3
The Multiple Linear Regression Model and Its Extensions

The main topic of this chapter is the linear regression model with more than one independent variables. The principles of *least squares* and *maximum likelihood* are used for the estimation of parameters. We present the algebraic, geometric, and statistical aspects of the problem, each of which has an intuitive appeal.

3.1 The Linear Model

Let y denotes the dependent (or study) variable that is linearly related to K independent (or explanatory) variables $X_1, \ldots, X_K$ through the parameters $\beta_1, \ldots, \beta_K$ and we write

$$y = X_1\beta_1 + \cdots + X_K\beta_K + e\,. \tag{3.1}$$

This is called as the multiple linear regression model. The parameters $\beta_1, \ldots, \beta_K$ are the regression coefficients associated with $X_1, \ldots, X_K$, respectively and e is the difference between the observed and the fitted linear relationship.

We have T sets of observations on y and $(X_1, \ldots, X_K)$, which we represent as follows:

$$(y, X) = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{K1} \\ \vdots & \vdots & & \vdots \\ y_T & x_{1T} & \cdots & x_{KT} \end{pmatrix} = (y,\ x_{(1)}, \ldots, x_{(K)}) = \begin{pmatrix} y_1,\ x_1' \\ \vdots \\ y_T,\ x_T' \end{pmatrix} \tag{3.2}$$

where $y = (y_1, \dots, y_T)'$ is a T-vector, $x_i = (x_{1i}, \dots, x_{Ki})'$ is a K-vector and $x_{(j)} = (x_{j1}, \dots, x_{jT})'$ is a T-vector. (Note that in (3.2), the first, third and fourth matrices are partitioned matrices.)

In such a case, there are T observational equations of the form (3.1):

$$y_t = x_t'\beta + e_t \,, \quad t = 1, \dots, T \,, \tag{3.3}$$

where $\beta' = (\beta_1, \dots, \beta_K)$, which can be written using the matrix notation,

$$y = X\beta + e \,, \tag{3.4}$$

where X is a $T \times K$ design matrix of T observations on each of the K explanatory variables and $e = (e_1, \dots, e_T)'$. If $x_1 = (1, \dots, 1)'$, then β_1 represents the intercept term in the model (3.4).

We consider the problems of estimation and testing of hypotheses on β under some assumptions. A general procedure for the estimation of β is to minimize

$$\sum_{t=1}^{T} M(e_t) = \sum_{t=1}^{T} M(y_t - x_t'\beta) \tag{3.5}$$

for a suitably chosen function M, some examples of which are $M(x) = |x|$ and $M(x) = x^2$ and more generally, $M(x) = |x|^p$. In general, one could minimize a global function of e such as $\max_t |e_t|$ over t. First we consider the case $M(x) = x^2$, which leads to the least-squares theory, and later introduce other functions that may be more appropriate in some situations.

Assumptions in Multiple Linear Regression Model

Some assumptions about the model (3.4) are needed for drawing the statistical inferences. For this purpose, we assume that e is observed as a random variable ϵ with the following assumptions:

(i) $\mathrm{E}(\epsilon) = 0$,

(ii) $\mathrm{E}(\epsilon\epsilon') = \sigma^2 I_T$,

(iii) $\mathrm{Rank}(X) = K$,

(iv) X is a non-stochastic matrix and

(v) $\epsilon \sim N(0, \sigma^2 I_T)$.

These assumptions are used to study the statistical properties of estimators of regression coefficients. The following assumption is required to study particularly the large sample properties of the estimators:

(vi) $\lim_{T\to\infty}(X'X/T) = \Delta$ exists and is a non-stochastic and nonsingular matrix (with finite elements).

The independent variables can also be stochastic in some cases. A case when X is stochastic is discussed later in Section 3.12 . We assume that X is non-stochastic in all further analysis.

3.2 The Principle of Ordinary Least Squares (OLS)

Let B be the set of all possible vectors β. If there is no further information, we have $B = \mathbb{R}^K$ (K-dimensional real Euclidean space). The object is to find a vector $b' = (b_1, \ldots, b_K)$ from B that minimizes the sum of squared residuals

$$S(\beta) = \sum_{t=1}^{T} e_t^2 = e'e = (y - X\beta)'(y - X\beta) \tag{3.6}$$

given y and X. A minimum will always exist, since $S(\beta)$ is a real-valued, convex, differentiable function. If we rewrite $S(\beta)$ as

$$S(\beta) = y'y + \beta'X'X\beta - 2\beta'X'y \tag{3.7}$$

and differentiate with respect to β (with the help of Theorems A.91–A.95), we obtain

$$\frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'y\,, \tag{3.8}$$

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X'X \quad \text{(at least nonnegative definite).} \tag{3.9}$$

Equating the first derivative to zero yields what are called the *normal equations*

$$X'Xb = X'y. \tag{3.10}$$

If X is of full rank K, then $X'X$ is positive definite and the unique solution of (3.10) is

$$b = (X'X)^{-1}X'y\,. \tag{3.11}$$

If X is not of full rank, equation (3.10) has a set of solutions

$$b = (X'X)^{-}X'y + (I - (X'X)^{-}X'X)w\,, \tag{3.12}$$

where $(X'X)^{-}$ is a g-inverse (generalized inverse) of $X'X$ and w is an arbitrary vector. [We note that a g-inverse $(X'X)^{-}$ of $X'X$ satisfies the properties $X'X(X'X)^{-}X'X = X'X$, $X(X'X)^{-}X'X = X$, $X'X(X'X)^{-}X' = X'$, and refer the reader to Section A.12 in Appendix A for the algebra of g-inverses and methods for solving linear equations, or to the books by Rao and Mitra (1971), and Rao and Rao (1998).] We prove the following theorem.

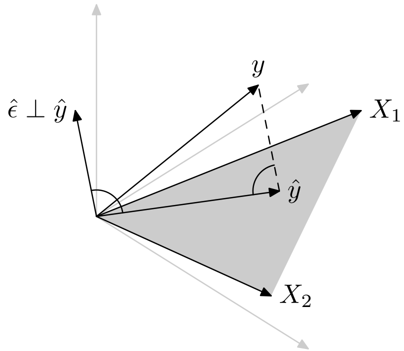

FIGURE 3.1. Geometric properties of OLS, $\theta \in \mathcal{R}(X)$ (for $T = 3$ and $K = 2$)

Theorem 3.1

(i) *$\hat{y} = Xb$, the empirical predictor of y, has the same value for all solutions b of $X'Xb = X'y$.*

(ii) *$S(\beta)$, the sum of squares defined in (3.6), attains the minimum for any solution of $X'Xb = X'y$.*

Proof: To prove (i), choose any b in the set (3.12) and note that

$$\begin{aligned} Xb &= X(X'X)^{-}X'y + X(I - (X'X)^{-}X'X)w \\ &= X(X'X)^{-}X'y \quad \text{(which is independent of } w\text{)}. \end{aligned}$$

Note that we used the result $X(X'X)^{-}X'X = X$ given in Theorem A.81.

To prove (ii), observe that for any β,

$$\begin{aligned} S(\beta) &= (y - Xb + X(b-\beta))'(y - Xb + X(b-\beta)) \\ &= (y-Xb)'(y-Xb) + (b-\beta)'X'X(b-\beta) + 2(b-\beta)'X'(y-Xb) \\ &= (y-Xb)'(y-Xb) + (b-\beta)'X'X(b-\beta)\,, \quad \text{using (3.10)} \\ &\geq (y-Xb)'(y-Xb) = S(b) \\ &= y'y - 2y'Xb + b'X'Xb = y'y - b'X'Xb = y'y - \hat{y}'\hat{y}\,. \end{aligned} \tag{3.13}$$

3.3 Geometric Properties of OLS

For the $T \times K$-matrix X, we define the *column space*

$$\mathcal{R}(X) = \{\theta : \ \theta = X\beta, \ \beta \in \mathbb{R}^K\}\,,$$

which is a subspace of $\mathbb{R}^T$. If we choose the norm $\|x\| = (x'x)^{1/2}$ for $x \in \mathbb{R}^T$, then the principle of least squares is the same as that of minimizing $\| \, y - \theta \, \|$ for $\theta \in \mathcal{R}(X)$. Geometrically, we have the situation as shown in Figure 3.1.

We then have the following theorem:

Theorem 3.2 *The minimum of* $\| y - \theta \|$ *for* $\theta \in \mathcal{R}(X)$ *is attained at* $\hat{\theta}$ *such that* $(y - \hat{\theta}) \perp \mathcal{R}(X)$*, that is, when* $y - \hat{\theta}$ *is orthogonal to all vectors in* $\mathcal{R}(X)$*, which is when* $\hat{\theta}$ *is the orthogonal projection of* y *on* $\mathcal{R}(X)$*. Such a* $\hat{\theta}$ *exists and is unique, and has the explicit expression*

$$\hat{\theta} = Py = X(X'X)^{-}X'y\,, \tag{3.14}$$

where $P = X(X'X)^{-}X'$ *is the orthogonal projection operator on* $\mathcal{R}(X)$*.*

Proof: Let $\hat{\theta} \in \mathcal{R}(X)$ be such that $(y - \hat{\theta}) \perp \mathcal{R}(X)$, that is, $X'(y - \hat{\theta}) = 0$. Then

$$\begin{aligned} \| y - \theta \|^2 &= (y - \hat{\theta} + \hat{\theta} - \theta)'(y - \hat{\theta} + \hat{\theta} - \theta) \\ &= (y - \hat{\theta})'(y - \hat{\theta}) + (\hat{\theta} - \theta)'(\hat{\theta} - \theta) \geq \| y - \hat{\theta} \|^2 \end{aligned}$$

since the term $(y - \hat{\theta})'(\hat{\theta} - \theta)$ vanishes using the orthogonality condition. The minimum is attained when $\theta = \hat{\theta}$. Writing $\hat{\theta} = X\hat{\beta}$, the orthogonality condition implies $X'(y - X\hat{\beta}) = 0$, that is, $X'X\hat{\beta} = X'y$. The equation $X'X\beta = X'y$ admits a solution, and $X\beta$ is unique for all solutions of β as shown in Theorem A.79. This shows that $\hat{\theta}$ exists.

Let $(X'X)^{-}$ be any g-inverse of $X'X$. Then $\hat{\beta} = (X'X)^{-}X'y$ is a solution of $X'X\beta = X'y$, and

$$X\hat{\beta} = X(X'X)^{-}X'y = Py\,,$$

which proves (3.14) of Theorem 3.2.

Note 1: If $\operatorname{rank}(X) = s < K$, it is possible to find a matrix U of order $(K - s) \times K$ and rank $K - s$ such that $\mathcal{R}(U') \cap \mathcal{R}(X') = \{0\}$, where 0 is the null vector. In such a case, $X'X + U'U$ is of full rank K, $(X'X + U'U)^{-1}$ is a g-inverse of $X'X$, and a solution of the normal equation $X'X\beta = X'y$ can be written as

$$\hat{\beta} = (X'X + U'U)^{-1}(X'y + U'u)\,, \tag{3.15}$$

where u is arbitrary. Also the projection operator P defined in (3.14) can be written as $P = X(X'X + U'U)^{-1}X'$. In some situations it is easy to find a matrix U satisfying the above conditions so that the g-inverse of $X'X$ can be computed as a regular inverse of a nonsingular matrix.

Note 2: The solution (3.15) can also be obtained as a conditional least-squares estimator when β is subject to the restriction $U\beta = u$ for a given arbitrary u. To prove this, we need only verify that $\hat{\beta}$ as in (3.15) satisfies the equation. Now

$$\begin{aligned} U\hat{\beta} &= U(X'X + U'U)^{-1}(X'y + U'u) \\ &= U(X'X + U'U)^{-1}U'u = u\,, \end{aligned}$$

which is true in view of result (iv) of Theorem A.81.

Note 3: It may be of some interest to establish the solution (3.15) using the calculus approach by differentiating

$$(y - X\beta)'(y - X\beta) + \lambda'(U\beta - u)$$

with respect to λ and β, where λ is a Lagrangian multiplier, which gives the equations

$$\begin{aligned} X'X\beta &= X'y + U'\lambda\,, \\ U\beta &= u\,, \end{aligned}$$

yielding the solution for β as in (3.15).

3.4 Best Linear Unbiased Estimation

3.4.1 Basic Theorems

In Sections 3.1 through 3.3, we viewed the problem of the linear model $y = X\beta + e$ as one of fitting the function $X\beta$ to y without making any assumptions on e. Now we consider e as a random variable denoted by ϵ, make some assumptions on its distribution, and discuss the estimation of β considered as an unknown vector parameter.

The usual assumptions made as in Section 3.1 are

$$\mathrm{E}(\epsilon) = 0\,, \quad \mathrm{E}(\epsilon\epsilon') = \sigma^2 I\,, \tag{3.16}$$

and X is a fixed or nonstochastic matrix of order $T \times K$, with full rank K.

When $\mathrm{E}(\epsilon\epsilon') = \sigma^2 I$, then ϵ's are termed as homoscedastic or spherical disturbances. When it does not hold true, then ϵ's are termed as heteroscedastic or non-spherical disturbances. Similarly, when X is a rank deficient matrix, then the problem is termed as multicollinearity (cf. Section 3.14).

We prove two lemmas that are of independent interest in estimation theory and use them in the special case of estimating β by linear functions of y.

Lemma 3.3 (Rao, 1973a, p. 317) *Let T be a statistic such that $\mathrm{E}(T) = \theta$, $\mathrm{V}(T) < \infty$, $\mathrm{V}(.)$ denotes the variance, and where θ is a scalar parameter. Then a necessary and sufficient condition that T is MVUE (minimum variance unbiased estimator) of the parameter θ is*

$$\mathrm{cov}(T, t) = 0 \quad \forall t \quad \textit{such that} \quad \mathrm{E}(t) = 0 \quad \textit{and} \quad \mathrm{V}(t) < \infty\,. \tag{3.17}$$

Proof of necessity: Let T be MVUE and t be such that $\mathrm{E}(t) = 0$ and $\mathrm{V}(t) < \infty$. Then $T + \lambda t$ is unbiased for θ for every $\lambda \in \mathbb{R}$, and

$$\begin{aligned} \mathrm{V}(T + \lambda t) &= \mathrm{V}(T) + \lambda^2\,\mathrm{V}(t) + 2\lambda\,\mathrm{cov}(T, t) \geq \mathrm{V}(T) \\ &\Rightarrow \lambda^2\,\mathrm{V}(t) + 2\lambda\,\mathrm{cov}(T, t) \geq 0 \quad \forall\lambda \\ &\Rightarrow \mathrm{cov}(T, t) = 0\,. \end{aligned}$$

Proof of sufficiency: Let $\tilde{T}$ be any unbiased estimator with finite variance. Then $\tilde{T} - T$ is such that $\mathrm{E}(\tilde{T} - T) = 0$, $\mathrm{V}(\tilde{T} - T) < \infty$, and

$$\begin{aligned} \mathrm{V}(\tilde{T}) = \mathrm{V}(T + \tilde{T} - T) &= \mathrm{V}(T) + \mathrm{V}(\tilde{T} - T) + 2\operatorname{cov}(T, \tilde{T} - T) \\ &= \mathrm{V}(T) + \mathrm{V}(\tilde{T} - T) \geq \mathrm{V}(T) \end{aligned}$$

if (3.17) holds.

Let $T' = (T_1, \ldots, T_k)$ be an unbiased estimate of the vector parameter $\theta' = (\theta_1, \ldots, \theta_k)$. Then the $k \times k$-matrix

$$D(T) = \mathrm{E}(T-\theta)(T-\theta)' = \begin{pmatrix} \mathrm{V}(T_1) & \operatorname{cov}(T_1, T_2) & \cdots & \operatorname{cov}(T_1, T_k) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(T_k, T_1) & \operatorname{cov}(T_k, T_2) & \cdots & \mathrm{V}(T_k) \end{pmatrix} \tag{3.18}$$

is called the dispersion matrix of T. We say T_0 is MDUE (minimum dispersion unbiased estimator) of θ if $D(T) - D(T_0)$ is nonnegative definite, or in our notation

$$D(T) - D(T_0) \geq 0 \tag{3.19}$$

for any T such that $\mathrm{E}(T) = \theta$.

Lemma 3.4 *If T_{i0} is MVUE of θ_i, $i = 1, \ldots, k$, then $T_0' = (T_{10}, \ldots, T_{k0})$ is MDUE of θ and vice versa.*

Proof: Consider $a'T_0$, which is unbiased for $a'\theta$. Since $\operatorname{cov}(T_{i0}, t) = 0$ for any t such that $\mathrm{E}(t) = 0$, it follows that $\operatorname{cov}(a'T_0, t) = 0$, which shows that

$$\mathrm{V}(a'T_0) = a'D(T_0)a \leq a'D(T)a\,, \tag{3.20}$$

where T is an alternative estimator to T_0. Then (3.20) implies

$$D(T_0) \leq D(T)\,. \tag{3.21}$$

The converse is true, since (3.21) implies that the i^{th} diagonal element of $D(T_0)$, which is $\mathrm{V}(T_{i0})$, is not greater than the i^{th} diagonal element of $D(T)$, which is $\mathrm{V}(T_i)$.

The lemmas remain true if the estimators are restricted to a particular class that is closed under addition, such as all linear functions of observations.

Combining Lemmas 3.3 and 3.4, we obtain the fundamental equation characterizing an MDUE t of θ at a particular value θ_0:

$$\operatorname{cov}(t, z|\theta_0) = 0 \quad \forall z \quad \text{such that} \quad \mathrm{E}(z|\theta) = 0 \quad \forall \theta\,, \tag{3.22}$$

which we exploit in estimating the parameters in the linear model. If there is a t for which (3.22) holds for all θ_0, then we have a globally optimum estimator. The basic theory of equation (3.22) and its applications is first given in Rao (1989).

We revert back to the linear model

$$y = X\beta + \epsilon \tag{3.23}$$

with $\mathrm{E}(\epsilon) = 0$, $D(\epsilon) = \mathrm{E}(\epsilon\epsilon') = \sigma^2 I$, and discuss the estimation of β. Let $a + b'y$ be a linear function with zero expectation, then

$$\begin{aligned} \mathrm{E}(a + b'y) &= a + b'X\beta = 0 \quad \forall\beta \\ &\Rightarrow a = 0\,, \quad b'X = 0 \quad \text{or} \quad b \in \mathcal{R}(Z)\,, \end{aligned}$$

where Z is the matrix whose columns span the space orthogonal to $\mathcal{R}(X)$ with $\operatorname{rank}(Z) = T - \operatorname{rank}(X)$. Thus, the class of all linear functions of y with zero expectation is

$$(Zc)'y = c'Z'y\,, \tag{3.24}$$

where c is an arbitrary vector.

Case 1: Rank$(X) = K$. Rank$(Z) = T - K$ and $(X'X)$ is nonsingular, admitting the inverse $(X'X)^{-1}$. The following theorem provides the estimate of β.

Theorem 3.5 *The MDLUE (minimum dispersion linear unbiased estimator) of β is*

$$\hat\beta = (X'X)^{-1}X'y\,, \tag{3.25}$$

which is the same as the least squares estimator of β, and the minimum dispersion matrix is

$$\sigma^2(X'X)^{-1}\,. \tag{3.26}$$

Proof: Let $a + By$ be an unbiased estimator of β. Then

$$\mathrm{E}(a + By) = a + BX\beta = \beta \quad \forall\beta \quad \Rightarrow \quad a = 0\,, BX = I\,. \tag{3.27}$$

If By is MDLUE, using equation (3.22), it is sufficient that

$$\begin{aligned} 0 &= \operatorname{cov}(By, c'Z'y) \quad \forall c \\ &= \sigma^2 BZc \quad \forall c \\ &\Rightarrow BZ = 0 \quad \Rightarrow \quad B = AX' \quad \text{for some } A\,. \end{aligned} \tag{3.28}$$

Thus we have two equations for B from (3.27) and (3.28):

$$BX = I\,, \quad B = AX'\,.$$

Substituting AX' for B in $BX = I$:

$$A(X'X) = I \quad \Leftrightarrow \quad A = (X'X)^{-1}\,, \quad B = (X'X)^{-1}X'\,, \tag{3.29}$$

giving the MDLUE

$$\hat\beta = By = (X'X)^{-1}X'y$$

with the dispersion matrix

$$\begin{aligned} D(\hat\beta) &= D((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'D(y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}, \end{aligned}$$

which proves Theorem 3.5.

Case 2: Rank$(X) = r < K$ (deficiency in rank) and rank$(Z) = T - r$, in which case $X'X$ is singular. We denote any g-inverse of $X'X$ by $(X'X)^-$. The consequences of deficiency in the rank of X, which arises in many practical applications, are as follows.

(i) The linear model, $y = X\beta + \epsilon$, is not identifiable in the sense that there may be several values of β for which $X\beta$ has the same value, so that no particular value can be associated with the model.

(ii) The condition of unbiasedness for estimating β is $BX = I$, as derived in (3.27). If X is deficient in rank, we cannot find a B such that $BX = I$, and thus β cannot be unbiasedly estimated.

(iii) Let $l'\beta$ be a given linear parametric function and let $a + b'y$ be an estimator. Then

$$\mathrm{E}(a + b'y) = a + b'X\beta = l'\beta \quad \Rightarrow \quad a = 0\,, \quad X'b = l\,. \tag{3.30}$$

The equation $X'b = l$ has a solution for b if $l \in \mathcal{R}(X')$. Thus, although the whole parameter is not unbiasedly estimable, it is possible to estimate all linear functions of the type $l'\beta$, $l \in \mathcal{R}(X')$. The following theorem provides the MDLUE of a given number s such linear functions

$$(l_1'\beta, \ldots, l_s'\beta) = (L'\beta)' \quad \text{with} \quad L = (l_1, \ldots, l_s)\,. \tag{3.31}$$

A linear function $m'\beta$ is said to be nonestimable if $m \notin \mathcal{R}(X')$.

Theorem 3.6 *Let $L'\beta$ be s linear functions of β such that $\mathcal{R}(L) \subset \mathcal{R}(X')$, implying $L = X'A$ for some A. Then the MDLUE of $L'\beta$ is $L'\hat\beta$, where $\hat\beta = (X'X)^-X'y$, and the dispersion matrix of $L'\hat\beta$ is $\sigma^2L'(X'X)^-L$, where $(X'X)^-$ is any g-inverse of $X'X$.*

Proof: Let Cy be an unbiased estimator of $L'\beta$. Then

$$\mathrm{E}(Cy) = CX\beta = L'\beta \quad \Rightarrow \quad CX = L'\,.$$

Now

$$\mathrm{cov}(Cy, Z'y) = \sigma^2CZ = 0 \quad \Rightarrow \quad C = BX' \quad \text{for some } B\,.$$

Then $CX = L' = BX'X = L'$, giving $B = L'(X'X)^-$ as one solution, and $C = BX' = L'(X'X)^-X'$. The MDLUE of $L'\beta$ is

$$Cy = L'(X'X)^-X'y = L'\hat\beta\,.$$

An easy computation gives $D(L'\hat{\beta}) = \sigma^2 L'(X'X)^- L$.

Note that $\hat{\beta}$ *is not an estimate of* β. However, it can be used to compute the best estimates of estimable parametric functions of β.

Case 3: Rank$(X) = r < K$, in which case not all linear parametric functions are estimable. However there may be additional information in the form of linear relationships

$$u = U\beta + \delta \tag{3.32}$$

where U is an $s \times K$-matrix, with $\mathrm{E}(\delta) = 0$ and $D(\delta) = \sigma_0^2 I$. Note that (3.32) reduces to a nonstochastic relationship when $\sigma_0 = 0$, so that the following treatment covers both the stochastic and nonstochastic cases. Let us consider the estimation of the linear function $p'\beta$ by a linear function of the form $a'y + b'u$. The unbiasedness condition yields

$$\mathrm{E}(a'y + b'u) = a'X\beta + b'U\beta = p'\beta \quad \Rightarrow \quad X'a + U'b = p\,. \tag{3.33}$$

Then

$$\mathrm{V}(a'y + b'u) = a'a\sigma^2 + b'b\sigma_0^2 = \sigma^2(a'a + \rho b'b)\,, \tag{3.34}$$

where $\rho = \sigma_0^2/\sigma^2$, and the problem is one of minimizing $(a'a + \rho b'b)$ subject to the condition (3.33) on a and b. Unfortunately, the expression to be minimized involves an unknown quantity, except when $\sigma_0 = 0$. However, we shall present a formal solution depending on ρ. Considering the expression with a Lagrangian multiplier

$$a'a + \rho b'b + 2\lambda'(X'a + U'b - p)\,,$$

the minimizing equations are

$$a = X\lambda\,, \quad \rho b = U\lambda\,, \quad X'a + U'b = p\,.$$

If $\rho \neq 0$, substituting for a and b in the last equation gives another set of equations:

$$(X'X + \rho^{-1}U'U)\lambda = p\,, \quad a = X\lambda\,, \quad b = U\lambda \tag{3.35}$$

which is easy to solve. If $\rho = 0$, we have the equations

$$a = X\lambda\,, \quad b = U\lambda\,, \quad X'a + U'b = p\,.$$

Eliminating a, we have

$$X'X\lambda + U'b = p\,, \quad U\lambda = 0\,. \tag{3.36}$$

We solve equations (3.36) for b and λ and obtain the solution for a by using the equation $a = X\lambda$. For practical applications, it is necessary to have some estimate of ρ when $\sigma_0 \neq 0$. This may be obtained partly from the available data and partly from previous information.

3.4.2 Linear Estimators

The statistician's task is now to estimate the true but unknown vector β of regression parameters in the model (3.23) on the basis of observations (y, X) and assumptions already stated. This will be done by choosing a suitable estimator $\hat{\beta}$, which then will be used to calculate the conditional expectation $\mathrm{E}(y|X) = X\beta$ and an estimate for the error variance σ^2. It is common to choose an estimator $\hat{\beta}$ that is linear in y, that is,

$$\hat{\beta} = Cy + d\,, \tag{3.37}$$

where $C : K \times T$ and $d : K \times 1$ are nonstochastic matrices to be determined by minimizing a suitably chosen risk function.

First we have to introduce some definitions.

Definition 3.7 *$\hat{\beta}$ is called a homogeneous estimator of β if $d = 0$; otherwise $\hat{\beta}$ is called heterogeneous.*

In Section 3.2, we have measured the model's goodness of fit by the sum of squared errors $S(\beta)$. Analogously we define, for the random variable $\hat{\beta}$, the quadratic loss function

$$L(\hat{\beta}, \beta, A) = (\hat{\beta} - \beta)'A(\hat{\beta} - \beta)\,, \tag{3.38}$$

where A is a symmetric and ≥ 0 (*i.e.*, at least nonnegative definite) $K \times K$-matrix. (See Theorems A.36–A.38 where the definitions of $A > 0$ for positive definiteness and $A \geq 0$ for nonnegative definiteness are given.)

Obviously the loss (3.38) depends on the sample. Thus we have to consider the average or expected loss over all possible samples, which we call the risk.

