Chapter 5

Estimation of Regression Parameters

§1 The Problem of Estimating Regression Parameters

In the previous section we considered methods for estimating regression under conditions when the sample size increases indefinitely. However, strictly speaking, the results were related to the problem of *estimating regression parameters* rather than the problem of *regression estimation.* This substitution (instead of approximating functions we estimate their parameters) is legitimate for samples of sufficiently large size. As the sample size increases, the estimated parameters approach the true values and hence the function constructed using these parameters tends to the regression function. However, for samples of limited size the estimation of the regression is not always equivalent to the estimation of its parameters.

Indeed, the quality of the estimator $\hat{\alpha}$ of the parameter α_0 of the regression $y(x) = F(x, \alpha_0)$ is determined by the proximity of the vectors α_0 and $\hat{\alpha}$:

$$\rho(\alpha_0, \hat{\alpha}) = \|\hat{\alpha} - \alpha_0\|, \tag{5.1}$$

whereas the quality of the approximation of a function $F(x, \hat{\alpha})$ to the regression $F(x, \alpha_0)$ is measured by the proximity of functions. In Chapter 1 we agreed to consider the mean-square measure of proximity

$$\rho_L(F(x, \alpha_0); F(x, \hat{\alpha})) = \left(\int (F(x, \hat{\alpha}) - F(x, \alpha_0))^2 P(x)\, dx\right)^{1/2}. \tag{5.2}$$

The criteria (5.1) and (5.2) are not identical, and it is possible that a solution which is the best according to one criterion may be the worst according to another.

Example. In the class of functions

$$F(x, \alpha) = \alpha^0 + \alpha^1 x + \alpha^2 x^2$$

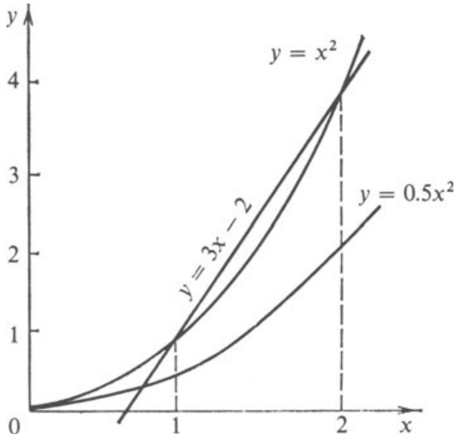

Figure 5

on the interval [1, 2], let the regression

$$y = x^2$$

be estimated. Consider two solutions (Figure 5): first the polynomial

$$F(x, \hat{\alpha}) = 0.5x^2$$

and second the polynomial

$$F(x, \hat{\hat{\alpha}}) = 3x - 2.$$

From the aspect of the parameter estimation criterion the first solution is better than the second (in any norm (5.1) the vector $\hat{\alpha} = (0, 0, 0.5)^{\mathrm{T}}$ is closer to the vector $\alpha_0 = (0, 0, 1)^{\mathrm{T}}$ than the vector $\hat{\hat{\alpha}} = (-2, 3, 0)^{\mathrm{T}}$ is).

However, from the form of the criterion (5.2) the second solution $F(x, \hat{\hat{\alpha}})$ is better. For any measure $P(x)$ the inequality

$$\rho_L(3x - 2, x^2) < \rho_L(0.5x^2, x^2)$$

is valid.

When then is the problem of estimation of parameters of a regression based on samples of finite size equivalent to the problem of regression estimation?

Assume that the class of functions to which the regression belongs is linear in its parameters

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x), \tag{5.3}$$

and let $\varphi_1(x), \ldots, \varphi_n(x)$ be a system of orthonormal functions with weight $P(x)$, i.e., functions such that

$$\int_a^b \varphi_p(x)\varphi_q(x)P(x)\,dx = \begin{cases} 1 & \text{for } p = q, \\ 0 & \text{for } p \neq q. \end{cases} \tag{5.4}$$

In this case the quantities which characterize the proximity of functions in the L_P^2 metric and the proximity of parameters in the Euclidean metric coincide, and the problem of approximating a function on $[a, b]$ to the regression becomes equivalent to the problem of parameter estimation. Indeed,

$$\begin{aligned}\rho_L^2(F(x, \hat{\alpha}), F(x, \alpha)) &= \int_a^b \left(\sum_{i=1}^n \hat{\alpha}_i \varphi_i(x) - \sum_{i=1}^n \alpha_i \varphi_i(x)\right)^2 P(x)\, dx \\ &= \sum_{i=1}^n (\hat{\alpha}_i - \alpha_i)^2. \end{aligned} \tag{5.5}$$

The conditions (5.3) and (5.4) are sufficient to replace the problem of estimating the regression with that of estimating its parameters. However, in order to construct an orthogonal system of functions the knowledge of $P(x)$ is needed. In this chapter we shall assume that the density $P(x)$ is known.

§2 The Theory of Normal Regression

The estimation theory of regression parameters based on samples of fixed size is developed for the case when the class of functions to which the regression belongs is linear in its parameters:

$$F(x, \alpha) = \sum_{i=1}^n \alpha_i \varphi_i(x), \tag{5.6}$$

and secondly the structure of the measurement follows the *Gauss–Markov model.* It is assumed that the measurements of functional dependence

$$y(x) = \sum_{i=1}^n \alpha_i^0 \varphi_i(x)$$

are carried out at l fixed points

$$x_1, \ldots, x_l.$$

(These points are not random.)

The measurements are subject to an additive noise which arises randomly according to the density $P(\xi)$, and has mean zero (i.e., $\int \xi P(\xi)\, d\xi = 0$) and finite variance ($\int \xi^2 P(\xi)\, d\xi < \infty$). The errors at points x_i and x_j ($i \neq j$) are uncorrelated.

The result of measurements of the function $\bar{y} = y(x)$ at points $x_1, \ldots, x_l$ is the random vector $Y = (y_1, \ldots, y_l)^{\mathrm{T}}$ whose coordinates are equal to

$$y_j = \sum_{i=1}^n \alpha_i^0 \varphi_i(x_j) + \xi_j = \bar{y}_j + \xi_j, \qquad j = 1, 2, \ldots, l.$$

Using vector notation, we have

$$Y = \Phi\alpha_0 + \xi, \tag{5.7}$$

where Φ is an $l \times n$ matrix with elements $\varphi_i(x_j)$ $(j = 1, 2, \ldots, l; i = 1, 2, \ldots, n)$, α_0 is the vector of parameters, and ξ is the noise vector. Thus the equalities

$$MY = \Phi\alpha_0, \qquad M\{(Y - MY)(Y - MY)^{\mathrm{T}}\} = \sigma^2 I, \tag{5.8}$$

where I is the unit matrix, define the Gauss–Markov model.

In the theory of estimating regression parameters, the special case of the Gauss–Markov model is considered for which the errors ξ are normally distributed.

For the normal distribution of the errors the so-called theory of *normal regression* is valid. It is based on the following fact: the extremal method of estimating parameters of normal regression is the least-squares method, according to which as an estimator of parameters α one should choose the vector α_{emp} which yields the minimum of the functional

$$I_{\mathrm{emp}}(\alpha) = \frac{1}{l}\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n}\alpha_i\varphi_i(x_j)\right)^2.$$

The following theorem is valid.

Theorem 5.1. *The least-squares estimators of parameters of a normal regression are jointly efficient.*

Below we shall prove this theorem and then construct a method estimating normal regression which is superior to the one based on the least-squares method.

Proof. We write the probability density of the error in the form

$$P(\xi_j) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left(y_j - \sum_{i=1}^{n}\alpha_i^0\varphi_i(x_j)\right)^2\right\}. \tag{5.9}$$

Here the problem of estimating regression parameters is equivalent to estimating the parameter of the distribution (5.9) based on the results of measuring the function $\bar{y} = y(x)$ at points $x_1, \ldots, x_l$.

We now write the likelihood function†

$$P(y_1, \ldots, y_l; \alpha) = P(\alpha)$$

$$= \frac{1}{(2\pi)^{l/2}\sigma^l}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n}\alpha_i\varphi_i(x_j)\right)^2\right]\right\}. \tag{5.10}$$

† For brevity we shall write $P(\alpha)$ in place of $P(y_1, \ldots, y_l; \alpha)$.

In view of the Cramèr–Rao inequality (cf. Chapter 3, Section 11) the Fisher information matrix $\|f_{ij}\|$ (the matrix with elements

$$f_{ij} = -M\frac{\partial^2 \ln P(\alpha)}{\partial\alpha_i\,\partial\alpha_j}\Bigg)$$

determines the limiting accuracy of the joint estimators of the vector of parameters α in the class of unbiased estimators. Namely, for any vector z the inequality

$$z^{\mathrm{T}}\|f_{ij}\|^{-1}z \le z^{\mathrm{T}}Bz$$

is valid, where B is the covariance matrix of unbiased estimators of the parameter vector. Thus the limiting accuracy in the class of unbiased estimators is attained for the estimation method for which

$$B = \|f_{ij}\|^{-1}. \tag{5.11}$$

We shall show that in the case of normal errors the equality (5.11) is attained when the regression parameters are estimated using the least-squares method. Indeed let us compute the elements f_{ij} of the Fisher matrix. Taking (5.10) into account we obtain

$$f_{ij} = -M\frac{\partial^2 \ln P(\alpha)}{\partial\alpha_i\,\partial\alpha_j} = \frac{1}{\sigma^2}M\sum_{r=1}^{l}\varphi_i(x_r)\varphi_j(x_r),$$

or in matrix form

$$\|f_{ij}\| = \frac{1}{\sigma^2}M\Phi^{\mathrm{T}}\Phi, \tag{5.12}$$

where Φ is an $l \times n$ matrix with elements $\varphi_i(x_j)$, $i = 1, \ldots, n, j = 1, \ldots, l$.

