

Part I

## **Point Estimation and Linear Regression**



# 1 Fundamentals

In this chapter, a brief introduction into the theory of linear regression models is given and a small numerical example is created, providing the opportunity to pose some of the central problems. In addition, because of the relevance for comparison of point estimators, an introduction into the basics of decision theory is delivered.

## 1.1 Linear Models

A short description of the linear regression model and its representation in matrix notation is given. Different types of linear models are introduced and the process of analysis via linear models is discussed.

### 1.1.1 Application of Linear Models

In a scientific investigation one may often be inclined to formulate a hypothesis about a linear relationship

$$y = \beta_1 x_1 + \cdots + \beta_p x_p$$

between a variable  $y$  on the one side and variables  $x_1, \dots, x_p$  on the other side, making it possible to explain  $y$  via  $x_1, \dots, x_p$ . This can be quite useful when values of  $x_1, \dots, x_p$  are rather easy to obtain, while this is not the case for the corresponding value of  $y$ . Then the above equation may be employed for predicting the outcome of  $y$  given values  $x_1, \dots, x_p$ .

Although one may be able to specify such a hypothesis for different reasons, the exact nature of the considered linear relationship will not be known, or, in other words, the parameters  $\beta_1, \dots, \beta_p$  will be unknown. Information about them can be drawn from a given set of observations of the variables  $y$  and  $x_1, \dots, x_p$ .

*Example 1.1.* Consider for a batch of cement the heat  $y$  evolving during the hardening of the cement. Suppose that the batch consists of amounts of four main ingredients denoted by  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$ . Now, suppose that it is desired to predict the heat on the basis of the amount of one or more of the ingredients. One possible hypothesis might be

$$y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 ,$$

postulating that the heat is a linear function of the two amounts of ingredients  $x_1$  and  $x_2$  (and a constant term). In order to obtain a formula for predicting the heat for given  $x_1$  and  $x_2$ , information about  $\beta_1, \beta_2$  and  $\beta_3$  is required. This may be given in form of point estimates of the three parameters, using the data from Table 1.1. It provides observations of the variable  $y$  (recorded in calories per gram) and of the variables  $x_1, \dots, x_4$  (recorded in addition percentages) for  $n = 13$  batches. See also Appendix C for an analysis of this data with the statistical-computing environment R.  $\square$

**Table 1.1.** Cement data, see [49, p. 647]

Batch	Heat	Ingredient 1	Ingredient 2	Ingredient 3	Ingredient 4
$i$	$y$	$x_1$	$x_2$	$x_3$	$x_4$
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

In the above example we cannot expect that the heat is an exact linear function of the amount of the two ingredients  $x_1$  and  $x_2$ . As a matter of fact, in practice there will be an *approximate* linear rather than an *exact* linear relationship between the variables of interest. One may account for this by adding a further non-observable variable, fancied as a collection of small errors. Then the hypothesis is altered to an equation of the form

$$y = \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon ,$$

letting  $y$  not only depend on  $x_1, \dots, x_p$ , but also on  $\varepsilon$  (the error variable), being random but non-observable.

If it is assumed that for given variables  $y$  and  $x_1, \dots, x_p$  the above hypothesis is correct, then statements about the parameters  $\beta_1, \dots, \beta_p$  can be derived, which should be as close as possible to the unknowns, but nonetheless can only be of stochastic nature. Such ‘statements’ may be given in form of *point estimates* based on (say  $n$ ) rows of observed values

$$(y_1, x_{1,1}, \dots, x_{1,p}), \dots, (y_n, x_{n,1}, \dots, x_{n,p}) .$$

The  $y_1, \dots, y_n$  are regarded as sample values of  $y$ , given fixed values  $(x_{1,1}, \dots, x_{1,p}), \dots, (x_{n,1}, \dots, x_{n,p})$ . Thus  $y_i$  is the observed realization of a random variable, while the observation  $x_{i,j}$  is assumed to be non-stochastic.

In the literature on linear models, there is usually no notational distinction between a realization  $y_i$  and the corresponding random variable standing behind the observation. The random variable  $y_i$  satisfies

$$y_i = \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i ,$$

where  $\varepsilon_i$  is a random non-observable variable. It is obvious that if for a fixed  $i$  one has observed not only one but (infinitely) many realizations of  $y_i$  for fixed  $(x_{i,1}, \dots, x_{i,p})$ , then the corresponding errors should have equalized. Hence it is assumed that the expectation of  $\varepsilon_i$  equals 0. Moreover, it is assumed that all  $\varepsilon_1, \dots, \varepsilon_n$  have the same, possibly unknown, variance  $\sigma^2$ . Eventually, two error variables  $\varepsilon_i$  and  $\varepsilon_k$  are assumed to be uncorrelated for  $i \neq k$ .

The  $n$  equations together with the assumptions about  $\varepsilon_i$  can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n) ,$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} .$$

In the following, this equation will be the considered description of a linear model. The variable  $y$  will be called *dependent* variable while the variables  $x_1, \dots, x_p$  will be called *independent* variables. The vector  $\mathbf{y}$  is the vector of sample variables (each one fancied as a random observable copy of  $y$  given a set of  $x$ -values), while  $\mathbf{X}$  is the matrix comprising the values of the independent variables.

*Example 1.2.* In example 1.1 the observation vector  $\mathbf{y}$  and the matrix  $\mathbf{X}$  are given as

$$\mathbf{y} = \begin{pmatrix} 78.5 \\ 74.3 \\ 104.3 \\ 87.6 \\ 95.9 \\ 109.2 \\ 102.7 \\ 72.5 \\ 93.1 \\ 115.9 \\ 83.8 \\ 113.3 \\ 109.4 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & 7 & 26 \\ 1 & 1 & 29 \\ 1 & 11 & 56 \\ 1 & 11 & 31 \\ 1 & 7 & 52 \\ 1 & 11 & 55 \\ 1 & 3 & 71 \\ 1 & 1 & 31 \\ 1 & 2 & 54 \\ 1 & 21 & 47 \\ 1 & 1 & 40 \\ 1 & 11 & 66 \\ 1 & 10 & 68 \end{pmatrix}.$$

The ordinary least squares estimator for  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$  is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 52.5774 \\ 1.4683 \\ 0.6623 \end{pmatrix}.$$

The estimated relationship, neglecting the error variable, is therefore

$$y = 52.5774 + 1.4683x_1 + 0.6623x_2 \\ (2.2862) \quad (0.1213) \quad (0.0459),$$

where the numbers in brackets are the standard errors (see also Sect. 2.4.1). This equation may be applied to predict the heat  $y$  for given  $x_1$  and  $x_2$ . For example for  $x_1 = 15$  and  $x_2 = 50$  the predicted heat is 107.72. See also Appendix C for a more detailed analysis.  $\square$

### 1.1.2 Types of Linear Models

Depending on the type of variables  $x_1, \dots, x_p$ , linear models can be classified in the following way:

- (1) *Analysis of Variance Model.* When the matrix  $\mathbf{X}$  consists only of zeros and ones, then the linear model will be called analysis of variance model. In this case the variables  $x_1, \dots, x_p$  are *qualitative*. If  $y_i$  denotes the  $i$ -th measurement and  $x_j$  denotes the  $j$ -th treatment, then  $x_{i,j}$  is either 1 or 0, depending on whether the  $i$ -th unit has received the  $j$ -th treatment or not.
- (2) *Regression Model.* If all variables  $x_1, \dots, x_p$  are *quantitative*, then the model will be called regression model. The term ‘regression’ has originally been introduced by Galton [42] in a paper on laws of heredity.

In econometrics, one will usually be confronted with such variables (e.g. when  $x_1$  stands for the total income from trading and  $x_2$  is the total income from wages). If there exists one variable which is a constant equal to 1 and all other variables are quantitative, then the model is called regression with *intercept*.

- (3) *Analysis of Covariance Model*. If some of the variables are (0,1) qualitative and others are quantitative, then the model is sometimes called analysis of covariance model.

In this book we assume that the underlying model is a linear regression model (with or without intercept). Nonetheless, every statement will be true under any type of model which satisfies the required assumptions.

### 1.1.3 Proceeding with Linear Models

For investigating the relationship between variables via linear models, an appropriate procedure is outlined by the following three stages:

*Stage 1: Model Building*. In a first step, the relationship one wishes to investigate must be specified. Having formulated a linear model equation

$$y_i = \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, \dots, n,$$

motivated from formal reasons, in a second step it should be clarified whether there exists some evidence in favor or against the model assumptions (*model diagnostics*). For example residual plots from computer outputs may be used to find some indication in favor or against the assumptions concerning the errors  $\varepsilon_i$ . If it is not sure whether a specific variable  $x_j$ ,  $j \in \{1, \dots, p\}$ , should be incorporated into the model, then a test of the hypothesis  $H_0 : \beta_j = 0$  can be performed. In addition, statistical software packages (as for example S-PLUS or R, cf. [27, 40, 122]) can be used to carry out a forward and/or backward variable elimination process, based e.g. on Akaike's 'An Information Criterion' (AIC). Other plots produced by such software will aid to find extreme values (outliers and/or high leverage points), which could be worth eliminating from the analysis. The stage of model building and diagnostics is rather delicate and complicated, since there are many aspects to consider. One should be aware, that most of the model assumptions cannot be proved, but only be more or less supported by diagnostic methods. Chapter 6 gives a review on regression diagnostics.

*Stage 2: Inference*. If a certain model has been fixed, then the relationship between the dependent variable  $y$  and the independent variables  $x_1, \dots, x_p$  can be examined. That is, statements about the unknown parameters (like point estimates) can be concluded from the observations. In practice, this stage cannot always be clearly distinguished from Stage 1, since for model building/diagnostic purposes it is already required

to derive statements about the unknowns. If Stage 1 is performed, then usually some statistical knowledge about the relevant parameters will automatically be given. (For example testing  $H_0 : \beta_j = 0$  usually requires a point estimate of  $\beta_j$ .) In addition, inference under a given model might lead to conclusions which question the choice of model, thus making it necessary to return to Stage 1.

*Stage 3: Prediction.* Frequently, the ultimate goal of an analysis via linear models is the prediction of  $y$  given some values  $(x_{m,1}, \dots, x_{m,p})$ . Such a prediction will be based on the results from Stages 1 and 2. The more adequate the chosen model and the more precise the obtained estimates, the better one can hope to predict  $y$ .

In the following Chapters 2 to 5 we assume that the stage of model building has been completed. Our interest focuses on inference about the unknown regression parameters under a given model (Stage 2). Here we are mainly interested in comparisons of the performance of different point estimators. Chapter 6 gives a short review on regression diagnostic methods and Appendix C demonstrates the analysis of a linear regression model with the statistical-computing environment R.

#### 1.1.4 A Preliminary Example

The following numerical example will give an impression about the main topic of this text, being the consideration of different estimators for regression parameters and the appropriate evaluation of their performance.