Definition 3.8 *The quadratic risk of an estimator $\hat{\beta}$ of β is defined as*

$$R(\hat{\beta}, \beta, A) = \mathrm{E}(\hat{\beta} - \beta)'A(\hat{\beta} - \beta). \tag{3.39}$$

The next step now consists of finding an estimator $\hat{\beta}$ that minimizes the quadratic risk function over a class of appropriate functions. Therefore we have to define a criterion to compare estimators:

Definition 3.9 *($R(A)$ superiority) An estimator $\hat{\beta}_2$ of β is called $R(A)$ superior or an $R(A)$-improvement over another estimator $\hat{\beta}_1$ of β if*

$$R(\hat{\beta}_1, \beta, A) - R(\hat{\beta}_2, \beta, A) \geq 0\,. \tag{3.40}$$

3.4.3 Mean Dispersion Error

The quadratic risk is closely related to the matrix-valued criterion of the mean dispersion error (MDE) of an estimator. The MDE is defined as the matrix

$$\mathrm{M}(\hat{\beta},\beta)=\mathrm{E}(\hat{\beta}-\beta)(\hat{\beta}-\beta)'. \tag{3.41}$$

We again denote the covariance matrix of an estimator $\hat{\beta}$ by $\mathrm{V}(\hat{\beta})$:

$$\mathrm{V}(\hat{\beta})=\mathrm{E}(\hat{\beta}-\mathrm{E}(\hat{\beta}))(\hat{\beta}-\mathrm{E}(\hat{\beta}))'. \tag{3.42}$$

If $\mathrm{E}(\hat{\beta})=\beta$, then $\hat{\beta}$ will be called unbiased (for β). If $\mathrm{E}(\hat{\beta})\neq\beta$, then $\hat{\beta}$ is called biased. The difference between $\mathrm{E}(\hat{\beta})$ and β is

$$\mathrm{Bias}(\hat{\beta},\beta)=\mathrm{E}(\hat{\beta})-\beta. \tag{3.43}$$

If $\hat{\beta}$ is unbiased, then obviously $\mathrm{Bias}(\hat{\beta},\beta)=0$.

The following decomposition of the mean dispersion error often proves to be useful:

$$\begin{aligned}\mathrm{M}(\hat{\beta},\beta) &= \mathrm{E}[(\hat{\beta}-\mathrm{E}(\hat{\beta}))+(\mathrm{E}(\hat{\beta})-\beta)][(\hat{\beta}-\mathrm{E}(\hat{\beta}))+(\mathrm{E}(\hat{\beta})-\beta)]' \\ &= \mathrm{V}(\hat{\beta})+(\mathrm{Bias}(\hat{\beta},\beta))(\mathrm{Bias}(\hat{\beta},\beta))'\,, \end{aligned} \tag{3.44}$$

that is, the MDE of an estimator is the sum of the covariance matrix and the squared bias (in its matrix version, *i.e.*, $(\mathrm{Bias}(\hat{\beta},\beta))(\mathrm{Bias}(\hat{\beta},\beta))'$).

The weighted mean dispersion error with the positive semidefinite matrix W is defined as the matrix

$$WM(\hat{\beta},\beta)=E(\hat{\beta}-\beta)W(\hat{\beta}-\beta)'\,. \tag{3.45}$$

When $W=X'X$, then the matrix in (3.45) is termed as predictive mean dispersion error.

MDE Superiority

As the MDE contains all relevant information about the quality of an estimator, comparisons between different estimators may be made on the basis of their MDE matrices.

Definition 3.10 (MDE I criterion) *Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be two estimators of β. Then $\hat{\beta}_2$ is called MDE-superior to $\hat{\beta}_1$ (or $\hat{\beta}_2$ is called an MDE-improvement to $\hat{\beta}_1$) if the difference of their MDE matrices is nonnegative definite, that is, if*

$$\Delta(\hat{\beta}_1,\hat{\beta}_2)=\mathrm{M}(\hat{\beta}_1,\beta)-\mathrm{M}(\hat{\beta}_2,\beta)\geq 0\,. \tag{3.46}$$

MDE superiority is a local property in the sense that (besides its dependency on σ^2) it depends on the particular value of β.

The quadratic risk function (3.39) is just a scalar-valued version of the MDE:

$$R(\hat{\beta},\beta,A)=\mathrm{tr}\{A\,\mathrm{M}(\hat{\beta},\beta)\}\,. \tag{3.47}$$

One important connection between $R(A)$ superiority and MDE superiority has been given by Theobald (1974) and Trenkler (1981):

Theorem 3.11 *Consider two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of β. The following two statements are equivalent:*

$$\begin{aligned} \Delta(\hat{\beta}_1, \hat{\beta}_2) &\geq 0, \qquad (3.48) \\ R(\hat{\beta}_1, \beta, A) - R(\hat{\beta}_2, \beta, A) = \operatorname{tr}\{A\Delta(\hat{\beta}_1, \hat{\beta}_2)\} &\geq 0 \qquad (3.49) \end{aligned}$$

for all matrices of the type $A = aa'$.

Proof: Using (3.46) and (3.47) we get

$$R(\hat{\beta}_1, \beta, A) - R(\hat{\beta}_2, \beta, A) = \operatorname{tr}\{A\Delta(\hat{\beta}_1, \hat{\beta}_2)\}. \qquad (3.50)$$

From Theorem A.43 it follows that $\operatorname{tr}\{A\Delta(\hat{\beta}_1, \hat{\beta}_2)\} \geq 0$ for all matrices $A = aa' \geq 0$ if and only if $\Delta(\hat{\beta}_1, \hat{\beta}_2) \geq 0$.

3.5 Estimation (Prediction) of the Error Term ϵ and σ^2

The linear model (3.23) may be viewed as the decomposition of the observation y into a nonstochastic part $X\beta$, also called the signal, and a stochastic part ϵ, also called the noise (or error), as discussed in Rao (1989). Since we have estimated $X\beta$ by $X\hat{\beta}$, we may consider the residual

$$\hat{\epsilon} = y - X\hat{\beta} = (I - P_X)y\,, \qquad (3.51)$$

where $P_X = X(X'X)^- X'$ is the projection operator on $\mathcal{R}(X)$, as an estimator (or predictor) of ϵ, with the mean prediction error

$$\begin{aligned} \mathrm{D}(\hat{\epsilon}) &= \mathrm{D}(y - X\hat{\beta}) = \mathrm{D}(I - P_X)y \\ &= \sigma^2(I - P_X)(I - P_X) = \sigma^2(I - P_X)\,. \qquad (3.52) \end{aligned}$$

However, the following theorem provides a systematic approach to the problem.

Theorem 3.12 *The MDLU predictor of ϵ is $\hat{\epsilon}$ as defined in (3.51).*

Proof: Let $C'y$ be an unbiased predictor of ϵ. Then

$$\mathrm{E}(C'y) = C'X\beta = 0 \quad \forall\beta \quad \Rightarrow \quad C'X = 0\,. \qquad (3.53)$$

The dispersion of error is

$$\mathrm{D}(\epsilon - C'y) = \mathrm{D}(\epsilon - C'\epsilon) = \sigma^2(I - C')(I - C)\,.$$

Putting $I - C' = M$, the problem is that of finding

$$\min MM' \quad \text{subject to} \quad MX = X\,. \qquad (3.54)$$

Since P_X and Z span the whole $\mathbb{R}^T$, we can write

$$M' = P_X A + ZB \quad \text{for some } A \text{ and } B\,,$$

giving

$$\begin{aligned} X' = X'M' &= X'A\,, \\ MM' &= A'P_X A + B'Z'ZB \\ &= A'X(X'X)^- X'A + B'Z'ZB \\ &= X(X'X)^- X' + B'Z'ZB \geq P_X \end{aligned}$$

with equality when $B = 0$. Then

$$M' = P_X A = X(X'X)^- X'A = X(X'X)^- X'\,,$$

and the best predictor of ϵ is

$$\hat{\epsilon} = C'y = (I - M)y = (I - P_X)y\,.$$

Using the estimate $\hat{\epsilon}$ of ϵ we can obtain an unbiased estimator of σ^2 as

$$s^2 = \frac{1}{T-r}\hat{\epsilon}'(I - P_X)\hat{\epsilon} = \frac{1}{T-r}y'(I - P_X)y \tag{3.55}$$

since (with rank $(X) = r$)

$$\begin{aligned} \mathrm{E}(s^2) &= \frac{1}{T-r}\,\mathrm{E}[y'(I - P_X)y] = \frac{1}{T-r}\mathrm{tr}(I - P_X)\,\mathrm{D}(y) \\ &= \frac{\sigma^2}{T-r}\mathrm{tr}(I - P_X) = \sigma^2\frac{T-r}{T-r} = \sigma^2\,. \end{aligned}$$

3.6 Classical Regression under Normal Errors

All results obtained so far are valid irrespective of the actual distribution of the random disturbances ϵ, provided that $\mathrm{E}(\epsilon) = 0$ and $\mathrm{E}(\epsilon\epsilon') = \sigma^2 I$. Now, we assume that the vector ϵ of random disturbances ϵ_t is distributed according to a T-dimensional normal distribution $N(0, \sigma^2 I)$, with the probability density

$$\begin{aligned} f(\epsilon; 0, \sigma^2 I) &= \prod_{t=1}^{T}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\epsilon_t^2\right) \\ &= (2\pi\sigma^2)^{-\frac{T}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\epsilon_t^2\right\}\,. \end{aligned} \tag{3.56}$$

Note that the components ϵ_t $(t = 1, \ldots, T)$ are independent and identically distributed as $N(0, \sigma^2)$. This is a special case of a general T-dimensional normal distribution $N(\mu, \Sigma)$ with density

$$f(\xi; \mu, \Sigma) = \{(2\pi)^T|\Sigma|\}^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\xi - \mu)'\Sigma^{-1}(\xi - \mu)\right\}\,. \tag{3.57}$$

The classical linear regression model under normal errors is given by

$$\left.\begin{array}{l} y = X\beta + \epsilon, \\ \epsilon \sim N(0, \sigma^2 I), \\ X \text{ nonstochastic, } \operatorname{rank}(X) = K. \end{array}\right\} \tag{3.58}$$

3.6.1 The Maximum-Likelihood (ML) Principle

Definition 3.13 *Let $\xi = (\xi_1, \dots, \xi_n)'$ be a random variable with density function $f(\xi; \Theta)$, where the parameter vector $\Theta = (\Theta_1, \dots, \Theta_m)'$ is an element of the parameter space Ω comprising all values that are a priori admissible.*

The basic idea of the maximum-likelihood principle is to consider the density $f(\xi; \Theta)$ for a specific realization of the sample ξ_0 of ξ as a function of Θ:

$$L(\Theta) = L(\Theta_1, \dots, \Theta_m) = f(\xi_0; \Theta).$$

$L(\Theta)$ will be referred to as the likelihood function of Θ given ξ_0.

The ML principle postulates the choice of a value $\hat{\Theta} \in \Omega$ that maximizes the likelihood function, that is,

$$L(\hat{\Theta}) \geq L(\Theta) \quad \text{for all } \Theta \in \Omega.$$

Note that $\hat{\Theta}$ may not be unique. If we consider all possible samples, then $\hat{\Theta}$ is a function of ξ and thus a random variable itself. We will call it the maximum-likelihood estimator of Θ.

3.6.2 Maximum Likelihood Estimation in Classical Normal Regression

Following Theorem A.82, we have for y from (3.58)

$$y = X\beta + \epsilon \sim N(X\beta, \sigma^2 I)\,, \tag{3.59}$$

so that the likelihood function of y is given by

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}. \tag{3.60}$$

Since the logarithmic transformation is monotonic, it is appropriate to maximize $\ln L(\beta, \sigma^2)$ instead of $L(\beta, \sigma^2)$, as the maximizing argument remains unchanged:

$$\ln L(\beta, \sigma^2) = -\frac{T}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta). \tag{3.61}$$

If there are no a priori restrictions on the parameters, then the parameter space is given by $\Omega = \{\beta; \sigma^2 : \beta \in \mathbb{R}^K; \sigma^2 > 0\}$. We derive the ML estimators of β and σ^2 by equating the first derivatives to zero (Theorems A.91–A.95):

$$\begin{aligned}
\text{(I)} \quad \frac{\partial \ln L}{\partial \beta} &= \frac{1}{2\sigma^2} 2X'(y - X\beta) = 0\,, && (3.62)\\
\text{(II)} \quad \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\beta)'(y - X\beta) = 0\,. && (3.63)
\end{aligned}$$

The *likelihood equations* are given by

$$\left.\begin{array}{ll}
\text{(I)} & X'X\hat{\beta} = X'y\,, \\
\text{(II)} & \hat{\sigma}^2 = \frac{1}{T}(y - X\hat{\beta})'(y - X\hat{\beta})\,.
\end{array}\right\} \qquad (3.64)$$

Equation (I) of (3.64) is identical to the well-known normal equation (3.10). Its solution is unique, as rank$(X) = K$ and we get the unique ML estimator

$$\hat{\beta} = b = (X'X)^{-1}X'y\,. \qquad (3.65)$$

If we compare (II) with the unbiased estimator s^2 (3.55) for σ^2, we see immediately that

$$\hat{\sigma}^2 = \frac{T-K}{T} s^2, \qquad (3.66)$$

so that $\hat{\sigma}^2$ is a biased estimator. The asymptotic expectation is given by (cf. Theorem A.102 (i))

$$\begin{aligned}
\lim_{T\to\infty} \mathrm{E}(\hat{\sigma}^2) &= \bar{\mathrm{E}}(\hat{\sigma}^2)\\
&= \mathrm{E}(s^2)\\
&= \sigma^2. && (3.67)
\end{aligned}$$

Thus we can state the following result.

Theorem 3.14 *The maximum-likelihood estimator and OLS estimator of β are identical in the model (3.59) of classical normal regression. The ML estimator $\hat{\sigma}^2$ of σ^2 is asymptotically unbiased.*

Note: The Cramér-Rao bound defines a lower bound (in the sense of definiteness of matrices) for the covariance matrix of unbiased estimators. In the model of normal regression, the Cramér-Rao bound is given by

$$\mathrm{V}(\tilde{\beta}) \geq \sigma^2(X'X)^{-1},$$

where $\tilde{\beta}$ is an arbitrary estimator. The covariance matrix of the ML estimator is just identical to this lower bound, so that b is the minimum dispersion unbiased estimator in the linear regression model under normal errors.

3.7 Consistency of Estimators

The OLSE of β under the model (3.4) with $e = \epsilon$ is $b = (X'X)^{-1}X'y$ and $\mathrm{V}(b) = \sigma^2(X'X)^{-1}$.

Under the assumption that $\lim_{T\to\infty}(X'X/T) = \Delta$ exists as a nonstochastic and nonsingular matrix (with finite elements),

$$\begin{aligned}\lim_{T\to\infty} \mathrm{V}(b) &= \sigma^2 \lim_{T\to\infty} \frac{1}{T}\left(\frac{X'X}{T}\right)^{-1} \\ &= \sigma^2 \lim_{T\to\infty} \frac{1}{T}\,\Delta^{-1} \\ &= 0. \end{aligned} \tag{3.68}$$

This implies that OLSE converges to β in quadratic mean and not only in probability. Thus OLSE is a consistent estimator of β.

Same conclusion can also be drawn using the notion of probability in limits. Consider a series $\{z^{(t)}\} = z^{(1)}, z^{(2)}, \ldots$ of random variables. Each random variable has a specific distribution, variance, and expectation. For example, $z^{(t)}$ could be the sample mean of a sample of size t of a given population. The series $\{z^{(t)}\}$ would then be the series of sample means of a successively increasing sample. Assume that $z^* < \infty$ exists, such that

$$\lim_{t\to\infty} P\{|z^{(t)} - z^*| \geq \delta\} = 0 \quad \text{for every} \quad \delta > 0.$$

Then z^* is called the *probability limit* of $\{z^{(t)}\}$, and we write $\operatorname{plim} z^{(t)} = z^*$ or $\operatorname{plim} z = z^*$ (cf. Definition A.101 and Goldberger, 1964, p. 115).

The consistency conclusion about OLSE can also be obtained under the weaker assumptions that

$$\operatorname*{plim}_{T\to\infty}\left(\frac{X'X}{T}\right) = \Delta_* \tag{3.69}$$

exists and is a nonsingular and nonstochastic matrix such that

$$\operatorname*{plim}_{T\to\infty}\left(\frac{X'\epsilon}{T}\right) = 0\ . \tag{3.70}$$

The assumptions (3.69) and (3.70) are denoted as $\operatorname{plim}(X'X/T) = \Delta_*$ and $\operatorname{plim}(X'\epsilon/T) = 0$, respectively.

Again, note that

$$\begin{aligned} b - \beta &= (X'X)^{-1}X'\epsilon \\ &= \left(\frac{X'X}{T}\right)^{-1}\frac{X'\epsilon}{T} \end{aligned}$$

and, therefore

$$\begin{aligned}\text{plim } (b-\beta) &= \text{plim}\left(\frac{X'X}{T}\right)^{-1} \text{plim}\left(\frac{X'\epsilon}{T}\right)\\ &= \Delta_*^{-1}\cdot 0\\ &= 0\ .\end{aligned}$$

Now we look at the consistency of an estimate of σ^2 as

$$\begin{aligned}s^2 &= \frac{1}{T-K}\hat{\epsilon}'\hat{\epsilon}\\ &= \frac{1}{T-K}\left[\epsilon'\epsilon-\epsilon'X(X'X)^{-1}X'\epsilon\right]\\ &= \frac{1}{T}\left(1-\frac{K}{T}\right)^{-1}\left[\epsilon'\epsilon-\epsilon'X(X'X)^{-1}X'\epsilon\right]\\ &= \left(1-\frac{K}{T}\right)^{-1}\left[\frac{\epsilon'\epsilon}{T}-\frac{\epsilon'X}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'\epsilon}{T}\right]\ . \end{aligned} \tag{3.71}$$

Note that $\epsilon'\epsilon/T$ consists of $\frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2$ and $\left\{\epsilon_t^2 : t=1,\ldots,T\right\}$ is a sequence of i.i.d. random variables with mean σ^2. Using the law of large number

$$\text{plim } \epsilon'\epsilon=\sigma^2\ . \tag{3.72}$$

Further

$$\begin{aligned}\text{plim}\left[\frac{\epsilon'X}{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'\epsilon}{T}\right] &= \left(\text{plim } \frac{\epsilon'X}{T}\right)\left[\text{plim}\left(\frac{X'X}{T}\right)^{-1}\right]\\ &\quad\times\left(\text{plim } \frac{X'\epsilon}{T}\right)\\ &= \left(\text{plim } \frac{\epsilon'X}{T}\right)\left[\text{plim}\left(\frac{X'X}{T}\right)\right]^{-1}\\ &\quad\times\left(\text{plim } \frac{X'\epsilon}{T}\right)\\ &= 0\cdot\Delta_*^{-1}\cdot 0 \qquad (3.73)\\ &= 0\ . \qquad (3.74)\end{aligned}$$

Using (3.72) and (3.74) in (3.71), we see that

$$\text{plim } s^2=\sigma^2\ .$$

Thus s^2 is a consistent estimator of σ^2.

3.8 Testing Linear Hypotheses

In this section, we consider the problem of testing a general linear hypothesis

$$H_0\colon R\beta = r \tag{3.75}$$

with R a $(K-s)\times K$–matrix and $\operatorname{rank}(R) = K - s$, against the alternative

$$H_1\colon R\beta \neq r \tag{3.76}$$

where it will be assumed that R and r are nonstochastic and known.

The hypothesis H_0 expresses the fact that the parameter vector β obeys $(K-s)$ exact linear restrictions, which are linearly independent, as it is required that $\operatorname{rank}(R) = K - s$. The general linear hypothesis (3.75) contains two main special cases:

Case 1: $s = 0$. The $K \times K$-matrix R is regular by the assumption $\operatorname{rank}(X) = K$, and we may express H_0 and H_1 in the following form:

$$H_0\colon \beta = R^{-1}r = \beta^*, \tag{3.77}$$

$$H_1\colon \beta \neq \beta^*. \tag{3.78}$$

Case 2: $s > 0$. We choose an $s \times K$-matrix G complementary to R such that the $K \times K$-matrix $\begin{pmatrix} G \\ R \end{pmatrix}$ is regular of rank K. Let

$$X\begin{pmatrix} G \\ R \end{pmatrix}^{-1} = \underset{T\times K}{\tilde{X}} = \left(\underset{T\times s}{\tilde{X}_1}\,, \quad \underset{T\times(K-s)}{\tilde{X}_2} \right),$$

$$\underset{s\times 1}{\tilde{\beta}_1} = G\beta\,, \qquad \underset{(K-s)\times 1}{\tilde{\beta}_2} = R\beta\,.$$

Then we may write

$$\begin{aligned} y = X\beta + \epsilon &= X\begin{pmatrix} G \\ R \end{pmatrix}^{-1}\begin{pmatrix} G \\ R \end{pmatrix}\beta + \epsilon \\ &= \tilde{X}\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} + \epsilon \\ &= \tilde{X}_1\tilde{\beta}_1 + \tilde{X}_2\tilde{\beta}_2 + \epsilon\,. \end{aligned}$$

The latter model obeys all assumptions (3.59). The hypotheses H_0 and H_1 are thus equivalent to

$$H_0\colon \tilde{\beta}_2 = r\,; \quad \tilde{\beta}_1 \text{ and } \sigma^2 > 0 \text{ arbitrary}, \tag{3.79}$$

$$H_1\colon \tilde{\beta}_2 \neq r\,; \quad \tilde{\beta}_1 \text{ and } \sigma^2 > 0 \text{ arbitrary}. \tag{3.80}$$

Ω stands for the whole parameter space (either H_0 or H_1 is valid) and $\omega \subset \Omega$ stands for the subspace in which only H_0 is true; thus

$$\left.\begin{array}{l} \Omega = \{\beta; \sigma^2 : \beta \in \mathbb{R}^K, \sigma^2 > 0\}\,, \\ \omega = \{\beta; \sigma^2 : \beta \in \mathbb{R}^K \text{ and } R\beta = r;\ \sigma^2 > 0\}\,. \end{array}\right\} \tag{3.81}$$

As a test statistic we will use the likelihood ratio

$$\lambda(y) = \frac{\max_\omega L(\Theta)}{\max_\Omega L(\Theta)}, \tag{3.82}$$

which may be derived in the following way.

Let $\Theta = (\beta, \sigma^2)$, then

$$\begin{aligned} \max_{\beta,\sigma^2} L(\beta, \sigma^2) &= L(\hat\beta, \hat\sigma^2) \\ &= (2\pi\hat\sigma^2)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2\hat\sigma^2}(y - X\hat\beta)'(y - X\hat\beta)\right\} \\ &= (2\pi\hat\sigma^2)^{-\frac{T}{2}} \exp\left\{-\frac{T}{2}\right\} \end{aligned} \tag{3.83}$$

and therefore

$$\lambda(y) = \left(\frac{\hat\sigma^2_\omega}{\hat\sigma^2_\Omega}\right)^{-\frac{T}{2}}, \tag{3.84}$$

where $\hat\sigma^2_\omega$ and $\hat\sigma^2_\Omega$ are ML estimators of σ^2 under H_0 and in Ω.

The random variable $\lambda(y)$ can take values between 0 and 1, which is obvious from (3.82). If H_0 is true, the numerator of $\lambda(y)$ gets closer to the denominator, so that $\lambda(y)$ should be close to 1 in repeated samples. On the other hand, $\lambda(y)$ should be close to 0 if H_1 is true.

Consider the linear transform of $\lambda(y)$:

$$\begin{aligned} F &= \left\{(\lambda(y))^{-\frac{2}{T}} - 1\right\}(T - K)(K - s)^{-1} \\ &= \frac{\hat\sigma^2_\omega - \hat\sigma^2_\Omega}{\hat\sigma^2_\Omega} \cdot \frac{T - K}{K - s}\,. \end{aligned} \tag{3.85}$$

If $\lambda \to 0$, then $F \to \infty$, and if $\lambda \to 1$, we have $F \to 0$, so that F *is close to 0* if H_0 is true and F *is sufficiently large* if H_1 is true.

Now we will determine F and its distribution for the two special cases of the general linear hypothesis.

Case 1: $s = 0$

The ML estimators under H_0 (3.77) are given by

$$\hat\beta = \beta^* \quad \text{and} \quad \hat\sigma^2_\omega = \frac{1}{T}(y - X\beta^*)'(y - X\beta^*)\,. \tag{3.86}$$

The ML estimators over Ω are available from Theorem 3.14:

$$\hat{\beta} = b \quad \text{and} \quad \hat{\sigma}^2_\Omega = \frac{1}{T}(y - Xb)'(y - Xb)\,. \tag{3.87}$$

Some rearrangements then yield

$$\left.\begin{array}{rcl} b - \beta^* &=& (X'X)^{-1}X'(y - X\beta^*)\,, \\ (b - \beta^*)'X'X &=& (y - X\beta^*)'X\,, \\ y - Xb &=& (y - X\beta^*) - X(b - \beta^*)\,, \\ (y - Xb)'(y - Xb) &=& (y - X\beta^*)'(y - X\beta^*) \\ && + (b - \beta^*)'X'X(b - \beta^*) \\ && - 2(y - X\beta^*)'X(b - \beta^*) \\ &=& (y - X\beta^*)'(y - X\beta^*) \\ && - (b - \beta^*)'X'X(b - \beta^*)\,. \end{array}\right\} \tag{3.88}$$

It follows that

$$T(\hat{\sigma}^2_\omega - \hat{\sigma}^2_\Omega) = (b - \beta^*)'X'X(b - \beta^*)\,, \tag{3.89}$$

leading to the test statistic

$$F = \frac{(b - \beta^*)'X'X(b - \beta^*)}{(y - Xb)'(y - Xb)} \cdot \frac{T - K}{K}\,. \tag{3.90}$$

Distribution of F

Numerator: The following statements are in order:

$b - \beta^* = (X'X)^{-1}X'[\epsilon + X(\beta - \beta^*)]$ [by (3.81)],
$\tilde{\epsilon} = \epsilon + X(\beta - \beta^*) \sim N(X(\beta - \beta^*), \sigma^2 I)$ [Theorem A.82],
$X(X'X)^{-1}X'$ idempotent and of rank K,
$(b - \beta^*)'X'X(b - \beta^*) = \tilde{\epsilon}'X(X'X)^{-1}X'\tilde{\epsilon}$
$\sim \sigma^2\chi^2_K(\sigma^{-2}(\beta - \beta^*)'X'X(\beta - \beta^*))$ [Theorem A.84]
and $\sim \sigma^2\chi^2_K$ under H_0.

Denominator:

$$\left.\begin{array}{ll} (y - Xb)'(y - Xb) = (T - K)s^2 = \epsilon'(I - P_X)\epsilon & \text{[cf. (3.55)],} \\ \epsilon'(I - P_X)\epsilon \sim \sigma^2\chi^2_{T-K} & \text{[Theorem A.87].} \end{array}\right\} \tag{3.91}$$

as $I - P_X = I - X(X'X)^{-1}X'$ is idempotent of rank $T - K$ (cf. Theorem A.61 (vi)).

We have

$$(I - P_X)X(X'X)^{-1}X' = 0 \quad \text{[Theorem A.61 (vi)],} \tag{3.92}$$

such that numerator and denominator are independently distributed (Theorem A.89).

Thus, the ratio F has the following properties (Theorem A.86):

- F is distributed as $F_{K,T-K}(\sigma^{-2}(\beta-\beta^*)'X'X(\beta-\beta^*))$ under H_1, and
- F is distributed as central $F_{K,T-K}$ under H_0: $\beta = \beta^*$.

If we denote by $F_{m,n,1-q}$ the $(1-q)$-quantile of $F_{m,n}$ (*i.e.*, $P(F \leq F_{m,n,1-q}) = 1-q$), then we may derive a uniformly most powerful test, given a fixed level of significance α (cf. Lehmann, 1986, p. 372):

$$\left.\begin{array}{lr} \text{Region of acceptance of } H_0\text{:} & 0 \leq F \leq F_{K,T-K,1-\alpha}\,, \\ \text{Critical region:} & F > F_{K,T-K,1-\alpha}\,. \end{array}\right\} \tag{3.93}$$

A selection of F-quantiles is provided in Appendix B.