We now compute the elements b_{ij} of the covariance matrix B of estimators obtained using the least-squares method. For this purpose we shall find the estimator of regression parameters using the least-squares method, i.e., the vector α_{emp} which minimizes the functional

$$I_{\mathrm{emp}}(\alpha) = \frac{1}{l}\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n}\alpha_i\varphi_i(x_j)\right)^2. \tag{5.13}$$

Minimization of $I_{\mathrm{emp}}(\alpha)$ with respect to α is equivalent to the solution of the following equation:

$$\Phi^{\mathrm{T}}\Phi\alpha = \Phi^{\mathrm{T}}Y. \tag{5.14}$$

Equation (5.14) is called the *normal equation.* A solution of the normal equation for the vector of parameters α equals†

$$\alpha = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y.$$

† It is assumed that $(\Phi^{\mathrm{T}}\Phi)$ is nonsingular; otherwise the generalized inverse $(\Phi^{\mathrm{T}}\Phi)^{+}$ is used in place of $(\Phi^{\mathrm{T}}\Phi)^{-1}$.

Observe that the least-squares estimator is unbiased:

$$M\alpha = M[(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y] = \alpha_0.$$

We now write the vector $\alpha - \alpha_0$ of deviations of estimators of regression parameters from the true value of parameters

$$\alpha - \alpha_0 = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\xi,$$

where ξ is the vector of errors in measurement.

Now we shall obtain the covariance matrix:

$$B = M(\alpha - \alpha_0)(\alpha - \alpha_0)^{\mathrm{T}} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}M\xi\xi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-1}.$$

Taking into account that $M\xi\xi^{\mathrm{T}} = \sigma^2 I$, we arrive at

$$B = \sigma^2(\Phi^{\mathrm{T}}\Phi)^{-1}.$$

Hence for the case of normally distributed errors the covariance matrix of vectors of estimators is equal to the inverse of the Fisher information matrix. We have thus shown the efficiency of the least-squares method for the problem of estimating regression parameters when the observations are assumed to follow the Gauss–Markov model. □

It should be mentioned that the least-squares method is an efficient method of estimating parameters only in the case of the Gauss–Markov model. In models with nonfixed measurement points x_i, even with normally distributed errors, the least-squares method is only asymptotically efficient. Thus even in the case of the estimation of one parameter,

$$\bar{y} = ax,$$

when measurements subject to additive normal error

$$y = ax + \xi$$

are taken at points $x_1, \ldots, x_l$ which are chosen randomly and independently according to distribution $P(x)$, the estimator of the parameter a is not efficient. Indeed, exactly as above one can find the value of the Fisher information quantity:

$$I_\Phi = \frac{M\sum_{i=1}^{l} x_i^2}{\sigma^2},$$

and compute the variance of the estimator of parameter a:

$$D(a) = M\frac{\sigma^2}{\sum_{i=1}^{l} x_i^2}.$$

Observe now that since the function $1/x^2$ is convex, the inequality

$$M \frac{1}{\sum_{i=1}^{l} x_i^2} \geq \frac{1}{M \sum_{i=1}^{l} x_i^2} \tag{5.15}$$

is valid. This implies that in the example under consideration

$$D(a) \geq I_{\Phi}^{-1}.$$

The only case when the inequality (5.15) becomes equality is when the observation points are fixed, which results in the Gauss–Markov model.

§3 Methods of Estimating the Normal Regression that are Uniformly Superior to the Least-Squares Method

Thus in the Gauss–Markov model the least-squares method is an efficient procedure for estimating parameters of a normal regression. This assertion required two stipulations:

(1) The observations are carried out with normal errors.
(2) The least-squares method is the best only among unbiased estimators.

The question arises: Are these stipulations essential? They are indeed. The least-squares method retains its extremal properties only in the case of normal errors ξ. When the number of observations $l \geq 2n + 1$ (n is the dimensionality of the basis), then the efficiency of the least-squares method implies that the errors are normally distributed [23].

No less important is the second stipulation: even under the conditions of normally distributed errors in a class of biased estimators, there exist estimators which are uniformly superior to the least-squares estimators.

Definition. We say that for the loss function

$$\|\alpha - \alpha_0\|^2 = (\alpha - \alpha_0)^{\mathrm{T}}(\alpha - \alpha_0),$$

the estimation method $\alpha_A(y_1, \ldots, y_l)$ of a vector of parameters α_0 is *uniformly better* than the estimation method $\alpha_B(y_1, \ldots, y_l)$ if for any α_0 the inequalities

$$M\|\alpha_A(y_1, \ldots, y_l) - \alpha_0\|^2 < M\|\alpha_B(y_1, \ldots, y_l) - \alpha_0\|^2$$

are satisfied.

In this section we shall construct algorithms for approximating regression which are uniformly better (i.e., better for any α_0) than those which result from the least-squares method. The bases for these algorithms are methods of estimating the mean vector of a multivariate normal distribution, and in particular the following

Theorem 5.2 (James–Stein). *Let x be an n-dimensional ($n \geq 3$) random vector distributed according to a normal distribution $N(\alpha, \sigma^2 I)$ with the mean vector α and covariance matrix $\sigma^2 I$. Let S be a random variable independent of x distributed according to the central $\sigma^2\chi^2$ distribution with q degrees of freedom. Then the estimator of the mean given by*

$$\hat{\alpha}(x, S) = \left(1 - \frac{n-2}{q+2}\frac{S}{\|x\|^2}\right)_+ x, \qquad (z)_+ = \begin{cases} z & \textit{for } z \geq 0, \\ 0 & \textit{for } z < 0 \end{cases} \tag{5.16}$$

is uniformly better than $\hat{\hat{\alpha}}(x) = x$.

In other words, the theorem asserts that the vector $\hat{\alpha}(x, S)$ collinear to the observed vector x but different from x in its absolute value should be chosen as the estimator of α. This theorem is a particular case of a more general assertion to be proven in the next section.

We shall now utilize Theorem 5.2 to construct an algorithm for estimating regression which is uniformly superior to the one based on the least-squares method. Let observations $y_1, \ldots, y_l$ be carried out at the points $x_1, \ldots, x_l$; our purpose is to construct an approximation of a normal regression superior to the least-squares one. As above, we shall define proximity of functions using the L_P^2 metric:

$$\rho_L(F(x, \hat{\alpha}), F(x, \alpha)) = \left(\int (F(x, \hat{\alpha}) - F(x, \alpha))^2 P(x)\, dx\right)^{1/2}.$$

We now proceed to a doubly orthogonal basis

$$\psi_1(x), \ldots, \psi_n(x), \tag{5.17}$$

i.e., a basis which satisfies

$$\int \psi_i(x)\psi_j(x)P(x)\, dx = \begin{cases} \lambda_i & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \qquad \sum_{r=1}^{l} \psi_i(x_r)\psi_j(x_r) = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \tag{5.18}$$

and seek the regression expanded with respect to the basis (5.17)†

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \psi_i(x).$$

† According to the theorem on simultaneous reduction of two quadratic forms to a diagonal form using a linear transformation, such a basis exists and may be constructed using linear algebra.

In the new basis the proximity of the function $F(x, \alpha)$ to the regression $F(x, \alpha_0)$ is given by

$$\rho_L^2(F(x, \alpha), F(x, \alpha_0)) \equiv \rho_L^2(\alpha, \alpha_0)$$
$$= \int \left(\sum_{i=1}^n (\alpha_i^0 - \alpha_i)\psi_i(x) \right)^2 P(x)\, dx = \sum_{i=1}^n \lambda_i(\alpha_i^0 - \alpha_i)^2.$$

Thus our purpose is to obtain an algorithm $\hat{\alpha}(y_1, \ldots, y_l)$ for estimating the parameter α_0 such that the quantity

$$M\rho_L^2(\hat{\alpha}(y_1, \ldots, y_l), \alpha_0) = M \sum_{i=1}^n \lambda_i(\hat{\alpha}_i(y_1, \ldots, y_l) - \alpha_i^0)^2 \tag{5.19}$$

is less than

$$M\rho_L^2(\alpha_{\text{lse}}, \alpha_0) = M \sum_{i=1}^n \lambda_i(\alpha_{\text{lse}}^i - \alpha_i^0)^2,$$

where $\alpha_{\text{lse}} = (\alpha_{\text{lse}}^1, \ldots, \alpha_{\text{lse}}^n)^{\text{T}}$ is the least-squares estimator.

Consider now the least-squares estimator of regression parameters. In the basis (5.17) this estimator becomes

$$\alpha_{\text{lse}} = \Phi^{\text{T}} Y,$$

where Φ is an $l \times n$ matrix with elements $\psi_i(x_j)$, $j = 1, \ldots, l$, $i = 1, \ldots, n$, and Y is the vector of observations. The vector α_{lse} is a random vector normally distributed with the mean vector

$$M\alpha_{\text{lse}} = M\Phi^{\text{T}} Y = \alpha_0$$

and the covariance matrix $\sigma^2 I$:

$$M(\alpha_{\text{lse}} - \alpha_0)(\alpha_{\text{lse}} - \alpha_0)^{\text{T}} = M\Phi^{\text{T}}\xi\xi^{\text{T}}\Phi = \sigma^2 I.$$

Thus the problem of estimating the parameter α_0 of the regression is reduced to the estimation of the mean vector α_0 of a normal distribution $N(\alpha_0, \sigma^2 I)$ based on its realization α_{lse}.