##### The Model

Consider a linear regression model with  $p = 3$  independent (explanatory) variables and  $n = 12$  observations, described by

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i, \quad i = 1, \dots, 12,$$

where  $E(\varepsilon_i) = 0$ ,  $\text{Var}(\varepsilon_i) = \sigma^2$  and  $\text{Cov}(\varepsilon_i, \varepsilon_k) = 0$  for  $i \neq k = 1, \dots, 12$ . The 12 respective observations of the dependent and the three independent variables are given in Table 1.2.

**Table 1.2.** Observed values of variables  $y$  and  $x_1, x_2, x_3$

$i$	1	2	3	4	5	6	7	8	9	10	11	12
$y_i$	2.275	3.658	.9561	3.329	2.462	3.843	.8117	.9014	2.715	3.882	3.196	.7634
$x_{i,1}$	1	1	1	1	1	1	1	1	1	1	1	1
$x_{i,2}$	2.965	2.839	3.466	2.538	1.993	3.670	1.011	2.972	3.504	1.872	2.160	3.218
$x_{i,3}$	1.776	2.365	.1971	1.442	1.419	3.215	1.140	.3691	.3799	.9427	.7657	.3075

Obviously, since  $x_{i,1} = 1$  for each  $i$ ,  $x_1$  is the intercept variable. In matrix notation we can write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_{12}),$$

where the elements  $y_i$  of  $\mathbf{y}$  are random variables for which observations are available from Table 1.2 and the elements  $x_{i,j}$  ( $i = 1, \dots, 12$ ,  $j = 1, 2, 3$ ) of  $\mathbf{X}$  are non-stochastic and determined by Table 1.2.

### Different Estimators

Since the matrix  $\mathbf{X}$  is of full column rank, the *ordinary least squares estimator* for  $\boldsymbol{\beta}$  exists and is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}.$$

By using the values from Table 1.2 we compute

$$\hat{\boldsymbol{\beta}} = (1.4570, -0.0233, 0.8423)'$$

as the least squares estimate for the parameter vector  $\boldsymbol{\beta}$ .

An alternative estimator for  $\boldsymbol{\beta}$  is the so-called *ridge estimator*, given by

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{y}, \quad k \geq 0.$$

We compute the scalar  $k$  from the formula

$$k = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}, \quad \hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

and obtain  $k = 1.2288$ . Then, the ridge estimate for  $\boldsymbol{\beta}$  is

$$\hat{\boldsymbol{\beta}}_k = (0.6715, 0.2518, 0.8242)'.$$

As can be seen, the ridge estimator  $\hat{\boldsymbol{\beta}}_k$  yields a fairly different estimate for the first element  $\beta_1$  of the vector  $\boldsymbol{\beta}$  compared to the ordinary least squares estimator  $\hat{\boldsymbol{\beta}}$ . Similarly, the estimates for  $\beta_2$  are different, while the estimates for  $\beta_3$  are almost the same.

As a second alternative one may consider the so-called *shrinkage estimator*

$$\hat{\boldsymbol{\beta}}(\varrho) = \frac{1}{1+\varrho}\hat{\boldsymbol{\beta}}, \quad \varrho \geq 0.$$

We compute  $\varrho$  from the formula

$$\varrho = \frac{\hat{\sigma}^2 \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}]}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}$$

and obtain  $\varrho = 0.6057$ . Then the shrinkage estimate for  $\beta$  is

$$\hat{\beta}(\varrho) = (0.9074, -0.0145, 0.5246)' .$$

As can be seen, the shrinkage estimator  $\hat{\beta}(\varrho)$  is designed such that it simply multiplies the estimates of  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  obtained from  $\hat{\beta}$  by the same factor  $(1+0.6057)^{-1} = 0.6228$  and thus shrinks the ordinary least squares estimates. This is similar to the behavior of the *Stein estimator* (also called *James-Stein estimator*, see [61]), given as

$$\hat{\beta}_S = \gamma \hat{\beta}, \quad \gamma = 1 - \frac{(p-2)(n-p)}{n-p+2} \frac{\hat{\sigma}^2}{\hat{\beta}' \mathbf{X}' \mathbf{X} \hat{\beta}} .$$

We compute  $\gamma = 0.9875$  and obtain

$$\hat{\beta}_S = (1.4387, -0.0230, 0.8317)'$$

as an estimate for  $\beta$ . Since the computed factor  $\gamma$  is close to 1, the estimates from the Stein estimator  $\hat{\beta}_S$  hardly differ from the estimates from the ordinary least squares estimator  $\hat{\beta}$ .

### Observed Loss

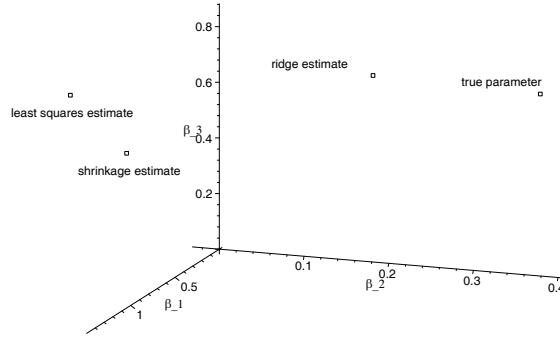
The data in Table 1.2 is obtained as a result from a small simulation process, where in a first step the values of  $x_{i,1}$ ,  $x_{i,2}$  and  $x_{i,3}$  had been fixed. Then 12 realizations of independent  $N(0,1)$  distributed random variables  $\varepsilon_i$  have been obtained by a random generator. The corresponding value of  $y_i$  has been computed from the model equation by using the vector

$$\beta = (0.2, 0.4, 0.7)' .$$

Hence, in distinction to a real-world application, the true parameter vector is known, and we are able to compare the true values with the estimated ones. Fig. 1.1 shows the vector  $\beta$  as a point in a 3-dimensional coordinate system together with the estimates  $\hat{\beta}$ ,  $\hat{\beta}_k$  and  $\hat{\beta}(\varrho)$ . The value of  $\hat{\beta}_S$  is not shown, since it cannot visually be distinguished from the least squares estimate  $\hat{\beta}$ .

Although, due to the graphic presentation, the differences in direction of  $\beta_1$  are not easy to make out, it can be seen that the ridge estimate and the shrinkage estimate are nearer to the true parameter(vector) than the least squares estimate. We compute the squared distances from the respective estimates to  $\beta = (0.2, 0.4, 0.7)'$ , and obtain

$$\begin{aligned} l_1 &= \|\hat{\beta} - \beta\|^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) = 1.7794 , \\ l_2 &= \|\hat{\beta}_k - \beta\|^2 = (\hat{\beta}_k - \beta)'(\hat{\beta}_k - \beta) = 0.2597 , \\ l_3 &= \|\hat{\beta}(\varrho) - \beta\|^2 = (\hat{\beta}(\varrho) - \beta)'(\hat{\beta}(\varrho) - \beta) = 0.7030 , \\ l_4 &= \|\hat{\beta}_S - \beta\|^2 = (\hat{\beta}_S - \beta)'(\hat{\beta}_S - \beta) = 1.7306 . \end{aligned}$$



**Fig. 1.1.** True parameter vector  $\beta = (0.2, 0.4, 0.7)'$  and least squares, ridge and shrinkage estimates

We may interpret these squared distances as *observed losses*, which have to be admitted when the respective estimator is chosen and  $\beta = (0.2, 0.4, 0.7)'$  is the true parameter.

If we compare the values  $l_1$ ,  $l_2$ ,  $l_3$  and  $l_4$ , then we will conclude that the observed loss of the ridge estimator  $\hat{\beta}_k$  is distinctly smaller than the observed loss of the least squares estimator  $\hat{\beta}$ . The shrinkage estimator  $\hat{\beta}(\rho)$  yields a smaller loss compared to the loss of least squares, but a greater loss compared to the loss of ridge. Even the Stein estimator  $\hat{\beta}_S$  produces a smaller observed loss than  $\hat{\beta}$ , although the difference is only marginal.

### Risk

Of course, the observed loss of an estimator depends on the observed realization of  $\mathbf{y}$  (the sample value). If we wish to assess the loss of an estimator  $\tilde{\beta}(\mathbf{y})$  independent of a given realization, then we can consider the *average loss*

$$E \left[ (\tilde{\beta}(\mathbf{y}) - \beta)' (\tilde{\beta}(\mathbf{y}) - \beta) \right] .$$

Here we do not consider a specific estimate  $\tilde{\beta}(\mathbf{y})$  depending on the observed value  $\mathbf{y}$ , but the random vector  $\tilde{\beta}(\mathbf{y})$  depending on the random vector  $\mathbf{y}$ . This expected loss is usually called the *risk* of the estimator  $\tilde{\beta}(\mathbf{y})$  with respect to  $\beta$ .

In the above situation an estimator can be called *better* than another, if it has a smaller risk with respect to  $\beta = (0.2, 0.4, 0.7)'$ . Of course, in practice we do not know the true parameter vector, so we are interested in risk comparisons which do not only hold for some fixed value of  $\beta$ , but for at least a certain set of possible values  $\beta \in \mathbb{R}^p$  and a certain set of possible values  $\sigma^2 > 0$ . Clearly it would be most helpful to have risk inequalities between estimators for  $\beta$  being valid for all possible parameter values  $\beta$ , but this will rarely turn out to be the case if we consider reasonable estimators.

### Some Questions

As we have demonstrated above, there are situations in which certain alternatives to the least squares estimator deliver smaller observed losses than  $\hat{\beta}$ . This, however, does not imply that expected losses must behave similarly, meaning that we cannot draw any reasonable conclusions about the behavior of the respective estimators from the above example. Nonetheless, the above results raise some questions:

- Is it possible that a better estimator than  $\hat{\beta}$  exists for all possible  $\beta \in \mathbb{R}^p$  and all  $\sigma^2 > 0$ ? In other words, does there exist an estimator which makes  $\hat{\beta}$  *inadmissible* for estimating  $\beta$ ?
- Do there exist different possibilities to define ‘better’? Are there different reasonable losses and risks?
- Do there exist estimators of  $\beta$  which are better than  $\hat{\beta}$  for certain sets of  $\beta \in \mathbb{R}^p$  and  $\sigma^2 > 0$ ? In other words, do there exist estimators which turn out to be *admissible* compared to  $\hat{\beta}$ ?
- If there existed such admissible estimators, under what conditions should they be used?
- Can we find estimators which are admissible compared to any other estimator? In such a case, we can never find a *uniformly* better estimator, i.e. an estimator which is better for all  $\beta \in \mathbb{R}^p$  and all  $\sigma^2 > 0$

Terms like ‘loss’, ‘risk’ or ‘admissibility’ are widely used in decision theory. Therefore, before trying to give some answers to the above questions, we present a short introduction into this theory.

## 1.2 Decision Theory and Point Estimation

Statistical decision theory has been established by Abraham Wald with a series of papers in the 1940s being combined 1950 in his famous book *Statistical Decision Functions* [124]. Further contributions on this topic are for example provided by Ferguson [38] and Berger [13].