Case 2: $s > 0$

Next we consider a decomposition of the model in order to determine the ML estimators under H_0 (3.79) and compare them with the corresponding ML estimator over Ω. Let

$$\beta' = \Big(\underset{1\times s}{\beta_1'}\,, \quad \underset{1\times(K-s)}{\beta_2'} \Big) \tag{3.94}$$

and, respectively,

$$y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon\,. \tag{3.95}$$

We set

$$\tilde{y} = y - X_2 r\,. \tag{3.96}$$

Because $\text{rank}(X) = K$, we have

$$\text{rank}\underset{T\times s}{(X_1)} = s\,, \quad \text{rank}\underset{T\times(K-s)}{(X_2)} = K - s\,, \tag{3.97}$$

such that the inverse matrices $(X_1'X_1)^{-1}$ and $(X_2'X_2)^{-1}$ do exist.

The ML estimators under H_0 are then given by

$$\hat{\beta}_2 = r, \quad \hat{\beta}_1 = (X_1'X_1)^{-1}X_1'\tilde{y} \tag{3.98}$$

and

$$\hat{\sigma}^2_\omega = \frac{1}{T}(\tilde{y} - X_1\hat{\beta}_1)'(\tilde{y} - X_1\hat{\beta}_1). \tag{3.99}$$

Separation of b

At first, it is easily seen that

$$\begin{aligned} b &= (X'X)^{-1}X'y \\ &= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}. \end{aligned} \tag{3.100}$$

Making use of the formulas for the inverse of a partitioned matrix yields (Theorem A.19)

$$\begin{pmatrix} (X_1'X_1)^{-1}[I + X_1'X_2D^{-1}X_2'X_1(X_1'X_1)^{-1}] & -(X_1'X_1)^{-1}X_1'X_2D^{-1} \\ -D^{-1}X_2'X_1(X_1'X_1)^{-1} & D^{-1} \end{pmatrix}, \tag{3.101}$$

where

$$D = X_2'M_1X_2 \tag{3.102}$$

and

$$M_1 = I - X_1(X_1'X_1)^{-1}X_1' = I - P_{X_1}\,. \tag{3.103}$$

M_1 is (analogously to $(I - P_X)$) idempotent and of rank $T - s$; further we have $M_1X_1 = 0$. The $(K - s) \times (K - s)$-matrix

$$D = X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2 \tag{3.104}$$

is symmetric and regular, as the normal equations are uniquely solvable. The estimators b_1 and b_2 of b are then given by

$$b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2D^{-1}X_2'M_1y \\ D^{-1}X_2'M_1y \end{pmatrix}. \tag{3.105}$$

Various relations immediately become apparent from (3.105):

$$\left.\begin{array}{rcl} b_2 & = & D^{-1}X_2'M_1y, \\ b_1 & = & (X_1'X_1)^{-1}X_1'(y - X_2b_2), \\ b_2 - r & = & D^{-1}X_2'M_1(y - X_2r) \\ & = & D^{-1}X_2'M_1\tilde{y} \\ & = & D^{-1}X_2'M_1(\epsilon + X_2(\beta_2 - r))\,, \end{array}\right\} \tag{3.106}$$

$$\left.\begin{array}{rcl} b_1 - \hat{\beta}_1 & = & (X_1'X_1)^{-1}X_1'(y - X_2b_2 - \tilde{y}) \\ & = & -(X_1'X_1)^{-1}X_1'X_2(b_2 - r) \\ & = & -(X_1'X_1)^{-1}X_1'X_2D^{-1}X_2'M_1\tilde{y}\,. \end{array}\right\} \tag{3.107}$$

Decomposition of $\hat{\sigma}^2_\Omega$

We write (using symbols u and v)

$$\begin{array}{rcccl} (y - Xb) & = & (y - X_2r - X_1\hat{\beta}_1) & - & \Big(X_1(b_1 - \hat{\beta}_1) + X_2(b_2 - r)\Big) \\ & = & u & - & v\,. \end{array} \tag{3.108}$$

Thus we may decompose the ML estimator $T\hat{\sigma}^2_\Omega = (y - Xb)'(y - Xb)$ as

$$(y - Xb)'(y - Xb) = u'u + v'v - 2u'v\,. \tag{3.109}$$

We have

$$\begin{aligned}
u &= y - X_2 r - X_1\hat{\beta}_1 = \tilde{y} - X_1(X_1'X_1)^{-1}X_1'\tilde{y} = M_1\tilde{y}\,, && (3.110)\\
u'u &= \tilde{y}'M_1\tilde{y}\,, && (3.111)\\
v &= X_1(b_1 - \hat{\beta}_1) + X_2(b_2 - r)\\
&= -X_1(X_1'X_1)^{-1}X_1'X_2D^{-1}X_2'M_1\tilde{y} \quad \text{[by (3.106)]}\\
&\quad\ +X_2D^{-1}X_2'M_1\tilde{y} \quad \text{[by (3.107)]}\\
&= M_1X_2D^{-1}X_2'M_1\tilde{y}\,, && (3.112)\\
v'v &= \tilde{y}'M_1X_2D^{-1}X_2'M_1\tilde{y}\\
&= (b_2 - r)'D(b_2 - r)\,, && (3.113)\\
u'v &= v'v\,. && (3.114)
\end{aligned}$$

Summarizing, we may state

$$\begin{aligned}
(y - Xb)'(y - Xb) &= u'u - v'v && (3.115)\\
&= (\tilde{y} - X_1\hat{\beta}_1)'(\tilde{y} - X_1\hat{\beta}_1) - (b_2 - r)'D(b_2 - r)
\end{aligned}$$

or,

$$T(\hat{\sigma}_\omega^2 - \hat{\sigma}_\Omega^2) = (b_2 - r)'D(b_2 - r)\,. \qquad (3.116)$$

We therefore get in case 2: $s > 0$:

$$F = \frac{(b_2 - r)'D(b_2 - r)}{(y - Xb)'(y - Xb)}\frac{T - K}{K - s}\,. \qquad (3.117)$$

Distribution of F

Numerator: We use the following relations:

$$\begin{aligned}
A &= M_1X_2D^{-1}X_2'M_1 \quad \text{is idempotent,}\\
\operatorname{rank}(A) &= \operatorname{tr}(A) = \operatorname{tr}\{(M_1X_2D^{-1})(X_2'M_1)\}\\
&= \operatorname{tr}\{(X_2'M_1)(M_1X_2D^{-1})\} \quad \text{[Theorem A.13 (iv)]}\\
&= \operatorname{tr}(I_{K-s}) = K - s\,,\\
b_2 - r &= D^{-1}X_2'M_1\tilde{\epsilon} \quad \text{[by (3.106)],}\\
\tilde{\epsilon} &= \epsilon + X_2(\beta_2 - r)\\
&\sim N(X_2(\beta_2 - r), \sigma^2 I) \quad \text{[Theorem A.82],}\\
(b_2 - r)'D(b_2 - r) &= \tilde{\epsilon}'A\tilde{\epsilon}\\
&\sim \sigma^2\chi^2_{K-s}(\sigma^{-2}(\beta_2 - r)'D(\beta_2 - r)) && (3.118)\\
&\sim \sigma^2\chi^2_{K-s} \quad \text{under } H_0. && (3.119)
\end{aligned}$$

Denominator: The denominator is equal in both cases; that is

$$(y - Xb)'(y - Xb) = \epsilon'(I - P_X)\epsilon \quad \sim \quad \sigma^2\chi^2_{T-K}\,. \qquad (3.120)$$

Because

$$(I - P_X)X = (I - P_X)(X_1, X_2) = ((I - P_X)X_1, (I - P_X)X_2) = (0, 0)\,, \tag{3.121}$$

we find

$$(I - P_X)M_1 = (I - P_X) \tag{3.122}$$

and

$$(I - P_X)A = (I - P_X)M_1 X_2 D^{-1} X_2' M_1 = 0\,, \tag{3.123}$$

so that the numerator and denominator of F (3.117) are independently distributed [Theorem A.89]. Thus [see also Theorem A.86] the test statistic F is distributed under H_1 as $F_{K-s,T-K}(\sigma^{-2}(\beta_2 - r)' D(\beta_2 - r))$ and as central $F_{K-s,T-K}$ under H_0.

The region of acceptance of H_0 at a level of significance α is then given by

$$0 \leq F \leq F_{K-s,T-K,1-\alpha}\,. \tag{3.124}$$

Accordingly, the critical area of H_0 is given by

$$F > F_{K-s,T-K,1-\alpha}\,. \tag{3.125}$$

3.9 Analysis of Variance

Assuming that

$$\epsilon \sim N(0, \sigma^2 I_T)\ ,$$

it follows from $y = X\beta + \epsilon$,

$$y \sim N(X\beta, \sigma^2 I_T) \tag{3.126}$$

and

$$b = (X'X)^{-1}X'y \sim N[\beta, \sigma^2 (X'X)^{-1}]\ . \tag{3.127}$$

We know that

$$s^2 = \frac{RSS}{T-K}$$

where

$$\begin{aligned} RSS &= (y - \hat{y})'(y - \hat{y}) \\ &= y'My \qquad (3.128) \\ &= y'y - b'X'y \qquad (3.129) \\ M &= I - X(X'X)^{-1}X'\ . \end{aligned}$$

Since $(X'X)^{-1}XM = 0$, so b and s^2 are independently distributed. Noting that M is an idempotent matrix, we see that

$$\frac{RSS}{\sigma^2} \sim \chi^2_{T-K}\left(\frac{\beta' X' M X \beta}{2\sigma^2}\right)$$

i.e., noncentral χ^2 distribution with $(T-K)$ degrees of freedom and noncentrality parameter $\beta' X' M X \beta / 2\sigma^2$, which becomes

$$(T-K)\frac{s^2}{\sigma^2} \sim \chi^2_{T-K}\left(\frac{\beta' X' M X \beta}{2\sigma^2}\right) . \tag{3.130}$$

Further, partitioning the total sum of squares gives

$$\begin{aligned} SYY &= y'y \\ &= b'X'y + (y'y - b'X'y) \\ &= SS_{Reg} + RSS \end{aligned} \tag{3.131}$$

where

$$SS_{Reg} = b'X'y = b'X'Xb = y'X(X'X)^{-1}X'y \tag{3.132}$$

is the sum of squares due to regression,

$$\begin{aligned} RSS &= y'y - b'X'y \\ &= SYY - SS_{Reg} \end{aligned} \tag{3.133}$$

is the sum of squares due to residuals and

$$\frac{SS_{Reg}}{\sigma^2} \sim \chi^2_K\left(\frac{\beta' X' P_X X \beta}{2\sigma^2}\right) , \tag{3.134}$$

$$\frac{SYY}{\sigma^2} \sim \chi^2_T\left(\frac{\beta' X' X \beta}{2\sigma^2}\right) . \tag{3.135}$$

Since $MP_X = 0$, so SS_{Reg} and RSS are independently distributed. The mean square due to regression is

$$MS_{Reg} = \frac{SS_{Reg}}{K}$$

and the mean square due to error is

$$MSE = \frac{RSS}{T-K} .$$

Then,

$$\frac{MS_{Reg}}{MSE} \sim F_{K,T-K}\left(\frac{\beta' X' X \beta}{2\sigma^2}\right) \tag{3.136}$$

which is the noncentral F distribution with $(K, T-K)$ degrees of freedom and noncentrality parameter $\beta' X' X \beta / 2\sigma^2$.

Under $H_0 : \beta_1 = \ldots = \beta_K$,

$$\frac{MS_{Reg}}{MSE} \sim F_{K,T-K} \ . \tag{3.137}$$

The calculation of F-statistic in (3.136) can be summarized in an analysis of variance table.

Source of variation	Sum of squares	*df*	Mean square
Regression on $X_1, \ldots, X_K$	SS_{Reg}	K	SS_{Reg}/K
Residual	RSS	$T-K$	$RSS/(T-K)$
Total	SYY	T	

Note that if the model $y = X\beta + \epsilon$ contains an additional intercept term, then K is replaced by $(K+1)$ in the whole analysis of variance.

3.10 Goodness of Fit

Consider the model

$$\begin{aligned} y &= \mathbf{1}\beta_0 + X\beta_* + \epsilon \\ &= \tilde{X}\beta + \epsilon \ , \end{aligned} \tag{3.138}$$

then β is estimated by

$$b = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_* \end{pmatrix}, \ \hat{\beta}_* = (X'X)^{-1}X'y, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_*'\bar{x} \ . \tag{3.139}$$

For such a model with an intercept term, the goodness of fit of a regression model is measured by the ratio

$$R^2 = \frac{SS_{Reg}}{SYY} \tag{3.140}$$

$$= 1 - \frac{RSS}{SYY} \tag{3.141}$$

where

$$\begin{aligned} RSS &= (y - \tilde{X}b)'(y - \tilde{X}b) \\ &= y'y - b'\tilde{X}'\tilde{X}b \\ &= (y - \mathbf{1}\,\bar{y})'(y - \mathbf{1}\,\bar{y}) - \hat{\beta}_*'(X'X)\hat{\beta}_* + T\bar{y}^2 \ , \end{aligned} \tag{3.142}$$

$$SYY = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2 = \hat{\epsilon}'\hat{\epsilon} \ , \tag{3.143}$$

$$SS_{Reg} = SYY - RSS \ . \tag{3.144}$$

If all observations are located on the hyperplane, we have obviously, $\sum_t (y_t - \hat{y}_t)^2 = 0$ and thus $SYY = SS_{Reg}$. The ratio SS_{Reg}/SYY in (3.140) describes the proportion of variability that is explained by the regression of y on $X_1, \ldots, X_K$ in relation to the total variability of y. The quantity in (3.141) is one minus the proportion of variability that is not covered by the regression.

The R^2 defined in (3.140) is termed as coefficient of determination and is not adjusted for the degrees of freedom. The $\sqrt{R^2}$ in (3.140) is also the multiple correlation coefficient between y and a set of regressors $X_1, \ldots, X_K$, which is shown in Section 3.12. So obviously

$$0 \le R^2 \le 1 \ . \tag{3.145}$$

Clearly, when the model fits the data well then R^2 is close to 1. In the absence of any linear relationship between y and $X_1, \ldots, X_K$, R^2 will be close to 0.

The coefficient of determination in (3.140) and (3.141) is adjusted for the degrees of freedom and is termed as adjusted R-squared. It is defined as

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{RSS/(T-K-1)}{SYY/(T-1)} \\ &= 1 - \frac{T-1}{T-K-1}(1-R^2) \quad \text{(cf. (3.141))} \ . \end{aligned} \tag{3.146}$$

Note that $\bar{R}^2$ is obtained from (3.141) by dividing RSS and SYY by their respective degrees of freedom.

One important point to be noted is that R^2 and $\bar{R}^2$ are defined in a linear or multiple linear model with an intercept term.

When the intercept term is absent, then the unadjusted coefficient of determination in the model $y = X\beta + \epsilon$ can be defined as follows. The square of the product moment correlation between y_t's and $\hat{y}_t$'s is

$$\begin{aligned} R_*^2 &= \frac{\left(\sum_{t=1}^T y_t \hat{y}_t\right)^2}{\left(\sum_{t=1}^T y_t^2\right)\left(\sum_{t=1}^T \hat{y}_t^2\right)} \\ &= \frac{(y'\hat{y})^2}{(y'y)(\hat{y}'\hat{y})} \\ &= \frac{b'X'y}{y'y} \quad (\text{using } \hat{y} = Xb,\ y'\hat{y} = \hat{y}'\hat{y} = y'P_X y) \\ &= \frac{SS_{Reg}}{SYY} \ . \end{aligned}$$

3.11 Checking the Adequacy of Regression Analysis

3.11.1 Univariate Regression

If model

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

is appropriate, the coefficient b_1 should be significantly different from zero. This is equivalent to the fact that X and y are significantly correlated.

Formally, we compare the models (cf. Weisberg, 1985, p. 17)

$$\begin{aligned} H_0\colon y_t &= \beta_0 + \epsilon_t\,, \\ H_1\colon y_t &= \beta_0 + \beta_1 x_t + \epsilon_t\,, \end{aligned}$$

by comparing testing $H_0\colon \beta_1 = 0$ against $H_1\colon \beta_1 \neq 0$.

We assume normality of the errors $\epsilon \sim N(0, \sigma^2 I)$. If we recall (3.104), that is

$$\begin{aligned} D &= x'x - x'\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'x\,, \quad \mathbf{1}' = (1, \ldots, 1) \\ &= \sum x_t^2 - \frac{(\sum x_t)^2}{T} = \sum (x_t - \bar{x})^2 = SXX\,, \end{aligned} \tag{3.147}$$

then the likelihood ratio test statistic (3.117) is given by

$$\begin{aligned} F_{1,T-2} &= \frac{b_1^2 SXX}{s^2} \\ &= \frac{SS_{\text{Reg}}}{RSS} \cdot (T-2) \\ &= \frac{MS_{\text{Reg}}}{s^2}\,. \end{aligned} \tag{3.148}$$

3.11.2 Multiple Regression

If we consider more than two regressors, still under the assumption of normality of the errors, we find the methods of analysis of variance to be most convenient in distinguishing between the two models $y = \mathbf{1}\beta_0 + X\beta_* + \epsilon = \tilde{X}\beta + \epsilon$ and $y = \mathbf{1}\beta_0 + \epsilon$. In the latter model we have $\hat{\beta}_0 = \bar{y}$, and the related residual sum of squares is

$$\sum (y_t - \hat{y}_t)^2 = \sum (y_t - \bar{y})^2 = SYY\,. \tag{3.149}$$

In the former model, $\beta = (\beta_0, \beta_*)'$ will be estimated by $b = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$.

The two components of the parameter vector β in the full model may be estimated by

$$b = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_* \end{pmatrix},\ \hat{\beta}_* = (X'X)^{-1}X'y,\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_*'\bar{x}\,. \tag{3.150}$$

Thus we have

$$RSS = (y - \mathbf{1}\bar{y})'(y - \mathbf{1}\bar{y}) - \hat{\beta}_*'(X'X)\hat{\beta}_* + T\bar{y}^2. \tag{3.151}$$

The proportion of variability explained by regression is (cf. (3.144))

$$SS_{\text{Reg}} = SYY - RSS \tag{3.152}$$

with RSS from (3.151) and SYY from (3.149). Then the ANOVA table is of the form

Source of variation	Sum of squares	*df*	Mean square
Regression on $X_1, \ldots, X_K$	SS_{Reg}	K	SS_{Reg}/K
Residual	RSS	$T - K - 1$	$RSS/(T - K - 1)$
Total	SYY	$T - 1$	

The F-test for

$$H_0\colon \beta_* = 0$$

versus

$$H_1\colon \beta_* \neq 0$$

(*i.e.*, $H_0\colon y = \mathbf{1}\beta_0 + \epsilon$ versus $H_1\colon y = \mathbf{1}\beta_0 + X\beta_* + \epsilon$) is based on the test statistic

$$F_{K,T-K-1} = \frac{SS_{\text{Reg}}/K}{s^2}\,. \tag{3.153}$$

Often, it is of interest to test for significance of single components of β. This type of a problem arises, for example, in stepwise model selection, with respect to the coefficient of determination.

Criteria for Model Choice

Draper and Smith (1998) and Weisberg (1985) have established a variety of criteria to find the right model. We will follow the strategy, proposed by Weisberg.

Ad Hoc Criteria

Denote by $X_1, \ldots, X_K$ all available regressors, and let $\{X_{i1}, \ldots, X_{ip}\}$ be a subset of $p \leq K$ regressors. We denote the respective residual sum of squares by RSS_K and RSS_p. The parameter vectors are

$$\begin{aligned} &\beta \text{ for } X_1, \cdots, X_K\,, \\ &\beta_1 \text{ for } X_{i1}, \cdots, X_{ip}\,, \\ &\beta_2 \text{ for } (X_1, \cdots, X_K) \backslash (X_{i1}, \cdots, X_{ip})\,. \end{aligned}$$

A choice between the two models can be examined by testing H_0: $\beta_2 = 0$. We apply the F-test since the hypotheses are nested:

$$F_{(K-p),T-K} = \frac{(RSS_p - RSS_K)/(K-p)}{RSS_K/(T-K)} \,. \tag{3.154}$$

We prefer the full model against the partial model if H_0: $\beta_2 = 0$ is rejected, that is, if $F > F_{1-\alpha}$ (with degrees of freedom $K-p$ and $T-K$).

Model choice based on an adjusted coefficient of determination

The coefficient of determination (see (3.140) and (3.152))

$$R_p^2 = 1 - \frac{RSS_p}{SYY} \tag{3.155}$$

is inappropriate to compare a model with K and one with $p < K$, because R_p^2 always increases if an additional regressor is incorporated into the model, irrespective of its values. The full model always has the greatest value of R_p^2.

Theorem 3.15 *Let $y = X_1\beta_1 + X_2\beta_2 + \epsilon = X\beta + \epsilon$ be the full model and $y = X_1\beta_1 + \epsilon$ be a submodel. Then we have*

$$R_X^2 - R_{X_1}^2 \geq 0. \tag{3.156}$$

Proof: Let

$$R_X^2 - R_{X_1} = \frac{RSS_{X_1} - RSS_X}{SYY},$$

so that the assertion (3.156) is equivalent to

$$RSS_{X_1} - RSS_X \geq 0.$$

Since

$$\begin{aligned} RSS_X &= (y - Xb)'(y - Xb) \\ &= y'y + b'X'Xb - 2b'X'y \\ &= y'y - b'X'y \end{aligned} \tag{3.157}$$

and, analogously,

$$RSS_{X_1} = y'y - \hat\beta_1' X_1' y \,,$$

where

$$b = (X'X)^{-1}X'y$$

and

$$\hat\beta_1 = (X_1'X_1)^{-1}X_1'y$$

are OLS estimators in the full and in the submodel, we have

$$RSS_{X_1} - RSS_X = b'X'y - \hat\beta_1' X_1' y \,. \tag{3.158}$$

Now with (3.100)–(3.106),

$$\begin{aligned} b'X'y &= (b_1', b_2') \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \\ &= (y' - b_2'X_2')X_1(X_1'X_1)^{-1}X_1'y + b_2'X_2'y \\ &= \hat{\beta}_1'X_1'y + b_2'X_2'M_1y\,. \end{aligned}$$

Thus (3.158) becomes

$$\begin{aligned} RSS_{X_1} - RSS_X &= b_2'X_2'M_1y \\ &= y'M_1X_2D^{-1}X_2'M_1y \geq 0\,, \end{aligned} \qquad (3.159)$$

which proves (3.156).

On the basis of Theorem 3.15 we define the statistic

$$F\text{-change} = \frac{(RSS_{X_1} - RSS_X)/(K-p)}{RSS_X/(T-K)}\,, \qquad (3.160)$$

which is distributed as $F_{K-p,T-K}$ under H_0: "submodel is valid." In model choice procedures, F-change tests for significance of the change of R_p^2 by adding additional $K-p$ variables to the submodel.

In multiple regression, the appropriate adjustment of the ordinary coefficient of determination is provided by the coefficient of determination adjusted by the degrees of freedom of the multiple model:

$$\bar{R}_p^2 = 1 - \left(\frac{T-1}{T-p}\right)(1 - R_p^2)\,. \qquad (3.161)$$

Note: If there is no constant β_0 present in the model, then the numerator is T instead of $T-1$, so that $\bar{R}_p^2$ may possibly take negative values. This cannot occur when using the ordinary R_p^2.

If we consider two models, the smaller of which is supposed to be fully contained in the bigger, and we find the relation

$$\bar{R}_{p+q}^2 < \bar{R}_p^2\,,$$

then the smaller model obviously shows a better goodness of fit.

Further criteria are, for example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mallows's C_p , or criteria based on the residual mean dispersion error $\hat{\sigma}_p^2 = RSS_p/(T-p)$. These are discussed in Section 7.8.

Confidence Intervals

As in bivariate regression, there is a close relation between the region of acceptance of the F-test and confidence intervals for β in the multiple regression model.

Confidence Ellipsoids for the Whole Parameter Vector β

Considering (3.90) and (3.93), we get for $\beta^* = \beta$ a confidence ellipsoid at level $1-\alpha$:

$$\frac{(b-\beta)'X'X(b-\beta)}{(y-Xb)'(y-Xb)} \cdot \frac{T-K}{K} \leq F_{K,T-K,1-\alpha} \,. \tag{3.162}$$

Confidence Ellipsoids for Subvectors of β

From (3.117) we have

$$\frac{(b_2-\beta_2)'D(b_2-\beta_2)}{(y-Xb)'(y-Xb)} \cdot \frac{T-K}{K-s} \leq F_{K-s,T-K,1-\alpha} \tag{3.163}$$

as a $(1-\alpha)$-confidence ellipsoid for β_2.

Further results may be found in Judge, Griffiths, Hill, Lütkepohl and Lee (1985); Goldberger (1964); Pollock (1979); Weisberg (1985); and Kmenta (1971).

3.11.3 A Complex Example

We now want to demonstrate model choice in detail by means of the introduced criteria on the basis of a data set. Consider the following model with $K = 4$ real regressors and $T = 10$ observations:

$$y = \mathbf{1}\beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \epsilon \,.$$

The data set (y, X) is

$$\begin{array}{c} \begin{array}{ccccc} y & X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 18 & 3 & 7 & 20 & -10 \\ 47 & 7 & 13 & 5 & 19 \\ 125 & 10 & 19 & -10 & 100 \\ 40 & 8 & 17 & 4 & 17 \\ 37 & 5 & 11 & 3 & 13 \\ 20 & 4 & 7 & 3 & 10 \\ 24 & 3 & 6 & 10 & 5 \\ 35 & 3 & 7 & 0 & 22 \\ 59 & 9 & 21 & -2 & 35 \\ 50 & 10 & 24 & 0 & 20 \end{pmatrix} \end{array}$$

The sample moments are displayed in the following table.

	Mean	Std. deviation	Variance
X_1	6.200	2.936	8.622
X_2	13.200	6.647	44.178
X_3	3.300	7.846	61.567
X_4	23.100	29.471	868.544
y	45.500	30.924	956.278

The following matrix contains the correlations, the covariances, the one-tailed p-values of the t-tests $t_{T-2} = r\sqrt{(T-2)/(1-r^2)}$ for H_0: "correlation equals zero," and the cross-products $\sum_{t=1}^{T} X_{1t}y_t$. For example, the upper right element has:

$$\begin{aligned} \text{Correlation}(X_1, y) &= 0.740 \\ \text{Covariance}(X_1, y) &= 67.222 \\ p\text{-value} &= 0.007 \\ \text{Cross-product} &= 605.000 \end{aligned}$$

	X_1	X_2	X_3	X_4	y
X_1	1.000	0.971	–0.668	0.652	0.740
	8.622	18.956	–15.400	56.422	67.222
		0.000	0.017	0.021	0.007
	77.600	170.600	–138.600	507.800	605.000
X_2	0.971	1.000	–0.598	0.527	0.628
	8.956	44.178	–31.178	103.000	129.000
	0.000		0.034	0.059	0.026
	170.600	397.600	–280.600	928.800	1161.000
X_3	–0.668	–0.598	1.000	–0.841	–0.780
	–15.400	–31.178	61.567	–194.478	–189.278
	0.017	0.034		0.001	0.004
	–138.600	–280.600	554.100	–1750.300	–1703.500
X_4	0.652	0.527	–0.841	1.000	0.978
	56.422	103.200	–194.478	868.544	890.944
	0.021	0.059	0.001		0.000
	507.800	928.800	–1750.300	7816.900	8018.500
y	0.740	0.628	–0.780	0.978	1.000
	67.222	129.000	–189.278	890.944	956.278
	0.007	0.026	0.004	0.000	
	605.000	1161.000	–1703.500	8018.500	8606.500

We especially recognize that

- X_1 and X_2 have a significant positive correlation ($r = 0.971$),
- X_3 and X_4 have a significant negative correlation ($r = -0.841$),
- all X-variables have a significant correlation with y.

The significance of the correlation between X_1 and X_3 or X_4, and between X_2 and X_3 or X_4 lies between 0.017 and 0.059, which is quite large as well.

We now apply a stepwise procedure for finding the best model.