If in (5.19) all the λ_i were equal, Theorem 5.2 could be used to construct an algorithm for estimating regression which is better than the least-squares one. Indeed, as will be shown below, the statistic

$$S = Y^{\text{T}} Y - \alpha_{\text{lse}}^{\text{T}} \alpha_{\text{lse}} \tag{5.20}$$

does not depend on α_{lse} and is distributed according to the central $\sigma^2\chi^2$ distribution with $l - n$ degrees of freedom. Therefore according to Theorem 5.2 the estimator

$$\hat{\alpha} = \left(1 - \frac{n-2}{l-n+2} \frac{Y^{\text{T}} Y - \alpha_{\text{lse}}^{\text{T}} \alpha_{\text{lse}}}{\alpha_{\text{lse}}^{\text{T}} \alpha_{\text{lse}}} \right)_+ \alpha_{\text{lse}} \tag{5.21}$$

is uniformly better than α_{lse}, i.e., yields a value of the criterion (5.19) (in the case when $\lambda_1 = \cdots \lambda_n$) smaller than α_{lse}. However, in the doubly orthogonal system (5.17) constructed above, not all λ_i are generally equal. Thus obtaining a better approximation to the regression in the case of unequal λ_i involves the determination of an estimation method yielding a value for the criterion (5.19) which is lower than that due to the least-squares method.

Construction of such an estimating algorithm is also based on the results of Theorem 5.2. We shall assume that the functions ψ_i are enumerated in increasing order of λ_i ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$). We shall introduce the following notation: let $\alpha_0(p)$ be a vector of dimensionality p, consisting of the first p coordinates of the vector $\alpha_0 = (\alpha_1^0, \ldots, \alpha_n^0)^{\mathrm{T}}$; let $\alpha_{\mathrm{lse}}(p)$ be the vector consisting of the first p coordinates of the vector of estimators obtained by the least-squares method α_{lse}.

Define n numbers $f_1, \ldots, f_n$:

$$f_1 = 1,$$
$$f_p = \left(1 - \frac{S}{\alpha_{\mathrm{lse}}^{\mathrm{T}}(p)\, \alpha_{\mathrm{lse}}(p)} \frac{p-2}{l-p+2}\right)_+, \qquad p = 2, \ldots, n.$$

Using these numbers, we construct n numbers h_p by the rule

$$h_p = \frac{\sum_{i=p}^{n} (\lambda_i - \lambda_{i+1}) f_i}{\lambda_p}, \quad \text{where } \lambda_{n+1} = 0, \qquad p = 1, 2, \ldots, n.$$

The following theorem is valid.

Theorem 5.3 (Bhattacharya). *For the risk function* (5.19) *the estimator*

$$\hat{\alpha}(y_1, \ldots, y_l) = (\alpha_{\mathrm{lse}}^1 h_1, \ldots, \alpha_{\mathrm{lse}}^n h_n)^{\mathrm{T}}, \qquad n \geq 3, \tag{5.22}$$

is uniformly better than the estimator $\alpha_{\mathrm{lse}} = (\alpha_{\mathrm{lse}}^1, \ldots, \alpha_{\mathrm{lse}}^n)^{\mathrm{T}}$.

Proof. The proof of Theorem 5.3 is based on Theorem 5.2, according to which for any p the inequality

$$M\|\alpha_{\mathrm{lse}}(p) f_p - \alpha_0(p)\|^2 \leq M\|\alpha_{\mathrm{lse}}(p) - \alpha_0(p)\|^2 \tag{5.23}$$

is valid.

Consider the randomized estimator

$$g\alpha_{\mathrm{lse}} = (\alpha_{\mathrm{lse}}^1 g_1, \ldots, \alpha_{\mathrm{lse}}^n g_n), \tag{5.24}$$

where g_k are random variables independent of S and y distributed according to

$$P\{(q_k = f_j)\} = \frac{\lambda_j - \lambda_{j+1}}{\lambda_k}, \quad k = 1, 2, \ldots, n, \quad j = k, \ldots, n; \qquad \lambda_{n+1} = 0.$$

The value of this risk (5.19) for this estimator is equal to

$$\rho_L^2(G\alpha_{\text{lse}}, \alpha_0) = M \sum_{k=1}^{n} \lambda_k (g_k \alpha_{\text{lse}}^k - \alpha_k^0)^2$$

$$= \sum_{k=1}^{n} \sum_{j=k}^{n} \frac{\lambda_j - \lambda_{j+1}}{\lambda_k} \lambda_k M(f_j \alpha_{\text{lse}}^k - \alpha_k^0)^2.$$

We now utilize the inequality (5.23):

$$\rho_L^2(G\alpha_{\text{lse}}, \alpha_0) = \sum_{k=1}^{n} \sum_{j=k}^{n} (\lambda_j - \lambda_{j+1}) M(\alpha_{\text{lse}}^k f_j - \alpha_k^0)^2$$

$$= \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \sum_{k=1}^{j} (\alpha_{\text{lse}}^k f_j - \alpha_k^0)^2$$

$$= \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \|\alpha_{\text{lse}}(j) f_j - \alpha_0(j)\|^2$$

$$\leq \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \|\alpha_{\text{lse}}(j) - \alpha_0(j)\|^2$$

$$\leq \sum_{j=1}^{n} \lambda_j M(\alpha_{\text{lse}}^j - \alpha_j^0)^2.$$

Thus the value of the risk for the randomized estimator of the parameters is less than the corresponding value for the least-squares estimator. On the other hand, it follows from the convexity of the loss function (5.19) that the nonrandomized estimator (5.22) is at least as good as the randomized estimator (5.24). Thus the approximation to the regression determined by the parameters (5.22) is uniformly better than the least-squares approximation. The theorem is proved. □

It remains to show that statistics $S = Y^{\mathrm{T}} Y - \alpha_{\text{lse}}^{\mathrm{T}} \alpha_{\text{lse}}$ does not depend on α_{lse} and is distributed according to the central $\sigma^2 \chi^2$ distribution with $l - n$ degrees of freedom. For this purpose we shall complete the system of n vectors $\psi_1, \ldots, \psi_n$, orthonormal on $x_1, \ldots, x_l$:

$$\psi_i = (\psi_i(x_1), \ldots, \psi_i(x_l))^{\mathrm{T}},$$

$$\psi_i^{\mathrm{T}} \psi_j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \qquad i, j = 1, 2, \ldots, n,$$

so that it becomes a complete orthonormal system consisting of l orthonormal vectors

$$\psi_1, \ldots, \psi_n, \psi_{n+1}, \ldots, \psi_l,$$

$$\psi_i^{\mathrm{T}} \psi_j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \qquad i, j = 1, 2, \ldots, l.$$

We now expand Y in terms of this system:

$$Y = \sum_{i=1}^{n} \gamma_i \psi_i + \sum_{j=n+1}^{l} \gamma_j \psi_j, \tag{5.25}$$

where

$$\gamma_i = Y^{\mathrm{T}}\psi_i = \alpha_{\mathrm{lse}}^i, \qquad i = 1, 2, \ldots, n,$$

$$\gamma_j = Y^{\mathrm{T}}\psi_j, \qquad j = n + 1, \ldots, l.$$

Substituting (5.25) into (5.20), we obtain

$$S = \sum_{j=n+1}^{l} \gamma_j^2, \tag{5.26}$$

and hence S does not depend on α_{lse}^i (but only on $\gamma_j, j = n + 1, \ldots, l$). Since by assumption $Y = Y_0 + \xi$ and the vector Y_0 can be expanded in terms of this incomplete system (5.17)

$$Y_0 = \sum_{i=1}^{n} \alpha_i^0 \psi_i,$$

we have the inequality

$$\gamma_j = \xi^{\mathrm{T}}\psi_j.$$

Substituting the value of γ_j into (5.26), we obtain

$$S = \sum_{j=n+1}^{l} \gamma_j^2 = \sum_{j=n+1}^{l} \left(\sum_{i=1}^{l} \xi_i\, \psi_j(x_i)\right)^2 = \sum_{j=n+1}^{l} \xi_j^2,$$

and hence the statistic S is distributed according to the central $\sigma^2\chi^2$ distribution with $l - n$ degrees of freedom.

§4 A Theorem on Estimating the Mean Vector of a Multivariate Normal Distribution

In this section we shall obtain a family of estimators of the mean vector which are uniformly better than the estimator $\alpha(x, S) = x$. The estimator (5.21) belongs to this class.

Let x be a random vector distributed according to $N(\alpha_0, \sigma^2 I)$, and S be a random variable independent of x distributed according to the central $\sigma^2\chi^2$ distribution with q degrees of freedom. We denote $F = x^{\mathrm{T}}x/S$.

The following theorem is valid.

Theorem 5.4 (Baranchik). *An estimator of the n-dimensional ($n \geq 3$) mean vector*

$$\hat{\alpha}(x, S) = \left(1 - \frac{r(F)}{F}\right)x,$$

where $r(F)$ is a monotonic nondecreasing function satisfying

$$0 \leq r(F) \leq 2\frac{n-2}{q+2}, \tag{5.27}$$

is uniformly better than the estimator $\alpha(x, S) = x$.

