROTHER, R.W. (1978). A Class of Shrinkage Estimators *Journal of the Royal Statistical Society* **B,40(1)** 47–49.

[14] FARRAR, D. E. AND GLAUBER, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited *The Review of Economics and Statistics* **49 (1)** 92–107.

[15] FOX, J. AND MONETTE, G. (1992). Generalized Collinearity Diagnostics *Journal of the American Statistical Association* **87 (417)** 178–183.

[16] GOLDSTEIN M. AND SMITH A. F. M. (1974). Ridge–type Estimators for Regression Analysis *Journal of the Royal Statistical Society* **B (36)** 284–291.

[17] GOLUB,G.H. (1996). *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, Baltimore, Maryland.

[18] GREENBERG E. (1975). Minimum Variance Properties of Principle Component Regression *Journal of the American Statistical Association* **70** 194–197.

[19] GRUBER, M.H.J. (1998). *Improving Efficiency by Shrinkage*, Marcel Dekker, Inc., New York.

[20] Guilkey, D. K. and Murphy, J. L. (1975). Directed Ridge Regression Techniques in Case of Multicollinearity *Journal of the American Statistical Association* **70 (352)** 769–775.

[21] Gunst R.F. (1983). Regression Analysis with Multicollinear Predictor Variables: Definition, Detection, and Effects. *Communications in Statistics- Theory and Methods* **12 (19)** 2217–2260.

[22] Chatterjee S., Hadi A.S. (2006). *Regression Analysis by Example*, 4th ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

[23] Harville, D. A. (1997). *Matrix Algebra from a Statistician´s Perspective*, Springer, New York.

[24] Hoerl,A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems *Technometrics* **12 (1)** 55–67.

[25] Hoerl,A. E. and Kennard, R. W. (1970b). Ridge Regression: Applications to Nonorthogonal Problems *Technometrics* **12 (1)** 69–82.

[26] Hoerl E.,Kennard R. W., Baldwin K. F. (1975). Ridge Regression: Some Simulations *Communications in Statistics* **4 (2)** 105–123.

[27] Hoerl E.,Kennard R. W., Baldwin K. F. (1976). Ridge Regression Iterative Estimation of the Biasing Parameter *Communications in Statistics- Theory and Methods* **A (5)** 77–88.

[28] Kibria,B. M. Golam (2003). Performance of Some New Ridge Regression Estimators *Communications in Statistics- Simulation and Computation* **32 (2)** 419–435.

[29] King G. (1986). How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science *American Journal of Political Science* **30(3)** 666–687.

[30] Lancaster, P. and Tismenetsky, M. (1985). *The Theory of Matrices*, 2nd ed. Academic Press Inc., San Diego.

[31] Lowerre, James M. (1974). On the Mean Squared Error of Parameter Estimates fo Some Biased Estimators*Technometrics* **16 (3)** 461–464.

[32] Marquardt D.W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation *Technometrics* **12 (3)** 591–612.

[33] Marquardt D.W., Snee R. D. (1975). Ridge Regression in Practice *The American Statistician* **29(1)** 3–20.

[34] Marquardt D.W. (1980). You Should Standardize the Predictor Variables in Your Regression Model. Comment to "‘A Critique of Some Ridge Regression Methods"’ *The American Statistical Association* **75** 87–91.

[35] Mason R.L., Gunst R.F., Webster J.T. (1975). Regression Analysis and Problems of Multicollinearity *Communications in Statistics* **4 (3)** 277–292.

[36] Mayer, L.S. and Willke, T.A. (1973). On Biased Estimation in Linear Models *Technometrics* **15 (3)** 497–508.

[37] McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo Evaluation of Some Ridge–Type Estimators *Journal of the American Statistical Association* **70 (350)** 407–416.

[38] Montgomery D. C., Peck E. A. Vining G. G. (2006). *Introduction to Linear Regression Analysis*, 4nd ed. Wiley & Sons Inc., Hoboken, New Jersey

[39] Newhouse, Joseph P. and Oman, Samuel D. (1971). An Evaluation of Ridge Estimators *RAND Report* **R-716-PR**.

[40] Obenchain, R. L. (1978). Good and Optimal Ridge Estimators *The Annals of Statistics* **6 (5)**, 1111-1121.

[41] Ohtani, K. (1996). Generalized Ridge Regression Estimators under the LINEX Loss Function *Statistical Papers* **36** 99–110.

[42] Ohtani, K. (2000). *Shrinkage Estimation of a Linear Regression Model in Econometrics* Nova Science Publishers, New York.

[43] Rao, C.R. (1975). Simultaneous Estimation of Parameters in Different Linear Models and Applications to Biometric Problems *Biometrics* **31** 545–554.

[44] Stewart, G.W. (1987). Collinearity and Least Squares Regression (with Comment)*Statistical Science* **2(1)** 68–100.

[45] Rao,C.R. and Toutenbourg, H. (2007) **3**. *Linear Models and Generalizations: Least Squares and Alternatives* Springer, Berlin.

[46] Silvey, S.D. (1969). Multicollinearity and Imprecise Estimation *Journal of the Royal Statistical Society* **B,31** 539–552.

[47] Smith L. (1984). *Linear Algebra*, 2nd ed. Springer, Berlin.

[48] Smith G., Campbell F. (1980). A Critique of Some Ridge Regression Methods *Journal of the American Statistical Association* **75(369)** 74–81.