Step 1 of the Procedure

The stepwise procedure first chooses the variable X_4, since X_4 shows the highest correlation with y (the p-values are X_4: 0.000, X_1: 0.007, X_2: 0.026, X_3: 0.004). The results of this step are listed below.

Multiple R	0.97760		
R^2	0.95571	R^2-change	0.95571
Adjusted R^2	0.95017	F-change	172.61878
Standard error	6.90290	Signif. F-change	0.00000

The ANOVA table is:

	df	Sum of squares	Mean square
Regression	1	8225.29932	8225.2993
Residual	8	381.20068	47.6500

with $F = 172.61878$ (Signif. F: 0.0000). The determination coefficient for the model $y = \mathbf{1}\hat{\beta}_0 + X_4\hat{\beta}_4 + \epsilon$ is

$$R_2^2 = \frac{SS_{\text{Reg}}}{SYY} = \frac{8225.29932}{8225.29932 + 381.20068} = 0.95571\,,$$

and the adjusted determination coefficient is

$$\bar{R}_2^2 = 1 - \left(\frac{10-1}{10-2}\right)(1 - 0.95571) = 0.95017\,.$$

The table of the estimates is as follows

			95% confidence interval	
	$\hat{\beta}$	SE$(\hat{\beta})$	lower	upper
X_4	1.025790	0.078075	0.845748	1.205832
Constant	21.804245	2.831568	15.274644	28.333845

Step 2 of the Procedure

Now the variable X_1 is included. The adjusted determination coefficient increases to $\bar{R}_3^2 = 0.96674$.

Multiple R	0.98698		
R^2	0.97413	R^2-change	0.01842
Adjusted R^2	0.96674	F-change	4.98488
Standard error	5.63975	Signif. F-change	0.06070

The ANOVA table is:

	df	Sum of squares	Mean square
Regression	2	8383.85240	4191.9262
Residual	7	222.64760	31.8068

with $F = 131.79340$ (Signif. F: 0.0000).

Step 3 of the Procedure

Now that X_3 is included, the adjusted determination coefficient increases to $\bar{R}_4^2 = 0.98386$.

Multiple R	0.99461		
R^2	0.98924	R^2-change	0.01511
Adjusted R^2	0.98386	F-change	8.42848
Standard error	3.92825	Signif. F-change	0.02720

The ANOVA table is:

	df	Sum of squares	Mean square
Regression	3	8513.91330	2837.9711
Residual	6	92.58670	15.4311

with $F = 183.91223$ (Signif. F: 0.00000).

The test statistic F-change was calculated as follows:

$$\begin{aligned} F_{1,6} &= \frac{RSS_{(X_4,X_1,1)} - RSS_{(X_4,X_1,X_3,1)}}{RSS_{(X_4,X_1,X_3,1)}/6} \\ &= \frac{222.64760 - 92.58670}{15.4311} \\ &= 8.42848. \end{aligned}$$

The 95% and 99% quantiles of the $F_{1,6}$-distribution are 5.99 and 13.71, respectively. The p-value of F-change is 0.0272 and lies between 1% and 5%. Hence, the increase in determination is significant on the 5% level, but not on the 1% level.

The model choice procedure stops at this point, and the variable X_2 is not taken into consideration. The model chosen is $y = \mathbf{1}\beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$ with the statistical quantities shown below.

			95% confidence interval	
	$\hat{\beta}$	$SE(\hat{\beta})$	lower	upper
X_4	1.079	0.084	0.873	1.285
X_1	2.408	0.615	0.903	3.913
X_3	0.937	0.323	0.147	1.726
Constant	2.554	4.801	–9.192	14.301

The Durbin-Watson test statistic is $d = 3.14$, which exceeds d_u^*. (Table 4.1 displays the values of d_u^* for T=15, 20, 30, ...), hence H_0: $\rho = 0$ cannot be rejected.

Note: The Durbin-Watson test is used for testing the presence of first order autocorrelation in the data and is discussed in Section 4.4.

3.11.4 Graphical Presentation

We now want to display the structure of the (y, X)-matrix by means of the bivariate scatterplots. The plots shown in Figures 3.2 to 3.5 confirm the relation between X_1, X_2 and X_3, X_4, and the X_i and y, but they also show the strong influence of single observations for specific data constellations. This influence is examined more closely with methods of the sensitivity analysis (Chapter 7).

The F-tests assume a normal distribution of the errors or y. This assumption is checked with the Kolmogorov-Smirnov test. The test statistic has a value of 0.77 (p-value .60). Hence, normality is not rejected at the 5% level.

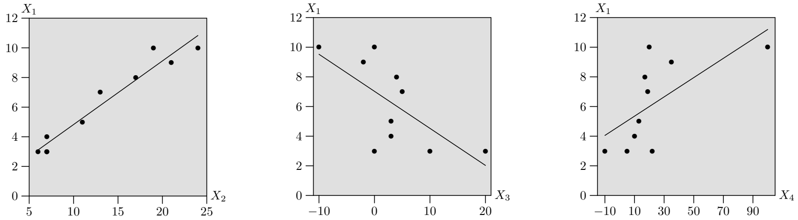

FIGURE 3.2. Scatterplots and regression for X_1 on X_2, X_3 and X_4, respectively

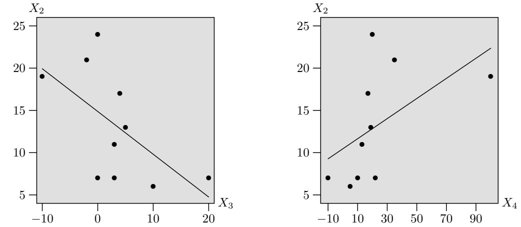

FIGURE 3.3. Scatterplots and regression for X_2 on X_3 and X_4, respectively

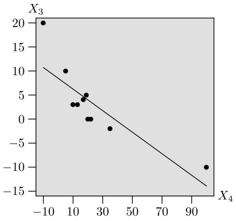

FIGURE 3.4. Scatterplot and regression for X_3 on X_4

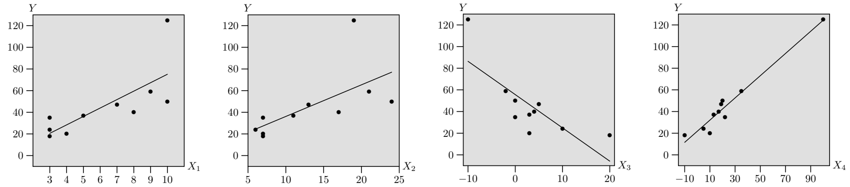

FIGURE 3.5. Scatterplot and regression for y on X_1, X_2, X_3 and X_4, respectively

3.12 Linear Regression with Stochastic Regressors

3.12.1 Regression and Multiple Correlation Coefficient

In many scientific and experimental studies, the regressors $X_1, \dots, X_K$ are often stochastic. In such a case, the multiple correlation coefficient can be related to regression problem because the central idea is to check the strength of dependency between study and explanatory variables. The given $(K+1)$ dimensional random vector $(y, X_1, \dots, X_K)$ is assumed to follow a multivariate distribution with mean vector μ and covariance matrix Σ. Partition $(y, X_1, \dots, X_K)$ into one-dimensional vector y and K-dimensional vector X as (y, X'). Further partition μ and Σ in submatrices accordingly as

$$\mu = \begin{pmatrix} \underset{1\times1}{\mu_y} \\ \\ \underset{K\times1}{\mu_X} \end{pmatrix}, \ \Sigma = \begin{pmatrix} \underset{1\times1}{\sigma_y^2} & \underset{1\times K}{\sigma'_{yX}} \\ \\ \underset{K\times1}{\sigma_{yX}} & \underset{K\times K}{\Sigma_{XX}} \end{pmatrix} . \tag{3.164}$$

Suppose a random sample (y_t, x'_t), $t = 1, \dots, T$ of size T is observed from the $(K+1)$ variate distribution.

Assume that there exists a linear dependency between y and the remaining set X of the variables. Such stochastic dependency can be measured by the correlation between y and a linear transformation $\beta' X$ of $X_1, \dots, X_K$ where β is a nonstochastic K-vector as

$$\text{Corr}(y, \beta' X) = \frac{\beta' \sigma_{yX}}{\sigma_y \sqrt{\beta' \Sigma_{XX} \beta}} . \tag{3.165}$$

To define (3.165) uniquely, find β such that

$$\max_{\beta} \text{Corr}(y, \beta' X) , \tag{3.166}$$

i.e., the correlation between y and given a linear function $\beta' X$ is maximum. Such a solution is called the multiple correlation coefficient between y and $\beta' X$.

Since the coefficient of correlation is invariant under the change of scale and location in y and X, so we apply the restriction $\beta' \Sigma_{XX} \beta = 1$ to have a unique solution. Now the problem (3.166) is restated as

$$\min_{\beta} \left[\beta' \sigma_{yX} - \frac{\lambda}{2} (\beta' \Sigma_{XX} \beta - 1) \right] = \min_{\beta} f(\beta) \ , \ \text{(say)} \tag{3.167}$$

where λ is a Lagrangian multiplier. Partially differentiating (3.167) with respect to β and λ (using Theorem A.91)

$$\frac{\partial}{\partial \beta} f(\beta) \ = \ \sigma_{yX} - \lambda \Sigma_{XX} \beta \tag{3.168}$$

$$\frac{\partial}{\partial \lambda} f(\beta) \ = \ \beta' \Sigma_{XX} \beta - 1 \ . \tag{3.169}$$

Solving

$$\frac{\partial}{\partial\beta} f(\beta) = 0,$$

we get

$$\beta = \frac{1}{\lambda} \Sigma_{XX}^{-1} \sigma_{yX} \ .$$

Using the restriction $\beta' \Sigma_{XX} \beta = 1$, it follows that $\lambda = 1$. Thus the unique solution is

$$\beta = \Sigma_{XX}^{-1} \sigma_{yX} \ . \tag{3.170}$$

In general without imposing the normalization rule $\beta' \Sigma_{XX} \beta = 1$, we get the multiple correlation coefficient between y and X as

$$\rho_{y.X} = \frac{\sqrt{\sigma_{yX}' \Sigma_{XX}^{-1} \sigma_{yX}}}{\sigma_y} = \rho_{1.2,3,\ldots,K+1} \tag{3.171}$$

and $0 \le \rho_{y.X} \le 1$.

If y is exactly linearly dependent on $X_1, \ldots, X_K$, *i.e.*, $y = \beta' X$ holds, then

$$\mathrm{Corr}(y, \beta' X) = \frac{\beta' \Sigma_{XX} \beta}{\beta' \Sigma_{XX} \beta} = 1 \tag{3.172}$$

and $\beta = \Sigma_{XX}^{-1} \sigma_{yX}$ is called as the vector of regression coefficients obtained by regressing y on $X_1, \ldots, X_K$.

It may be noted that when a set of variables $Y = (y_1, \ldots, y_P)$ is regressed on another set of variables $X_1, \ldots, X_K$, then the set of parameters $\beta_i = \Sigma_{XX}^{-1} \sigma_{y_i X}$, $(i = 1, \ldots, P)$ of all the regression coefficients expressed as $(P \times K)$ matrix

$$B = \begin{pmatrix} \beta_1' \\ \beta_2' \\ \vdots \\ \beta_P' \end{pmatrix} = \Sigma_{YX} \Sigma_{XX}^{-1} \tag{3.173}$$

where Σ_{YX} results from the partition

$$\Sigma = \begin{pmatrix} \underset{P \times P}{\Sigma_{YY}} & \underset{P \times K}{\Sigma_{YX}} \\ \underset{K \times P}{\Sigma_{XY}} & \underset{K \times K}{\Sigma_{XX}} \end{pmatrix} \tag{3.174}$$

is the covariance matrix of $(y_1, \ldots, y_P, X_1, \ldots, X_K)$.

3.12.2 Heterogenous Linear Estimation without Normality

Let α be any non-stochastic K-vector, $\tilde{X} = X - \mu_X$ and $\tilde{y} = y - \mu_y$ be the centered variables measured around their means. The mean squared error of an estimate $\alpha'\tilde{X}$ of the variable $\tilde{y}$ is

$$\begin{aligned} & \mathrm{E}(\tilde{y} - \alpha'\tilde{X})'(\tilde{y} - \alpha'\tilde{X}) \\ = \quad & \mathrm{E}_{y,X}\left[\tilde{y} - \mathrm{E}_y(\tilde{y}|\tilde{X}) + \mathrm{E}_y(\tilde{y}|\tilde{X}) - \alpha'X\right]' \\ & \times \left[\tilde{y} - \mathrm{E}_y(\tilde{y}|\tilde{X}) + \mathrm{E}_y(\tilde{y}|\tilde{X}) - \alpha'X\right] \\ = \quad & \mathrm{E}_X\left[\mathrm{E}_y\left\{\tilde{y} - \mathrm{E}_y(\tilde{y}|\tilde{X})\right\}'\left\{\tilde{y} - \mathrm{E}_y(\tilde{y}|\tilde{X})\right\}|\tilde{X}\right] \\ & + \; \mathrm{E}_X\left[\left\{\mathrm{E}_y(\tilde{y}|\tilde{X}) - \alpha'\tilde{X}\right\}'\left\{\mathrm{E}_y(\tilde{y}|\tilde{X}) - \alpha'\tilde{X}\right\}|\tilde{X}\right]. \end{aligned} \tag{3.175}$$

The mean squared error (3.175) is minimum with respect to α iff

$$\mathrm{E}_y(\tilde{y}|\tilde{X}) = \alpha'\tilde{X} \; ,$$

i.e., iff

$$\mathrm{E}(y|X) = \mu_y + \alpha'(X - \mu_X) \tag{3.176}$$

holds.

On the other hand, the minimizing α is found from

$$\begin{aligned} \mathrm{E}(\tilde{y} - \alpha'\tilde{X})'(\tilde{y} - \alpha'\tilde{X}) \quad = \quad & (\alpha - \Sigma_{XX}^{-1}\sigma_{yX})'\Sigma_{XX}(\alpha - \Sigma_{XX}^{-1}\sigma_{yX}) \\ & + \sigma_y^2 - \sigma_{yX}'\Sigma_{XX}^{-1}\sigma_{yX} \end{aligned} \tag{3.177}$$

as

$$\hat{\alpha} = \Sigma_{XX}^{-1}\sigma_{yX} \; . \tag{3.178}$$

Thus

$$\begin{aligned} \hat{y} \quad & = \quad \mu_y + \hat{\alpha}'(X - \mu_X) \\ & = \quad \mu_y + \sigma_{yX}'\Sigma_{XX}^{-1}(X - \mu_X) \end{aligned}$$

can be interpreted as the best linear estimate of $\mathrm{E}(y|X)$ in the class of heterogeneous linear estimators $\{\mu_y + \alpha'(X - \mu_X)\}$, whereas

$$\min_{\alpha} \mathrm{E}\left[\{y - \mu_y - \alpha'(X - \mu_X)\}'\{y - \mu_y - \alpha'(X - \mu_X)\}\right]$$

is obtained for $\hat{\alpha} = \Sigma_{XX}^{-1}\sigma_{yX}$ as in (3.178). This optimality is not dependent of the assumption of normal distribution, see Srivastava and Khatri (1979) for more details.

3.12.3 Heterogeneous Linear Estimation under Normality

Continuing further, now we assume $(y, X') \sim N_{K+1}(\mu, \Sigma)$ where μ and Σ are given by (3.164). Let $f(y, X)$ denote the joint density $N_{K+1}(\mu, \Sigma)$ and $h(X)$ denote the marginal density of $(X_1, \ldots, X_K)$ which is $N_K(\mu_X, \Sigma_{XX})$. The conditional density of y given X is

$$\begin{aligned} g(y|X) &= \frac{f(y,X)}{h(X)} \\ &= \frac{1}{\sqrt{2\pi d^2}} \exp\left[-\frac{1}{2d^2} q_y' q_y\right] \end{aligned}$$

where

$$\begin{aligned} q_y &= y - \mu_y - \sigma_{yX}' \Sigma_{XX}^{-1}(x - \mu_X) && (3.179) \\ d^2 &= \sigma_y^2 - \sigma_{yX}' \Sigma_{XX}^{-1} \sigma_{yX} \ . && (3.180) \end{aligned}$$

Thus $y|X \sim N_1(\mu_y + \sigma_{yX}' \Sigma_{XX}^{-1}(X - \mu_X), d^2)$ which is same as $N_1(\mathrm{E}(y|X), d^2)$.

Define the residual vector

$$\begin{aligned} e_{y.X} &= y - \mathrm{E}(y|X) \\ &= (y - \mu_y) - \sigma_{yX}' \Sigma_{XX}^{-1}(X - \mu_X) \ , && (3.181) \end{aligned}$$

then $e_{y.X}$ represents the difference of vector y and its predicted value from the linear relationship of the conditional mean vector given X. We have

$$\begin{aligned} \mathrm{E}\left[(y - \mu_y) e_{y.X}'\right] &= \mathrm{E}\left[y e_{y.X}'\right] \\ &= \sigma_y^2 - \sigma_{yX}' \Sigma_{XX}^{-1} \sigma_{yX} \\ &= d^2 \\ &= \mathrm{var}(y|X) && (3.182) \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}\left[(X - \mu_X) e_{y.X}'\right] &= \sigma_{yX} - \Sigma_{XX} \Sigma_{XX}^{-1} \sigma_{yX} \\ &= 0 \ . && (3.183) \end{aligned}$$

Thus nonstochastic variables X and residual $e_{y.X}$ are independently distributed.

The mean and variance of the residual are

$$\begin{aligned} \mathrm{E}(e_{y.X}) &= \mathrm{E}(y-\mu_y) - \sigma'_{yX}\Sigma_{XX}^{-1}\,\mathrm{E}(X-\mu_X) \\ &= 0 \qquad (3.184) \\ \mathrm{var}(e_{y.X}) &= \mathrm{E}(y-\mu_y)'(y-\mu_y) + \sigma'_{yX}\Sigma_{XX}^{-1}\Sigma_{XX}\Sigma_{XX}^{-1}\sigma_{yX} \\ &\quad -2\sigma'_{yX}\Sigma_{XX}^{-1}\sigma_{yX} \\ &= \sigma_y^2 - \sigma'_{yX}\Sigma_{XX}^{-1}\sigma_{yX} \\ &= d^2 \\ &= \mathrm{var}(y|X)\ . \qquad (3.185) \end{aligned}$$

Note that (3.181) can be rewritten as

$$y = \mathrm{E}(y|X) + e_{y.X}\ , \qquad (3.186)$$

i.e., $\mathrm{E}(y|X)$ and $e_{y.X}$ determine y linearly. This result provides alternative interpretations to the similarities between the regression coefficient vector β and the conditional expectation $\mathrm{E}(y|X)$.

So the problem of estimation of regression coefficients in a linear regression model with stochastic regressors can be reformulated and solved similarly as in the case of non-stochastic regressors.

Theorem 3.16 *Let (y_t, x'_t), $t = 1, \dots, T$ be an independent sample from*

$$N_{K+1}\left(\begin{pmatrix} \mu_y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma'_{yX} \\ \sigma_{yX} & \Sigma_{XX} \end{pmatrix}\right) \qquad \text{(cf. (3.164))}\ .$$

The conditional distribution of $(y_t|x_t)$ is

$$N_1(\beta_0 + x'_t\beta, d^2)$$

with

$$\begin{aligned} \beta_0 &= \mu_y - \mu'_X\beta \\ \beta &= \Sigma_{XX}^{-1}\sigma_{yX} \\ d^2 &= \sigma_y^2 - \sigma'_{yX}\Sigma_{XX}^{-1}\sigma_{yX} \\ &= \sigma_y^2(1-\rho^2_{1.2,3,\dots,K+1})\ . \end{aligned}$$

The regression function of interest is

$$\mathrm{E}(y_t|x_t) = \beta_0 + x'_t\beta\ . \qquad (3.187)$$

Define the sample mean vector and sample covariance matrix as

$$\begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix} \text{ and } S = \begin{pmatrix} s_y^2 & s'_{yX} \\ s_{yX} & S_{XX} \end{pmatrix}$$

respectively where $S_{XX} = \sum_t x_t x'_t - T\bar{x}\bar{x}'$, $s_{yX} = \sum_t \sum_t x_t y_t - T\bar{x}\bar{y}$ and $s_y^2 = \sum_t y_t^2 - T\bar{y}^2$.

Then the maximum likelihood estimators of β, β_0 and $\sigma^2 = d^2$ are

$$\begin{aligned} \hat\beta &= S_{XX}^{-1} s_{yX} && (3.188) \\ \hat\beta_0 &= \bar y - \bar x' \hat\beta && (3.189) \\ \hat\sigma^2 &= \hat{d^2} = \frac{1}{T}(s_y^2 - s'_{yX} S_{XX}^{-1} s_{yX}) \ , && (3.190) \end{aligned}$$

respectively.

For further references, see Morrison (1967, Chapter 3) and Fomby, Hill and Johnson (1984, p. 71).

An interesting similarity is as follows, see, (Dhrymes (1974, p. 23).

Let the regression model be $y = X\beta + \epsilon$ where $X_1, \ldots, X_K$ are stochastic and independently distributed of $\epsilon \sim N(0, \sigma^2 I)$, *i.e.*, $\mathrm{E}(\epsilon' X) = 0$. Assume that y and X are measured from their respective sample means. The least squares estimate of β is

$$\hat\beta = (X'X)^{-1} X'y \tag{3.191}$$

which can be written as

$$\hat\beta = \left(\frac{X'X}{T}\right)^{-1} \left(\frac{X'y}{T}\right) \tag{3.192}$$

where $(X'X/T)^{-1}$ and $(X'y/T)$ are the sample analogues of Σ_{XX}^{-1} and σ_{yX}, respectively. The sample analog of (3.183) is

$$(y - X\hat\beta)'X = y'X - \hat\beta' X'X = y'X - y'X = 0 \ . \tag{3.193}$$

The coefficient of determination

$$\begin{aligned} R^2 &= \frac{y'y - (y - X\hat\beta)'(y - X\hat\beta)}{y'y} \\ &= \frac{\hat\beta' X' y}{y'y} \quad \text{(cf. (3.193))} \\ &= \frac{y'X(X'X)^{-1}X'y}{y'y} \end{aligned}$$

which is a sample analog of coefficient of maximum correlation $\rho^2_{1.2,3,\ldots,K+1}$ (cf. (3.171)). The maximum likelihood estimators $\hat\beta$, $\hat\beta_0$ and $\hat\sigma^2$ (cf. (3.188) – (3.190)) coincide with the solution of least squares estimation in the model $y = X\beta + \epsilon$ with stochastic regressors when minimization is done conditional on X. Under some general conditions, the maximum likelihood estimates of parameters from a regular distribution are consistent, asymptotically normal and asymptotically efficient. Based on corollary to the Cramér-Rao Theorem Theil (1971, p. 395) and on investigations of Dhrymes (1974, Lemma 14, p. 122-123), Fomby et al. (1984, pp. 56) have concluded that this holds for the maximum likelihood estimates $\hat\beta$, $\hat\beta_0$ and $\hat\sigma^2$. Further, Fomby et al. (1984, pp. 72) state: "In summary, the inferential framework of the classical normal linear regression with fixed X is equally applicable to the multivariate normal regression model". This also concerns the usual F-tests as well as the confidence interval estimation.

3.13 The Canonical Form

To simplify considerations about the linear model—especially when X is deficient in rank, leading to singularity of $X'X$—the so-called canonical form is frequently used (Rao, 1973a, p. 43).

The spectral decomposition (Theorem A.30) of the symmetric matrix $X'X$ is

$$X'X = P\Lambda P' \tag{3.194}$$

with $P = (p_1, \ldots, p_K)$ and $PP' = I$. Model (3.58) can then be written as

$$\begin{aligned} y &= XPP'\beta + \epsilon \\ &= \tilde{X}\tilde{\beta} + \epsilon \end{aligned} \tag{3.195}$$

with $\tilde{X} = XP$, $\tilde{\beta} = P'\beta$, and $\tilde{X}'\tilde{X} = P'X'XP = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$, so that the column vectors of $\tilde{X}$ are orthogonal. The elements of $\tilde{\beta}$ are called regression parameters of the principal components.

Let $\hat{\beta} = Cy$ be a linear estimator of β with the MDE matrix $M(\hat{\beta}, \beta)$. In the transformed model we obtain for the linear estimator $P'\hat{\beta} = P'Cy$ of the parameter $\tilde{\beta} = P'\beta$

$$\begin{aligned} M(P'\hat{\beta}, \tilde{\beta}) &= \mathrm{E}(P'\hat{\beta} - P'\beta)(P'\hat{\beta} - P'\beta)' \\ &= P'M(\hat{\beta}, \beta)P. \end{aligned} \tag{3.196}$$

Hence, relations between two estimates remain unchanged. For the scalar MDE (cf. Chapter 5) we have

$$\operatorname{tr}\{M(P'\hat{\beta}, \tilde{\beta})\} = \operatorname{tr}\{M(\hat{\beta}, \beta)\}, \tag{3.197}$$

so that the scalar MDE is independent of the parametrization (3.195).

For the covariance matrix of the OLS estimate b of β in the original model, we have

$$\mathrm{V}(b) = \sigma^2 (X'X)^{-1} = \sigma^2 \sum \lambda_i^{-1} p_i p_i'. \tag{3.198}$$

The OLS estimate b^* of $\tilde{\beta}$ in the model (3.195) is

$$\begin{aligned} b^* &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y \\ &= \Lambda^{-1}\tilde{X}'y \end{aligned} \tag{3.199}$$

with the covariance matrix

$$\mathrm{V}(b^*) = \sigma^2 \Lambda^{-1}. \tag{3.200}$$

Hence the components of b^* are uncorrelated and have the variances $\mathrm{var}(b_i^*) = \sigma^2\lambda_i^{-1}$. If $\lambda_i > \lambda_j$, then $\tilde{\beta}_i$ is estimated more precisely than $\tilde{\beta}_j$:

$$\frac{\mathrm{var}(b_i^*)}{\mathrm{var}(b_j^*)} = \frac{\lambda_j}{\lambda_i} < 1 \,. \tag{3.201}$$

The geometry of the reparameterized model (3.195) is examined extensively in Fomby et al. (1984, pp. 289–293). Further remarks can be found in Vinod and Ullah (1981, pp. 5–8). In the case of problems concerning multicollinearity, reparametrization leads to a clear representation of dependence on the eigenvalues λ_i of $X'X$. Exact or strict multicollinearity means $|X'X| = 0$ in the original model and $|\tilde{X}'\tilde{X}| = |\Lambda| = 0$ in the reparameterized model, so that at least one eigenvalue is equal to zero. For weak multicollinearity in the sense of $|\tilde{X}'\tilde{X}| \approx 0$, the smallest eigenvalue or the so-called

$$\text{condition number} \quad k = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{\frac{1}{2}} \tag{3.202}$$

is used for diagnostics (cf. Weisberg, 1985, p. 200; Chatterjee and Hadi, 1988, pp. 157–178).

Belsley, Kuh and Welsch (1980, Chapter 3) give a detailed discussion about the usefulness of these and other measures for assessing weak multicollinearity.

3.14 Identification and Quantification of Multicollinearity

In this section, we want to introduce more algebraically oriented methods: principal components regression, ridge estimation, and shrinkage estimators which are used to solve the problem of multicollinearity. Other methods using exact linear restrictions and procedures with auxiliary information are considered in Chapter 5.

The readers may note that when X is rank deficient (which we define as the problem of multicollinearity), then $X_1, \ldots, X_K$ are not independent. Such violation increases the variance of least squares estimators depending on the degree of linear relationship.