Remark. Theorem 5.2 is a particular case of Theorem 5.4 obtained by setting

$$r(F) = \begin{cases} \dfrac{n-2}{q+2} & \text{for } F \geq \dfrac{n-2}{q+2}, \\[2ex] F & \text{for } F < \dfrac{n-2}{q+2}. \end{cases}$$

Proof. In the proof of Theorem 5.4 the following fact is used: the mathematical expectation of a random variable $f(\chi^2(n, b))$ taken with respect to the measure $\mu(\chi^2(n, b))$, where $\chi^2(n, b)$ is a random variable with the noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter b, can be represented as

$$Mf(\chi^2(n, b)) = Mf(\chi^2_{n+2k}),$$

where χ^2_{n+2k} is a random variable with the central χ^2 distribution with $n + 2k$ degrees of freedom, and k is a random variable distributed according to the Poisson distribution with parameter b:

$$P(k) = \exp\{-b\}\frac{b^k}{k!}.$$

(The mathematical expectation on the right-hand side is evaluated with respect to x as well as with respect to k.)

Thus

$$Mf(\chi^2(n, b)) = Mf(\chi^2_{n+2k}) = \exp\{-b\}\sum_{t=0}^{\infty}\frac{b^t}{t!}Mf(\chi^2_{n+2t}). \tag{5.28}$$

We now proceed directly to the proof of the theorem: Our purpose is to show that the difference

$$H = M\|\hat{\alpha}(x, S) - \alpha_0\|^2 - M\|x - \alpha_0\|^2 \tag{5.29}$$

is nonnegative. Denote

$$g(F) = 1 - \frac{r(F)}{F}$$

and transform (5.29)

$$H = M[x^{\mathrm{T}}xg^2(F)] - 2\alpha_0^{\mathrm{T}}Mg(F)x + \|\alpha_0\|^2 - n\sigma^2. \tag{5.30}$$

The expressions (5.31)–(5.34) below are derived under the assumption that S is fixed. According to (5.28) we have

$$M\left[x^{\mathrm{T}}xg^2\left(\frac{x^{\mathrm{T}}x}{S}\right)\right]$$

$$= \exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}\frac{\|\alpha_0\|^{2t}}{t!(2\sigma^2)^t}M\left[\sigma^2\chi^2_{n+2t}\,g^2\left(\frac{\sigma^2\chi^2_{n+2t}}{S}\right)\right]. \qquad (5.31)$$

We now transform the expression

$$\alpha_0^{\mathrm{T}}Mg(F)x = \alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x.$$

For this purpose we shall perform an orthogonal transformation of vectors x into vectors z such that in the new coordinate system the mean vector is equal to $(\|\alpha_0\|, 0, \ldots, 0)$ (only the first coordinate does not vanish, and it is equal to the norm of the mean vector). This transformation leaves S unaltered. We obtain

$$\alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x = \|\alpha_0\|Mg\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1,$$

where z is the first coordinate of the vector $z = (z_1, \ldots, z_n)^{\mathrm{T}}$.

Observe now that

$$M\left[g\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1\right] = \frac{\sigma^2}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}$$

$$\times\frac{d}{d\|\alpha_0\|}\int g\left(\frac{\sum_{i=1}^{n}z^2}{S}\right)\exp\left\{-\frac{\sum_{i=1}^{n}z_i^2 - 2\|\alpha_0\|z_1}{2\sigma^2}\right\}dz_1\cdots dz_n.$$

Thus we obtain

$$\|\alpha_0\|Mg\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1 = \frac{\sigma^2\|\alpha_0\|}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}$$

$$\times\frac{d}{d\|\alpha_0\|}\int g\left(\frac{\sum_{i=1}^{n}z_i^2}{S}\right)\exp\left\{-\frac{\sum_{i=1}^{n}z_i^2 - 2\|\alpha_0\|z_1}{2\sigma^2}\right\}dz_1\cdots dz_n$$

$$= \sigma^2\|\alpha_0\|\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\frac{d}{d\|\alpha_0\|}\exp\left\{\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}Mg\left(\frac{\sigma^2\chi^2_{n+2k}}{S}\right),$$

where k is a random variable distributed according to the Poisson distribution with the mean $\|\alpha_0\|^2/(2\sigma^2)$. Finally we obtain

$$\alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x = 2\sigma^2\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}t\left(\frac{\|\alpha_0\|^2}{2\sigma^2}\right)^t\frac{Mg(\sigma^2\chi^2_{n+2t}|S)}{t!}. \qquad (5.32)$$

Now taking into account that $\|\alpha_0\|^2/(2\sigma^2)$ is the mean of the random variable k distributed according to the Poisson distribution, we express the third summand in the sum (5.30) in the form

$$\|\alpha_0\|^2 = 2\sigma^2 \exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\} \sum_{t=0}^{\infty} t\, \frac{\left(\dfrac{\|\alpha_0\|^2}{2\sigma^2}\right)^t}{t!}. \tag{5.33}$$

We can thus represent the expression (5.30) in the form

$$\begin{aligned} H = {} & \sigma^2 \exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\} \sum_{t=0}^{\infty} \frac{\left(\dfrac{\|\alpha_0\|^2}{2\sigma^2}\right)^t}{t!} \\ & \times \left[M\chi^2_{n+2t}\, g^2\left(\frac{\sigma^2\chi^2_{n+2t}}{S}\right) - 4tMg\left(\frac{\sigma^2\chi^2_{n+2t}}{S}\right) - n + 2t\right]. \end{aligned} \tag{5.34}$$

Now let $S = \sigma^2\chi^2_q$ be a random variable distributed according to the central $\sigma^2\chi^2$ distribution with q degrees of freedom. The theorem will be proved if we verify that the expression

$$h = M\left[\chi^2_{n+2t} g^2\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right) - 4tg\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right) - n + 2t\right] \tag{5.35}$$

is nonpositive for all t.

Denote $\chi^2_{n+2t}/\chi^2_q = u$, and observe that

$$u(1 - g(u)) = r(u). \tag{5.36}$$

Therefore condition (5.27) implies that

$$g(u) > 1 - 2\frac{n-2}{q+2}u^{-1}. \tag{5.37}$$

We transform the expression (5.35) utilizing notation (5.36) and the fact that $M\chi^2_{n+2t} = n + 2t$:

$$\begin{aligned} h &= M\left[-2r(u)\chi^2_q + r(u)(1 - g(u))\chi^2_q + 4t\frac{r(u)}{u}\right] \\ &= M\left[r(u)\chi^2_q\left(-1 - g(u) + \frac{4t}{\chi^2_{n+2t}}\right)\right]. \end{aligned}$$

Taking (5.37) into account, we obtain that the quantity h does not exceed

$$\hat{h} = M(r(u)\zeta) = M\left[M\left\{r\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right)\zeta \,\middle|\, \chi^2_q\right\}\right],$$

where

$$\zeta = \chi^2_q\left[-2 + \left(4t + 2\frac{n-2}{q+2}\chi^2_q\right)\frac{1}{\chi^2_{n+2t}}\right].$$

For any fixed χ_q^2 we determine a constant a such that

$$-2 + \left(4t + 2\frac{n-2}{q+2}\chi_q^2\right)\frac{1}{a} = 0. \tag{5.38}$$

Observe that for any $\chi_{n+2t}^2 > a$ the inequality $\zeta < 0$ is valid. Therefore taking into account that in view of the condition of the theorem the function $r(u)$ is nondecreasing, we obtain the bound

$$\begin{aligned} &M\left\{r\left(\frac{\chi_{n+2t}^2}{\chi_q^2}\right)\zeta \,\middle|\, \chi_q^2\right\} \\ &\quad \le r\left(\frac{a}{\chi_q^2}\right)M\{\zeta \mid \chi_{n+2t}^2 \le a\}P\{\chi_{n+2t}^2 \le a\} \\ &\qquad + r\left(\frac{a}{\chi_q^2}\right)M\{\zeta \mid \chi_{n+2t}^2 > a\}P\{\chi_{n+2t}^2 > a\} \\ &\quad = r\left(\frac{a}{\chi_q^2}\right)M\{\zeta \mid \chi_q^2\} \\ &\quad = r\left(\frac{a}{\chi_q^2}\right)\chi_q^2\left[-2 + \left(4t + 2\frac{n-2}{q+2}\chi_q^2\right)\frac{1}{n+2t-2}\right] \\ &\quad = 2\frac{n-2}{n+2t-2}r\left(\frac{a}{\chi_q^2}\right)\chi_q^2\left(-1 + \frac{\chi_q^2}{q+2}\right). \end{aligned} \tag{5.39}$$

(We have used the equality $M(1/\chi_m^2) = 1/(m-2)$ $(m \ge 3)$.)

Substitute now into (5.39) the value of a satisfying (5.38), and compute the mathematical expectation of the last term in (5.39), which is

$$2\frac{n-2}{n+2t-2}M\left\{r\left(\frac{2t}{\chi_q^2} + \frac{n-2}{q+2}\right)\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right]\right\}.$$

Taking into account that $r(u)$ is a nondecreasing function we find the bound

$$\begin{aligned} &M\left\{r\left(\frac{2t}{\chi_q^2} + \frac{n-2}{q+2}\right)\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right]\right\} \\ &\quad \le r\left(\frac{n+2t-2}{q+2}\right)M\left\{\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right] \,\middle|\, \chi_q^2 \le q+2\right\} \\ &\qquad + r\left(\frac{n+2t-2}{q+2}\right)M\left\{\chi_q^2\left(-1 + \frac{\chi_q^2}{q+2}\right) \,\middle|\, \chi_q^2 > q+2\right\} \\ &\quad = r\left(\frac{n+2t-2}{q+2}\right)M\left\{\chi_q^2\left[-1 + \frac{\chi_q^2}{q+2}\right]\right\} = 0. \end{aligned}$$

(For a central χ^2 distribution we have $M\chi_q^2 = q$, $M(\chi_q^2)^2 = q(q+2)$.)