### 1.2.1 Decision Rule

Suppose we have to reach a decision  $d$  depending on  $p$  unknown quantities  $\theta_1, \dots, \theta_p$ . These quantities are combined in the parameter vector  $\theta = (\theta_1, \dots, \theta_p)'$ . Then:

- the set of all possible parameter vectors  $\theta$  is called *parameter space* and is denoted by  $\Theta$ ;
- the set of all possible decisions  $d$  is called *decision space* and is denoted by  $D$ .

*Example 1.3.* Suppose that we want to decide whether to use a certain coin for a game of chance or not. The set of decisions  $D$  consists of the two elements ‘use coin’ and ‘reject coin’. The decision depends on the unknown probability  $p$  for the appearance of the event ‘face side’. Hence the parameter space is  $\Theta = [0, 1]$ .  $\square$

To make a specific decision, claims about the unknown parameters  $\theta_1, \dots, \theta_p$  are required. To obtain such claims, a statistical experiment is conducted, being designed such that some conclusions about the vector  $\boldsymbol{\theta}$  are possible. More precisely, the experiment will yield an observation  $\mathbf{y} = (y_1, \dots, y_n)'$  of a random vector  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  whose distribution depends on  $\boldsymbol{\theta}$ . Given a certain *decision rule*  $\delta$  and given the observation  $\mathbf{y}$ , a decision  $d$  is chosen out of the decision space  $D$ .

**Definition 1.1.** *A mapping*

$$\begin{aligned} \delta : \mathbb{R}^n &\rightarrow D \\ \mathbf{y} &\mapsto \delta(\mathbf{y}) = d \end{aligned}$$

which assigns each observed value  $\mathbf{y}$  exactly one decision  $d$  is called decision rule.

Since  $\mathbf{Y}$  is a random vector, it is quite reasonable to demand the same property for the function  $\delta$ , i.e.  $\delta(\mathbf{Y})$  should be a random vector/variable, too. Hence, the decision space  $D$  must come along with an appropriate  $\sigma$ -algebra  $\mathcal{D}$ , so that  $(D, \mathcal{D})$  is a measurable space and  $\delta$  is  $(\mathcal{B}^n, \mathcal{D})$  measurable, where  $\mathcal{B}^n$  stands for the  $n$ -dimensional Borel algebra.

**Assumption 1.1.** *In the following we will only consider decision rules  $\delta$  such that  $\delta(\mathbf{Y})$  is a random vector/variable.*

*Remark 1.1.* In point estimation theory, the decision simply consists in accepting one of the possible values from  $\Theta$ . Thus  $D = \Theta$  and decision rules  $\delta$  are point estimators of  $\boldsymbol{\theta}$ .

*Example 1.4.* We can transform the decision problem from Example 1.3 to a point estimation problem simply by requiring a decision for a specific value  $p$  in  $[0, 1]$ . Then,  $D$  becomes  $D = \Theta = [0, 1]$ .

To obtain some information about the unknown  $p$  we can flip the coin 1000 times. Each time we note a 1 for the event ‘face side’ and a 0 otherwise. By this, we get 1000 realizations  $\mathbf{y} = (y_1, \dots, y_{1000})'$  of  $n = 1000$  independent random variables  $\mathbf{Y} = (Y_1, \dots, Y_n)'$ , each one being Bernoulli distributed with parameter  $p$ .

Let the set of decision rules  $\Delta$  be the set of all point estimators for  $p$ . We choose the rule

$$\delta(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n Y_i$$

and come to the decision  $\delta(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^n y_i = \frac{516}{1000} = 0.516$  (say).  $\square$

### 1.2.2 Non-operational Decision Rule

Decision rules  $\delta(\mathbf{Y})$  are random variables/vectors which yield a decision from the decision space  $D$  as soon as the sample value  $\mathbf{y}$  has been observed. The value  $\mathbf{y}$  is the realization of the random vector  $\mathbf{Y}$  whose distribution depends on some unknown vector of parameters  $\boldsymbol{\theta} \in \Theta$ . Sometimes, the distribution of  $\mathbf{Y}$  does not only depend on the vector  $\boldsymbol{\theta}$  of interest, but also on some additional unknown vector  $\boldsymbol{\xi} \in \Xi$ .

If we are solely interested in point estimation of the vector  $\boldsymbol{\theta}$ , then we can consider functions

$$\delta(\mathbf{Y}, \boldsymbol{\xi}) ,$$

which should have been reasonable point estimators for  $\boldsymbol{\theta}$ , if  $\boldsymbol{\xi}$  had been known. But since this is not the case, these functions do not yield a specific value from the parameter space  $\Theta$  and hence are *non-operational*.

If we replace the vector  $\boldsymbol{\xi}$  in  $\delta(\mathbf{Y}, \boldsymbol{\xi})$  by an appropriate point estimator  $\delta_0(\mathbf{Y})$  of  $\boldsymbol{\xi}$ , then we obtain an *operational variant*

$$\delta(\mathbf{Y}, \delta_0(\mathbf{Y})) ,$$

which (given an appropriate choice of  $\delta_0$ ) yields a value from  $\Theta$  when  $\mathbf{Y}$  is replaced by the sample value  $\mathbf{y}$ .

*Example 1.5.* We wish to estimate the unknown variance  $\sigma^2$  of a normally distributed random variable. Suppose that an experiment delivers the realizations of  $n$  independent  $N(\mu, \sigma^2)$  distributed random variables  $Y_1, \dots, Y_n$ , where neither  $\mu \in \mathbb{R}$  nor  $\sigma^2 \in (0, \infty)$  are known. For some reasons it appears that

$$\delta(\mathbf{Y}, \mu) = \frac{1}{n} \sum_{i=1}^n (Y_i - \mu)^2$$

could be an appropriate ‘point estimator’ of  $\sigma^2$ . But since  $\mu$  is not known, this ‘decision rule’ does not yield a specific value from  $(0, \infty)$  when we replace  $\mathbf{Y}$  by its observation. Hence, by this procedure we cannot come to the desired decision, so that  $\delta(\mathbf{Y}, \mu)$  is not operational as a decision rule. When we replace the unknown parameter  $\mu$  in  $\delta$  by the point estimator  $\delta_0(\mathbf{Y}) = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ , then we obtain

$$\delta(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

as an operational variant of  $\delta(\mathbf{Y}, \mu)$ . Note that  $\delta(\mathbf{Y})$  is a random variable which yields a specific value from the parameter space  $\Theta = (0, \infty)$  of  $\sigma^2$  when the observed  $y_1, \dots, y_n$  are given. This shows that in contrast to  $\delta(\mathbf{Y}, \mu)$  the function  $\delta(\mathbf{Y})$  can be called a point estimator. The two functions  $\delta(\mathbf{Y}, \mu)$  for known  $\mu$  and  $\delta(\mathbf{Y})$  do not necessarily have similar properties. For example, if  $\mu$  is known, then  $\delta(\mathbf{Y}, \mu)$  is unbiased for  $\sigma^2$ , while the operational variant  $\delta(\mathbf{Y})$  is always biased.  $\square$

*Remark 1.2.* The above example shows that there might exist situations when it can be useful to consider a non-operational decision rule as the basis for some operational rule.

In the following we will not use the terms ‘non-operational decision rule’ or ‘non-operational estimator’, since they comprise some kind of contradiction. The term ‘non-operational’ indicates that no decision can be made, while the terms ‘decision rule’ or ‘estimator’ imply by definition that a specific decision or estimate is assigned to a given sample value  $\mathbf{y}$ . Moreover, if we extend non-operational decision rules to functions not only depending on some unknown  $\xi \in \Xi$ , but also on the parameter of interest  $\theta \in \Theta$ , then the most appropriate decision rule for  $\theta$  would be  $\theta$  itself, which obviously is not very helpful. (Note that the parameter  $\theta$  is in fact an element from the parameter space, but not a specific value unless known and thus not a decision.)

### 1.2.3 Loss and Risk

A concomitant of any decision rule is a *loss function*, assigning a value to the parameter  $\theta$  and the decision  $d = \delta(\mathbf{y})$ . This value determines the extent of loss we have to admit when  $\theta$  is the true parameter and the decision  $d = \delta(\mathbf{y})$  is taken. A loss function is usually designed such that the loss is zero when the decision is correct, while it is growing with an increasing level of incorrectness. The choice of a specific loss function depends on the regarded decision problem.

**Definition 1.2.** *A mapping*

$$L : \Theta \times D \rightarrow \mathbb{R} ,$$

where  $L(\theta, d)$  gives the loss when  $\theta$  is the true parameter and the decision  $d$  is taken, is called *loss function*.

**Assumption 1.2.** *In the following we only consider loss functions  $L$  such that  $L(\theta, \delta(\mathbf{Y}))$  is a random variable (in the second component).*

If an observation  $\mathbf{y}$  is given and a specific value of  $\theta$  is considered, then the loss  $L(\theta, \delta(\mathbf{y}))$ , arising when this specific value of  $\theta$  is the true parameter and the decision  $\delta(\mathbf{y})$  is taken, can be computed (*observed loss*).

If we are interested in assessing a decision rule  $\delta(\mathbf{Y})$  independent of a given realization of  $\mathbf{Y}$ , then it is nearby to consider the average loss with respect to all possible realizations of  $\mathbf{Y}$ .

**Definition 1.3.** *The expected loss*

$$\rho(\theta, \delta) = \mathbb{E} [L(\theta, \delta(\mathbf{Y}))] ,$$

of a decision rule  $\delta(\mathbf{Y})$  is called the *risk of  $\delta(\mathbf{Y})$  when  $\theta$  is the true parameter*.

*Example 1.6.* In Example 1.3 we can compute the loss of an estimator  $\delta(\mathbf{Y})$  by using the function

$$L(p, \delta(\mathbf{Y})) = (\delta(\mathbf{Y}) - p)^2.$$

Clearly this loss function is a random variable in the second component. If  $p = 0.5$  (say) had been the true parameter and  $\delta(\mathbf{Y})$  had delivered the decision  $\delta(\mathbf{y}) = 0.5$ , then the observed loss would have been 0. The farther a decision had been away from 0.5, the greater would have been the corresponding loss.

If we make the decision  $\delta(\mathbf{y}) = 0.516$ , as before, then the loss will be  $L(0.5, 0.516) = (0.516 - 0.5)^2 = 0.000256$  when  $p = 0.5$  is the true parameter.