[49] Stahlecker, P. and Trenkler, G. (1983). On Heterogeneous Versions of the Best Linear and Ridge Estimator*Proc. First Tampere Sem. Linear Models* 301–322.

[50] Stewart, G. W. (1987). Collinearity and Least Squares Regression *Statistical Science* **2(1)** 68–84.

[51] Stewart, G. W. (1969). On the Continuity of the Generalized Inverse *SIAM Journal of Applied Mathematics* **17** 33–45.

[52] Stewart, G. W. (1973). *Introduction to Matrix Computations*, Academic Press, New York.

[53] Strang, G. (1976). *Linear Algebra and its Applications*, Academic Press, New York.

[54] Theil, H. (1971). *Principles of Econometrics*, John Wiley & Sons, Inc., New York.

[55] Theobald, C.M. (1974). Generalizations of Mean Squared Error Applied to Ridge Regression *Journal of the Royal Statistical Society* **B,36** 103–106.

[56] Thomas, G.B. and Finney, R. L. (1998). *Calculus and Analytic Geometry*, 9th ed. Addison–Wesley, inc.

[57] Trenkler, G. (1980). Generalized Mean Squared Error Comparisons of Biased Regression Estimators *Communications in Statistics–Theory and Methods* **A 9 (12)** 1247–1259.

[58] Trenkler Goetz (1981). *Biased Estimators in the Linear Regression Model*, Mathematical systems in economics, 58, Oelschlager Gunn & Hain.

[59] Trenkler, G. and Trenkler, D. (1983). A Note on Superiority Comparisons of Homogeneous Linear Estimators *Communications in Statistics–Theory and Methods* **12 (7)** 799–808.

[60] Trenkler, D. and Trenkler, G. (1984). A Simulation Study Comparing Biased Estimators in the Linear Model *Computational Statistics Quaterly* **1(1)** 45–60.

[61] Trenkler, D. and Trenkler, G. (1984). Minimum Mean Squared Error Ridge Estimation *The Indian Journal of Statistics* **46 A** 94–101.

[62] Trenkler, D. and Trenkler, G. (1984). On the Euclidean Distance between Biased Estimators *Communications in Statistics–Theory and Methods* **13(3)** 273–284.

[63] Trenkler, G. and Ihorst, G. (1990). Matrix Mean Squared Error Comparisons Based on a Certain Covariance Structure *Communications in Statistics–Simulation and Computation* **19(3)** 1035–1043.

[64] Trenkler, D. and Trenkler, G. (1995). An Objective Stability Criterion for Selecting the Biasing Parameter From the Ridge Trace *The Journal of the Industrial Mathematics Society* **45(2)** 93–104.

[65] Ueberhuber, C. (1995). *Comuter–Numerik 2*, Springer, Berlin.

[66] Varian, H. R. (1975). A Bayesian Approach to Real Estate Assessment *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, North Holland, Amsterdam 196–208.

[67] Vinod, H. D. and Ullah, A. (1981). *Recent Advances in Regression Methods*, Marcel Dekker, inc., New York.

[68] Wan, Alan T. K. (2002). On Generalized Ridge Regression Estimators under Collinearity and Balanced Loss *Applied Mathematics and Computation* **129** 455–467.

[69] Wichern, D. W. and Churchill, G. A. (1978). A Comparison of Ridge Estimators *Technometrics* **20 (3)** 301–311.

[70] Willan, A. R. and Watts, D. G. (1978). Meaningful Multicollinearity Measures *Technometrics* **20 (4)** 407–412.

[71] Zellner, A. (1986). Bayesian Estimation and Prediction using Asymmetric Loss Functions *Journal of the American Statistical Association* **81** 446–451.

[72] Zellner, A. (1994). Bayesian and Non–Bayesian Estimation Using Balanced Loss Functions *Statistical Decision Theory and Related Topics*, Springer, New York, 377-390.

Acknowledgment

An dieser Stelle möchte ich mich bei allen Personen bedanken, die zum Entstehen dieser Arbeit beigetragen haben.
Besonders möchte ich mich bei Herrn Prof. Dr. Falk für die freundliche Betreuung, sein stets offenes Ohr und die vielen motivierenden Worte bedanken. Auch bin ich ihm für das mir entgegengebrachte Vertrauen in den letzten Jahren zu Dank verpflichtet.
Mein weiterer Dank gilt Herrn Prof. Dr. Trenkler, der sich sofort als Zweitgutachter zur Verfügung gestellt und mir darüber hinaus konstruktive Anregungen für diese Arbeit gegeben hat.
Mein besonderer Dank gilt auch Herrn Dr. Heins, Herrn Dr. Wensauer und Frau Distler, die mir die Zusammenarbeit mit AREVA ermöglicht haben und mir immer hilfreich zur Seite standen.
Desweiteren möchte ich mich bei Matthias Hümmer, Daniel Hofmann und Dominik Streit für die moralische Unterstützung und die schöne Zeit in Erlangen bedanken.
Nicht zuletzt sind noch meine Eltern Hiltraud und Manfred zu nennen, denen ich einfach alles verdanke.

Ich versichere, diese Arbeit eigenhändig angefertigt und dazu nur die angegebenen Quellen benutzt zu haben.

Mömbris, im Februar 2008

Julia Wissel

Eingereicht im Februar 2009
an der Fakultät für Mathematik und Informatik.

1. Gutachter: Prof. Dr. M. Falk
2. Gutachter: Prof. Dr. G. Trenkler

Tag der mündlichen Prüfung: 16.06.2009