Thus the quantity (5.35) is nonpositive and the theorem is proved. □

§5 The Gauss–Markov Theorem

Up until now, when estimating regression it was assumed that the errors are distributed according to the normal distribution. We shall now relax this assumption. It will be assumed that the distribution of errors is unknown but has a bounded variance. Under these conditions it is required to construct the best algorithm for the regression estimation.

Above, when developing the theory of normal regression we first established that in the class of algorithms leading to unbiased estimators of the parameters the least-squares method was optimal, but for a wider class of algorithms procedures which are better than the least-squares method were obtained. We shall now proceed analogously. First we shall show that in some narrow class of estimating algorithms the least-squares method is the. best, and then we obtain estimation methods in a wider class of algorithms which are superior to the least-squares method.

Under the assumption of normal errors the least-squares method is the best in the class of unbiased methods of estimation. In this section we shall show that in a narrower class of estimates which are both linear and unbiased, the least-squares method yields the best estimating algorithms independently of the distribution of the errors.

Definition. We say that an estimator of the parameter α is *linear* in the observations $Y = (y_1, \ldots, y_l)^{\mathrm{T}}$ if it can be represented in the form

$$\alpha = LY \qquad \left(\alpha_j = \sum_{i=1}^{l} \beta_{ij} y_i\right), \tag{5.40}$$

where L is a matrix with the entries β_{ij} $(i = 1, \ldots, l; j = 1, \ldots, n)$.

The following theorem is valid:

Theorem (Gauss–Markov). *Among all the linear unbiased estimators the least-squares estimator possesses the minimal variances of the coordinates.*

We shall prove the Gauss–Markov theorem in its more general form for the case of linear biased estimators. Denote by α_0 the vector of parameters of the linear regression

$$MY = \Phi\alpha_0 \qquad (Y = \Phi\alpha_0 + \xi). \tag{5.41}$$

Define the estimator $\alpha(B)$ as the solution of the equation

$$(\Phi^{\mathrm{T}}\Phi + B)\alpha(B) = \Phi^{\mathrm{T}}Y, \tag{5.42}$$

where B is a symmetric nonnegative definite $n \times n$ matrix which defines the bias vector μ_0. We shall show that the estimator $\alpha(B)$ possesses extremal properties. Namely, the following theorem is valid.

Theorem 5.5. *Among all the linear estimators of the vector of parameters α with the bias vector equaling μ_0, the estimator $\alpha(B)$ possesses the minimal variance of coordinates.*

Proof. We obtain from (5.42)

$$M\alpha(B) = M(\Phi^{\mathrm{T}}\Phi + B)^{-1}\Phi^{\mathrm{T}}Y = (\Phi^{\mathrm{T}}\Phi + B)^{-1}\Phi^{\mathrm{T}}\Phi\alpha_0. \qquad (5.43)$$

Let $\hat{\alpha} = LY$ be an arbitrary linear estimator such that

$$M\hat{\alpha} = M\alpha(B) = \mu_0 + \alpha_0 = \mu. \qquad (5.44)$$

Then we obtain from (5.42)

$$MLY = L\Phi\alpha_0 = (\Phi^{\mathrm{T}}\Phi + B)^{-1}\Phi^{\mathrm{T}}\Phi\alpha_0. \qquad (5.45)$$

Since the equality (5.45) is valid for any α_0, then

$$L\Phi = (\Phi^{\mathrm{T}}\Phi + B)^{-1}\Phi^{\mathrm{T}}\Phi. \qquad (5.46)$$

We now write the variance of the ith coordinate of estimator $\hat{\alpha}$:

$$\begin{aligned} M(\hat{\alpha}_i - \mu_i)^2 &= M(\hat{\alpha}_i - \alpha_i(B) + \alpha_i(B) - \mu_i)^2 \\ &\geq M(\alpha_i(B) - \mu_i)^2 + 2M(\hat{\alpha}_i - \alpha_i(B))(\alpha_i(B) - \mu_i), \end{aligned} \qquad (5.47)$$

where μ_i is the ith coordinate of the vector μ.

We shall show that the second summand on the right-hand side of (5.47) vanishes. Indeed, utilizing (5.44) and (5.46), we obtain

$$\begin{aligned} &M(\hat{\alpha}_i - \alpha_i(B))(\alpha_i(B) - \mu_i) \\ &\quad = M(\hat{\alpha}_i - \alpha_i(B))\alpha_i(B) \\ &\quad = \sigma^2 \|(L - (\Phi^{\mathrm{T}}\Phi + B)^{-1}\Phi^{\mathrm{T}})\Phi(\Phi^{\mathrm{T}}\Phi + B)^{-1}\|_{ii} \\ &\quad = \sigma^2 \|(L\Phi - (\Phi^{\mathrm{T}}\Phi + B)^{-1}\Phi^{\mathrm{T}}\Phi)(\Phi^{\mathrm{T}}\Phi + B)^{-1}\|_{ii} = 0, \end{aligned}$$

where $\|A\|_{ii}$ denotes the element A_{ii} of the matrix $\|A\|$.

Thus

$$M(\hat{\alpha}_i - \mu_i)^2 \geq M(\alpha_i(B) - \mu_i)^2.$$

The theorem is thus proved. □

The Gauss–Markov theorem follows from the theorem just proved by setting $\|B\| = \|0\|$ in (5.42). In that case $\mu_0 = 0$.

Further, in Chapter 8 to construct the regression estimators from small samples we shall make use of this theorem. We shall search for the best estimators among the estimators of the class $\alpha(\gamma B)$ (where $\gamma > 0$ is a constant specifying the estimator of the class). The estimator $\alpha(\gamma^* B)$ is called a *ridge-regression estimator.*

§6 Best Linear Estimators

Thus, among linear unbiased estimators, the least-squares estimators are the best regardless of the distribution of the errors. In the next sections we shall consider a wider class of estimators—the class of linear estimators (not necessarily unbiased), and we shall obtain the best estimators in this class. These estimators will differ from the least-squares estimators provided nontrivial prior information concerning the estimated parameters is available. In cases when no nontrivial prior information is available the best linear estimator is still the least-squares method.

Let the parameters of the regression

$$\bar{y} = y(x) = \sum_{i=1}^{n} \alpha_i^0 \psi_i(x) \tag{5.48}$$

in a Gauss–Markov model be estimated from empirical data $x_1, y_1, \ldots, x_l, y_l$. Let $\hat{\psi}_1(x), \ldots, \hat{\psi}_n(x)$ be a doubly orthogonal basis

$$\begin{aligned} \int \hat{\psi}_i(x)\hat{\psi}_j(x)P(x)\,dx &= \begin{cases} \lambda_i & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \\ \sum_{r=1}^{l} \hat{\psi}_i(x_r)\hat{\psi}_j(x_r) &= \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases} \end{aligned} \tag{5.49}$$

Consider the class of linear estimators:

$$\hat{\alpha}_p = \theta_p^{\mathrm{T}} Y + \beta_0^p, \tag{5.50}$$

where

$$\theta_p = (\theta_1^p, \ldots, \theta_l^p)^{\mathrm{T}}, \qquad Y = (y_1, \ldots, y_l)^{\mathrm{T}}.$$

We introduce the system of orthogonal vectors

$$\chi_1, \ldots, \chi_l; \qquad \chi_i^{\mathrm{T}}\chi_j = \begin{cases} l & \text{for } i = j, \\ 0 & \text{for } i \neq j, \end{cases} \tag{5.51}$$

in which the first n vectors are

$$\chi_i = (\hat{\psi}_i(x_1), \ldots, \hat{\psi}_i(x_l))^{\mathrm{T}}, \qquad i = 1, \ldots, n.$$

We represent the vector θ_p in the expansion in terms of (5.51):

$$\theta_p = \sum_{i=1}^{l} \beta_i^p \chi_i. \tag{5.52}$$

Then the equality (5.50) can be rewritten as

$$\hat{\alpha}_p = \sum_{i=1}^{l} \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p. \tag{5.53}$$

We express the amount of deviation $M(\hat{\alpha}_p - \alpha_p^0)^2$ in terms of the parameters β. For this purpose we shall utilize the identity

$$M(\hat{\alpha}_p - \alpha_p^0)^2 = (M(\hat{\alpha}_p - \alpha_p^0))^2 + M(\hat{\alpha}_p - M\hat{\alpha}_p)^2. \tag{5.54}$$

The first summand on the right-hand side equals

$$(M(\hat{\alpha}_p - \alpha_p^0))^2 = \left(l \sum_{i=1}^{n} \beta_i^p \alpha_i^0 + \beta_0^p - \alpha_p^0\right)^2.$$

The second summand equals

$$M(\hat{\alpha}_p - M\hat{\alpha}_p)^2 = l\sigma^2 \sum_{i=1}^{l} (\beta_i^p)^2.$$

Thus

$$M(\hat{\alpha}_p - \alpha_p^0)^2 = \sigma^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left(l \sum_{i=1}^{n} \beta_i^p \alpha_i^0 + \beta_0^p - \alpha_p^0\right)^2 = \mathscr{D}^p(\beta \mid \alpha, \sigma). \tag{5.55}$$

The best linear estimator is the estimator which minimizes (5.55).

§7 Criteria for the Quality of Estimators

The best linear estimator can be obtained by directly minimizing with respect to $\beta_1, \ldots, \beta_l$ the right-hand side of the equality (5.55). The minimum of (5.55) is attained at $\beta_1^p = \beta_2^p = \cdots = \beta_l^p = 0$ and $\beta_0^p = \alpha_p^0$, and this minimum is zero.