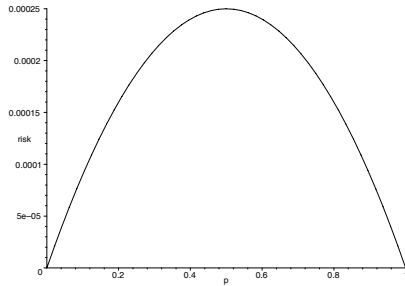
What about the loss independent of a given realization  $\mathbf{y}$ ? For our decision rule  $\delta(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n Y_i$  with  $n = 1000$  we have

$$E(\delta(\mathbf{Y})) = p$$

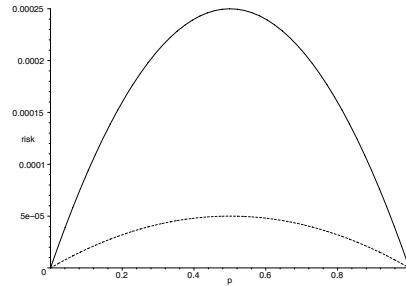
and therefore

$$\rho(p, \delta) = E[L(p, \delta(\mathbf{Y}))] = E[(\delta(\mathbf{Y}) - p)^2] = \text{Var}(\delta(\mathbf{Y})) = \frac{p(1-p)}{1000}.$$

Hence  $\rho(0.5, \delta) = 0.00025$  is the risk of  $\delta$  when  $p = 0.5$  is the true parameter. Of course we can determine the risk for every possible value  $p \in [0, 1]$ , see also Fig. 1.2.  $\square$



**Fig. 1.2.** Risk of the decision rule  $\delta(\mathbf{Y}) = \frac{1}{1000} \sum_{i=1}^{1000} Y_i$  in Example 1.6



**Fig. 1.3.** Risk of the decision rule  $\delta(\mathbf{Y}) = \frac{1}{1000} \sum_{i=1}^{1000} Y_i$  (solid line) and risk of  $\delta_*(\mathbf{Y})$  (dotted line) in Example 1.7

### 1.2.4 Choosing a Decision Rule

In Example 1.6 we have considered only *one* decision rule to come to a decision (point estimation of  $p$ ). In many cases, however, we can choose between many different decision rules to solve a specific decision problem. Then the problem is to find the most appropriate one. A possible way is to compare the risks of the respective rules for all possible values from the parameter space.

**Definition 1.4.** Let  $(\Theta, D, L, \mathbf{Y})$  denote a decision problem and let  $\delta_1(\mathbf{Y})$  and  $\delta_2(\mathbf{Y})$  be two decision rules. Then:

- (i) The rule  $\delta_1(\mathbf{Y})$  is called *uniformly not worse than* the rule  $\delta_2(\mathbf{Y})$  if

$$\rho(\boldsymbol{\theta}, \delta_1) \leq \rho(\boldsymbol{\theta}, \delta_2)$$

is satisfied for all  $\boldsymbol{\theta} \in \Theta$ .

- (ii) The rule  $\delta_1(\mathbf{Y})$  is called *uniformly better than* the rule  $\delta_2(\mathbf{Y})$  if  $\delta_1(\mathbf{Y})$  is uniformly not worse than  $\delta_2(\mathbf{Y})$  and in addition

$$\rho(\boldsymbol{\theta}, \delta_1) < \rho(\boldsymbol{\theta}, \delta_2)$$

for at least one  $\boldsymbol{\theta} \in \Theta$ .

In general, the term ‘uniformly better’ could also be interpreted in the sense that the risk of one rule is strictly smaller than the risk of another rule for *all* possible values from the parameter space. From the above definition, however, ‘uniformly better’ essentially means ‘uniformly not worse’, but ‘better’ for at least one parameter. One may call  $\delta_1(\mathbf{Y})$  *uniformly strictly better* than  $\delta_2(\mathbf{Y})$  if

$$\rho(\boldsymbol{\theta}, \delta_1) < \rho(\boldsymbol{\theta}, \delta_2)$$

for all  $\boldsymbol{\theta} \in \Theta$ .

*Example 1.7.* Suppose that in Example 1.3 we have a second decision rule  $\delta_*(\mathbf{Y})$  at hand with a risk function  $\rho(p, \delta_*)$  as shown in Fig. 1.3 by the dotted line. In this case the rule  $\delta_*$  is a uniformly better estimator of  $p$  than  $\delta$ . (The risk of  $\delta_*$  is strictly smaller than the risk of  $\delta$  for all possible values of  $p$ , except for  $p = 0$  and  $p = 1$ , where both risks are equal to zero.)  $\square$

Note that in Fig. 1.3 the dotted risk line  $\rho(p, \delta_*)$  is given by  $\rho(p, \delta_*) = p(1-p)/5000$ . Therefore, in fact, Fig. 1.3 actually does not show a comparison of the risks of two decision functions given a single decision problem, but one decision rule  $\delta(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n Y_i$  given two decision problems (flipping a coin  $n = 1000$  times and flipping a coin  $n = 5000$  times). The motivation for Fig. 1.3 is simply to show two visually distinct risk functions, one dominating the other.

*Remark 1.3.* In the following we only consider the comparison of different decision rules given a *specific* decision problem and not the comparison of different decision problems. This also implies that we compare different decision rules for a fixed sample size  $n$ .

*Example 1.8.* Suppose that in Example 1.3 we wish to compare the two decision rules

$$\delta_1(\mathbf{Y}) = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{and} \quad \delta_2(\mathbf{Y}) = \frac{n\bar{Y} + \sqrt{n/4}}{n + \sqrt{n}}$$

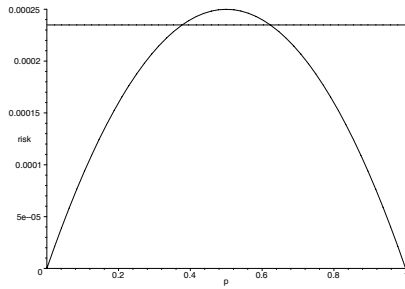
for  $n = 1000$ . As noted before, the risk of  $\delta_1$  is given by

$$\rho(p, \delta_1) = \mathbb{E}[(\delta_1(\mathbf{Y}) - p)^2] = \frac{p(1-p)}{n},$$

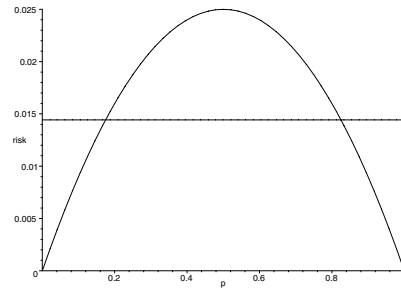
while the risk of  $\delta_2$  comes up to

$$\rho(p, \delta_2) = \mathbb{E}[(\delta_2(\mathbf{Y}) - p)^2] = \frac{n}{4(n + \sqrt{n})^2}$$

(Problem 1.3). Here we neither have  $\rho(p, \delta_1) \leq \rho(p, \delta_2)$  nor  $\rho(p, \delta_2) \leq \rho(p, \delta_1)$  for *all* values  $p \in [0, 1]$ , so that none of the two estimators is uniformly not worse than the other, see also Fig. 1.4.  $\square$



**Fig. 1.4.** Risk of  $\delta_1$  (parabola) and risk of  $\delta_2$  (horizontal line) in Example 1.8 for  $n = 1000$



**Fig. 1.5.** Risk of  $\delta_1$  (parabola) and risk of  $\delta_2$  (horizontal line) in Example 1.8 for  $n = 10$

In the above example none of the two decision rules  $\delta_1$  and  $\delta_2$  is preferable to the other on the basis of the squared error risk. However, we can see that  $\delta_1$  has a smaller risk for all those values of  $p$  which are not close to 0.5. Thus, if we have no prior knowledge that  $p$  is somewhere near to 0.5, we are inclined to use  $\delta_1$  rather than  $\delta_2$ . The situation had been quite different if the number  $n$  would have been much smaller than  $n = 1000$ . In this case the risk of  $\delta_2$  can be seen to be smaller than the risk of  $\delta_1$  for a broad range of values of  $p$ , see Fig. 1.5, and one is then inclined to use  $\delta_2$  as a decision function. This shows that even in those cases when two estimators are compared with respect to their risks and none of them turns out to be uniformly not worse (or better) than the other, the risk comparison can give valuable information about the decision rules in question, see also [19, p. 332].

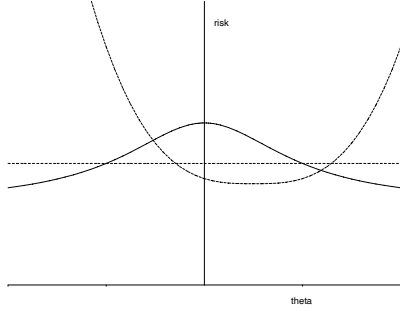
### 1.2.5 Admissibility

The choice of a decision rule depends on the given loss and corresponding risk function as well as on the set of given decision rules we can choose from.

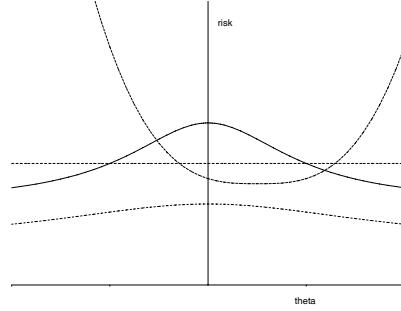
This set will be denoted by  $\Delta$  in the following. If we have a set  $\Delta_1$  at hand, then the most appropriate rule from this set can be different from the most appropriate rule from another set  $\Delta_2$ .

*Example 1.9.* If in a given decision problem  $(\Theta, D, L, \mathbf{Y})$  the set  $\Delta$  contains the three decision rules  $\delta_1(\mathbf{Y})$ ,  $\delta_2(\mathbf{Y})$  and  $\delta_3(\mathbf{Y})$  with corresponding risk functions as shown in Fig. 1.6, then none of the rules is uniformly better than another. Hence we can say that in  $\Delta = \{\delta_1(\mathbf{Y}), \delta_2(\mathbf{Y}), \delta_3(\mathbf{Y})\}$  all elements are *admissible* with respect to the risk.

If, however,  $\Delta$  contains the decision rules  $\delta_1(\mathbf{Y})$ ,  $\delta_2(\mathbf{Y})$ ,  $\delta_3(\mathbf{Y})$  as well as a fourth rule  $\delta_4(\mathbf{Y})$  with risk functions as shown in Fig. 1.7, then  $\delta_4(\mathbf{Y})$  is uniformly better than each of the three others. If we can choose from this set of decision rules, it is not reasonable to choose one of the rules  $\delta_1(\mathbf{Y})$ ,  $\delta_2(\mathbf{Y})$  and  $\delta_3(\mathbf{Y})$  when the risk is a criterion. Hence we can say that these three rules are *inadmissible* in  $\Delta = \{\delta_1(\mathbf{Y}), \delta_2(\mathbf{Y}), \delta_3(\mathbf{Y}), \delta_4(\mathbf{Y})\}$  with respect to the risk.  $\square$



**Fig. 1.6.** Risks of the rules  $\delta_1(\mathbf{Y})$ ,  $\delta_2(\mathbf{Y})$  and  $\delta_3(\mathbf{Y})$



**Fig. 1.7.** Risks of the rules  $\delta_1(\mathbf{Y})$ ,  $\delta_2(\mathbf{Y})$ ,  $\delta_3(\mathbf{Y})$  and  $\delta_4(\mathbf{Y})$

**Definition 1.5.** Let us be given a decision problem  $(\Theta, D, L, \mathbf{Y})$  and a set of decision functions  $\Delta$ . A decision rule  $\delta_0(\mathbf{Y}) \in \Delta$  is called *admissible* in  $\Delta$ , if there does not exist a rule in  $\Delta$  which is uniformly better than  $\delta_0(\mathbf{Y})$ . Otherwise  $\delta_0(\mathbf{Y})$  is called *inadmissible* in  $\Delta$ .