3.14.1 Principal Components Regression

The starting point of this procedure is the reparameterized model (3.195)

$$y = XPP'\beta + \epsilon = \tilde{X}\tilde{\beta} + \epsilon \,.$$

Let the columns of the orthogonal matrix $P = (p_1, \ldots, p_K)$ of the eigenvectors of $X'X$ be numbered according to the magnitude of the eigenvalues

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K$. Then $\tilde{x}_i = X p_i$ is the i^{th} principal component and we get

$$\tilde{x}_i' \tilde{x}_i = p_i' X' X p_i = \lambda_i \,. \tag{3.203}$$

We now assume exact multicollinearity. Hence $\operatorname{rank}(X) = K - J$ with $J \geq 1$. We get (A.31 (vii))

$$\lambda_{K-J+1} = \cdots = \lambda_K = 0 \,. \tag{3.204}$$

According to the subdivision of the eigenvalues into the groups $\lambda_1 \geq \cdots \geq \lambda_{K-J} > 0$ and the group (3.204), we define the subdivision

$$P = (P_1, P_2)\,, \quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{X} = (\tilde{X}_1, \tilde{X}_2) = (XP_1, XP_2)\,,$$

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} P_1'\beta \\ P_2'\beta \end{pmatrix}$$

with $\tilde{X}_2 = 0$ according to (3.203). We now obtain

$$\begin{aligned} y &= \tilde{X}_1\tilde{\beta}_1 + \tilde{X}_2\tilde{\beta}_2 + \epsilon & (3.205)\\ &= \tilde{X}_1\tilde{\beta}_1 + \epsilon. & (3.206) \end{aligned}$$

The OLS estimate of the $(K-J)$-vector $\tilde{\beta}_1$ is $b_1 = (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1' y$. The OLS estimate of the full vector $\tilde{\beta}$ is

$$\begin{aligned} \begin{pmatrix} b_1 \\ 0 \end{pmatrix} &= (X'X)^{-}X'y \\ &= (P\Lambda^{-}P')X'y, & (3.207) \end{aligned}$$

with Theorem A.63

$$\Lambda^{-} = \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix} \tag{3.208}$$

being a g-inverse of Λ.

Remark: The handling of exact multicollinearity by means of principal components regression corresponds to the transition from the model (3.205) to the reduced model (3.206) by putting $\tilde{X}_2 = 0$. This transition can be equivalently achieved by putting $\tilde{\beta}_2 = 0$ and hence by a linear restriction

$$0 = (0, I) \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix}.$$

The estimate b_1 can hence be represented as a restricted OLS estimate (cf. Section 5.2).

A cautionary note on PCR. In practice, zero eigenvalues can be distinguished only by the small magnitudes of the observed eigenvalues. Then, one may be tempted to omit all the principal components with the corresponding eigenvalues below a certain threshold value. But then, there is a possibility that a principal component with a small eigenvalue is a good predictor of the response variable and its omission may decrease the efficiency of prediction drastically.

3.14.2 Ridge Estimation

In case of rank$(X) = K$, the OLS estimate has the minimum-variance property in the class of all unbiased, linear, homogeneous estimators. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$ denote the eigenvalues of S. Then we have for the scalar MDE of b

$$\operatorname{tr}\{M(b,\beta)\} = \operatorname{tr}\{\mathrm{V}(b)\} = \sigma^2 \sum_{i=1}^{K} \lambda_i^{-1}. \tag{3.209}$$

In the case of weak multicollinearity, at least one eigenvalue λ_i is relatively small, so that $\operatorname{tr}\{\mathrm{V}(b)\}$ and the variances of all components b_j of $b = (b_1, \ldots, b_K)'$ are large:

$$\begin{aligned} b_j &= e_j' b\,, \\ \operatorname{var}(b_j) &= e_j' \,\mathrm{V}(b) e_j, \quad \text{and, hence,} \\ \operatorname{var}(b_j) &= \sigma^2 \sum_{i=1}^{K} \lambda_i^{-1} e_j' p_i p_i' e_j \\ &= \sigma^2 \sum_{i=1}^{K} \lambda_i^{-1} p_{ij}^2 \qquad (3.210) \end{aligned}$$

with the j^{th} unit vector e_j and the i^{th} eigenvector $p_i' = (p_{i1}, \ldots, p_{ij}, \ldots, p_{iK})$.

The scalar MDE

$$\operatorname{tr}\{M(b,\beta)\} = \mathrm{E}(b-\beta)'(b-\beta)$$

can be interpreted as the mean Euclidean distance between the vectors b and β, hence multicollinearity means a global unfavorable distance to the real parameter vector. Hoerl and Kennard (1970) used this interpretation as a basis for the definition of the ridge estimate

$$b(k) = (X'X + kI)^{-1} X'y, \tag{3.211}$$

with $k \geq 0$, the nonstochastic quantity, being the control parameter. Of course, $b(0) = b$ is the ordinary LS estimate.

Using the abbreviation

$$G_k = (X'X + kI)^{-1}, \tag{3.212}$$

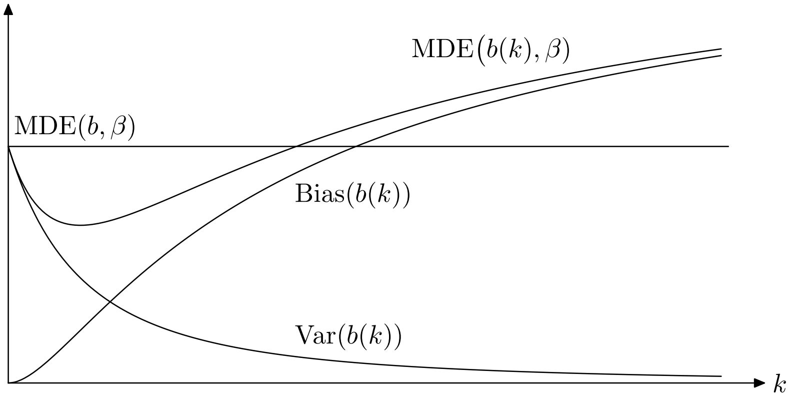

FIGURE 3.6. Scalar MDE function for $b = (X'X)^{-1}X'y$ and $b(k) = G_kX'y$ in dependence on k for $K = 1$

Bias$(b(k),\beta)$ and V$(b(k))$ can be expressed as follows:

$$\begin{aligned} \mathrm{E}(b(k)) &= G_kX'X\beta = \beta - kG_k\beta\,, & (3.213)\\ \mathrm{Bias}(b(k),\beta) &= -kG_k\beta\,, & (3.214)\\ \mathrm{V}(b(k)) &= \sigma^2 G_kX'XG_k\,. & (3.215)\end{aligned}$$

Hence the MDE matrix is

$$M(b(k),\beta) = G_k(\sigma^2X'X + k^2\beta\beta')G_k \tag{3.216}$$

and using $X'X = P\Lambda P'$, we get

$$\mathrm{tr}\{M(b(k),\beta)\} = \sum_{i=1}^{K} \frac{\sigma^2\lambda_i + k^2\beta_i^2}{(\lambda_i + k)^2} \tag{3.217}$$

(cf. Goldstein and Smith, 1974).

Proof: Let $X'X = P\Lambda P'$ be the spectral decomposition of $X'X$. We then have (Theorems A.30, A.31)

$$\begin{aligned} X'X + kI = G_k^{-1} &= P(\Lambda + kI)P'\,,\\ G_k &= P(\Lambda + kI)^{-1}P'\,,\end{aligned}$$

and in general

$$\mathrm{tr}\{\mathrm{diag}(l_1,\cdots,l_k)\beta\beta'\,\mathrm{diag}(l_1,\cdots,l_k)\} = \sum \beta_i^2\, l_i^2\,.$$

With $l_i = (\lambda_i + k)^{-1}$, we obtain relation (3.217).

The scalar MDE of $b(k)$ for fixed σ^2 and a fixed vector β is a function of the ridge parameter k, which starts at $\sum \sigma^2/\lambda_i = \mathrm{tr}\{\mathrm{V}(b)\}$ for $k = 0$, takes its minimum for $k = k_{\mathrm{opt}}$ and then it increases monotonically, provided that $k_{\mathrm{opt}} < \infty$ (cf. Figure 3.6).

We now transform $M(b,\beta)=M(b)=\sigma^2(X'X)^{-1}$ as follows:

$$\begin{aligned} M(b) &= \sigma^2 G_k(G_k^{-1}(X'X)^{-1}G_k^{-1})G_k \\ &= \sigma^2 G_k(X'X+k^2(X'X)^{-1}+2kI)G_k\,. \end{aligned} \tag{3.218}$$

From Definition 3.10 we obtain the interval $0<k<k^*$ in which the ridge estimator is MDE-I-superior to the OLS b, according to

$$\begin{aligned} \Delta(b,b(k)) &= M(b)-M(b(k),\beta) \\ &= kG_k[\sigma^2(2I+k(X'X)^{-1})-k\beta\beta']G_k. \end{aligned} \tag{3.219}$$

Since $G_k>0$, we have $\Delta(b,b(k))\geq 0$ if and only if

$$\sigma^2(2I+k(X'X)^{-1})-k\beta\beta'\geq 0\,, \tag{3.220}$$

or if the following holds (Theorem A.57):

$$\sigma^{-2}k\beta'(2I+k(X'X)^{-1})^{-1}\beta\leq 1\,. \tag{3.221}$$

As a sufficient condition for (3.220), independent of the model matrix X, we obtain

$$2\sigma^2 I-k\beta\beta'\geq 0 \tag{3.222}$$

or—according to Theorem A.57—equivalently,

$$k\leq\frac{2\sigma^2}{\beta'\beta}\,. \tag{3.223}$$

The range of k, which ensures the MDE-I superiority of $b(k)$ compared to b, is dependent on $\sigma^{-1}\beta$ and hence unknown.

If auxiliary information about the length (norm) of β is available in the form

$$\beta'\beta\leq r^2\,, \tag{3.224}$$

then

$$k\leq\frac{2\sigma^2}{r^2} \tag{3.225}$$

is sufficient for (3.223) to be valid. Hence possible values for k, in which $b(k)$ is better than b, can be found by estimation of σ^2 or by specification of a lower limit or by a combined a priori estimation $\sigma^{-2}\beta'\beta\leq\tilde{r}^2$.

Swamy, Mehta and Rappoport (1978) and Swamy and Mehta (1977) investigated the following problem:

$$\min_\beta\{\sigma^{-2}(y-X\beta)'(y-X\beta)|\beta'\beta\leq r^2\}\,.$$

The solution of this problem

$$\hat\beta(\mu)=(X'X+\sigma^2\mu I)^{-1}X'y\,, \tag{3.226}$$

is once again a ridge estimate and $\hat\beta'(\mu)\hat\beta(\mu) = r^2$ is fulfilled. Replacing σ^2 by the estimate s^2 provides a practical solution for the estimator (3.226) but its properties can be calculated only approximately.

Hoerl and Kennard (1970) derived the ridge estimator by the following reasoning. Let $\hat\beta$ be any estimator and $b = (X'X)^{-1}X'y$ the OLS. Then the error sum of squares estimated with $\hat\beta$ can be expressed, according to the property of optimality of b, as

$$\begin{aligned} S(\hat\beta) &= (y - X\hat\beta)'(y - X\hat\beta) \\ &= (y - Xb)'(y - Xb) + (b - \hat\beta)'X'X(b - \hat\beta) \\ &= S(b) + \Phi(\hat\beta)\,, \end{aligned} \tag{3.227}$$

since the term

$$\begin{aligned} 2(y - Xb)'X(b - \hat\beta) &= 2y'(I - X(X'X)^{-1}X')X(b - \hat\beta) \\ &= 2MX(b - \hat\beta) = 0 \end{aligned}$$

since $MX = 0$.

Let $\Phi_0 > 0$ be a fixed given value for the error sum of squares. Then a set $\{\hat\beta\}$ of estimates exists that fulfill the condition $S(\hat\beta) = S(b) + \Phi_0$. In this set $\{\hat\beta\}$ we look for the estimate $\hat\beta$ with minimal length:

$$\min_{\hat\beta} \left\{\hat\beta'\hat\beta + \frac{1}{k}[(b - \hat\beta)'X'X(b - \hat\beta) - \Phi_0]\right\}, \tag{3.228}$$

where $1/k$ is a Lagrangian multiplier. Differentiation of this function with respect to $\hat\beta$ and $1/k$ leads to the normal equations

$$\hat\beta + \frac{1}{k}(X'X)(\hat\beta - b) = 0\,,$$

and hence

$$\begin{aligned} \hat\beta &= (X'X + kI)^{-1}(X'X)b \\ &= G_k X'y\,, \end{aligned} \tag{3.229}$$

as well as

$$\Phi_0 = (b - \hat\beta)'X'X(b - \hat\beta)\,. \tag{3.230}$$

Hence, the solution of the problem (3.228) is the ridge estimator $\hat\beta = b(k)$ (3.229). The ridge parameter k is to be determined iteratively so that (3.230) is fulfilled.

For further representations about ridge regression see Vinod and Ullah (1981) and Trenkler and Trenkler (1983).

3.14.3 Shrinkage Estimates

Another class of biased estimators, which was very popular in research during the 1970s, is defined by the so-called shrinkage estimator (Mayer and Wilke, 1973):

$$\hat{\beta}(\rho) = (1+\rho)^{-1} b\,, \quad \rho \geq 0 \quad (\rho \text{ known})\,, \tag{3.231}$$

which "shrinks" the OLS estimate:

$$\begin{aligned} \mathrm{E}\left(\hat{\beta}(\rho)\right) &= (1+\rho)^{-1}\beta, \\ \mathrm{Bias}\left(\hat{\beta}(\rho),\beta\right) &= -\rho(1+\rho)^{-1}\beta\,, \\ \mathrm{V}\left(\hat{\beta}(\rho)\right) &= \sigma^2(1+\rho)^{-2}(X'X)^{-1}\,, \end{aligned}$$

and

$$M\big(\hat{\beta}(\rho),\beta\big) = (1+\rho)^{-2}(\mathrm{V}(b) + \rho^2\beta\beta')\,. \tag{3.232}$$

The MDE-I comparison with the OLS leads to

$$\Delta(b,\hat{\beta}(\rho)) = (1+\rho)^{-2}\rho\sigma^{-2}\big[(\rho+2)(X'X)^{-1} - \sigma^{-2}\rho\beta\beta'\big] \geq 0$$

if and only if (Theorem A.57)

$$\frac{\sigma^{-2}\rho}{(\rho+2)}\beta'X'X\beta \leq 1\,.$$

Then

$$\sigma^{-2}\beta'X'X\beta \leq 1 \tag{3.233}$$

is a sufficient condition for the MDE-I superiority of $\hat{\beta}(\rho)$ compared to b.

This form of restriction will be used as auxiliary information for the derivation of minimax-linear estimates in Section 3.17.

Note: Results about the shrinkage estimator in the canonical model can be found in Farebrother (1978).

Stein-Rule Shrinkage Estimators

The family of Stein-rule estimators are shrinkage estimators which shrink all the regression coefficients towards zero. The Stein-rule estimator improves on the OLSE under quadratic risk in the context of $y = X\beta + \epsilon$, $\epsilon \sim N(0,\sigma^2 I)$, see Stein (1956). The Stein-rule estimator can be written in the following form:

$$\hat{\beta}_S = \left[1 - \frac{c\sigma^2}{b'X'Xb}\right] b \tag{3.234}$$

where $b = (X'X)^{-1}X'y$ is the OLSE of β, σ^2 is known and $c > 0$ is a non-stochastic characterizing scalar. The Stein-rule estimator is nonlinear in y and biased for β but dominates OLSE under quadratic risk when

$$0 < c < 2(K-2)\ . \tag{3.235}$$

The quadratic risk of (3.234) is minimum when $c = (K-2)$ and the optimum family of Stein-rule estimators is proposed by James and Stein (1961) as

$$\hat{\beta}_{JS} = \left[1 - \frac{(K-2)\sigma^2}{b'X'Xb}\right] b \ ; \ K \geq 3 \tag{3.236}$$

which is called as James-Stein estimator.

When σ^2 in (3.234) and (3.236) is unknown, then σ^2 can be substituted by its estimate

$$s^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{T-p} = \frac{(y-Xb)'(y-Xb)}{T-p} \ .$$

There has been tremendous development in the area of Stein-type estimation. More recently Ohtani (2000) and Saleh (2006) compile many of the developments in different directions.

3.14.4 Partial Least Squares

Univariate partial least squares is a particular method of analysis in models with possibly more explanatory variables than samples. In spectroscopy one aim may be to predict a chemical composition from spectra of some material. If all wavelengths are considered as explanatory variables, then traditional stepwise OLS procedure soon runs into collinearity problems caused by the number of explanatory variables and their interrelationships (cf. Helland, 1988).

The aim of partial least squares is to predict the response by a model that is based on linear transformations of the explanatory variables. Partial least squares (PLS) is a method of constructing regression models of type

$$\hat{y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p \,, \tag{3.237}$$

where the T_i are linear combinations of the explanatory variables $X_1, X_2, \ldots, X_K$ such that the sample correlation for any pair T_i, T_j $(i \neq j)$ is 0. We follow the procedure given by Garthwaite (1994). First, all the data are centered. Let $\bar{y}, \bar{x_1}, \ldots, \bar{x_k}$ denote the sample means of the columns of the $T \times (K+1)$-data matrix

$$(y, X) = (y_1, x_1, \ldots, x_k) \,,$$

and define the variables

$$\begin{aligned} U_1 &= Y - \bar{x}_i \,, & (3.238) \\ V_{1i} &= X_i - \bar{x}_i \quad (i = 1, \ldots, K) \,. & (3.239) \end{aligned}$$

Then the data values are the T-vectors

$$\begin{aligned} u_1 &= y - \bar{y}1 \,, \quad (\bar{u}_1 = 0) \,, & (3.240) \\ v_{i1} &= x_i - \bar{x}_i 1 \,, \quad (\bar{v}_{1i} = 0) \,. & (3.241) \end{aligned}$$

The linear combinations T_j, called factors, latent variables, or *components*, are then determined sequentially. The procedure is as follows:

(i) U_1 is first regressed against V_{11}, then regressed against V_{12}, ..., then regressed against V_{1K}. The K univariate regression equations are

$$\begin{aligned} \hat{U}_{1i} &= b_{1i} V_{1i} \quad (i = 1, \dots, K), \qquad (3.242) \\ \text{where} \quad b_{1i} &= \frac{v'_{1i} u_1}{v'_{1i} v_{1i}}. \qquad (3.243) \end{aligned}$$

Then each of the K equations in (3.243) provides an estimate of U_1. To have one resulting estimate, one may use a simple average $\sum_{i=1}^{K} b_{1i} V_{1i}/K$ or a weighted average such as

$$T_1 = \sum_{i=1}^{K} w_{1i} b_{1i} V_{1i} \qquad (3.244)$$

with the data value

$$t_1 = \sum_{i=1}^{K} w_{1i} b_{1i} v_{1i}. \qquad (3.245)$$

(ii) The variable T_1 should be a useful predictor of U_1 and hence of Y. The information in the variable X_i that is not in T_1 may be estimated by the residuals from a regression of X_i on T_1, which are identical to the residuals, say Y_{2i}, if V_{1i} is regressed on T_1, that is,

$$V_{2i} = V_{1i} - \frac{t'_1 v_{1i}}{t'_1 t_1} T_1. \qquad (3.246)$$

To estimate the amount of variability in Y that is not explained by the predictor T_1, one may regress U_1 on T_1 and take the residuals, say U_2.

(iii) Define now the individual predictors

$$\hat{U}_{2i} = b_{2i} V_{2i} \quad (i = 1, \dots, K), \qquad (3.247)$$

where

$$b_{2i} = \frac{v'_{2i} u_2}{v'_{2i} v_{2i}} \qquad (3.248)$$

and the weighted average

$$T_2 = \sum_{i=1}^{K} w_{2i} b_{2i} V_{2i}. \qquad (3.249)$$

(iv) *General iteration step.* Having performed this algorithm k times, the remaining residual variability in Y is U_{k+1} and the residual information in X_i is $V_{(k+1)i}$, where

$$U_{k+1} = U_k - \frac{t'_k u_k}{t'_k t_k} T_k \tag{3.250}$$

and

$$V_{(k+1)i} = V_{ki} - \frac{t'_k v_{ki}}{t'_k t_k} T_k \,. \tag{3.251}$$

Regressing U_{k+1} against $V_{(k+1)i}$ for $I = 1, \ldots, K$ gives the individual predictors

$$\hat{U}_{(k+1)i} = b_{(k+1)i} V_{(k+1)i} \tag{3.252}$$

with

$$b_{(k+1)i} = \frac{v'_{(k+1)i} u_{k+1}}{v'_{(k+1)i} v_{(k+1)i}}$$

and the $(k+1)^{th}$ component

$$T_{k+1} = \sum_{i=1}^{K} w_{(k+1)i} b_{(k+1)i} V_{(k+1)i} \,. \tag{3.253}$$

(v) Suppose that this process has stopped in the p^{th} step, resulting in the PLS regression model given in (3.237). The parameters $\beta_0, \beta_1, \ldots, \beta_p$ are estimated by univariate OLS. This can be proved as follows. In matrix notation we may define

$$\begin{aligned}
V_{(k)} &= (V_{k1}, \ldots, V_{kK}) \quad (k = 1, \ldots, p)\,, && (3.254)\\
\hat{U}_{(k)} &= (b_{k1}V_{k1}, \ldots, b_{kK}V_{kK}) \quad (k = 1, \ldots, p)\,, && (3.255)\\
w_{(k)} &= (w_{k1}, \ldots, w_{kK})' \quad (k = 1, \ldots, p)\,, && (3.256)\\
T_{(k)} &= \hat{U}_{(k)} w_{(k)} \quad (k = 1, \ldots, p)\,, && (3.257)\\
V_{(k)} &= V_{(k-1)} - \frac{v'_{(k-1)} t_{k-1}}{t'_{k-1} t_{k-1}} T_{k-1}\,. && (3.258)
\end{aligned}$$

By construction (cf. (3.251)) the sample residuals $v_{(k+1)i}$ are orthogonal to $v_{ki}, v_{(k-1)i}, \ldots, v_{1i}$, implying that $v'_{(k)} v_{(j)} = 0$ for $k \neq j$, hence, $\hat{u}'_{(k)} \hat{u}_{(j)} = 0$ for $k \neq j$, and finally,

$$t'_k t_j = 0 \quad (k \neq j)\,. \tag{3.259}$$

This is the well-known feature of the PLS (cf. Wold, Wold, Dunn and Ruhe, 1984; Helland, 1988) that the sample components t_i are pairwise uncorrelated. The simple consequence is that parameters β_k in equation (3.237) may be estimated by simple univariate regressions of Y against T_k. Furthermore, the preceding estimates $\hat{\beta}_k$ stay unchanged if a new component is added.

Specification of the Weights

In the literature, two weighting policies are discussed. First, one may set $w_{ij} = 1/K$ to give each predictor $\hat{U}_{ki}$ $(i = 1, \ldots, K)$ the same weight in any k^{th} step. The second policy in practice is the choice

$$w_{ki} = v'_{ki} v_{ki} \quad \text{(for all } k, i) \,. \tag{3.260}$$

As $\bar{v}_{ki} = 0$, the sample variance of V_{ki} is $\widehat{\text{var}}(V_{ki}) = v'_{ki} v_{ki}/(T-1)$. Using w_{ki} defined in (3.260) gives $w_{ki} b_{ki} = v'_{ki} u_k$ and

$$T_k = \sum_{i=1}^{K} (v'_{ki} u_k) V_{(k)i} \,. \tag{3.261}$$

The T-vector v_{ki} is estimating the amount of information in X_i that was not included in the preceding component T_{k-1}. Therefore, its vector norm $v'_{ki} v_{ki}$ is a measure for the contribution of X_i to T_k.

Size of the Model

Deciding the number of components (p) usually is done via some *cross-validation* (Stone, 1974; Geisser, 1974). The data set is divided into groups. At each step k, the model is fitted to the data set reduced by one of the groups. Predictions are calculated for the deleted data, and the sum of squares of predicted minus observed values for the deleted data is calculated. Next, the second data group is left out, and so on, until each data point has been left out once and only once. The total sum of squares (called PRESS) of predictions minus observations is a measure of the predictive power of the k^{th} step of the model. If for a chosen constant

$$\text{PRESS}_{(k+1)} - \text{PRESS}_{(k)} < \text{constant} \,,$$

then the procedure stops. In simulation studies, Wold et al. (1984) and Garthwaite (1994) have compared the predictive power of PLS, stepwise OLS, principal components estimator (PCR), and other methods. They found PLS to be better than OLS and PCR and comparable to, for example, ridge regression.

Multivariate extension of PLS is discussed by Garthwaite (1994). Helland (1988) has discussed the equivalence of alternative univariate PLS algorithms.

3.15 Tests of Parameter Constancy

One of the important assumptions in regression analysis is that the parameter vector β is invariant against the changes in data matrix within the sample, *i.e.*, the parameters remain constant, see e.g., Johnston and DiNardo (1997). In practice, this assumption may be violated over time and it gives rise to the problem of structural change. For example, the annual economic data may exhibit a structural change in the consumption pattern if there is a war and the point of structural change will be the year of war. Consequently the parameters of the model before and after the war will not remain same. There are various statistical and graphical methods to test the presence of structural change and parameter constancy in the data.

3.15.1 The Chow Forecast Test

The idea behind the test of Chow (1960) is to divide the complete regression model into two independent regression models such that the sample of size T is divided into two subsamples of sizes T_1 and T_2 and $T_1 + T_2 = T$. Partition $\underset{T\times 1}{y} = \underset{T\times K}{X}\beta + \epsilon$ into two independent regression models as

$$\begin{pmatrix} \underset{T_1\times 1}{y_1} \\ \underset{T_2\times 1}{y_2} \end{pmatrix} = \begin{pmatrix} \underset{T_1\times K}{X_1} \\ \underset{T_2\times K}{X_2} \end{pmatrix}\beta + \begin{pmatrix} \underset{T_1\times 1}{\epsilon_1} \\ \underset{T_2\times 1}{\epsilon_2} \end{pmatrix} \tag{3.262}$$

with $\mathrm{E}(\epsilon_1\epsilon_2') = 0$. The test of Chow for testing the constancy of parameters through $\mathrm{H}_0 : \beta_1 = \beta_2$ has the following steps:

(i) Estimate β using ordinary least squares estimator (OLSE) from the first submodel $y_1 = X_1\beta + \epsilon_1$ based on a sample of size T_1

$$b_1 = (X_1'X_1)^{-1}X_1'y_1.$$

(ii) Calculate the classical prediction from the second submodel $y_2 = X_2\beta + \epsilon_2$ according to

$$\hat{y}_2 = X_2b_1.$$

(iii) Now find the prediction error of $\hat{y}_2$, *i.e.*, assuming that the parameter vector β remains constant for both the submodels.

$$\begin{aligned} \Delta &= y_2 - \hat{y}_2 \qquad (3.263)\\ &= y_2 - X_2b_1 \\ &= \epsilon_2 - X_2(b_1 - \beta) \end{aligned}$$

where $b_1 = \beta + (X_1'X_1)^{-1}X_1'\epsilon_1$ and using $\mathrm{E}(\epsilon_1\epsilon_2') = 0$, we get

$$\mathrm{E}(\Delta) = 0$$

and

$$\begin{aligned} \mathrm{V}(\Delta) = \mathrm{E}(\Delta\Delta') &= \sigma^2 I_{T_2} + X_2\,\mathrm{V}(b_1)X_2' \\ &= \sigma^2(I_{T_2} + X_2(X_1'X_1)^{-1}X_2'). \qquad (3.264) \end{aligned}$$

Assuming that $\epsilon \sim N(0, \sigma^2 I)$, we have

$$\Delta \sim N(0, \mathrm{V}(\Delta))\,,$$

and

$$\Delta' \mathrm{V}^{-1}(\Delta)\Delta \sim \chi^2_{T_2} . \qquad \text{(cf. A.85(i))} \tag{3.265}$$

The residual from the first submodel is

$$\hat{\epsilon}_1 = y_1 - X_1 b_1 = (I_{T_1} - P_{X_1})y_1 \tag{3.266}$$

where $P_{X_1} = X_1(X_1'X_1)^{-1}X_1'$ is the hat matrix and

$$\hat{\epsilon}_1'\hat{\epsilon}_1 \sim \sigma^2\chi^2_{T_1-K} . \tag{3.267}$$

Further, $\Delta' \mathrm{V}^{-1}(\Delta)\Delta$ and $\hat{\epsilon}_1'\hat{\epsilon}_1$ are independently distributed. Therefore under the null hypothesis $\mathrm{H}_0 : \beta_1 = \beta_2$ (*i.e.*, β remains same in both submodels), the Chow's statistic is

$$\begin{aligned} F &= \frac{\Delta'(I_{T_2} + X_2(X_1'X_1)^{-1}X_2')^{-1}\Delta/T_2}{\hat{\epsilon}_1'\hat{\epsilon}_1/(T_1-K)} \\ &\sim F_{T_2,T_1-K} \end{aligned} \tag{3.268}$$

under H_0. The decision rule is to reject H_0 when $F \geq F_{T_2,T_1-K,1-\alpha}$.