Thus for each specific problem (specific values of α_0 and σ) a trivial biased estimator can be found which yields the minimum of the square of deviations. Now we wish to construct a linear estimator which will be suitable for a solution of a class of problems rather than for a single one.

Let us define a class of problems $R(a, \sigma)$, to which the algorithm is applicable, by means of the inequalities

$$\begin{gathered} a_p \le \alpha_p \le b_p, \\ d \le \sigma \le e. \end{gathered} \tag{5.56}$$

We shall now determine the quality of an algorithm for estimating the parameter α_p in the class $R(\alpha, \sigma)$. As usual in such situations, we shall consider two approaches: Bayesian and minimax. For each approach a different notion of the quality of a linear estimator will be introduced.

According to Bayes's principle the best method for estimation is that for which the mean value of the criterion over the set of problems belonging to $R(\alpha, \sigma)$ is minimal (the measure on this set is given by the distribution $P(\alpha, \sigma)$).

Definition. The estimator

$$\alpha_p^{(1)} = \sum \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p$$

is called *linearly best in the mean* if among all linear estimators it yields the minimum of the functional

$$\mathscr{D}_1^p(\beta) = \int \mathscr{D}^p(\beta\,|\,\alpha, \sigma)P(\alpha, \sigma)\,d\alpha_1 \cdots d\alpha_n\,d\sigma. \tag{5.57}$$

Below we shall compute a Bayesian estimator for the case when the parameters α and σ are distributed independently according to the uniform distribution on the corresponding intervals, i.e.,

$$P(\alpha, \sigma) = \begin{cases} \displaystyle\prod_{i=1}^{n} \frac{1}{b_i - a_i}\frac{1}{e - d} & \text{if } a_p \le \alpha_p \le b_p,\ d \le \sigma \le e, \\ 0 & \text{otherwise.} \end{cases} \tag{5.58}$$

Thus the quality of the estimator is determined by the functional

$$\mathscr{D}_1^p(\beta) = \int \mathscr{D}^p(\beta\,|\,\alpha, \sigma) \prod_{i=1}^{n} \frac{d\alpha_i}{b_i - a_i}\frac{d\sigma}{e - d}. \tag{5.59}$$

In accordance with the minimax principle the best method of estimation is considered to be the one which yields the minimum of $\mathscr{D}^p(\beta\,|\,\alpha, \sigma)$ for the least favorable problem (pair α, σ).

Definition. The estimator

$$\alpha_p^{(2)} = \sum \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p$$

is called the *best linear minimax estimator* in the class $R(\alpha, \sigma)$ if it yields the minimum of the functional

$$\mathscr{D}_2^p(\beta) = \sup_{\alpha, \sigma} \mathscr{D}^p(\beta\,|\,\alpha, \sigma) \tag{5.60}$$

in the class of linear estimators.

In general there may exist problems belonging to the class $R(\alpha, \sigma)$ for which the estimators $\alpha_p^{(1)}$ and $\alpha_p^{(2)}$ introduced above are worse than the least-squares estimators $\beta_{\mathrm{lse}}^p = (0, \ldots, 1/l, \ldots, 0)^{\mathrm{T}}$, $\beta_0^p = 0$ (only the pth coordinate of the vector β_{lse}^p is nonvanishing). Therefore we shall define the third optimal estimator in such a manner that it will be uniformly better than the least-squares estimator. For this purpose we introduce the loss function

$$\mathscr{D}_3^p(\beta) = \sup_{\alpha, \sigma}(\mathscr{D}^p(\beta\,|\,\alpha, \sigma) - \mathscr{D}^p(\beta_{\mathrm{lse}}\,|, \alpha, \sigma)) \tag{5.61}$$

and require that the optimal estimator minimize the expression (5.61).

Definition. The estimator

$$\alpha_p^{(3)} = \sum \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p$$

is called *linearly uniformly better* than the least-squares estimator if it yields the minimum of the functional (5.61) in the class of linear estimators and $\min_\beta D_3^p(\beta) < 0$.

§8 Evaluation of the Best Linear Estimators

The following three theorems constitute the basic content of the theory of the best linear estimator.

Theorem 5.6 (Koshcheev). *The best linear estimator of parameter α_p in the class $R(\alpha, \sigma)$ is of the form*

$$\alpha_p^{(i)} = \frac{\alpha_{\mathrm{lse}}^p + \dfrac{c_p}{l}\rho_p^{(i)}}{1 + \dfrac{1}{l}\rho_p^{(i)}}, \qquad i = 1, 2, 3, \tag{5.62}$$

where $c_p = (a_p + b_p)/2$, α_{lse}^p is the least-squares estimator, $\alpha_p^{(1)}$ is the best in the mean estimator,

$$\rho_p^{(1)} = 4\frac{d^2 + de + e^2}{(a_p - b_p)^2}, \tag{5.63}$$

$\alpha_p^{(2)}$ is the best minimax estimator,

$$\rho_p^{(2)} = 4\frac{e^2}{(a_p - b_p)^2}, \tag{5.64}$$

$\alpha_p^{(3)}$ is the uniformly best estimator, and

$$\rho_p^{(3)} = 4\frac{d^2}{(a_p - b_p)^2}. \tag{5.65}$$

It thus turns out that the best linear estimators are biased. The structure of the estimators is given by the expression (5.62), where $\rho_p^{(i)}$ are defined in (5.63)–(5.65), depending on the specific notion of the quality of an estimator. There exists a simple relationship which shows by how much a Bayes or minimax estimator is superior to a least-squares estimator.

Theorem 5.7 (Koshcheev). *The equality*

$$\mathscr{D}_i^p(\alpha_p^{(i)}) = \frac{1}{1 + \dfrac{1}{l}\rho_p^{(i)}}\mathscr{D}_i^p(\alpha_{\mathrm{lse}}^p), \qquad i = 1, 2 \tag{5.66}$$

is valid.

According to Theorem 5.7 the optimal estimators $\alpha_p^{(i)}$ are superior to the least-squares estimator by the factor $(1 + (1/l)\rho_p^{(i)})$. Hence the smaller the sample size l, the better the estimators $\alpha_p^{(i)}$.

Below we shall present the proof of Theorem 5.6. The validity of Theorem 5.7 follows from a more general theorem considered in the next section.

Proof of Theorem 5.6

(1) *Derivation of the best linear estimator in the mean.* We write the functional whose minimum determines under our conditions the best estimator in the mean:

$$\mathscr{D}_1^p(\beta) = \int_{a_1}^{b_1}\cdots\int_{a_n}^{b_n}\int_d^e \left[l\sigma^2\sum_{i=1}^{l}(\beta_i^p)^2 + \left(l\sum_{i=1}^{n}\beta_i^p\alpha_i + \beta_0^p - \alpha_p\right)^2\right]\prod_{i=1}^{n}\frac{d\alpha_i}{b_i - a_i}\frac{d\sigma}{e-d}. \tag{5.67}$$

This integral can be easily evaluated:

$$\begin{aligned}\mathscr{D}_1^p(\beta) &= \frac{l}{3}\frac{e^3 - d^3}{e-d}\sum_{i=1}^{l}(\beta_i^p)^2\\ &\quad + \prod_{j=1}^{n}\frac{1}{(b_j - a_j)}\int_{a_1}^{b_1}\cdots\int_{a_n}^{b_n}\left(l\sum_{i=1}^{n}\beta_i^p\alpha_i + \beta_0^p - \alpha_p\right)^2 d\alpha_1\cdots d\alpha_n.\end{aligned}$$

Denoting $(a_i + b_i)/2 = c_i$, $(a_i - b_i)/2 = \mathscr{M}_i$, $t_i = \alpha_i - c_i$, and substituting the variables, we obtain

$$\begin{aligned}\mathscr{D}_1^p(\beta) &= \frac{l}{3}\frac{e^3 - d^3}{e-d}\sum_{i=1}^{l}(\beta_i^p)^2\\ &\quad + \prod_{j=1}^{n}\frac{1}{2\mathscr{M}_j}\int_{-\mathscr{M}_1}^{\mathscr{M}_1}\cdots\int_{-\mathscr{M}_n}^{\mathscr{M}_n}\left(l\sum_{i=1}^{n}\beta_i^p(t_i + c_i) + \beta_0^p - (t_p + c_p)\right)^2 dt_1\cdots dt_n.\end{aligned} \tag{5.68}$$

Since the integration is carried out over the symmetric intervals $[-\mathscr{M}, \mathscr{M}]$, the terms linear in t vanish. We thus obtain

$$\begin{aligned}\mathscr{D}_1^p(\beta) &= \frac{l}{3}(e^2 + ed + d^2)\sum_{i=1}^{l}(\beta_i^p)^2 + \left(\beta_0^p + \sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})c_i\right)^2\\ &\quad + \prod_{j=1}^{n}\frac{1}{2\mathscr{M}_j}\int_{-\mathscr{M}_1}^{\mathscr{M}_1}\cdots\int_{-\mathscr{M}_n}^{\mathscr{M}_n}\sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})^2 t_i^2\, dt_1\cdots dt_n.\end{aligned} \tag{5.69}$$

Here the notation

$$\delta_{ip} = \begin{cases}1 & \text{for } i = p,\\ 0 & \text{for } i \neq p\end{cases}$$

is utilized. Finally we arrive at

$$\begin{aligned}\mathscr{D}_1^p(\beta) &= \frac{l}{3}(e^2 + ed + d^2)\sum_{i=1}^{l}(\beta_i^p)^2\\ &\quad + \left(\beta_0^p + \sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})c_i\right)^2 + \sum_{i=1}^{n}\frac{\mathscr{M}_i^2}{3}(l\beta_i^p - \delta_{ip})^2.\end{aligned} \tag{5.70}$$

In order to obtain the best linear estimator in the mean it remains only to minimize the expression (5.70) with respect to parameters β.