*Remark 1.4.* Admissibility is not a criterion for choosing a decision rule. An admissible rule is not necessarily a reasonable rule.

Problem 1.3 gives an example for a hardly reasonable but nonetheless admissible estimator in a set  $\Delta$ . On the other hand, if we want to use a specific decision rule which appears to be practical for some reason, then this rule should be admissible within some set  $\Delta$ , since otherwise there exists a more practical (i.e. uniformly better) rule in this set.

### 1.2.6 Squared Error Loss

Remark 1.1 implies that in point estimation problems the set of all possible decisions  $D$  coincides with the set of all possible parameters  $\Theta$ . In this case, decision rules  $\delta(\mathbf{Y})$  are point estimators of the unknown parameter vector  $\boldsymbol{\theta}$ , i.e. functions of the random vector  $\mathbf{Y}$  which assign an observation  $\mathbf{y}$  of  $\mathbf{Y}$  to a specific value  $\delta(\mathbf{y})$  from the parameter space  $\Theta$ . In point estimation problems we will assume that the  $p \times 1$  parameter vector  $\boldsymbol{\theta}$  is real-valued.

A frequently used criterion to evaluate the performance of an estimator  $\delta(\mathbf{Y})$  is the *weighted squared error loss function*

$$L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = (\delta(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta(\mathbf{Y}) - \boldsymbol{\theta})$$

with corresponding weighted squared error risk

$$\rho(\boldsymbol{\theta}, \delta) = \mathbb{E} [(\delta(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta(\mathbf{Y}) - \boldsymbol{\theta})] .$$

The  $p \times p$  weight matrix  $\mathbf{W}$  is assumed to be symmetric nonnegative definite.

### Squared Error Loss as Distance

If  $\mathbf{W}$  is a  $p \times p$  symmetric nonnegative definite matrix, then the mapping

$$\begin{aligned} \|\cdot\|_{\mathbf{W}} : \mathbb{R}^p &\rightarrow \mathbb{R} \\ \mathbf{x} &\mapsto \|\mathbf{x}\|_{\mathbf{W}} = \sqrt{\mathbf{x}' \mathbf{W} \mathbf{x}} \end{aligned}$$

satisfies the conditions of a *vector seminorm*, see Sect. A.3.4. If in addition  $\mathbf{W}$  is nonsingular, then  $\|\cdot\|_{\mathbf{W}}$  satisfies the conditions of a *vector norm*. Hence, the weighted squared error loss is the squared distance between  $\delta(\mathbf{Y})$  and  $\boldsymbol{\theta}$  with respect to a vector (semi)norm  $\|\cdot\|_{\mathbf{W}}$ .

### Different Weight Matrices

The application of weight matrices  $\mathbf{W} \neq \mathbf{I}_p$  can sometimes be useful under specific estimation problems. For example,  $\mathbf{W}$  could be a diagonal matrix with different weights on its main diagonal such that individual squared distances  $(\delta(\mathbf{Y})_i - \boldsymbol{\theta}_i)^2$  enter the total loss with higher or lower magnitude.

Different weight matrices can cause seriously different losses. For example the observed unweighted squared error loss ( $\mathbf{W} = \mathbf{I}_p$ ) of an estimator  $\delta_1$  can be smaller than the observed unweighted squared error loss of an estimator  $\delta_2$ , while this relation can be reversed when an alternative weight matrix  $\mathbf{W} \neq \mathbf{I}_p$  is used, see Problem 1.5.

### Weighted Squared Error Loss and Admissibility

Although the choice of the weight matrix  $\mathbf{W}$  can be of great importance when the performance of two estimators is compared, it turns out to be reasonable to consider the choice  $\mathbf{W} = \mathbf{I}_p$  when admissibility of an estimator is investigated, see the following theorem. For this, it is assumed that the set of decision rules  $\Delta$  containing the estimators of a  $p \times 1$  parameter vector  $\boldsymbol{\theta}$  satisfies

$$\delta_0, \delta_1 \in \Delta \quad \Rightarrow \quad \delta_0 + \mathbf{A}(\delta_1 - \delta_0) \in \Delta$$

for any  $p \times p$  matrix  $\mathbf{A}$ . Compare also Problems 1.7 and 1.8.

**Theorem 1.1.** *Let us be given an estimation problem  $(\Theta, L, \mathbf{Y})$  with weighted squared error loss function  $L$ . If  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  in the case  $\mathbf{W} = \mathbf{I}_p$ , then  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  in the case  $\mathbf{W} = \mathbf{W}_*$ , where  $\mathbf{W}_*$  is an arbitrary  $p \times p$  symmetric nonnegative definite matrix.*

*Proof.* For simplicity we will write  $\delta$  to denote an estimator  $\delta(\mathbf{Y})$  of  $\boldsymbol{\theta}$ .

Let  $\delta_0$  be admissible for  $\boldsymbol{\theta}$  with  $\mathbf{W} = \mathbf{I}_p$ . We assume that  $\delta_0$  is not admissible for  $\boldsymbol{\theta}$  with  $\mathbf{W} = \mathbf{W}_*$  and show that this assumption leads to a contradiction.

If  $\delta_0$  is not admissible for  $\boldsymbol{\theta}$  in the case  $\mathbf{W} = \mathbf{W}_*$ , then there exists an estimator  $\delta_1$ , which is uniformly better than  $\delta_0$  in the case  $\mathbf{W} = \mathbf{W}_*$ . This means that

$$c(\boldsymbol{\theta}, \mathbf{W}_*) := \mathbb{E}[(\delta_1 - \boldsymbol{\theta})' \mathbf{W}_* (\delta_1 - \boldsymbol{\theta})] - \mathbb{E}[(\delta_0 - \boldsymbol{\theta})' \mathbf{W}_* (\delta_0 - \boldsymbol{\theta})]$$

is smaller than or equal to 0 for all  $\boldsymbol{\theta} \in \Theta$  and strictly smaller than 0 for at least one  $\boldsymbol{\theta} \in \Theta$ .

Let  $\mathbf{F} := (1/\lambda_{\max})\mathbf{W}_*$ , where  $\lambda_{\max}$  denotes the largest eigenvalue of  $\mathbf{W}_*$ . It follows that

$$c(\boldsymbol{\theta}, \mathbf{F}) \begin{cases} \leq 0 & \text{for all } \boldsymbol{\theta} \in \Theta \\ < 0 & \text{for at least one } \boldsymbol{\theta} \in \Theta \end{cases}.$$

Now, define

$$\delta := \delta_0 + \mathbf{F}(\delta_1 - \delta_0).$$

Then

$$\begin{aligned} & \mathbb{E}[(\delta - \boldsymbol{\theta})' (\delta - \boldsymbol{\theta})] \\ &= \mathbb{E}[(\delta_0 + \mathbf{F}(\delta_1 - \delta_0) - \boldsymbol{\theta})' (\delta_0 + \mathbf{F}(\delta_1 - \delta_0) - \boldsymbol{\theta})] \\ &= \mathbb{E}[(\delta_0 - \boldsymbol{\theta})' + (\delta_1 - \delta_0)' \mathbf{F}] [(\delta_0 - \boldsymbol{\theta}) + \mathbf{F}(\delta_1 - \delta_0)] \\ &= \mathbb{E}[(\delta_0 - \boldsymbol{\theta})' (\delta_0 - \boldsymbol{\theta})] + q, \end{aligned}$$

where  $q := \mathbb{E}[(\delta_1 - \delta_0)' \mathbf{F}^2 (\delta_1 - \delta_0) + 2(\delta_1 - \delta_0)' \mathbf{F} (\delta_0 - \boldsymbol{\theta})]$ . Since

$$\mathbf{x}' \mathbf{F}^2 \mathbf{x} \leq \mathbf{x}' \mathbf{F} \mathbf{x}$$

is satisfied for all  $p \times 1$  vectors  $\mathbf{x}$  (see Problem 1.6), it follows

$$\mathbf{E}[(\delta_1 - \delta_0)' \mathbf{F}^2 (\delta_1 - \delta_0)] \leq \mathbf{E}[(\delta_1 - \delta_0)' \mathbf{F} (\delta_1 - \delta_0)] ,$$

see e.g. [79, p. 70]. Thus,

$$q \leq \mathbf{E}[(\delta_1 - \delta_0)' \mathbf{F} (\delta_1 - \delta_0) + 2(\delta_1 - \delta_0)' \mathbf{F} (\delta_0 - \boldsymbol{\theta})] .$$

By some simple transformations it can be shown that the right-hand side of this inequality equals  $c(\boldsymbol{\theta}, \mathbf{F})$ . Therefore

$$\mathbf{E}[(\delta - \boldsymbol{\theta})' (\delta - \boldsymbol{\theta})] \leq \mathbf{E}[(\delta_0 - \boldsymbol{\theta})' (\delta_0 - \boldsymbol{\theta})] + c(\boldsymbol{\theta}, \mathbf{F}) ,$$

and

$$\mathbf{E}[(\delta - \boldsymbol{\theta})' (\delta - \boldsymbol{\theta})] \begin{cases} \leq \mathbf{E}[(\delta_0 - \boldsymbol{\theta})' (\delta_0 - \boldsymbol{\theta})] & \text{for all } \boldsymbol{\theta} \in \Theta \\ < \mathbf{E}[(\delta_0 - \boldsymbol{\theta})' (\delta_0 - \boldsymbol{\theta})] & \text{for at least one } \boldsymbol{\theta} \in \Theta \end{cases} .$$

This shows that  $\delta$  is uniformly better than  $\delta_0$  in case  $\mathbf{W} = \mathbf{I}_p$ , which is a contradiction to the assumed admissibility of  $\delta_0$  in the case  $\mathbf{W} = \mathbf{I}_p$ . Hence  $\delta_0$  must also be admissible for the case  $\mathbf{W} = \mathbf{W}_*$ , which concludes the proof.  $\square$

*Remark 1.5.* If  $\mathbf{W}_*$  is symmetric positive definite, then the assertion in the above theorem can also be reversed: If  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  in the case  $\mathbf{W} = \mathbf{W}_*$ , then  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  in the case  $\mathbf{W} = \mathbf{I}_p$ , see Problem 1.9(b).

*Remark 1.6.* If we consider admissibility with respect to some weighted squared error loss, then the relevant claims can be deduced from the un-weighted case, irrespective of the choice of weight matrix.