Remark: The OLSE b, tests and measures of fit are invariant with respect to the permutation of rows of the data matrix (y, X). Therefore the division of the whole sample into two subsamples is arbitrary. In case of time series data, the observations can follow the natural order, *i.e.*, the first T_1 observations in the first subsample and remaining in the second subsample. In general, the size of the second sample T_2 should not be more than $5\% - 15\%$ of the total sample size T.

The Chow–Test as a Mean–Shift Test

The Chow-test can also be derived using the idea of mean–shift outlier model (cf. (7.49)). Assuming that the observations follow the model

$$y_1 = X_1\beta + \epsilon_1 , \tag{3.269}$$

and the period of forecasting follow another linear model with parameter vector α as

$$\begin{aligned} y_2 &= X_2\alpha + \epsilon_2 \\ &= X_2\beta + X_2(\alpha - \beta) + \epsilon_2 \\ &= X_2\beta + \delta + \epsilon_2 \end{aligned} \tag{3.270}$$

where $\delta = X_2(\alpha - \beta)$. Since $\delta = 0$ is equivalent to $\alpha = \beta$, so the hypothesis of parameter constancy can be formulated as $\mathrm{H}_0 : \delta = 0$.

The two models (3.269) and (3.270) can be written as mixed model:

$$\begin{aligned} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= \begin{pmatrix} X_1 & 0 \\ X_2 & I_{T_2} \end{pmatrix}\begin{pmatrix} \beta \\ \delta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \\ &= Z\tilde{\beta} + \epsilon . \end{aligned} \tag{3.271}$$

Since

$$Z'Z = \begin{pmatrix} X_1'X_1 + X_2'X_2 & X_2' \\ X_2 & I_{T_2} \end{pmatrix},$$

using Theorem A.19 of the Appendix A and further simplifying using (A.18) gives

$$(Z'Z)^{-1} = \begin{pmatrix} (X_1'X_1)^{-1} & -(X_1'X_1)^{-1}X_2' \\ -X_2(X_1'X_1)^{-1} & I_{T_2} + X_2(X_1'X_1)^{-1}X_2' \end{pmatrix}.$$

So the OLSE of $\tilde{\beta}$ in (3.271) is

$$\begin{aligned} \begin{pmatrix} b \\ d \end{pmatrix} &= (Z'Z)^{-1} \begin{pmatrix} X_1'y_1 + X_2'y_2 \\ y_2 \end{pmatrix} \\ &= \begin{pmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ y_2 - X_2(X_1'X_1)^{-1}X_1'y_1 \end{pmatrix} = \begin{pmatrix} b_1 \\ \Delta \end{pmatrix} \end{aligned}$$

where $\Delta = y_2 - \hat{y}_2$ (cf. (3.263)). This means, that the T_2 coefficients δ in the second equation of the model (3.271) are estimated by the prediction error Δ in (3.263).

Therefore the residuals in model (3.271) are estimated by

$$\begin{aligned} \hat{\epsilon} &= y - Z \begin{pmatrix} b_1 \\ \Delta \end{pmatrix} \\ &= \begin{pmatrix} y_1 - X_1 b_1 \\ y_2 - X_2 b_1 - I_{T_2}\Delta \end{pmatrix} = \begin{pmatrix} \hat{\epsilon}_1 \\ 0 \end{pmatrix}, \end{aligned}$$

which clearly shows that $\hat{\epsilon}_2 = 0$.

The hypothesis of parameter constancy $\mathrm{H}_0 : \alpha = \beta$ is equivalent to $\mathrm{H}_0 : \delta = 0$ and can be rewritten as $\mathrm{H}_0 : R\tilde{\beta} = 0$ with $R = (0, I)$ following the structure of model (3.271). Therefore

$$R \begin{pmatrix} b_1 \\ \Delta \end{pmatrix} = \Delta \sim N(0,\ \mathrm{V}(\Delta))$$

under H_0 where $\mathrm{V}(\Delta)$ is given in (3.264) and $\Delta' \,\mathrm{V}^{-1}(\Delta)\Delta \sim \chi^2_{T_2}$.

This means that the statistic for testing the constancy of parameters is equivalent to testing the linear restriction $\delta = 0$. Thus $\mathrm{H}_0 : \delta = 0$ can be tested using the statistic F of Chow from (3.268) (cf. also (3.117)).

Alternatively, we may interpret testing $\mathrm{H}_0 : \delta = 0$ in model (3.271) as equivalent to choosing one of the models (3.262) or (3.271). Therefore we again use the test statistic (3.85) as

$$F = \frac{(\hat{\epsilon}'\hat{\epsilon} - \hat{\epsilon}_1'\hat{\epsilon}_1)/T_2}{\hat{\epsilon}_1'\hat{\epsilon}_1/(T_1 - K)} \tag{3.272}$$

where $\hat{\epsilon}_1'\hat{\epsilon}_1$ is the RSS in the regression of y_1 on X_1 based on T_1 observations and $\hat{\epsilon}'\hat{\epsilon}$ is the RSS in the regression based on all T observations. To use this test in practice, one has to calculate

- the RSS $\hat{\epsilon}_1'\hat{\epsilon}_1$ from the regression of y_1 on X_1 based on T_1 observations,
- the RSS $\hat{\epsilon}'\hat{\epsilon}$ from the regression of y on X based on all T observations

and substitute these values in (3.272).

3.15.2 The Hansen Test

The sample was divided into two subsamples arbitrarily in the test of Chow. The Hansen test does not consider an arbitrary division of the sample. This test is based on cumulative observations. Considering the model $y = X\beta+\epsilon$, the residuals based on the OLS estimation are

$$\hat{\epsilon} = y - Xb = (I - P_X) = M\epsilon$$

where $b = (X'X)^{-1}X'y$ and $P_X = X(X'X)^{-1}X'$ is the hat matrix. It holds that $X'M = 0$. Therefore we have

$$X'\hat{\epsilon} = 0,$$

which can be written componentwise as

$$x_i'\hat{\epsilon} = \sum_{t=1}^{T} x_{it}\hat{\epsilon}_t = 0, \ (i = 1, \dots, K). \tag{3.273}$$

The maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_t^2$, (cf. (3.64)) which can be expressed as

$$\sum_{t=1}^{T}(\hat{\epsilon}_t^2 - \hat{\sigma}^2) = 0. \tag{3.274}$$

Hansen (1992) defined a function

$$f_{it} = \begin{cases} x_{it}\hat{\epsilon}_t & i = 1, \dots, K \\ \hat{\epsilon}_t^2 - \hat{\sigma}^2 & i = K+1. \end{cases} \tag{3.275}$$

Then using (3.273) and (3.274), we note that

$$\sum_{t=1}^{T} f_{it} = 0, \qquad i = 1, \dots, K+1. \tag{3.276}$$

The Hansen test statistic is based on the cumulative sums of f_{it}, defined as

$$S_{it} = \sum_{j=1}^{t} f_{ij} \ . \tag{3.277}$$

The statistic S_{it} can be used for constructing the test statistic for testing the stability of individual parameters as well as the stability of several parameters.

For testing the stability of individual parameters, the test statistic is defined as

$$L_i = \frac{1}{T v_i} \sum_{t=1}^{T} S_{it}^2 \qquad (i = 1, \dots, K+1) \tag{3.278}$$

where

$$v_i = \sum_{t=1}^{T} f_{it}^2 \, . \tag{3.279}$$

Let $f_t = (f_{1t}, \dots, f_{K+1,t})'$ and $s_t = (S_{1t}, \dots, S_{K+1,t})'$, then the test statistic for testing the stability of several parameters is defined as

$$L_c = \frac{1}{T} \sum_{t=1}^{T} s_t' \, \mathrm{V}^{-1} \, s_t \tag{3.280}$$

where

$$\mathrm{V} = \sum_{t=1}^{T} f_t f_t' \, . \tag{3.281}$$

Under the null hypothesis of the parameter constancy, the test statistics L_i in (3.278) or L_c in (3.280) are expected to be distributed around zero. The distributions of L_i and L_c are not standard and tables are available for their critical values, see e.g., Johnston and DiNardo (1997, Table 7, Appendix). Then large values of L_i or L_c suggest the rejection of null hypothesis meaning thereby the parameters are not stable.

3.15.3 Tests with Recursive Estimation

The t^{th} row of the model $y = X\beta + \epsilon$ is

$$y_t = x_t'\beta + \epsilon_t \qquad (t = 1, \dots, T) \, . \tag{3.282}$$

Let $X_i = (x_1', \dots, x_i')'$ be the matrix of the first i rows corresponding to the observation vector y_i and $b_i = (X_i'X_i)^{-1}X_i'y_i$ is the corresponding OLSE.

Now we fit the model successively starting with $i = K$ and obtain an estimate b_K based on first K observations. In the second step, use the first $K+1$ observations to estimate β and obtain b_{K+1}. This procedure is continued with $K+2, \dots, T$ observations and the OLSEs of β are obtained. This generates a sequence $b_K, b_{K+1}, \dots, b_T$ where $b_i = (X_i'X_i)^{-1}X_i'y_i$ $(i = K, \dots, T)$. This sequence can be plotted ($b_i \pm 2\times$ standard deviations) and a visual inspection can give a good idea about the possible parameter inconstancy.

Some other available procedures are CUSUM and CUSUMSQ–Tests, see e.g., Brown, Durbin and Evans (1975) for more details.

3.15.4 Test for Structural Change

In the Chow test, the model $y = X\beta + \epsilon$ was partitioned such that β remains same over the two independent submodels. Now we partition the model into two different submodels with different parameters as follows:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = X \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \epsilon \tag{3.283}$$

where each β_i is $K \times 1$, $X = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}$ and assume that $\epsilon = (\epsilon_1 \ , \ \epsilon_2)' \sim N(0, \sigma^2 I)$.

Thus the null hypothesis about the structural change is

$$H_0 : \beta_1 = \beta_2 \ , \tag{3.284}$$

which means there is no structural change. There are three test procedures to test $\mathrm{H}_0 : \beta_1 = \beta_2$.

Two-Sample-Test

The model (3.283) is the aggregation of two independent regression models. The OLSE of the whole vector $(\beta_1, \beta_2)'$ is

$$\begin{aligned} b &= \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y_1 \\ X_2'y_2 \end{pmatrix} \\ &= \begin{pmatrix} (X_1'X_1)^{-1}X_1'y_1 \\ (X_2'X_2)^{-1}X_2'y_2 \end{pmatrix} , \end{aligned} \tag{3.285}$$

where b_1 and b_2 are independent. Under $\mathrm{H}_0 : \beta_1 = \beta_2$, we have

$$b_1 - b_2 \sim N(0, \sigma^2[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]) \, .$$

Hence we get the two-sample test statistic:

$$F = \frac{(b_1 - b_2)'[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-1}(b_1 - b_2)}{s^2} \cdot \frac{T - 2K}{K} \tag{3.286}$$

with $(T-K)s^2 = \hat\epsilon'\hat\epsilon$ and $\hat\epsilon'\hat\epsilon = \hat\epsilon_1'\hat\epsilon_1 + \hat\epsilon_2'\hat\epsilon_2$ where $\hat\epsilon_i = y_i - X_i b_i$ for $i = 1, 2$.

The test statistic F in (3.286) can also be derived in an alternative way.

The hypothesis (3.284) can also be rewritten as $\mathrm{H}_0 : \beta_1 - \beta_2 = 0$, *i.e.*,

$$R \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = r \text{ with } r = 0 \text{ and } R = (I_K, -I_K) \ . \tag{3.287}$$

Therefore under H_0,

$$Rb \sim N(0, \sigma^2 R(X'X)^{-1}R')$$

and

$$(Rb - r)'[R(X'X)^{-1}R']^{-1}(Rb - r) \sim \sigma^2 \chi^2_K$$

which is same as the numerator of F in (3.286).

Tests Using Restricted Least Squares Estimator

If we interpret (3.283) as a unrestricted classical regression model, then the hypothesis $\mathrm{H}_0 : \beta_1 = \beta_2 = \beta$ is equivalent to the following submodel:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \epsilon \ . \tag{3.288}$$

The OLSE of β in model (3.288) is

$$\hat{\beta} = (X_1'X_1 + X_2'X_2)^{-1}(X_1'y_1 + X_2'y_2) \ . \tag{3.289}$$

The model (3.288) may be interpreted as submodel of (3.283) under the linear restriction (3.287). The corresponding restricted least squares estimator for model (3.283) is :

$$b(R) = b + (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(r - Rb) \tag{3.290}$$

with X and b as in (3.283) and (3.285), respectively.
The equivalence of the estimators (3.290) and (3.289) may be proved as follows:
Let $S_i = (X_i'X_i)$, $i = 1, 2$. Then we may write

$$\begin{aligned} b(R) &= \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} S_1^{-1} & 0 \\ 0 & S_2^{-1} \end{pmatrix} \begin{pmatrix} I \\ -I \end{pmatrix} \cdot \\ & \quad \left((I, -I) \begin{pmatrix} S_1^{-1} & 0 \\ 0 & S_2^{-1} \end{pmatrix} \begin{pmatrix} I \\ -I \end{pmatrix} \right)^{-1} (0 - b_1 + b_2) \, . \end{aligned}$$

Let

$$A = (I, -I) \begin{pmatrix} S_1^{-1} & 0 \\ 0 & S_2^{-1} \end{pmatrix} \begin{pmatrix} I \\ -I \end{pmatrix} = (S_1^{-1} + S_2^{-1}) \ ,$$

then (cf. A.18(iii))

$$\begin{aligned} A^{-1} &= S_1 - S_1(S_1 + S_2)^{-1}S_1 \\ &= S_2 - S_2(S_1 + S_2)^{-1}S_2 \, . \end{aligned}$$

Using this we get

$$\begin{aligned} b(R) &= \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} S_1^{-1} \\ -S_2^{-1} \end{pmatrix} A^{-1}(b_2 - b_1) \\ &= \begin{pmatrix} b_2 - (S_1 + S_2)^{-1}S_1b_2 + (S_1 + S_2)^{-1}X_1'y_1 \\ b_1 - (S_1 + S_2)^{-1}S_2b_1 + (S_1 + S_2)^{-1}X_2'y_2 \end{pmatrix} \\ &= \begin{pmatrix} (S_1 + S_2)^{-1}(X_1'y_1 + X_2'y_2) \\ (S_1 + S_2)^{-1}(X_1'y_1 + X_2'y_2) \end{pmatrix} \end{aligned} \tag{3.291}$$

as

$$\begin{aligned} (S_1 + S_2)^{-1}S_1b_2 &= (S_1 + S_2)^{-1}(S_1 + S_2 - S_2)b_2 \\ &= b_2 - (S_1 + S_2)^{-1}X_2'y_2 \ . \end{aligned}$$

The restricted estimator follows the restrictions, *i.e.*, $Rb(R) = 0$. (But this is just the relation (3.291), which as a consequence, corresponds to $\hat{\beta}$ (3.289).)

The model (3.283) corresponds to the whole parameter space Ω and model (3.288) corresponds to the subset of parameter space $\omega \subseteq \Omega$. Therefore the test statistic F in (3.85) is

$$F = \frac{\hat{\epsilon}'_R \hat{\epsilon}_R - \hat{\epsilon}'\hat{\epsilon}}{\hat{\epsilon}'\hat{\epsilon}} \cdot \frac{T-2K}{K} \sim F_{K,T-2K} \tag{3.292}$$

where

$$\hat{\epsilon}_R = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \hat{\beta}$$

with $\hat{\beta}$ from (3.289) and

$$\hat{\epsilon} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} .$$

In practice both the models can be used with any statistical software and give the test statistic F as in (3.292).

Alternative Test in Unrestricted Model

Following Johnston and DiNardo (1997, p. 127), we use the unrestricted model

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ X_2 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 - \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} . \tag{3.293}$$

Now $\mathrm{H}_0 : \beta_1 = \beta_2$ can be tested by checking the significance of the last K regressors. This procedure can also be used with any statistical software.

Testing the Slope Parameter

For testing the slope parameter, use partition

$$X = \begin{pmatrix} X_1 & 0 \\ X_2 & X_2 \end{pmatrix} , \ \beta' = (\alpha, \ \beta^*)$$

and further partition

$$X_1 = (\mathbf{1}_1, \ X_1^*), \quad X_2 = (\mathbf{1}_2, \ X_2^*) \text{ with } X_i^* : T_i \times (K-1).$$

The test of hypothesis

$$H_0 : \ \beta_1^* = \beta_2^* \tag{3.294}$$

is now based on the unrestricted model

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_1 & 0 & X_1^* & 0 \\ 0 & \mathbf{1}_2 & 0 & X_2^* \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1^* \\ \beta_2^* \end{pmatrix} + \epsilon \ . \tag{3.295}$$

On the other hand, the restricted model becomes

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_1 & 0 & X_1^* \\ 0 & \mathbf{1}_2 & X_2^* \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta^* \end{pmatrix} + \epsilon \,. \tag{3.296}$$

The test of H_0 in (3.294) is now based on the residual sum of squares from both the models (3.295) and (3.296).

Alternatively, the model (3.295) can be written as

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_1 & 0 & X_1^* & 0 \\ \mathbf{1}_2 & \mathbf{1}_2 & X_2^* & X_2^* \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 - \alpha_1 \\ \beta_1^* \\ \beta_2^* - \beta_1^* \end{pmatrix} + \epsilon \,. \tag{3.297}$$

Now to test $\mathrm{H}_0 : \beta_1^* = \beta_2^*$ as in (3.294), we simply have to test the significance of the parameters associated with the last $(K-1)$ regressors.

3.16 Total Least Squares

In contrast to our treatment in the other chapters, we now change assumptions on the independent variables, that is, we allow the X_i to be measured with errors also. The method of fitting such models is known as orthogonal regression or errors-in-variables regression, also called total least squares. The idea is as follows (cf. van Huffel and Zha, 1993).

Consider an overdetermined set of $m > n$ linear equations in n unknowns x ($A : m \times n$, $x : n \times 1$, $a : m \times 1$)

$$Ax = a \,. \tag{3.298}$$

Then the ordinary least-squares problem may be written as

$$\min_{\hat{a} \in \mathbb{R}^m} \|a - \hat{a}\|_2 \quad \text{subject to} \quad \hat{a} \in \mathcal{R}(A) \,, \tag{3.299}$$

where $\|x\|_2$ is the L_2-norm or Euclidean norm of a vector x. Let $\hat{a}$ be a solution of (3.299), then any vector x satisfying $Ax = \hat{a}$ is called a LS solution (LS = least squares). The difference

$$\Delta a = a - \hat{a} \tag{3.300}$$

is called the LS correction. The assumptions are that errors occur only in the vector a and that A is exactly known.

If we also allow for perturbations in A, we are led to the following definition.

The *total least-squares (TLS) problem* for solving an overdetermined linear equation $Ax = a$ is defined by

$$\min_{(\hat{A},\hat{a}) \in \mathbb{R}^{m \times (n+1)}} \|(A, a) - (\hat{A}, \hat{a})\|_F \tag{3.301}$$

subject to

$$\hat{a} \in R(\hat{A}), \tag{3.302}$$

where

$$\|Q\|_F = [\mathrm{tr}(QQ')]^{\frac{1}{2}} \tag{3.303}$$

is the Frobenius norm of a matrix Q.

If a minimizer $(\hat{A}, \hat{a})$ is found, then any x satisfying $\hat{A}x = \hat{a}$ is called a TLS solution, and

$$[\Delta\hat{A}, \Delta\hat{a}] = (A, a) - (\hat{A}, \hat{a}) \tag{3.304}$$

is called the TLS correction.

Indeed, the TLS problem is more general than the LS problem, for the TLS solution is obtained by approximating the columns of the matrix A by $\hat{A}$ and a by $\hat{a}$ until $\hat{a}$ is in the space $\mathcal{R}(\hat{A})$ and $\hat{A}x = \hat{a}$.

Basic Solution to TLS

We rewrite $Ax = a$ as

$$(A, a)\begin{pmatrix} x \\ -1 \end{pmatrix} = 0. \tag{3.305}$$

Let the singular value decomposition (SVD; cf. Theorem A.32) of the $(m, n+1)$-matrix (A, a) be

$$\begin{aligned}(A, a) &= ULV' \\ &= \sum_{i=1}^{n+1} l_i u_i v_i', \end{aligned} \tag{3.306}$$

where $l_1 \geq \ldots \geq l_{n+1} \geq 0$. If $l_{n+1} \neq 0$, then (A, a) is of rank $n+1$, $\mathcal{R}\big((A, a)'\big) = \mathbb{R}^{n+1}$, and (3.305) has no solution.

Lemma 3.17 (Eckart-Young-Mirsky matrix approximation theorem) *Let A : $n \times n$ be a matrix of $\mathrm{rank}(A) = r$, and let $A = \sum_{i=1}^{r} l_i u_i v_i'$, $l_i > 0$, be the singular value decomposition of A. If $k < r$ and $A_k = \sum_{i=1}^{k} l_i u_i v_i'$, then*

$$\min_{\mathrm{rank}(\hat{A})=k} \|A - \hat{A}\|_2 = \|A - \hat{A}_k\|_2 = l_{k+1}$$

and

$$\min_{\mathrm{rank}(\hat{A})=k} \|A - \hat{A}\|_F = \|A - \hat{A}_k\|_F = \sqrt{\sum_{i=k+1}^{p} l_i^2},$$

where $p = \min(m, n)$.

Proof: See Eckart and Young (1936), Mirsky (1960), Rao (1979; 1980).

Based on this theorem, the best rank n approximation $(\hat{A}, \hat{a})$ of (A, a) in the sense of minimal deviation in variance is given by

$$(\hat{A}, \hat{a}) = U\hat{L}V' , \quad \text{where} \quad \hat{L} = (l_1, \ldots, l_n, 0) . \tag{3.307}$$

The minimal TLS correction is then given by

$$l_{n+1} = \min_{\operatorname{rank}(\hat{A},\hat{a})=n} \|(A, a) - (\hat{A}, \hat{a})\|_F . \tag{3.308}$$

So we have

$$(A, a) - (\hat{A}, \hat{a}) = (\Delta\hat{A}, \Delta\hat{a}) = l_{n+1}u_{n+1}v'_{n+1} . \tag{3.309}$$

Then the approximate equation (cf. (3.305))

$$(\hat{A}, \hat{a}) \begin{pmatrix} x \\ -1 \end{pmatrix} = 0 \tag{3.310}$$

is compatible and has solution

$$\begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} = \frac{-1}{v_{n+1,n+1}} v_{n+1} , \tag{3.311}$$

where $v_{n+1,n+1}$ is the $(n+1)^{th}$ component of the vector v_{n+1}. Finally, $\hat{x}$ is solution of the TLS equation $\hat{A}x = \hat{a}$.

On the other hand, if l_{n+1} is zero, then $\operatorname{rank}(A, a) = n$, $v_{n+1} \in \mathcal{N}\{(A, a)\}$, and the vector $\hat{x}$ defined in (3.311) is the exact solution of $Ax = a$.

3.17 Minimax Estimation

3.17.1 Inequality Restrictions

Minimax estimation is based on the idea that the quadratic risk function for the estimate $\hat{\beta}$ is not minimized over the entire parameter space $\mathbb{R}^K$, but only over an area $B(\beta)$ that is restricted by a priori knowledge. For this, the supremum of the risk is minimized over $B(\beta)$ in relation to the estimate (minimax principle).

In many of the models used in practice, the knowledge of a priori restrictions for the parameter vector β may be available in a natural way. Stahlecker (1987) shows a variety of examples from the field of economics (such as input-output models), where the restrictions for the parameters are so-called workability conditions of the form $\beta_i \geq 0$ or $\beta_i \in (a_i, b_i)$ or $\mathrm{E}(y_t|X) \leq a_t$ and more generally

$$A\beta \leq a . \tag{3.312}$$

Minimization of $S(\beta) = (y - X\beta)'(y - X\beta)$ under inequality restrictions can be done with the simplex algorithm. Under general conditions we obtain a numerical solution. The literature deals with this problem under the generic term *inequality restricted least squares* (cf. Judge and Takayama, 1966; Dufour, 1989; Geweke, 1986; Moors and van Houwelingen, 1987). The advantage of this procedure is that a solution $\hat{\beta}$ is found that fulfills the restrictions. The disadvantage is that the statistical properties of the estimates are not easily determined and no general conclusions about superiority can be made. If all restrictions define a convex area, this area can often be enclosed in an ellipsoid of the following form:

$$B(\beta) = \{\beta : \beta' T \beta \leq k\} \tag{3.313}$$

with the origin as center point or in

$$B(\beta, \beta_0) = \{\beta : (\beta - \beta_0)' T (\beta - \beta_0) \leq k\} \tag{3.314}$$

with the center point vector β_0.

For example, (3.312) leads to $\beta' A' A \beta \leq a^2$, and hence to the structure $B(\beta)$.

Inclusion of Inequality Restrictions in an Ellipsoid

We assume that for all components β_i of the parameter vector β, the following restrictions in the form of intervals are given a priori:

$$a_i \leq \beta_i \leq b_i \quad (i = 1, \ldots, K)\,. \tag{3.315}$$

The empty restrictions ($a_i = -\infty$ and $b_i = \infty$) may be included. The limits of the intervals are known. The restrictions (3.315) can alternatively be written as

$$\frac{|\beta_i - (a_i + b_i)/2|}{1/2(b_i - a_i)} \leq 1 \quad (i = 1, \ldots, K)\,. \tag{3.316}$$

We now construct an ellipsoid $(\beta - \beta_0)' T (\beta - \beta_0) = 1$, which encloses the cuboid (3.316) and fulfills the following conditions:

(i) The ellipsoid and the cuboid have the same center point, $\beta_0 = \frac{1}{2}(a_1 + b_1, \ldots, a_K + b_K)$.

(ii) The axes of the ellipsoid are parallel to the coordinate axes, that is, $T = \text{diag}(t_1, \ldots, t_K)$.

(iii) The corner points of the cuboid are on the surface of the ellipsoid, which means we have

$$\sum_{i=1}^{K} \left(\frac{a_i - b_i}{2} \right)^2 t_i = 1\,. \tag{3.317}$$

(iv) The ellipsoid has minimal volume:

$$V = c_K \prod_{i=1}^{K} t_i^{-\frac{1}{2}}, \tag{3.318}$$

with c_K being a constant dependent on the dimension K.