Equating the partial derivatives of (5.70) to zero, we obtain that

$$\beta_i^p = 0 \quad \text{for } i \neq p,$$

$$\beta_0^p = -c_p(l\beta_p^p - 1),$$

$$\beta_p^p = \frac{\dfrac{\mathcal{M}_p^2}{e^2 + ed + d^2}}{1 + \dfrac{l\mathcal{M}_p^2}{e^2 + ed + d^2}}. \tag{5.71}$$

Substituting the values (5.71) obtained into (5.53), we have

$$\alpha_p^{(1)} = \frac{\dfrac{\mathcal{M}_p^2}{e^2 + ed + d^2}}{1 + \dfrac{l\mathcal{M}_p^2}{e^2 + ed + d^2}} \chi_p^{\mathrm{T}} Y + \frac{c_p}{1 + \dfrac{l.\mathcal{M}_p^2}{e^2 + ed + d^2}}.$$

Introduce the notation $\rho_p^{(1)} = (e^2 + ed + d^2)/\mathcal{M}_p^2$. Then

$$\alpha_p^{(1)} = \frac{\dfrac{1}{l}\chi_p^{\mathrm{T}} Y + \dfrac{c_p}{l}\rho_p^{(1)}}{1 + \dfrac{1}{l}\rho_p^{(1)}}.$$

Observe that the quantity $(1/l)\chi_p^{\mathrm{T}} Y$ is the least-squares estimator of the parameter α_p^0. Thus

$$\alpha_p^{(1)} = \frac{\alpha_{\mathrm{lse}}^p + \dfrac{c_p}{l}\rho_p^{(1)}}{1 + \dfrac{1}{l}\rho_p^{(1)}}.$$

The first part of the theorem is proved.

(2) *Derivation of the best minimax estimator.* The functional whose minimum determines the best minimax estimator is equal to

$$\mathscr{D}_2^p(\beta) = \sup_{\sigma, \alpha}\left[\sigma^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left(l \sum_{i=1}^{n} (\beta_i^p \alpha_i + \beta_0^p - \alpha_p)\right)^2\right]. \tag{5.72}$$

Utilizing the notation

$$c_i = \frac{b_i + a_i}{2}, \qquad \mathcal{M}_i = \frac{b_i - a_i}{2}, \qquad t_i = \alpha_i - c_i,$$

and substituting the variables in (5.72), we have

$$\begin{aligned}
\mathscr{D}_2^p(\beta) &= e^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \sup_{|t_i| \leq \mathcal{M}_i}\left[\sum_{i=1}^{n} (l\beta_i^p - \delta_{ip})(t_i + c_i) + \beta_0^p\right]^2 \\
&= e^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \sup_{|t_i| \leq \mathcal{M}_i}\left[\sum_{i=1}^{n} (l\beta_i^p - \delta_{ip})t_i + \sum_{i=1}^{n} (l\beta_i^p - \delta_{ip})c_i + \beta_0^p\right]^2 \\
&= e^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left[\sum_{i=1}^{n} |l\beta_i^p - \delta_{ip}|\mathcal{M}_i + \left|\sum_{i=1}^{n} (l\beta_i^p - \delta_{ip})c_i + \beta_0^p\right|\right]^2.
\end{aligned}$$

Thus

$$\mathcal{D}_2^p(\beta) = e^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left[\sum_{i=1}^{n} |l\beta_i^p - \delta_{ip}|\mathcal{M}_i + \left|\sum_{i=1}^{n} (l\beta_i^p - \delta_{ip})c_i + \beta_0^p\right|\right]^2. \quad (5.73)$$

We shall now obtain the minimum of (5.73). By choosing β_0^p to be equal to

$$\beta_0^p = -\sum_{i=1}^{n} (l\beta_i^p - \delta_{ip})c_i,$$

the second term of the sum in the square brackets becomes zero. Therefore it is sufficient to minimize

$$\mathcal{D}_2^p(\beta) = e^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left(\sum_{i=1}^{n} |l\beta_i^p - \delta_{ip}|\mathcal{M}_i\right)^2. \quad (5.74)$$

The minimum of (5.74) is attained for

$$\beta_i^p = 0 \quad \text{for } i \neq p, \quad (5.75)$$

whence for $\beta_i^p = 0$ $(i \neq p)$ the functional (5.74) becomes

$$\mathcal{D}_2^p(\beta)|_{\beta_i^p = 0\,(i \neq p)} = le^2(\beta_p^p)^2 + (l\beta_p^p - 1)^2\mathcal{M}_p^2. \quad (5.76)$$

The minimum of this expression is attained at

$$\beta_p^p = \frac{\mathcal{M}_p^2}{e^2 + l\mathcal{M}_p^2}. \quad (5.77)$$

Substituting (5.75) and (5.77) into (5.53), we obtain the best minimax estimator

$$\alpha_p^{(2)} = \frac{\mathcal{M}_p^2}{e^2 + l\mathcal{M}_p^2}\chi_p^{\mathrm{T}}Y + \left(\frac{l\mathcal{M}_p^2}{e^2 + l\mathcal{M}_p^2} - 1\right)c_p = \frac{l\mathcal{M}_p^2\alpha_{\mathrm{lse}}^p + c_p e^2}{e^2 + l\mathcal{M}_p^2}.$$

Introducing the notation $\rho_p^{(2)} = e^2/\mathcal{M}_p^2$, we arrive at

$$\alpha_p^{(2)} = \frac{\alpha_{\mathrm{lse}}^p + \dfrac{c_p}{l}\rho_p^{(2)}}{1 + \dfrac{1}{l}\rho_p^{(2)}}.$$

(3) *Derivation of the uniformly best linear estimator.* To evaluate the uniformly best estimator it is required to minimize the functional

$$\mathcal{D}_3^p(\beta) = \sup_{\alpha,\sigma}(\mathcal{D}^p(\beta|\alpha,\sigma) - \mathcal{D}^p(\beta_{\mathrm{lse}}|\alpha,\sigma)),$$

or explicitly,

$$\mathcal{D}_3^p(\beta) = \sup_{d \le \sigma \le e}\left[l\sigma^2\left(\sum_{i=1}^{l}(\beta_i^p)^2 - 1\right)\right]$$

$$+ \sup_{a_i \le \alpha_i \le b_i}\left[\sum_{i=1}^{n}(l\beta_i^p - \delta_{ip})\alpha_i + \beta_0^p\right]^2. \quad (5.78)$$

It is easy to verify that in this case all the calculations are the same as those carried out in the preceding subsection, except that if

$$\sum_{i=1}^{l} (\beta_i^p)^2 - 1 < 0, \tag{5.79}$$

then $d = \inf \sigma$ should be taken instead of $e = \sup \sigma$.

Consequently

$$\beta_0^p = -\sum_{i=1}^{l} (l\beta_i^p - \delta_{ip})c_i,$$

$$\beta_i^p = \begin{cases} 0 & \text{for } i \neq p, \\ \dfrac{\mathcal{M}_i^2}{s^2 + l\mathcal{M}_i^2} & \text{for } i = p, \end{cases} \tag{5.80}$$

where s is either $\inf \sigma$ or $\sup \sigma$, depending on the sign of $\sum_{i=1}^{n} (\beta_i^p)^2 - 1$. However, for β_i^p as given by (5.80) the expression (5.79) is negative:

$$\sum_{i=1}^{l} (\beta_i^p)^2 - 1 = \left(\frac{\mathcal{M}_p^2}{s^2 + \mathcal{M}_p^2 l}\right)^2 - 1 < 0.$$

Hence $s = \inf \sigma = d$. Thus the uniformly best linear estimator is equal to

$$\alpha_p^{(3)} = \frac{\alpha_{\text{lse}}^p + \dfrac{c_p}{l}\rho_p^{(3)}}{1 + \dfrac{1}{l}\rho_p^{(3)}},$$

where in this case

$$\rho_p^{(3)} = \frac{d^2}{\mathcal{M}_p^2}.$$ □

§9 Utilizing Prior Information

According to Theorem 5.6 the availability of the following prior information:

(1) the interval $[a_i, b_i]$ to which the estimated parameter α_p belongs,
(2) the interval $[d, e]$ to which the variance of the noise σ belongs,

allows us to construct the best linear estimators. According to Theorem 5.7 the functional defining the quality of the best linear estimator is $1 + (\rho_p^{(i)}/l)$ times smaller than the functional corresponding to the least-squares estimator.

Usually it is not too difficult to obtain this prior information for solving practical problems within the Gauss–Markov model. As a rule the intervals in which the measured values of y are situated,

$$\tau_i \leq y_i \leq T_i \tag{5.81}$$

are known. This knowledge results from long experience or from the knowledge of the laws of nature. For example, when constructing the regression for the temperature forecast in Moscow on the 166th day of the year, it is known *a priori* that the forecast value of t lies within the given limits $+5°\text{C} \leq t \leq 35°\text{C}$. The knowledge of the bounds (5.81) allows us to obtain intervals for the estimated parameters. The equality $\alpha_p^0 = M(1/l)\chi_p^{\mathrm{T}} Y$ implies that

$$b_p = \sup_Y \frac{1}{l}\chi_p^{\mathrm{T}} Y \leq \frac{1}{l}\left(\sum_{i=1}^{l}{}' T_i \hat{\psi}_p(x_i) + \sum_{i=1}^{l}{}'' \tau_i \hat{\psi}_p(x_i)\right).$$

Here the first sum $\sum'$ contains the positive coordinates of the vector $\chi_p = (\hat{\psi}_p(x_1), \ldots, \hat{\psi}_p(x_l))^{\mathrm{T}}$, while the second contains the negative ones. Analogously the bounds

$$a_p = \inf_Y \frac{1}{l}\chi_p^{\mathrm{T}} Y \geq \frac{1}{l}\left(\sum_{i=1}^{l}{}' \tau_i \hat{\psi}_p(x_i) + \sum_{i=1}^{l}{}'' T_i \hat{\psi}_p(x_i)\right)$$

are obtained.