### 1.2.7 Matrix Valued Squared Error Loss

According to Definition 1.2, a loss function takes its values in the set of real numbers  $\mathbb{R}$ . In connection with estimating a  $p \times 1$  parameter vector  $\boldsymbol{\theta}$ , the *matrix-valued squared error loss function*

$$L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = (\delta(\mathbf{Y}) - \boldsymbol{\theta})(\delta(\mathbf{Y}) - \boldsymbol{\theta})'$$

is often considered, taking its values in  $\mathbb{R}^{p \times p}$ . The corresponding risk is given by

$$\text{MSE}(\boldsymbol{\theta}, \delta) = \mathbf{E}[(\delta(\mathbf{Y}) - \boldsymbol{\theta})(\delta(\mathbf{Y}) - \boldsymbol{\theta})'] .$$

The  $i$ -th diagonal element of this matrix is the *mean squared error*

$$\text{mse}(\theta_i, \delta(\mathbf{Y})_i) = \mathbf{E}[(\delta(\mathbf{Y})_i - \theta_i)^2]$$

of the estimator  $\delta(\mathbf{Y})_i$  for the  $i$ -th element  $\theta_i$  of the vector  $\boldsymbol{\theta}$ ,  $i = 1, \dots, p$ . Furthermore, for any  $\boldsymbol{\theta} \in \Theta$ , the trace

$$\text{tr}[\mathbf{W} \text{MSE}(\boldsymbol{\theta}, \delta)] = \rho(\boldsymbol{\theta}, \delta) = \mathbf{E}[(\delta(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta(\mathbf{Y}) - \boldsymbol{\theta})]$$

is the risk of  $\delta(\mathbf{Y})$  with respect to the weighted squared error loss with weight matrix  $\mathbf{W}$ .

### Comparison of Estimators

If we use  $\text{MSE}(\boldsymbol{\theta}, \delta)$  for a comparison of estimators, then the corresponding previous definitions must be altered accordingly.

**Definition 1.6.** *Let us be given an estimation problem  $(\Theta, L, \mathbf{Y})$  with matrix-valued squared error loss function  $L$ . Let  $\delta_1(\mathbf{Y})$  and  $\delta_2(\mathbf{Y})$  be two estimators for  $\boldsymbol{\theta}$ . Then:*

- (i) *The estimator  $\delta_1(\mathbf{Y})$  is called uniformly not worse than  $\delta_2(\mathbf{Y})$ , if the symmetric matrix*

$$\text{MSE}(\boldsymbol{\theta}, \delta_2) - \text{MSE}(\boldsymbol{\theta}, \delta_1)$$

*is nonnegative definite for all  $\boldsymbol{\theta} \in \Theta$ .*

- (ii) *The estimator  $\delta_1(\mathbf{Y})$  is called uniformly better than  $\delta_2(\mathbf{Y})$ , if  $\delta_1(\mathbf{Y})$  is uniformly not worse than  $\delta_2(\mathbf{Y})$  and in addition*

$$\text{MSE}(\boldsymbol{\theta}, \delta_1) \neq \text{MSE}(\boldsymbol{\theta}, \delta_2)$$

*for at least one  $\boldsymbol{\theta} \in \Theta$ .*

Note that some authors define  $\delta_1(\mathbf{Y})$  to be uniformly better than  $\delta_2(\mathbf{Y})$  if the difference  $\text{MSE}(\boldsymbol{\theta}, \delta_2) - \text{MSE}(\boldsymbol{\theta}, \delta_1)$  is positive definite for all  $\boldsymbol{\theta} \in \Theta$ . Correspondingly, we would then call  $\delta_1(\mathbf{Y})$  *uniformly strictly better* than  $\delta_2(\mathbf{Y})$ , similarly to the real-valued case. It is essential not to mix up both definitions since they are not equivalent. Clearly, if the difference  $\text{MSE}(\boldsymbol{\theta}, \delta_2) - \text{MSE}(\boldsymbol{\theta}, \delta_1)$  is positive definite for some  $\boldsymbol{\theta} \in \Theta$ , then this difference is also nonnegative definite and in addition nonzero, but the converse is not true, i.e. from ‘uniformly better’ in our sense does not necessarily follow ‘uniformly strictly better’.

### Matrix Valued and Real Valued Risks

To denote the nonnegative definiteness of the symmetric matrix  $\text{MSE}(\boldsymbol{\theta}, \delta_2) - \text{MSE}(\boldsymbol{\theta}, \delta_1)$  we will also write

$$\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2),$$

where  $\leq_L$  is the Löwner partial ordering in the set of square matrices, see also Sect. A.5. The following theorem, see [115], shows the connection between a relationship of this type and a corresponding relationship between weighted real-valued risks.

**Theorem 1.2.** *For an arbitrary  $\boldsymbol{\theta} \in \Theta$ , the inequality*

$$\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2)$$

*is satisfied if and only if*

$$\mathbb{E}[(\delta_1(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_1(\mathbf{Y}) - \boldsymbol{\theta})] \leq \mathbb{E}[(\delta_2(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_2(\mathbf{Y}) - \boldsymbol{\theta})]$$

*is valid for every  $p \times p$  symmetric nonnegative definite matrix  $\mathbf{W}$ .*

*Proof.* For an arbitrary estimator  $\delta(\mathbf{Y})$  of  $\boldsymbol{\theta}$  we have

$$\begin{aligned} & \mathbb{E}[(\delta(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta(\mathbf{Y}) - \boldsymbol{\theta})] \\ &= \mathbb{E}[\text{tr}\{\mathbf{W}(\delta(\mathbf{Y}) - \boldsymbol{\theta})(\delta(\mathbf{Y}) - \boldsymbol{\theta})'\}] = \text{tr}[\mathbf{W} \text{MSE}(\boldsymbol{\theta}, \delta(\mathbf{Y}))] . \end{aligned}$$

This shows

$$\begin{aligned} & \text{tr}[\mathbf{W}\{\text{MSE}(\boldsymbol{\theta}, \delta_2) - \text{MSE}(\boldsymbol{\theta}, \delta_1)\}] \\ &= \mathbb{E}[(\delta_2(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_2(\mathbf{Y}) - \boldsymbol{\theta})] - \mathbb{E}[(\delta_1(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_1(\mathbf{Y}) - \boldsymbol{\theta})] , \end{aligned}$$

and the assertion follows from Theorem A.54.  $\square$

Note that in the above theorem, given  $\boldsymbol{\theta} \in \Theta$ , it is not enough that

$$\mathbb{E}[(\delta_1(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_1(\mathbf{Y}) - \boldsymbol{\theta})] \leq \mathbb{E}[(\delta_2(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_2(\mathbf{Y}) - \boldsymbol{\theta})]$$

for *some* symmetric nonnegative definite  $\mathbf{W}$  to guarantee

$$\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2) ,$$

but the real-valued inequality must hold for *all* weight matrices  $\mathbf{W}$  to imply the matrix-valued inequality.

Moreover, note that if for some  $\boldsymbol{\theta} \in \Theta$  we have

$$\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2), \quad \text{MSE}(\boldsymbol{\theta}, \delta_1) \neq \text{MSE}(\boldsymbol{\theta}, \delta_2) ,$$

then we do not necessarily have

$$\mathbb{E}[(\delta_1(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_1(\mathbf{Y}) - \boldsymbol{\theta})] < \mathbb{E}[(\delta_2(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta_2(\mathbf{Y}) - \boldsymbol{\theta})]$$

for an arbitrary  $p \times p$  symmetric nonnegative definite matrix  $\mathbf{W}$ .

*Remark 1.7.* If an estimator  $\delta_1$  is uniformly not worse than  $\delta_2$  with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ , then it is always uniformly not worse than  $\delta_2$  with respect to an arbitrary real-valued weighted squared error risk. If an estimator  $\delta_1$  is uniformly better than  $\delta_2$  with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ , then it is always uniformly not worse *but not necessarily uniformly better* than  $\delta_2$  with respect to an arbitrary real-valued weighted squared error risk.

Nonetheless, from the relation  $\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2)$  together with  $\text{MSE}(\boldsymbol{\theta}, \delta_1) \neq \text{MSE}(\boldsymbol{\theta}, \delta_2)$  we can always conclude that  $\text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta_1)] < \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta_2)]$ .

*Remark 1.8.* If an estimator  $\delta_1$  is uniformly not worse/better than  $\delta_2$  with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ , then it is uniformly not worse/better with respect to the real-valued *unweighted* squared error risk.

### Comparison of Elements of Estimators

If for some  $\boldsymbol{\theta} \in \Theta$  the inequality

$$\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2)$$

holds true, then for every  $i \in \{1, \dots, p\}$  it follows

$$\mathbb{E}[(\delta_1(\mathbf{Y})_i - \theta_i)^2] \leq \mathbb{E}[(\delta_2(\mathbf{Y})_i - \theta_i)^2] .$$

On the other hand, if for some  $\boldsymbol{\theta} \in \Theta$  the relations

$$\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2), \quad \text{MSE}(\boldsymbol{\theta}, \delta_1) \neq \text{MSE}(\boldsymbol{\theta}, \delta_2)$$

hold true, then the inequality

$$\mathbb{E}[(\delta_1(\mathbf{Y})_i - \theta_i)^2] < \mathbb{E}[(\delta_2(\mathbf{Y})_i - \theta_i)^2]$$

does not necessarily hold for every  $i \in \{1, \dots, p\}$ . The latter would have been only the case if the difference  $\text{MSE}(\boldsymbol{\theta}, \delta_2) - \text{MSE}(\boldsymbol{\theta}, \delta_1)$  had been *positive definite*.

*Remark 1.9.* If an estimator  $\delta_1$  is uniformly not worse than  $\delta_2$  with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ , then each element of  $\delta_1$  is uniformly not worse than the corresponding element of  $\delta_2$  with respect to the usual mean squared error. If an estimator  $\delta_1$  is uniformly better than  $\delta_2$  with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ , then each element of  $\delta_1$  is uniformly not worse *but not necessarily uniformly better* than the corresponding element of  $\delta_2$  with respect to the usual mean squared error.

### Admissibility of Estimators

According to Remark 1.8, for the comparison of two estimators, the matrix-valued squared risk  $\text{MSE}(\boldsymbol{\theta}, \delta)$  is a stronger criterion than the real-valued unweighted squared error risk  $\rho(\boldsymbol{\theta}, \delta) = \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta)]$ . Nonetheless, with regard to admissibility of estimators it is reasonable to consider the latter rather than the former, as shown by the following theorem.

**Theorem 1.3.** *If an estimator  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  with respect to  $\rho(\boldsymbol{\theta}, \delta) = \mathbb{E}[(\delta(\mathbf{Y}) - \boldsymbol{\theta})'(\delta(\mathbf{Y}) - \boldsymbol{\theta})]$ , then  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ .*

*Proof.* We assume that  $\delta_0$  is admissible with respect to  $\rho(\boldsymbol{\theta}, \delta)$  but not with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$  in a set  $\Delta$ , and show that this assumption leads to a contradiction.