We now include the linear restriction (3.317) for the t_i by means of Lagrangian multipliers λ and solve (with $c_K^{-2}V_K^2 = \prod t_i^{-1}$)

$$\min_{\{t_i\}} \tilde{V} = \min_{\{t_i\}} \left\{ \prod_{i=1}^{K} t_i^{-1} - \lambda \left[\sum_{i=1}^{K} \left(\frac{a_i - b_i}{2} \right)^2 t_i - 1 \right] \right\}. \tag{3.319}$$

The normal equations are then

$$\frac{\partial \tilde{V}}{\partial t_j} = -t_j^{-2} \prod_{i \neq j} t_i^{-1} - \lambda \left(\frac{a_j - b_j}{2} \right)^2 = 0 \tag{3.320}$$

and

$$\frac{\partial \tilde{V}}{\partial \lambda} = \sum \left(\frac{a_i - b_i}{2} \right)^2 t_i - 1 = 0\,. \tag{3.321}$$

From (3.320) we get

$$\begin{aligned} \lambda &= -t_j^{-2} \prod_{i \neq j} t_i^{-1} \left(\frac{2}{a_j - b_j} \right)^2 \quad (\text{for all } j = 1, \ldots, K) \\ &= -t_j^{-1} \prod_{i=1}^{K} t_i^{-1} \left(\frac{2}{a_j - b_j} \right)^2, \end{aligned} \tag{3.322}$$

and for any two i, j we obtain

$$t_i \left(\frac{a_i - b_i}{2} \right)^2 = t_j \left(\frac{a_j - b_j}{2} \right)^2, \tag{3.323}$$

and hence—after summation—according to (3.321),

$$\sum_{i=1}^{K} \left(\frac{a_i - b_i}{2} \right)^2 t_i = K t_j \left(\frac{a_j - b_j}{2} \right)^2 = 1\,. \tag{3.324}$$

This leads to the required diagonal elements of T:

$$t_j = \frac{4}{K}(a_j - b_j)^{-2} \quad (j = 1, \ldots, K)\,.$$

Hence, the optimal ellipsoid $(\beta - \beta_0)'T(\beta - \beta_0) = 1$, which contains the cuboid, has the center point vector

$$\beta_0' = \frac{1}{2}(a_1 + b_1, \ldots, a_K + b_K) \tag{3.325}$$

and the following matrix, which is positive definite for finite limits a_i, b_i $(a_i \neq b_i)$,

$$T = \text{diag} \frac{4}{K} \left((b_1 - a_1)^{-2}, \ldots, (b_K - a_K)^{-2} \right). \tag{3.326}$$

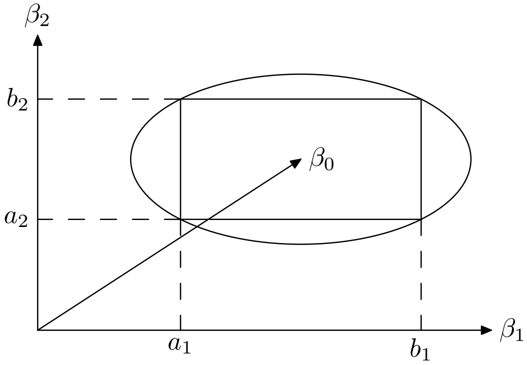

FIGURE 3.7. A priori rectangle and enclosing ellipsoid

Interpretation: The ellipsoid has a larger volume than the cuboid. Hence, the transition to an ellipsoid as a priori information represents a weakening, but comes with an easier mathematical handling.

Example 3.1: (Two real regressors) The center-point equation of the ellipsoid is (cf. Figure 3.7)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

or

$$(x, y) \begin{pmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 1$$

with

$$T = \operatorname{diag}\left(\frac{1}{a^2}, \frac{1}{b^2}\right) = \operatorname{diag}(t_1, t_2)$$

and the area $F = \pi ab = \pi t_1^{-\frac{1}{2}} t_2^{-\frac{1}{2}}$.

3.17.2 The Minimax Principle

Consider the quadratic risk $R_1(\hat\beta, \beta, A) = \operatorname{tr}\{AM(\hat\beta, \beta)\}$ and a class $\{\hat\beta\}$ of estimators. Let $B(\beta) \subset \mathbb{R}^K$ be a convex region of a priori restrictions for β. The criterion of the minimax estimator leads to the following.

Definition 3.18 *An estimator $b^* \in \{\hat\beta\}$ is called a minimax estimator of β if*

$$\min_{\{\hat\beta\}} \sup_{\beta \in B} R_1(\hat\beta, \beta, A) = \sup_{\beta \in B} R_1(b^*, \beta, A)\,. \tag{3.327}$$

Linear Minimax Estimators

We now confine ourselves to the class of linear homogeneous estimators $\{\hat{\beta} = Cy\}$. For these estimates the risk can be expressed as (cf. (4.16))

$$R_1(Cy, \beta, A) = \sigma^2 \mathrm{tr}(ACC') + \beta' T^{\frac{1}{2}} \tilde{A} T^{\frac{1}{2}} \beta \tag{3.328}$$

with

$$\tilde{A} = T^{-\frac{1}{2}} (CX - I)' A (CX - I) T^{-\frac{1}{2}}, \tag{3.329}$$

and $T > 0$ is the matrix of the a priori restriction

$$B(\beta) = \{\beta : \beta' T \beta \leq k\}. \tag{3.330}$$

Using Theorem A.44 we get

$$\sup_{\beta} \frac{\beta' T^{\frac{1}{2}} \tilde{A} T^{\frac{1}{2}} \beta}{\beta' T \beta} = \lambda_{\max}(\tilde{A})$$

and hence

$$\sup_{\beta' T \beta \leq k} R_1(Cy, \beta, A) = \sigma^2 \mathrm{tr}(ACC') + k \lambda_{\max}(\tilde{A}). \tag{3.331}$$

Since the matrix $\tilde{A}$ (3.329) is dependent on the matrix C, the maximum eigenvalue $\lambda_{\max}(\tilde{A})$ is dependent on C as well, but not in an explicit form that could be used for differentiation. This problem has received considerable attention in the literature. In addition to iterative solutions (Kuks, 1972; Kuks and Olman, 1971, 1972) the suggestion of Trenkler and Stahlecker (1987) is of great interest. They propose to use the inequality $\lambda_{\max}(\tilde{A}) \leq \mathrm{tr}(\tilde{A})$ to find an upper limit of $R_1(Cy, \beta, A)$ that is differentiable with respect to C, and hence find a substitute problem with an explicit solution. A detailed discussion can be found in Schipp (1990).

An explicit solution can be achieved right away if the weight matrices are confined to matrices of the form $A = aa'$ of rank 1, so that the $R_1(\hat{\beta}, \beta, A)$ risk equals the weaker $R_2(\hat{\beta}, \beta, a)$ risk (cf. (4.5)).

Linear Minimax Estimates for Matrices $A = aa'$ of Rank 1

In the case where $A = aa'$, we have

$$\tilde{A} = [T^{-\frac{1}{2}} (CX - I)' a][a' (CX - I) T^{-\frac{1}{2}}] = \tilde{a} \tilde{a}', \tag{3.332}$$

and according to the first Corollary to Theorem A.28 we obtain $\lambda_{\max}(\tilde{A}) = \tilde{a}' \tilde{a}$. Therefore, (3.331) becomes

$$\sup_{\beta' T \beta \leq k} R_2(Cy, \beta, a) = \sigma^2 a' CC' a + k a' (CX - I) T^{-1} (CX - I)' a. \tag{3.333}$$

Differentiation with respect to C leads to (Theorems A.91, A.92)

$$\frac{1}{2} \frac{\partial}{\partial C} \{ \sup_{\beta' T \beta \leq k} R_2(Cy, \beta, a) \} = (\sigma^2 I + k X T^{-1} X') C' aa' - k X T^{-1} aa'. \tag{3.334}$$

Since a is any fixed vector, (3.334) equals zero for all matrices aa' if and only if

$$C'_* = k(\sigma^2 I + kXT^{-1}X')^{-1}XT^{-1}. \tag{3.335}$$

After transposing (3.335) and multiplying from the left with $(\sigma^2 T + kS)$, we obtain

$$\begin{aligned}(\sigma^2 T + kS)C_* &= kX'[\sigma^2 I + kXT^{-1}X'][\sigma^2 I + kXT^{-1}X']^{-1} \\ &= kX',\end{aligned}$$

which leads to the solution $(S = X'X)$

$$C_* = (S + k^{-1}\sigma^2 T)^{-1}X'. \tag{3.336}$$

Using the abbreviation

$$D_* = (S + k^{-1}\sigma^2 T), \tag{3.337}$$

we have the following theorem.

Theorem 3.19 (Kuks, 1972) *In the model* $y = X\beta + \epsilon$, $\epsilon \sim (0, \sigma^2 I)$, *with the restriction* $\beta' T\beta \leq k$ *with* $T > 0$, *and the risk function* $R_2(\hat{\beta}, \beta, a)$, *the linear minimax estimator is of the following form:*

$$\begin{aligned}b_* &= (X'X + k^{-1}\sigma^2 T)^{-1}X'y \\ &= D_*^{-1}X'y \end{aligned}\tag{3.338}$$

with

$$\begin{aligned}\text{Bias}(b_*, \beta) &= -k^{-1}\sigma^2 D_*^{-1}T\beta, & (3.339)\\ \text{V}(b_*) &= \sigma^2 D_*^{-1}SD_*^{-1} & (3.340)\end{aligned}$$

and the minimax risk

$$\sup_{\beta' T\beta \leq k} R_2(b_*, \beta, a) = \sigma^2 a' D_*^{-1} a. \tag{3.341}$$

Theorem 3.20 *Given the assumptions of Theorem 3.19 and the restriction* $(\beta - \beta_0)' T(\beta - \beta_0) \leq k$ *with center point* $\beta_0 \neq 0$, *the linear minimax estimator is of the following form:*

$$b_*(\beta_0) = \beta_0 + D_*^{-1}X'(y - X\beta_0) \tag{3.342}$$

with

$$\begin{aligned}\text{Bias}(b_*(\beta_0), \beta) &= -k^{-1}\sigma^2 D_*^{-1}T(\beta - \beta_0), & (3.343)\\ \text{V}(b_*(\beta_0)) &= V(b_*), & (3.344)\end{aligned}$$

and

$$\sup_{(\beta-\beta_0)' T(\beta-\beta_0) \leq k} R_2(b_*(\beta_0), \beta, a) = \sigma^2 a' D_*^{-1} a. \tag{3.345}$$

Proof: The proof is similar to that used in Theorem 3.19, with $\beta - \beta_0 = \tilde{\beta}$.

Interpretation: A change of the center point of the a priori ellipsoid has an influence only on the estimator itself and its bias. The minimax estimator is not operational, because of the unknown σ^2. The smaller the value of k, the stricter is the a priori restriction for fixed T. Analogously, the larger the value of k, the smaller is the influence of $\beta' T\beta \le k$ on the minimax estimator. For the borderline case we have

$$B(\beta) = \{\beta : \beta' T\beta \le k\} \to \mathbb{R}^K \quad \text{as} \quad k \to \infty$$

and

$$\lim_{k\to\infty} b_* \to b = (X'X)^{-1}X'y\,. \tag{3.346}$$

Comparison of b_* and b

(i) Minimax Risk Since the OLS estimator is unbiased, its minimax risk is

$$\sup_{\beta' T\beta\le k} R_2(b,\cdot,a) = R_2(b,\cdot,a) = \sigma^2 a' S^{-1} a\,. \tag{3.347}$$

The linear minimax estimator b_* has a smaller minimax risk than the OLS estimator, because of its optimality, according to Theorem 3.19. Explicitly, this means (Toutenburg, 1976)

$$\begin{aligned} R_2(b,\cdot,a) \;&-\; \sup_{\beta' T\beta\le k} R_2(b_*,\beta,a) \\ &= \sigma^2 a'(S^{-1} - (k^{-1}\sigma^2 T + S)^{-1})a \ge 0\,, \end{aligned} \tag{3.348}$$

since $S^{-1} - (k^{-1}\sigma^2 T + S)^{-1} \ge 0$ (cf. Theorem A.40 or Theorem A.52).

(ii) MDE-I Superiority With (3.343) and (3.344) we get

$$\begin{aligned} M(b_*,\beta) &= V(b_*) + \text{Bias}(b_*,\beta)\,\text{Bias}(b_*,\beta)' \\ &= \sigma^2 D_*^{-1}(S + k^{-2}\sigma^2 T\beta\beta' T')D_*^{-1}\,. \end{aligned} \tag{3.349}$$

Hence, b_* is MDE-I-superior to b if

$$\Delta(b,b_*) = \sigma^2 D_*^{-1}[D_* S^{-1} D_* - S - k^{-2}\sigma^2 T\beta\beta' T']D_*^{-1} \ge 0\,, \tag{3.350}$$

hence if and only if

$$\begin{aligned} B &= D_* S^{-1} D_* - S - k^{-2}\sigma^2 T\beta\beta' T' \\ &= k^{-2}\sigma^4 T[\{S^{-1} + 2k\sigma^{-2}T^{-1}\} - \sigma^{-2}\beta\beta']T \ge 0 \\ &= k^{-2}\sigma^4 T C^{\frac{1}{2}}[I - \sigma^{-2}C^{-\frac{1}{2}}\beta\beta' C^{-\frac{1}{2}}]C^{\frac{1}{2}}T \ge 0 \end{aligned} \tag{3.351}$$

with $C = S^{-1} + 2k\sigma^{-2}T^{-1}$. This is equivalent (Theorem A.57) to

$$\sigma^{-2}\beta'(S^{-1} + 2k\sigma^{-2}T^{-1})^{-1}\beta \le 1\,. \tag{3.352}$$

Since $(2k\sigma^{-2}T^{-1})^{-1} - (S^{-1} + 2k\sigma^{-2}T^{-1}) \ge 0$,

$$k^{-1} \le \frac{2}{\beta'\beta} \tag{3.353}$$

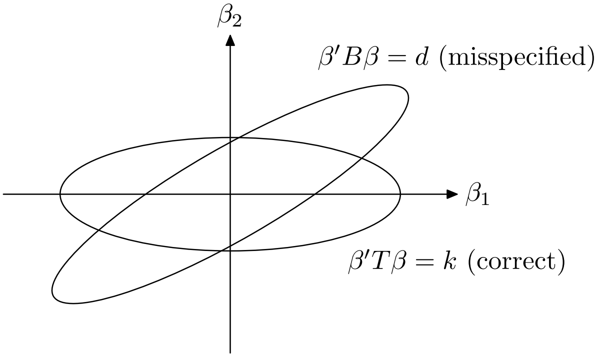
β_2
$\beta' B\beta = d$ (misspecified)
β_1
$\beta' T\beta = k$ (correct)

FIGURE 3.8. Misspecification by rotation and distorted length of the axes

is sufficient for the MDE-I superiority of the minimax estimator b_* compared to b. This condition corresponds to the condition (3.223) for the MDE-I superiority of the ridge estimator $b(k)$ compared to b.

We now have the following important interpretation: The linear minimax estimator b_* is a ridge estimate $b(k^{-1}\sigma^2)$. Hence, the restriction $\beta' T\beta \leq k$ has a stabilizing effect on the variance. The minimax estimator is operational if σ^2 can be included in the restriction $\beta' T\beta \leq \sigma^2 k = \tilde{k}$:

$$b_* = (X'X + \tilde{k}^{-1}T)^{-1}X'y\,.$$

Alternative considerations, as in Chapter 6, when σ^2 is not known in the case of mixed estimators, have to be made (cf. Toutenburg, 1975a; 1982, pp. 95–98).

From (3.352) we can derive a different sufficient condition: $kT^{-1} - \beta\beta' \geq 0$, equivalent to $\beta' T\beta \leq k$. Hence, the minimax estimator b_* is always MDE-I-superior to b, in accordance with Theorem 3.19, if the restriction is satisfied, that is, if it is chosen correctly.

The problem of robustness of the linear minimax estimator relative to misspecification of the a priori ellipsoid is dealt with in Toutenburg (1984; 1990)

Figures 3.8 and 3.9 show typical situations for misspecifications.

3.18 Censored Regression

3.18.1 Overview

Consider the regression model (cf. (3.23))

$$y_t = x_t'\beta + \epsilon_t\,, \quad t = 1, \ldots, T\,. \tag{3.354}$$

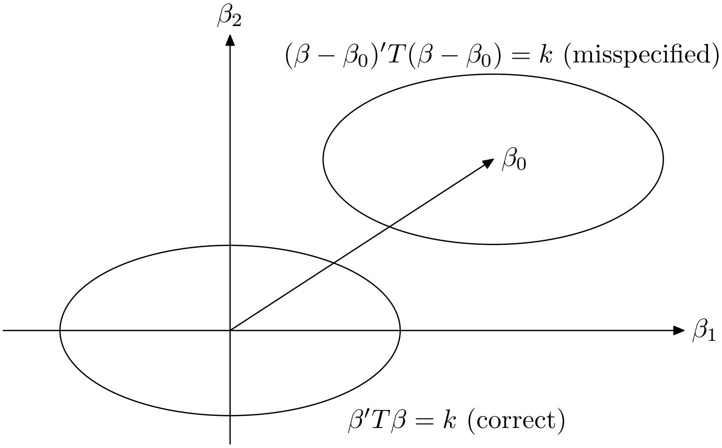
β_2
$(\beta - \beta_0)'T(\beta - \beta_0) = k$ (misspecified)
β_0
β_1
$\beta'T\beta = k$ (correct)

FIGURE 3.9. Misspecification by translation of the center point

There are numerous examples in economics where the dependent variable y_t is censored, and what is observable is, for example,

$$\begin{aligned} y_t^* &= 1 \quad \text{if} \quad y_t \geq 0\,, \\ y_t^* &= 0 \quad \text{if} \quad y_t < 0\,, \end{aligned} \tag{3.355}$$

or

$$\begin{aligned} y_t^* &= y \quad \text{if} \quad y > 0\,, \\ y_t^* &= 0 \quad \text{if} \quad y \leq 0\,. \end{aligned} \tag{3.356}$$

Model (3.355) is called the binary choice model, and model (3.356), the Tobit model. The problem is to estimate β from such models, generally referred to as limited dependent variable models. For specific examples of such models in economics, the reader is referred Maddala (1983). A variety of methods have been proposed for the estimation of β under models (3.354, 3.355) and (3.354, 3.356) when the e_t's have normal and unknown distributions.

Some of the well-known methods in the case of the Tobit model (3.354, 3.356) are the maximum likelihood method under a normality assumption (as described in Maddala, 1983, pp. 151–156; Amemiya, 1985, Chapter 10; Heckman, 1976), distribution-free least-squares type estimators by Buckley and James (1979) and Horowitz (1986); quantile including the LAD (least absolute deviations) estimators by Powell (1984); and Bayesian computing methods by Polasek and Krause (1994). A survey of these methods and Monte Carlo comparisons of their efficiencies can be found in the papers by Horowitz (1988) and Moon (1989). None of these methods provides closed-form solutions. They are computationally complex and their efficiencies depend on the distribution of the error component in the model and the intensity of censoring. No clear-cut conclusions emerge from these studies on the relative merits of various methods, especially when the sample size is small. Much work remains to be done in this area.

In the present section, we consider some recent contributions to the asymptotic theory of estimation of the regression parameters and tests of linear hypotheses based on the LAD method, with minimal assumptions.

3.18.2 LAD Estimators and Asymptotic Normality

We consider the Tobit model (3.354, 3.356), which can be written in the form

$$y_t^+ = (x_t'\beta + \epsilon_t)^+\,, \quad t = 1, \ldots, T\,, \tag{3.357}$$

where $y_t^+ = y_t I$, $(y_t > 0)$, and $I(\cdot)$ denotes the indicator function of a set, and assume that

(A.1) $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. random variables such that the distribution function F of ϵ_1 has median zero and positive derivative $f(0)$ at zero.

(A.2) The parameter space B to which β_0, the true value of β, belongs is a bounded open set of $\mathbb{R}^K$ (with a closure $\bar{B}$).

Based on the fact $\mathrm{med}(y_t^+) = (x_t'\beta_0)^+$, Powell (1984) introduced and studied the asymptotic properties of the LAD estimate $\hat{\beta}_T$ of β_0, which is a Borel-measurable solution of the minimization problem

$$\sum_{t=1}^{T} |y_t^+ - (x_t'\hat{\beta}_T)^+| = \min \left\{ \sum_{t=1}^{T} |y_t^+ - (x_t'\beta)^+| \;:\; \beta \in \bar{B} \right\}. \tag{3.358}$$

Since $\sum_{t=1}^{T} |y_t - (x_t'\beta)^+|$ is not convex in β, the analysis of $\hat{\beta}_T$ is quite difficult. However, by using uniform laws of large numbers, Powell established the strong consistency of $\hat{\beta}_T$ when x_t's are independent variables with $\mathrm{E}\,\|x_t\|^3$ being bounded, where $\|\cdot\|$ denotes the Euclidean norm of a vector. He also established its asymptotic normal distribution under some conditions.

With the help of the maximal inequalities he developed, Pollard (1990) improved the relevant result of Powell on asymptotic normality by relaxing Powell‘s assumptions and simplified the proof to some extent. Pollard permitted vectors $\{x_t\}$ to be deterministic. We investigate the asymptotic behavior of $\hat{\beta}_T$ under weaker conditions. We establish the following theorem, where we write

$$\mu_t = x_t'\beta_0 \quad \text{and} \quad S_T = \sum_{t=1}^{T} I(\mu_t > 0) x_t x_t'\,. \tag{3.359}$$

Theorem 3.21 *Assume that (A.1), (A.2) hold, and the following assumptions are satisfied:*

(A.3) For any $\sigma > 0$, there exists a finite $\alpha > 0$ such that

$$\sum_{t=1}^{T} \|x_t\|^2 I(\|x_t\| > \alpha) < \sigma \lambda_{\min}(S_T) \quad \textit{for } T \textit{ large}\,,$$

where $\lambda_{\min}(S_T)$ is the smallest eigenvalue of S_T.

(A.4) For any $\sigma > 0$, there is a $\delta > 0$ such that

$$\sum_{t=1}^{T} \|x_t\|^2 I(|\mu_t| \le \delta) \le \sigma \lambda_{\min}(S_T) \quad \textit{for } T \textit{ large}\,.$$

(A.5)

$$\lambda_{\min} \frac{(S_T)}{(\log T)^2} \to \infty\,, \quad \textit{as} \quad T \to \infty\,.$$

Then

$$2f(0) S_T^{\frac{1}{2}} (\hat{\beta}_T - \beta_0) \xrightarrow{L} N(0, I_K)$$

where I_K denotes the identity matrix of order K.

Note: If (A.1)–(A.4) and (A.5*): $\lambda_{\min}(S_T)/\log T \to \infty$ hold, then

$$\lim_{T\to\infty} \hat{\beta}_T = \beta_0 \quad \text{in probability}\,.$$

For a proof of Theorem 3.21, the reader is referred to Rao and Zhao (1993).

3.18.3 Tests of Linear Hypotheses

We consider tests of linear hypotheses such as

$$H_0\colon H'(\beta - \beta_0) = 0 \quad \text{against} \quad H_1\colon H'(\beta - \beta_0) \neq 0\,, \tag{3.360}$$

where H is a known $K \times q$-matrix of rank q, and β_0 is a known K-vector $(0 < q < K)$. Let

$$\beta_T^* = \arg \inf_{H'(\beta-\beta_0)=0} \sum_{t=1}^{T} |(x_t'b)^+ - y_t^+|\,, \tag{3.361}$$

$$\hat{\beta}_T = \arg \inf_{b} \sum_{t=1}^{T} |(x_t'b)^+ - y_t^+|\,, \tag{3.362}$$

where all the infima are taken over $b \in \bar{B}$. Define the likelihood ratio, Wald and Rao's score statistics:

$$
\begin{aligned}
&M_T = \sum_{t=1}^{T} |(x_t'\beta_T^*)^+ - y_t^+| - \sum_{t=1}^{T} |(x_t'\hat{\beta}_T)^+ - y_t^+| \,, && (3.363)\\
&W_T(\hat{\beta}_T - \beta_0)'H(H'S_T^{-1}H)^{-1}H'(\hat{\beta}_T - \beta_0)\,, && (3.364)\\
&R_T = \xi(\beta_T^*)'S_T^{-1}\xi(\beta_T^*)\,, && (3.365)
\end{aligned}
$$

where S_T is as defined in (3.359) and

$$
\begin{aligned}
\xi(b) \quad &= \quad \sum_{t=1}^{T} I(x_i'b > 0)\,\mathrm{sgn}(x_t'b - y_t^+)x_t \\
&= \quad \sum_{t=1}^{T} I(x_t'b > 0)\,\mathrm{sgn}(x_t b - y_t)x_t \,.
\end{aligned}
$$

The main theorem concerning tests of significance is as follows, where we write

$$
x_{tT} = S_T^{-\frac{1}{2}}x_t\,, \quad H_T = S_T^{-\frac{1}{2}}H(H'S_T^{-1}H)^{-\frac{1}{2}}\,,
$$

$$
\sum_{t=1}^{T} I(\mu_t > 0)x_{tT}x_{tT}' = I_K\,, \quad H_T'H_T = I_q\,.
$$

Theorem 3.22 *Suppose that the assumptions (A.1)–(A.5) are satisfied. If β is the true parameter and H_0 holds, then each of $4f(0)M_T$, $4[f(0)]^2W_T$, and R_T can be expressed as*

$$
\left\| \sum_{t=1}^{T} I(\mu_t > 0)\,\mathrm{sgn}(e_t)H_T'x_{tT} \right\|^2 + o_K(1)\,. \tag{3.366}
$$

Consequently, $4f(0)M_T$, $4f(0)^2W_T$, and R_T have the same limiting chi-square distribution with the degrees of freedom q.

In order for the results of Theorem 3.22 to be useful in testing the hypothesis H_0 against H_1, some "consistent" estimates of S_T and $f(0)$ should be obtained. We say that $\hat{S}_T$ is a "consistent" estimate of the matrix S_T if

$$
S_T^{-\frac{1}{2}}\hat{S}_T S_T^{-\frac{1}{2}} \to I_K \quad \text{as} \quad T \to \infty\,. \tag{3.367}
$$

It is easily seen that

$$
\hat{S}_T = \sum_{t=1}^{T} I(x_t'\hat{\beta}_T > 0)x_t x_t'
$$

can be taken as an estimate of S_T. To estimate $f(0)$, we take $h = h_T > 0$ such that $h_T \to 0$ and use

$$\begin{aligned}\hat{f}_T(0) \quad = \quad & h \sum_{t=1}^{T} I(x_t'\hat{\beta}_T > 0)^{-1} \\ & \times \sum_{t=1}^{T} I(x_t'\hat{\beta}_T > 0)\, I(x_t'\beta_T < y_t^+ \leq x_t'\hat{\beta}_T + h) \quad (3.368)\end{aligned}$$

as an estimate of $f(0)$, which is similar to that suggested by Powell (1984). Substituting $\hat{S}_T$ for S_T and $\hat{f}_T$ for $f(0)$ in (3.363), (3.364), and (3.365), we denote the resulting statistics by $\hat{M}_T$, $\hat{W}_T$, and $\hat{R}_T$, respectively. Due to consistency of $\hat{S}_T$ and $\hat{f}_T(0)$, all the statistics

$$4\hat{f}_T(0)\hat{M}_T\,, \quad 4[\hat{f}_T(0)]^2\hat{W}_T\,, \quad \text{and} \quad \hat{R}_T \tag{3.369}$$

have the same asymptotic chi-square distribution on q degrees of freedom.

Note: It is interesting to observe that the nuisance parameter $f(0)$ does not appear in the definition of $\hat{R}_T$. We further note that

$$4\hat{f}_T(0)\hat{M}_T = 4[\hat{f}_T(0)]^2\hat{W}_T + o_K(1)\,, \tag{3.370}$$

and under the null hypothesis, the statistic

$$U_T = 4\left(\frac{\hat{M}_T}{\hat{W}_T}\right)^2 \hat{W}_T = 4\frac{\hat{M}_T^2}{\hat{W}_T} \overset{L}{\Rightarrow} \chi_q^2\,. \tag{3.371}$$

We can use U_T, which does not involve $f(0)$, to test H_0. It would be of interest to examine the relative efficiencies of these tests by Monte Carlo simulation studies.