To estimate the interval for the variance we can also utilize our experience and knowledge of the laws which govern errors. However, if the interval obtained for the variance is too wide, we can then use alternatively the probabilistic approach, which consists of choosing the interval which contains the true value of the variance with the highest probability.

It is known that the quantity

$$\sigma_{\text{emp}}^2 = \frac{\sum_{i=1}^{l} y_i^2 - l\sum_{p=1}^{n} (\alpha_{\text{lse}}^p)^2}{l - n}$$

is an unbiased estimator of the error variance. We shall utilize Chebyshev's inequality

$$P\left\{\sigma_{\text{emp}}^2 \geq \frac{\sigma^2}{\eta}\right\} \leq \eta,$$

which implies that with probability $1 - \eta$

$$\sigma^2 \geq \sigma_{\text{emp}}^2 \eta. \tag{5.82}$$

The bound (5.82) may be refined if the nature of the error distribution is known.

Based on the interval for the variance $d \leq \sigma \leq e$ and the interval to which the parameter α_p belongs, the parameters $\rho_p^{(i)}$ and $c_p^{(i)}$ are found by means of which optimal linear estimators are constructed. Note that the more indefinite the prior information is (the wider the interval is), the smaller the value of $\rho_p^{(i)}$ will be and the closer the best linear estimator will be to the least-squares estimator. It can be shown that for trivial prior information ($-\infty < \alpha_p < \infty$, $0 < \sigma < \infty$) the best linear estimator coincides with the least-squares one.

To complete the theory of the best linear estimation it remains to clarify how sensitive the methods of linear estimation are to the precision of prior information. Theorem 5.8 answers this question.

Theorem 5.8 (Koshcheev). *Let $\hat{\alpha}^i_p = \alpha_p(\hat{\rho}^{(i)}_p, \hat{c}_p)$ be the best linear estimator computed from approximate values of the parameters $\hat{\rho}^{(i)}_p$, $\hat{c}_p$, $\hat{\mathscr{M}}_p$, while the true values of the parameters equal $\rho^{(i)}_p$, c_p, $\mathscr{M}_p$. Then the quality of the estimator obtained is given by*

$$\mathscr{D}^p_i(\hat{\alpha}_p(\hat{\rho}^{(i)}_p, \hat{c}_p)) = \frac{1 + v_i \dfrac{(\hat{\rho}^{(i)}_p)^2}{\rho^{(i)}_p}}{(1 + \hat{\rho}^{(i)}_p)^2} \mathscr{D}^p_i(\alpha^p_{\text{lse}}) \qquad (i = 1, 2), \tag{5.83}$$

where

$$v_1 = 1 + 3\left(\frac{\hat{c}_p - c_p}{\hat{\mathscr{M}}_p}\right)^2, \qquad v_2 = \left(1 + \frac{|c_p - \hat{c}_p|}{\hat{\mathscr{M}}_p}\right)^2. \tag{5.84}$$

Observe that Theorem 5.7 is a particular case of Theorem 5.8 for $\hat{c}_p = c_p$ and $\hat{\rho}^{(i)}_p = \rho^{(i)}_p$.

It follows from the equality (5.83) that if the value of parameter $\hat{\rho}^{(i)}_p$ is related to $\rho^{(i)}_p$ and v_i by the inequality

$$\rho^{(i)}_p > \frac{\hat{\rho}^{(i)}_p v_i}{2 + \hat{\rho}^{(i)}_p}, \tag{5.85}$$

then the estimator obtained using $\hat{\rho}^{(i)}_p$, $\hat{c}_p$ will be better than the least-squares estimator. Consequently the choice of $\hat{\rho}^{(i)}_p$ is based on two contradictory considerations. To obtain an estimator at least as good as the least-squares one, the value of $\hat{\rho}^{(i)}_p$ should be reduced (so that (5.85) is fulfilled). But the gain, which is approximately equal to $\mathscr{D}_i(\alpha^p_{\text{lse}})/(1 + \hat{\rho}^{(i)})$, is decreased.

Proof of Theorem 5.8. First we shall compute the value of the criterion (5.55) for the estimator $\hat{\alpha}_p(\hat{\rho}^{(i)}_p c_p)$:

$$\begin{aligned} &M(\alpha_p(\hat{\rho}^{(i)}_p, \hat{c}_p) - \alpha^0_p)^2 \\ &= M\left(\frac{\alpha^p_{\text{lse}} + \dfrac{\hat{c}_p}{l}\hat{\rho}_p{}^{(i)}}{1 + \dfrac{\hat{\rho}^{(i)}_p}{l}} - \frac{\alpha^0_p + \dfrac{\hat{c}_p}{l}\hat{\rho}^{(i)}_p}{1 + \hat{\rho}^{(i)}_p \dfrac{1}{l}}\right)^2 + \left(\frac{\alpha^0_p + \dfrac{\hat{c}_p}{l}\hat{\rho}^{(i)}_p}{1 + \hat{\rho}^{(i)}_p \dfrac{1}{l}} - \alpha^0_p\right)^2 \\ &= \frac{\dfrac{\sigma^2}{l}}{\left(1 + \dfrac{\hat{\rho}^{(i)}_p}{l}\right)^2} + \frac{\left(\dfrac{\hat{\rho}^{(i)}_p}{l}\right)^2 (\hat{c}_p - \alpha^0_p)^2}{\left(1 + \dfrac{\hat{\rho}^{(i)}_p}{l}\right)^2}. \end{aligned}$$

The two relations (5.83) claimed in the theorem are verified by elementary calculations

$$\mathscr{D}_1^p(\hat{\alpha}) = \int_{c_p - \mathscr{M}_p}^{c_p + \mathscr{M}_p} \int_d^e \frac{\dfrac{\sigma^2}{l} + \left(\dfrac{\hat{\rho}_p^{(1)}}{l}\right)^2 (\hat{c}_p - \alpha_p^0)^2}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(1)}\right)^2} \frac{d\sigma}{e - d} \frac{d\alpha_p^0}{\mathscr{M}_p}$$

$$= \frac{\dfrac{1}{3}\dfrac{e^3 - d^3}{e - d} + \left(\dfrac{\hat{\rho}_p^{(1)}}{l}\right)^2 \left(\dfrac{\mathscr{M}_p^2}{3} + (c_p - \hat{c}_p)^2\right)}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(1)}\right)^2},$$

$$\mathscr{D}_1^p(\alpha_{\text{lse}}^p) = \int_{c_p - \mathscr{M}_p}^{c_p + \mathscr{M}_p} \frac{d\alpha}{2\mathscr{M}_p} \int_d^e \frac{\sigma^2}{l} \frac{d\sigma}{e - d} = \frac{e^2 + ed + d^2}{3l},$$

hence

$$\frac{\mathscr{D}_1^p(\hat{\alpha})}{\mathscr{D}_1^p(\alpha_{\text{lse}})} = \frac{1 + \dfrac{1}{\rho_p^{(1)}}\left(\dfrac{\hat{\rho}_p^{(1)}}{l}\right)^2 v_1}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(1)}\right)^2}, \qquad v_1 = 1 + 3\left(\frac{\hat{c}_p - c_p}{\mathscr{M}_p}\right)^2.$$

We now compute

$$\mathscr{D}_2^p(\hat{\alpha}) = \sup_{\alpha, \sigma} \frac{\dfrac{\sigma^2}{l} + \left(\dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2 (c_p - \alpha_p^0)^2}{\left(1 + \dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2} = \frac{\dfrac{e^2}{l} + \left(\dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2 (|\hat{c}_p - c_p| + \mathscr{M}_p)^2}{\left(1 + \hat{\rho}_p^{(2)}\dfrac{1}{l}\right)^2}.$$

On the other hand,

$$\mathscr{D}_2^p(\alpha_{\text{lse}}) = \sup_{\sigma} \frac{\sigma^2}{l} = \frac{e^2}{l},$$

hence

$$\frac{\mathscr{D}_2^p(\hat{\alpha}_p)}{\mathscr{D}_2^p(\alpha_{\text{lse}})} = \frac{1 + \dfrac{1}{\rho_p^{(2)}}\left(\dfrac{\hat{\rho}_p^{(2)}}{l}\right)^2 v_2}{\left(1 + \dfrac{1}{l}\hat{\rho}_p^{(2)}\right)^2}, \qquad v_2 = \left(1 + \frac{|c_p - \hat{c}_p|}{\mathscr{M}_p}\right)^2.$$

The theorem is proved. □

We have thus studied the theory of estimating regression parameters. This theory is based on the fact that in a certain narrow class of estimators the least-squares method is optimal (for normal regression this class is the class of unbiased estimators, and for general regression theory it is the class of linear unbiased estimators). It then turned out that in a class of biased estimators, better estimators than those arising from the least-squares method may be constructed. Such nonlinear biased methods of estimation were obtained for estimating parameters of normal regression, while linear biased methods arise in the general model of regression estimation.

Estimation methods presented in this chapter can be utilized for regression estimation provided the density $P(x)$ is known and the regression is indeed a linear function in the parameters.