If  $\delta_0 \in \Delta$  is not admissible with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ , then there exists a better estimator  $\delta_1 \in \Delta$  with respect to this criterion. By Definition 1.6, this means that the difference

$$\mathbf{D} = \mathbf{E}[(\delta_0 - \boldsymbol{\theta})(\delta_0 - \boldsymbol{\theta})'] - \mathbf{E}[(\delta_1 - \boldsymbol{\theta})(\delta_1 - \boldsymbol{\theta})']$$

is nonnegative definite for all  $\boldsymbol{\theta} \in \Theta$  and nonzero for at least one  $\boldsymbol{\theta} \in \Theta$ . Now, the trace of a nonnegative definite matrix is nonnegative and equals zero if and only if the matrix is the zero matrix. Hence,  $0 \leq \text{tr}(\mathbf{D})$  for all  $\boldsymbol{\theta} \in \Theta$  and  $0 \neq \text{tr}(\mathbf{D})$  for at least one  $\boldsymbol{\theta} \in \Theta$ . This implies

$$\text{tr}[\mathbf{E}[(\delta_1 - \boldsymbol{\theta})(\delta_1 - \boldsymbol{\theta})']] \leq \text{tr}[\mathbf{E}[(\delta_0 - \boldsymbol{\theta})(\delta_0 - \boldsymbol{\theta})']]$$

for all  $\boldsymbol{\theta} \in \Theta$  and

$$\text{tr}[\mathbf{E}[(\delta_1 - \boldsymbol{\theta})(\delta_1 - \boldsymbol{\theta})']] < \text{tr}[\mathbf{E}[(\delta_0 - \boldsymbol{\theta})(\delta_0 - \boldsymbol{\theta})']]$$

for at least one  $\boldsymbol{\theta} \in \Theta$ . Since  $\text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta)] = \rho(\boldsymbol{\theta}, \delta)$ , the two inequalities show that  $\delta_1$  is uniformly better than  $\delta_0$  with respect to  $\rho(\boldsymbol{\theta}, \delta)$ . This contradicts the assumption and shows that  $\delta_0 \in \Delta$  must also be admissible with respect to  $\text{MSE}(\boldsymbol{\theta}, \delta)$ .  $\square$

We have shown that if an estimator is admissible with respect to the real-valued unweighted squared error loss, then it is also admissible with respect to the matrix-valued squared error loss. The converse is not true in general, i.e. there might exist an estimator which is admissible in some set  $\Delta$  with respect to the matrix-valued squared error risk  $\text{MSE}(\boldsymbol{\theta}, \delta)$ , but which is not admissible in the same set  $\Delta$  with respect to the real-valued unweighted squared error risk  $\rho(\boldsymbol{\theta}, \delta) = \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta)]$ .

### 1.2.8 Alternative Loss Functions

Traditionally, the previously described real-valued (weighted) squared error loss and the matrix-valued squared error loss are considered for investigation of the performance of an estimator  $\delta(\mathbf{Y})$  for a parameter vector  $\boldsymbol{\theta} \in \Theta$ . Nonetheless it is sometimes reasonable to take alternative loss functions into account. To give an impression of the possibilities, we introduce some of these functions in the following.

#### Vector Valued Squared Error Loss

The matrix-valued loss function  $L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = (\delta(\mathbf{Y}) - \boldsymbol{\theta})(\delta(\mathbf{Y}) - \boldsymbol{\theta})'$  might be replaced by the vector-valued loss function

$$L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = \text{dg}[(\delta(\mathbf{Y}) - \boldsymbol{\theta})(\delta(\mathbf{Y}) - \boldsymbol{\theta})'] ,$$

where  $\text{dg}(\mathbf{A})$  denotes the vector whose  $i$ -th element is the  $i$ -th element of the main diagonal of the matrix  $\mathbf{A}$ . Then the expected loss is given by  $\text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta)]$ , and a decision rule  $\delta_1(\mathbf{Y})$  is called uniformly not worse than a rule  $\delta_2(\mathbf{Y})$  if the difference

$$\text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_2)] - \text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_1)]$$

is a vector whose every element is nonnegative for all  $\boldsymbol{\theta} \in \Theta$ . We may write

$$\text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_1)] \leq \text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_2)] \quad \forall \boldsymbol{\theta} \in \Theta .$$

If in addition  $\text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_2(\mathbf{Y}))] \neq \text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_1(\mathbf{Y}))]$  for at least one  $\boldsymbol{\theta} \in \Theta$ , then  $\delta_1(\mathbf{Y})$  is called uniformly better than  $\delta_2(\mathbf{Y})$ .

This criterion is stronger than  $\rho(\boldsymbol{\theta}, \delta) = \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta)]$  but weaker than  $\text{MSE}(\boldsymbol{\theta}, \delta)$ . For an arbitrary given  $\boldsymbol{\theta} \in \Theta$  it follows

$$\begin{aligned} \text{MSE}(\boldsymbol{\theta}, \delta_1) &\leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2) \\ \Rightarrow \text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_1)] &\leq \text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_2)] \\ \Rightarrow \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta_1)] &\leq \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta_2)] . \end{aligned}$$

Similarly we can conclude

$$\begin{aligned} \text{MSE}(\boldsymbol{\theta}, \delta_1) &\neq \text{MSE}(\boldsymbol{\theta}, \delta_2) \\ \Rightarrow \text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_1)] &\neq \text{dg}[\text{MSE}(\boldsymbol{\theta}, \delta_2)] \\ \Rightarrow \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta_1)] &\neq \text{tr}[\text{MSE}(\boldsymbol{\theta}, \delta_2)] , \end{aligned}$$

showing that the vector-valued square loss is in between the matrix-valued and the real-valued squared error loss.

### Alternative Distances as Loss Functions

As noted before, a real-valued weighted squared error loss gives the squared distance between  $\delta(\mathbf{Y})$  and  $\boldsymbol{\theta}$  with respect to a vector (semi)norm  $\|\cdot\|_{\mathbf{W}}$ . Of course it is also possible to consider alternative measures of distance as loss functions. For example we may consider the *absolute distance*

$$L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = \sum_{i=1}^p |\delta(\mathbf{Y})_i - \theta_i| .$$

Another example is the *maximal distance*

$$L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = \max\{|\delta(\mathbf{Y})_i - \theta_i|, i = 1, \dots, p\} .$$

Although such measures could be reasonable loss functions, they will, on the whole, cause more mathematical difficulties than the usual squared error distance and thus are rarely applied.

### Balanced Loss Function

Suppose that the distribution of the random vector  $\mathbf{Y}$  depends on  $\boldsymbol{\theta}$  only via  $E(\mathbf{Y}) = \boldsymbol{\theta}$ . For this case Zellner [132] proposes a *balanced loss function*

$$L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = \alpha \|\mathbf{Y} - \delta(\mathbf{Y})\|^2 + (1 - \alpha) \|\boldsymbol{\theta} - \delta(\mathbf{Y})\|^2, \quad 0 \leq \alpha < 1.$$

Here  $\|\mathbf{Y} - \delta(\mathbf{Y})\|^2$  measures how good  $\delta(\mathbf{Y})$  fits  $\mathbf{Y}$ . If we have an observation  $\mathbf{y}$  of  $\mathbf{Y}$ , then  $\|\mathbf{y} - \delta(\mathbf{y})\|^2$  gives the squared distance between the realization  $\mathbf{y}$  and the decision  $\delta(\mathbf{y})$ . Given  $\mathbf{y}$ , this function yields the same value irrespective of  $\boldsymbol{\theta} \in \Theta$ . Thus, standing alone,  $\|\mathbf{Y} - \delta(\mathbf{Y})\|^2$  is not a loss function in our sense. Therefore, the case  $\alpha = 1$  is excluded in the above definition.

The balanced loss function consists of two components. First, a measure of *goodness of fit*  $\|\mathbf{Y} - \delta(\mathbf{Y})\|^2$  with a relative weight  $\alpha$ , and, second, a measure of *goodness of estimation*  $\|\boldsymbol{\theta} - \delta(\mathbf{Y})\|^2$  with a relative weight  $1 - \alpha$ . The latter is already known as the unweighted squared error loss of the estimator  $\delta(\mathbf{Y})$  for  $\boldsymbol{\theta}$ . Dey, Ghosh and Strawderman [29] study the performance of point estimators with respect to the balanced loss function and compare it with their performance under the unweighted squared error loss.

### Asymmetric Loss Function

If we want to estimate a real-valued parameter  $\theta$ , then

$$L(\theta, \delta(\mathbf{Y})) = b(\delta(\mathbf{Y}) - \theta)^2, \quad b > 0,$$

is a squared loss of the point estimator  $\delta(\mathbf{Y})$ . Using this loss function, it doesn't matter whether the error

$$\Phi = \delta(\mathbf{Y}) - \theta$$

is positive or negative, i.e. over- and underestimation of the parameter  $\theta$  are equally treated. Now it is easy to fancy a situation in which overestimation (say) could do more harm than underestimation, e.g. when overestimation is more expensive. To take this into account, one may consider asymmetric loss functions, which assign more loss to overestimation than to the same amount of underestimation (or vice versa). For example the so-called LINEX loss function may be used, given as

$$L(\theta, \delta(\mathbf{Y})) = b[e^{a\Phi} - a\Phi - 1], \quad a \neq 0, b > 0, \quad \Phi = \delta(\mathbf{Y}) - \theta,$$

cf. [131]. The two parameters  $a$  and  $b$  can be chosen according to a specific problem. With the parameter  $b$  we can set an appropriate scale, while we can use the parameter  $a$  to determine the shape of the function. Fig. 1.8 shows the LINEX loss function for  $b = 0.5$  and  $a = 1$  depending on  $\Phi = \delta(\mathbf{Y}) - \theta$ .

In this case overestimation of  $\theta$  yields a greater loss than underestimation. If for example  $\Phi = 3$ , then the loss comes up to  $L(\theta, \delta(\mathbf{Y})) = 8.0428$ , while for  $\Phi = -3$  the loss is only  $L(\theta, \delta(\mathbf{Y})) = 1.0249$ , compare also Fig. 1.8.

For small values of  $|a|$  the LINEX loss function is almost symmetric and yields similar results as a corresponding squared loss function. For estimating a *vector* of parameters  $\boldsymbol{\theta}$ , one may use the LINEX loss to evaluate the loss of the single elements  $\theta_i$ , or one might consider the estimation of a linear combination  $\boldsymbol{\lambda}'\boldsymbol{\theta}$ , where  $\boldsymbol{\lambda}$  is a known vector.

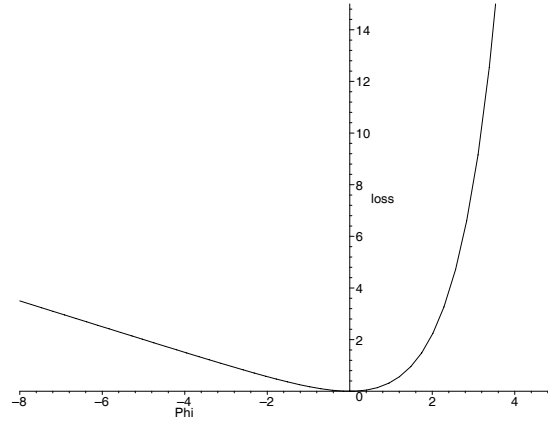


Fig. 1.8. LINEX loss function with  $b = 0.5$  and  $a = 1$

### 1.3 Problems

**1.1.** Let us be given  $p \geq 3$  independent normally distributed random variables  $Y_i \sim N(\mu_i, 1)$ ,  $i = 1, \dots, p$ . Let  $\mathbf{Y} = (Y_1, \dots, Y_p)'$  and  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$ . Let the loss of a decision rule  $\delta(\mathbf{Y})$  for estimating the parameter vector  $\boldsymbol{\mu}$  be  $L(\boldsymbol{\mu}, \delta(\mathbf{Y})) = (\delta(\mathbf{Y}) - \boldsymbol{\mu})'(\delta(\mathbf{Y}) - \boldsymbol{\mu}) = \sum_{i=1}^p (\delta(Y)_i - \mu_i)^2$ .

- Determine the risk of the estimator  $\delta_1(\mathbf{Y}) = \mathbf{Y}$ .
- Show that the risk of the estimator

$$\delta_2(\mathbf{Y}) = \left(1 - \frac{(p-2)}{Z}\right)\mathbf{Y}, \quad Z = \mathbf{Y}'\mathbf{Y},$$

is given as

$$\rho(\boldsymbol{\mu}, \delta_2) = p - (p-2)^2 \mathbb{E}(1/Z).$$

[Hint: Use the identity  $\mathbb{E}[(\mathbf{Y} - \boldsymbol{\mu})'\mathbf{Y}/Z] = (p-2) \mathbb{E}(1/Z)$ .]

- Compare both risks. What can be said?

**1.2.** Consider the situation from Problem 1.1 for the case  $p = 10$ . Let

$$\mathbf{y} = (1.9665, .5456, 2.2983, .6172, 1.9213, -2.1017, -.2727, 1.3510, 2.6029, 1.4598)'$$

be a given observation of  $\mathbf{Y}$ .

- Compute the two possible decisions  $\delta_1(\mathbf{y})$  and  $\delta_2(\mathbf{y})$  (estimates of  $\boldsymbol{\mu}$ ) corresponding to the decision rules  $\delta_1(\mathbf{Y})$  and  $\delta_2(\mathbf{Y})$  from Problem 1.1.
- Compute the observed losses of  $\delta_1(\mathbf{y})$  and  $\delta_2(\mathbf{y})$  when  $\mu_i = 1$ ,  $i = 1, \dots, 10$ , are the true parameters.
- The  $i$ -th element of  $\delta_1(\mathbf{y})$  and  $\delta_2(\mathbf{y})$  can respectively be seen as an estimate of  $\mu_i$ . Check whether each element of  $\delta_2(\mathbf{y})$  has a smaller observed loss than the corresponding element of  $\delta_1(\mathbf{y})$  when  $\mu_i = 1$  is the true parameter.

**1.3.** Consider the situation from Example 1.3 for an arbitrary sample size  $n$ . Assume that the set  $\Delta$  contains the three rules

$$\delta_1(\mathbf{Y}) = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \delta_2(\mathbf{Y}) = \frac{n\bar{Y} + \sqrt{n/4}}{n + \sqrt{n}}$$

and  $\delta_{0.5}(\mathbf{Y}) = 0.5$ .

- (a) Determine the risks of  $\delta_2$  and  $\delta_{0.5}$ .  
 (b) Is  $\delta_{0.5}(\mathbf{Y})$  admissible in  $\Delta = \{\delta_1(\mathbf{Y}), \delta_2(\mathbf{Y}), \delta_{0.5}(\mathbf{Y})\}$ ? (Plot the three risk functions for  $n = 10$ .) Is  $\delta_{0.5}(\mathbf{Y})$  a reasonable decision rule?

**1.4.** Let  $Y_i \sim N(\mu, 1)$ ,  $i = 1, \dots, n$ , be  $n$  independent distributed random variables. Show that  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$  is admissible for  $\mu$  in the class

$$\Delta = \{\delta(\mathbf{Y}) : \delta(\mathbf{Y}) = c\bar{Y}, c \in \mathbb{R}\}$$

with respect to the loss function  $L(\mu, \delta(\mathbf{Y})) = (\delta(\mathbf{Y}) - \mu)^2$ .

**1.5.** Consider the situation from Problem 1.2 and compute the observed losses of the estimates  $\delta_1(\mathbf{y})$  and  $\delta_2(\mathbf{y})$  when  $\mu_i = 1$ ,  $i = 1, \dots, 10$  are the true parameters. Use the weighted squared error loss function

$$L(\boldsymbol{\theta}, \delta(\mathbf{Y})) = (\delta(\mathbf{Y}) - \boldsymbol{\theta})' \mathbf{W} (\delta(\mathbf{Y}) - \boldsymbol{\theta}) ,$$

where

$$\mathbf{W} = \text{diag}(0.01, 0.46, 0.01, 0.46, 0.01, 0.01, 0.01, 0.01, 0.01, 0.01) .$$

**1.6.** (See the proof of Theorem 1.1) Let  $\mathbf{W}$  be a  $p \times p$  symmetric nonnegative definite matrix. Let  $\lambda_{\max}$  denote the largest eigenvalue of  $\mathbf{W}$ .

- (a) Show that all eigenvalues of  $\mathbf{F} = (1/\lambda_{\max})\mathbf{W}$  lie in the closed interval  $[0, 1]$ .  
 (b) Show that  $\mathbf{F} - \mathbf{F}^2$  is symmetric nonnegative definite.

[Hint: Use the spectral decomposition of  $\mathbf{W}$ .]

**1.7.** Explain why the condition

$$\delta_0, \delta_1 \in \Delta \quad \Rightarrow \quad \delta_0 + \mathbf{A}(\delta_1 - \delta_0) \in \Delta$$

is necessary for the proof of Theorem 1.1. Explain in addition, why it is actually enough to demand the condition to be true only for symmetric matrices  $\mathbf{A}$  with all eigenvalues in  $[0, 1]$ .

**1.8.** Is the condition

$$\delta_0, \delta_1 \in \Delta \quad \Rightarrow \quad \delta_0 + \mathbf{A}(\delta_1 - \delta_0) \in \Delta ,$$

where  $\mathbf{A}$  is an arbitrary  $p \times p$  matrix, satisfied for the following sets of decision rules?

- (i)  $\Delta = \{\delta(\mathbf{Y}) : \delta(\mathbf{Y}) = \mathbf{F}\mathbf{Y} + \mathbf{f}\}$ , where  $\mathbf{F}$  is a  $p \times n$  matrix and  $\mathbf{f}$  is a  $p \times 1$  vector;  
(ii)  $\Delta = \{\delta(\mathbf{Y}) : E[\delta(\mathbf{Y})] = \boldsymbol{\theta}\}$ .

**1.9.** Let  $(\Theta, L, \mathbf{Y})$  be an estimation problem with a weighted squared error loss function  $L$  and symmetric nonnegative definite weight matrix  $\mathbf{W}$ . Let  $\mathbf{W}_*$  denote an arbitrary but fixed symmetric positive definite matrix.

- (a) Show that if  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  for the case  $\mathbf{W} = \mathbf{W}_*$ , then  $\mathbf{W}_*^{1/2}\delta_0(\mathbf{Y})$  is admissible for  $\mathbf{W}_*^{1/2}\boldsymbol{\theta}$  for the case  $\mathbf{W} = \mathbf{I}_p$ . Note that  $\mathbf{W}_*^{1/2}$  is the (uniquely determined) symmetric positive matrix satisfying  $\mathbf{W}_*^{1/2}\mathbf{W}_*^{1/2} = \mathbf{W}_*$ .  
(b) Use part (a) and Theorem 1.1 to derive the following claim: If  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  for the case  $\mathbf{W} = \mathbf{W}_*$ , then  $\delta_0(\mathbf{Y})$  is admissible for  $\boldsymbol{\theta}$  for the case  $\mathbf{W} = \mathbf{I}_p$ .

**1.10.** Show that

$$\text{MSE}(\boldsymbol{\theta}, \delta) = \text{Cov}(\delta(\mathbf{Y})) + \text{bias}(\delta(\mathbf{Y})) \text{bias}(\delta(\mathbf{Y}))',$$

where  $\text{bias}(\delta(\mathbf{Y})) = E(\delta(\mathbf{Y})) - \boldsymbol{\theta}$ .

**1.11.** Let  $\delta_1(\mathbf{Y})$  and  $\delta_2(\mathbf{Y})$  be two decision rules such that the matrix-valued inequality  $\text{MSE}(\boldsymbol{\theta}, \delta_1) \leq_L \text{MSE}(\boldsymbol{\theta}, \delta_2)$  is satisfied for all  $\boldsymbol{\theta} \in \Theta$ . Show that if for some given  $\boldsymbol{\theta} \in \Theta$  the difference  $\text{MSE}(\boldsymbol{\theta}, \delta_2) - \text{MSE}(\boldsymbol{\theta}, \delta_1)$  is positive definite, then for this  $\boldsymbol{\theta}$  the following two statements hold true:

- (i)  $E[(\delta_1(\mathbf{Y})_i - \theta_i)^2] < E[(\delta_2(\mathbf{Y})_i - \theta_i)^2]$  for every  $i \in \{1, \dots, p\}$ .  
(ii)  $E[(\delta_1(\mathbf{Y}) - \boldsymbol{\theta})'\mathbf{W}(\delta_1(\mathbf{Y}) - \boldsymbol{\theta})] < E[(\delta_2(\mathbf{Y}) - \boldsymbol{\theta})'\mathbf{W}(\delta_2(\mathbf{Y}) - \boldsymbol{\theta})]$  for every  $p \times p$  symmetric positive definite matrix  $\mathbf{W}$ .

**1.12.** Consider the situation from Problem 1.1. Check whether  $\delta(\mathbf{Y}) = \mathbf{0}$  is admissible for  $\boldsymbol{\mu}$  in

$$\Delta = \{\delta(\mathbf{Y}) : \delta(\mathbf{Y}) = \mathbf{F}\mathbf{Y} + \mathbf{f}\}$$

with respect to the unweighted squared error loss. Here  $\mathbf{F}$  denotes a  $p \times p$  matrix and  $\mathbf{f}$  denotes a  $p \times 1$  vector.



<http://www.springer.com/978-3-540-40178-0>

Linear Regression

Groß, J.

2003, XII, 398 p., Softcover

ISBN: 978-3-540-40178-0