3.19 Simultaneous Confidence Intervals

In the regression model

$$\underset{T\times 1}{y} = \underset{T\times K}{X}\ \underset{K\times T}{\beta} + \underset{T\times 1}{\epsilon}$$

with $\mathrm{E}(\epsilon) = 0, \mathrm{E}(\epsilon\epsilon') = \sigma^2 I$, the least squares estimator of β is $\hat{\beta} = (X'X)^{-1}X'y$ and $\mathrm{V}(\hat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2 H$ (say). To test the hypothesis $\beta = \beta_0$, we have seen that the test criterion is

$$F = \frac{(\hat{\beta} - \beta_0)H^{-1}(\hat{\beta} - \beta_0)}{Ks^2} \sim F_{K,T-K} \tag{3.372}$$

where $(T-K)s^2 = y'y - \hat{\beta}'X'y$, and $F_{K,T-K}$ is the F-statistic with K and $T-K$ degrees of freedom.

We give a characterization of the above F-test, which leads to the construction of Scheffé's simultaneous confidence intervals on linear functions of β. Consider a single linear function $l'\beta$ of β. The least squares estimator of $l'\beta$ is $l'\hat{\beta}$ with covariance $\sigma^2 l'Hl$. Then the t-statistic to test a hypothesis on $l'\beta$ is

$$t = \frac{l'\hat{\beta} - l'\beta}{\sqrt{s^2 l'Hl}}. \tag{3.373}$$

Now we choose l to maximize

$$t^2 = \frac{l'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'l}{s^2 l'Hl}. \tag{3.374}$$

Using the Cauchy-Schwarz inequality (see Theorem A.54), we see the maximum value of t^2 is

$$\frac{(\hat{\beta} - \beta)'H^{-1}(\hat{\beta} - \beta)}{s^2},$$

which is KF, where F is as defined in (3.372). Thus, we have

$$\frac{(\hat{\beta} - \beta)'H^{-1}(\hat{\beta} - \beta)}{Ks^2} = \frac{1}{Ks^2} \max_l \frac{l'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'l}{l'Hl} \sim F_{K,T-K}.$$

If $F_{1-\alpha}$ is the $(1 - \alpha)$ quantile of $F_{K,T-K}$, we have

$$P\left\{ \max_l \frac{|l'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'l|}{\sqrt{l'Hl}} \leq s\sqrt{KF_{1-\alpha}} \right\} = 1 - \alpha$$

that is,

$$P\left\{ \left|l'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'l\right| \leq s\sqrt{KF_{1-\alpha} l'Hl} \text{ for all } l \right\} = 1 - \alpha$$

or

$$P\left\{ l'\beta \in l'\hat{\beta} \pm s\sqrt{KF_{1-\alpha} l'Hl} \text{ for all } l \right\} = 1 - \alpha. \tag{3.375}$$

Equation (3.375) provides confidence intervals for all linear functions $l'\beta$. Then, as pointed out by Scheffé (1959),

$$P\left\{ l'\beta \in l'\hat{\beta} \pm s\sqrt{KF_{1-\alpha} l'Hl} \text{ for any given subset of } l \right\} \geq 1 - \alpha, \tag{3.376}$$

which ensures that the simultaneous confidence intervals for linear functions $l'\beta$ where l belongs to any set (finite or infinite) has a probability not less than $1 - \alpha$.

3.20 Confidence Interval for the Ratio of Two Linear Parametric Functions

Let $\theta_1 = P_1'\beta$ and $\theta_2 = P_2'\beta$ be two linear parametric functions and we wish to find a confidence interval of $\lambda = \frac{\theta_1}{\theta_2}$.

The least squares estimators of θ_1 and θ_2 are

$$\hat{\theta}_1 = P_1'\hat{\beta} \quad \text{and} \quad \hat{\theta}_2 = P_2'\hat{\beta}$$

with the variance-covariance matrix

$$\sigma^2 \begin{pmatrix} P_1'HP_1 & P_1'HP_2 \\ P_2'HP_1 & P_2'HP_2 \end{pmatrix} = \sigma^2 \begin{pmatrix} a & b \\ b' & c \end{pmatrix}, \text{ say.}$$

Then

$$\mathrm{E}(\hat{\theta}_1 - \lambda\hat{\theta}_2) = 0\,, \quad \mathrm{var}(\hat{\theta}_1 - \lambda\hat{\theta}_2) = \sigma^2(a - 2\lambda b + \lambda^2 c)\,.$$

Hence

$$F = \frac{(\hat{\theta}_1 - \lambda\hat{\theta}_2)^2}{s^2(a - 2\lambda b + \lambda^2 c)} \sim F_{1,T-K}$$

and

$$P\left\{(\hat{\theta}_1 - \lambda\hat{\theta}_2)^2 - F_{1-\alpha}s^2(a - 2\lambda b + \lambda^2 c) \le 0\right\} = 1 - \alpha\,. \tag{3.377}$$

The inequality within the brackets in (3.377) provides a $(1-\alpha)$ confidence region for λ. Because the expression in (3.377) is quadratic in λ, the confidence region is the interval between the roots of the quadratic equation or outside the interval, depending on the nature of the coefficients of the quadratic equation.

3.21 Nonparametric Regression

The nonparametric regression model describes the dependence of study variable on explanatory variables without specifying the function that relates them. The general nonparametric regression model is expressed as

$$y = \psi(X) + \epsilon \tag{3.378}$$

where y is a study variable, X is a vector of explanatory variables, ϵ is disturbance term and $\psi(x)$ is the unspecified real valued function of X at some fixed value x given by

$$\psi(x) = E(y|X = x)\,. \tag{3.379}$$

The first derivative of $\psi(x)$ indicates the response or regression coefficient of y with respect to x and the second derivative of $\psi(x)$ indicates the curvature of $\psi(x)$.

We assume that both y and $X_1, \ldots, X_K$ are stochastic and related as

$$y = \psi(X_1, \ldots, X_K) + \epsilon \tag{3.380}$$

with $\mathrm{E}(\epsilon|X = x) = 0$ and $\mathrm{V}(\epsilon|X = x) = \sigma^2 I$. We observe T identically and independently distributed observations $(y_t, x_{t1}, \ldots, x_{tK})$, $t = 1, \ldots, T$ from an absolutely continuous $(K + 1)$-variate distribution with density $f(y, X_1, \ldots, X_K) = f(y, X)$. If $\mathrm{E}(|y|) < \infty$, then the conditional mean of y given $X = x$ exists as in (3.379).

The regression coefficient related to x_j $(j = 1, \ldots, K)$ is

$$\beta_j(x) = \beta(x) = \frac{\partial\psi(x)}{\partial x_j} = \lim_{h\to 0} \frac{\psi(x+h) - \psi(x-h)}{2h} \tag{3.381}$$

where $\psi(x-h) = \psi(x_1, \ldots, x_j - h, \ldots, x_k)$. When $\psi(x)$ is linear, then $\beta_j(x)$ is the j^{th} regression coefficient and is fixed for all x. When $\psi(x)$ is non–linear, then $\beta_j(x)$ depends on x and $\beta_j(x)$ is a varying regression coefficient. The fixed regression coefficient can be defined as $\beta(\bar{x})$, *i.e.*, $\beta(x)$ evaluated at $x = \bar{x} = (\bar{x}_1, \ldots, \bar{x}_K)$. Similarly the second order partial derivative is

$$\beta^{(2)}(x) = \frac{\partial^2}{\partial x_j^2}\psi(x) = \lim_{h\to 0} \frac{\psi(x+2h) - 2\psi(x) - \psi(x-2h)}{(2h)^2}, \tag{3.382}$$

and, in general, the p^{th} order partial derivative $(p = 1, 2, \ldots)$ is

$$\begin{aligned}\beta^{(p)}(x) &= \frac{\partial^p}{\partial x_j^p}\psi(x) \\ &= \lim_{h\to 0}\left[\left(\frac{1}{2h}\right)^p \sum_{m=0}^{p}(-1)^m \binom{p}{m}\psi\Big(x + (p-2m)h\Big)\right]\end{aligned} \tag{3.383}$$

and the cross partial derivative is

$$\begin{aligned}&\frac{\partial^{p_1+\ldots+p_r}}{\partial x_{j1}^{p_1}, \ldots, \partial x_{jr}^{p_r}} = \\ &\quad\lim_{h\to 0}\left[\left(\frac{1}{2h}\right)^{p_1+\ldots+p_r} \sum_{m_1=0}^{p_1} \cdots \sum_{m_r=0}^{p_r} (-1)^{m_1+\ldots+m_r}\right. \\ &\quad\left.\times \binom{p_1}{m_1} \cdots \binom{p_r}{m_r}\psi\Big(x + (p_1 - 2m_1)h, \ldots, x + (p_r - 2m_r)h\Big)\right]\end{aligned} \tag{3.384}$$

respectively, where each of $j_1, \ldots, j_r = 1, \ldots, K\,(j_1 \neq \ldots \neq j_r)$, $x + (p_1 - 2m_1)h = (x_1, \ldots, x_{j1} + (p_1 - 2m_1)h, \ldots, x_K)$, $x + (p_r - 2m_r)h = (x_1, \ldots, x_{jr} + (p_r - 2m_r)h, \ldots, x_K)$.

Now we consider the nonparametric estimation of partial derivatives of $\psi(x)$ without specifying its form. We first consider a nonparametric estimator of $\psi(x)$ and then take its partial derivatives.

3.21.1 Estimation of the Regression Function

Most of the estimation methods of nonparametric regression assume that the regression function is smooth in some sense.

Since $\psi(x)$ depends on unknown densities, so we use the data of $(K+1)$ variables as $z_t = (x_t, y_t)$, $t = 1, \dots T$ and note that

$$g_T(z) = \frac{1}{Th^{K+1}} \sum_{t=1}^{T} \mathcal{K}\Big(\frac{z_t - z}{h}\Big) \,, \tag{3.385}$$

$$g_{1T}(x) = \int g_T(z) dy = \frac{1}{Th^K} \sum_{t=1}^{T} \mathcal{K}_1\Big(\frac{x_t - x}{h}\Big) \,, \tag{3.386}$$

where h is the window width (also called as band width or smoothing parameter) which is a positive function of T that goes to zero as $T \rightarrow \infty$, $\mathcal{K}$ is a kernel or a weight function such that $\int \mathcal{K}(z)dz = 1$ and $\mathcal{K}_1(x) = \int \mathcal{K}(z)dy$. The kernel $\mathcal{K}$ determines the shape of the curve and h determines their width. See, Prakasa-Rao (1983), Silverman (1986), Ullah and Vinod (1988) and Pagan and Ullah (1999) for the details on kernel density estimation. Substituting (3.385) and (3.386) in (3.379), we have

$$\psi_T(x) = \psi_T = \int y \frac{g_T(z)}{g_{1T}(x)} dy = \sum_{t=1}^{T} y_t w_t(x) \tag{3.387}$$

where

$$w_t(x) = \frac{\mathcal{K}_1\Big(\frac{x_t - x}{h}\Big)}{\sum_{t=1}^{T} \mathcal{K}_1\Big(\frac{x_t - x}{h}\Big)} \,. \tag{3.388}$$

The estimator ψ_T of ψ in (3.387) is known as Nadaraya–Watson type estimator due to Nadaraya (1964) and Watson (1964) and is a kernel nonparametric regression estimate. Note that (3.387) is a weighted average of the observed values y_t where the weight of t^{th} observation depends on the distance x_t to x through the kernel $\mathcal{K}$. The fitted nonparametric regression model that is obtained by without making any assumption about the functional form of $\psi(x)$ is

$$y = \psi_T(x) + \hat{\epsilon} \tag{3.389}$$

where $\hat{\epsilon}$ is the nonparametric residual.

This estimator (3.387) is also the weighted least squares estimator of $\psi(x)$ because $\psi_T(x)$ is the value of $\psi(x)$ for which the weighted squared error

$$\sum_{t=1}^{T} \mathcal{K}\Big(\frac{x_t - x}{h}\Big)(y_t - \psi(x))^2$$

is minimum. The method of moments also yields the same estimator of $\psi(x)$ as in (3.387).

When the window width h is not the same for all data points, then some alternative estimators of $\psi(x)$ are suggested. The recursive regression estimator of $\psi(x)$ in such a case is

$$\hat{\psi}_T(x) = \frac{\displaystyle\sum_{t=1}^{T} \frac{y_t}{h_t^K}\mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)}{\displaystyle\sum_{t=1}^{T} \frac{1}{h_t^K}\mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)} \tag{3.390}$$

where h_t denotes a sequence of positive numbers, assumed to satisfy $\sum h_t^K \to \infty$ as $T \to \infty$. An alternative estimator is

$$\tilde{\psi}_T(x) = \frac{\displaystyle\sum_{t=1}^{T} y_t\,\mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)}{\displaystyle\sum_{t=1}^{T} \mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)}\,. \tag{3.391}$$

Both (3.390) and (3.391) are recursive as

$$\hat{\psi}_T(x) = \hat{\psi}_{T-1}(x) + \frac{y_T - \hat{\psi}_{T-1}(x)}{1 + (T-1)\frac{\hat{f}_{T-1}(x)}{h_T^K}\mathcal{K}\Big(\frac{x_T - x}{h_T}\Big)} \tag{3.392}$$

$$\tilde{\psi}_T(x) = \tilde{\psi}_{T-1}(x) + \vartheta_T^{-1}\left[y_T - \tilde{\psi}_{T-1}(x)\mathcal{K}\left(\frac{x_T - x}{h_T}\right)\right] \tag{3.393}$$

where

$$\vartheta_T = \vartheta_{T-1} + \mathcal{K}\Big(\frac{x_T - x}{h_T}\Big),\ \vartheta_0 = 0\,.$$

Both (3.392) and (3.393) can be updated as additional data points are available.

When ϵ's are such that $\mathrm{V}(\epsilon) = \Sigma$ $(\neq \sigma^2 I)$, a $T \times T$ positive definite matrix, then the generalized least squares estimator of $\psi(x)$ is obtained by minimizing $\epsilon' \mathcal{K}^{1/2}\Sigma^{-1}\mathcal{K}^{1/2}\epsilon$ with respect to $\psi(x)$ as

$$\psi_T^* = (\mathbf{1}'\mathcal{K}^{\frac{1}{2}}\Sigma^{-1}\mathcal{K}^{\frac{1}{2}}\mathbf{1})^{-1}\mathbf{1}'\mathcal{K}^{\frac{1}{2}}\Sigma^{-1}y \tag{3.394}$$

where $\mathbf{1} = (1, \ldots, 1)'$, $\mathcal{K} = \mathrm{diag}(\mathcal{K}_1, \ldots, \mathcal{K}_T)$ is a diagonal matrix with $\mathcal{K}_T = \mathcal{K}\Big(\frac{x_t - x}{h}\Big)$.

An operational version of a consistent estimator of $\beta(x)$ in (3.381) is

$$b_T(x) = \frac{\psi_T(x+h) - \psi_T(x-h)}{2h} \tag{3.395}$$

where $\psi_T(x)$ is given by (3.387). Similarly, the estimators of the p^{th} order partial derivative and cross partial derivatives can be obtained by replacing $\psi(\cdot)$ with $\psi_T(\cdot)$ in (3.383) and (3.384), respectively.

Ullah and Vinod (1988) analytically derived the estimator of $\beta(x) = \partial\psi(x)/\partial x$ as

$$\hat{\beta}_T(x) = \frac{\partial}{\partial x}\psi_T(x) = \sum_{t=1}^{T} y_t(\omega_{1t} - \omega_{2t})$$

where

$$\omega_{1t} = \frac{\mathcal{K}'\left(\frac{x_t - x}{h}\right)}{\sum_{t=1}^{T} \mathcal{K}\left(\frac{x_t - x}{h}\right)}$$

and $\omega_{2t} = \omega_t(x)\sum \omega_{1t}$; $\omega_t(x)$ is as in (3.388) and

$$\mathcal{K}'\left(\frac{x_t - x}{h}\right) = \frac{\partial}{\partial x_j}\mathcal{K}\left(\frac{x_t - x}{h}\right).$$

Alternatively, $\hat{\beta}_T(x)$ and its generalization for p^{th} order derivatives of $\psi(x)$, $\hat{\beta}^{(p)}(x)$ can be obtained as a solution of

$$\sum_{m=o}^{p} \binom{p}{m} \beta_T^{(m)}(x) f_T^{(p-1)}(x) = g_T^{(p)}(x), \quad (p = 1, 2, \ldots)$$

where $g^{(p)}(x)$ is the p^{th} order partial derivative of $g(x) = \int yf(y,x)dy$ with respect to x_j.

The restricted least squares estimator of $\psi(x)$ under the exact linear restrictions $R\beta(x) = r$ is

$$\hat{\beta}_T(x) = b_T(x) - R'(RR')^{-1}[Rb_T(x) - r] \tag{3.396}$$

where $b_T(x)$ is given by (3.395).

The regression function can also be estimated by using various nonparametric procedures like nearest neighbor kernel estimation, local polynomial regression, and smoothing splines.

The method of nearest neighborhood kernel estimation is based on defining a symmetric unimodal weight function $W(x)$ which is centered on the focal observation and goes to zero at the boundaries of the neighborhood around the focal value. Let x_{fo} be a focal x-value at which $\psi(x)$ is to be estimated. Now find ν nearest x-neighbors of x_{fo} where ν/x_{fo} is the span of the kernel smoother. The larger the span, smoother is the estimated regression function. Using the weights defined by $W(x)$, calculate the weighted average of y and obtain the fitted value

$$\hat{y}_{fo} = \hat{f}(x_{fo}) = \frac{\sum_{t=1}^{T} y_t W(x_t)}{\sum_{t=1}^{T} W(x_t)}. \tag{3.397}$$

Repetition of this procedure at a range of x-values spanning the data and connecting the fitted values produces an estimate of the regression function.

In local polynomial regression, the fitted values are produced by locally weighted regression rather than by locally weighted averaging. Another method of nonparametric regression is smoothing splines which are the solution to the penalized regression problem. Additive regression models are an alternate to nonparametric regression with several explanatory variables.

The readers are referred to Prakasa-Rao (1983), Silverman (1986), Ullah (1989a), Ullah (1989b), Härdle (1990) and Pagan and Ullah (1999) for the asymptotic properties of the estimators, related testing of hypothesis and other aspects on nonparametric regression.

3.22 Classification and Regression Trees (CART)

Nonparametric regression with multiple explanatory variables suffers from the problem of *curse of dimensionality.* This means that if the number of explanatory variables is high, then it may be difficult to catch the relevant features of the problem in hand, e.g. the influence of interactions of explanatory variables on the study variable may be difficult to study. We have only a finite sample available, but there may be big volumes in the space of explanatory variables where there may be no observation or only a few observations are obtained (*sparseness problem*). Therefore a reliable statistical estimation is not possible in these volumes. Parametric models, such as simple linear models, or additive models as proposed by Hastie and Tibshirani (1990) try to catch at least the *main effects* of the explanatory variables and discard any global or local interactions of the explanatory variables on the study variable. Furthermore, the results from the approaches like Projection Pursuit Regression (see Section 3.24) or Neural Networks (see Section 3.25) may be hard to interpret. In such situations, CART is more useful. CART tries to catch the relevant interactions of the explanatory variables in their influence on the study variable and present the results in a simple way.

Consider a general regression setup in which the study variable y is either real-valued or categorical and $X_1, \dots, X_K$ are the explanatory variables. In the usual nonparametric regression setup, we assume

$$y = \psi(X_1, \dots, X_K) + \epsilon \ , \tag{3.398}$$

with $\mathrm{E}(\epsilon|X) = 0$. If the function ψ is unknown and not parameterized by a finite dimensional parameter, Breiman, H., Olshen and Stone (1984) suggested a recursive partitioning algorithm of the covariate space which results in a tree structure. If y is real-valued, as in (3.398), then the resulting tree is called as a regression tree. If y is e.g., binary, then we may assume

$$\log \frac{P(y=1)}{1-P(y=1)} = \psi(X_1, \ldots, X_K) \ , \tag{3.399}$$

where the left hand side of (3.399) is the logit transformation which is used in logistic regression. In this case, the tree is called as a classifcation tree. The idea is to recursively partition the covariate space using binary splits of the form $\xi_1 = \{x : X_j \leq x\}$ (left node) and $\xi_2 = \{x : X_j > x\}$ (right node), where $j \in \{1, \ldots, K\}$ is the actual chosen splitting variable. The splitting criteria depends on the situation (regression or classification) and measurement scale of the explanatory variables $X_1, \ldots, X_K$. The main objective is to find a split such that the response values are as homogeneous as possible within the splitting sets ξ_1 and ξ_2; and as heterogeneous as possible between the two sets. Then the binary partitioning proceeds to each of the sets ξ_1 and ξ_2. The left node ξ_1 and the right node ξ_2 are again partitioned into left and right nodes using the same or another splitting variable, and so on. This leads to a partitioning of the space of explanatory variables into rectangular regions. At the final stage in a classical regression tree, the responses of all cases in the leaf nodes of the tree are averaged after the tree has grown until a certain stopping criteria is fulfilled. The leaf nodes are the final nodes in the tree where no further splitting is sensible according to a chosen criterion. The average can, e.g., be the sample mean or sample median. This leads to a piecewise constant regression function. Typical stopping criteria for a specific node are when the number of cases in a leaf node would become lower than a predetermined number n_0 or when certain p-values in the splitting criteria are greater than a predetermined p-value p_0. Alternatively, a tree may be grown to a high complexity and pruned afterwards using some cost-complexity measure and cross-validation as proposed in Breiman et al. (1984).

In the following we focus on regression trees. Chaudhuri, Huang, Loh, and Yao (1994), Chaudhuri, Lo, Loh and Yang (1995), Loh (2002) and Kim, Loh, Shih and Chaudhuri (2007) have presented many extensions to the original regression tree procedure of Breiman et al. (1984). For illustration, we present an approach called GUIDE (Generalized, Unbiased, Interaction Detection and Estimation) proposed by Loh (2002) and extended in Kim et al. (2007). Such an approach also detects the local pairwise interactions among explanatory variables.

Algorithm of GUIDE

1. Let s denote the current node. Use stepwise regression to find two *quantitative* explanatory variables to fit a linear model to the data in s.

2. Do not split the node if its model $R^2 > 0.99$ or if the number of observations is less than $2n_0$, where n_0 is a small user-specified constant. Otherwise go to step 3.

3. For each observation, define the class variable $Z = 1$ if it is associated with a positive residual. Otherwise, define $Z = 0$.

4. For each explanatory variable X_j, $j = 1, \ldots, K$:

   (a) Construct a $2 \times m$ contingency table. The rows are formed by the values of Z (0 or 1). If X_j is a categorical variable, its values define the columns, *i.e.*, m is the number of distinct categories of X_j. If X_j is quantitative, its values are grouped into four intervals at the sample quartiles and the four intervals constitute the columns, *i.e.*, $m = 4$.

   (b) Compute the significance probability of the χ^2-test of association between the rows and columns of the table.

5. Select the explanatory variable X_j with the smallest significance probability to split s. Let s_L and s_R denote the left and right sub-nodes of s.

   (a) If X_j is quantitative, search for a split of the form $X_j \leq x$. For each chosen x, both s_L and s_R should contain at least n_0 observations:

      i. Use stepwise regression to choose two quantitative explanatory variables to fit a model with two explanatory variables to each of the data sets in s_L and s_R.

      ii. Compute S, the total sum of squared residuals in s_L and s_R. Select the smallest value of x that minimizes S.

   (b) If X_j is categorical, then search for a split of the form $X_j \in C$, where C is a subset of the values taken by X_j. For every C such that each of the s_L and s_R has at least n_0 observations, calculate the sample variance of Z in s_L and s_R. Choose the set C for which the weighted sum of the variances is minimum, with weights proportional to sample sizes.

6. After splitting is stopped, prune the tree as described in Breiman et al. (1984) with ten-fold cross-validation. Let E_0 be the smallest cross-validation estimate of prediction mean square error (PMSE) and let α be a positive number. Select the smallest subtree whose cross-validation estimate is within α times the standard error of E_0. They use the default value of $\alpha = 0.5$ and call it '0.5-SE rule'. Truncate all predicted values to avoid large prediction errors caused by extrapolation, so that they lie within the range of the training sample values in their nodes.

Remark 1. A modification in the algorithm is to use only one quantitative explanatory variable in step one. A very important and desirable consequence of the proposed algorithm is that conditional unbiasedness in the selection of split variable is achieved. Unbiasedness means that if the explanatory variables are statistically independent of the study variable then each of the explanatory variable has the same chance of being selected. This does not hold, e.g., for CART.

Remark 2. In Chaudhuri et al. (1995), a multiple linear model instead of a two predictor model was allowed for a Poisson response model in the step 1 of the GUIDE algorithm. In step 2, instead of ordinary residuals, the adjusted Anscombe residuals were used. In step 4, a t-test on the ungrouped explanatory variable X_j was used. But this is only meaningful if X_j is a quantitative variable. In step 5 the explanatory variable selected for a split is the one with the largest absolute t-statistic.

Remark 3. Bayesian CART approaches using similar ideas and are proposed by Denison, Mallick and Smith (1998), Chipman, George and McCulloch (1998; 2002).

The advantages of tree-structured approach can be summarized in the following statements:

- The tree structure handles most of the complexities like interactions of explanatory variables, nonlinear influence of explanatory variables on study variable. The models in each partition represented by going from the root node to the leafs can be kept at a low order and can therefore interpreted easily.

- Interactions among explanatory variables are made visible by the structure of the decision tree. Local interactions can be included, as in the approach of Kim et al. (2007).

- Theoretical consistency results can be obtained, see, e.g. Chaudhuri et al. (1995).

Some disadvantages are:

- The recursive structure can be dangerous, as bad splits in upper nodes near to the root node or at the root node itself can lead to bad results in nodes that are near to the leaf nodes.

- To avoid over-fitting and over-complexity of the tree, no unique best strategy exists.

- Many algorithms are like black boxes and it is often very difficult to find out what splitting and fitting criteria are used. This may change as Open Source programs like R and algorithms implemented in this language allow a deeper insight into the detailed estimation processes.

- In the original GUIDE algorithms proposed by Loh (2002), bootstrap aggregating (bagging) is needed to avoid variable selection bias. This increases the black-box factor of the method. Whether biasedness can be avoided without bagging in the GUIDE algorithm is not clear from the available literature but it seems so.

3.23 Boosting and Bagging

As in Section 3.22, we consider a general regression setup in which the study variable y is either real-valued or categorical and $X = (X_1, \dots, X_K)$ is the matrix of observations on K explanatory variables. In the usual regression setup, we assume

$$y = \psi(X_1, \dots, X_K) + \epsilon = \psi(X) + \epsilon \ , \tag{3.400}$$

with $\mathrm{E}(\epsilon|X) = 0$. The function ψ is unknown and not parameterized by a finite dimensional parameter. Bagging (Breiman, 1996) and boosting (Freund and Schapire, 1996) were recently proposed in the context of machine learning, where the main objective is to improve upon the accuracy in classification problems. An earlier reference on boosting is Drucker, Schapire and Simard (1993). As stated by Borra and Di Ciaccio (2002), less attention was, at least initially, paid to nonparametric regression methods. A training data set $\mathcal{T}$ with T cases (y_t, x_t'), $t = 1, \dots, T$ is drawn from the population. Now $\psi(X)$ is to be estimated by an approximating function $\eta(X)$ for a given $\mathcal{T}$. For example, regression trees (see Section 3.22) are potential candidates for η. The prediction capability of an approximation η based on a training sample $\mathcal{T}$, $\eta(x|\mathcal{T})$ is defined by the prediction mean squared error

$$\mathrm{MSEP}(\eta|\mathcal{T}) = \mathrm{M}_{y,X}(y - \eta(x|\mathcal{T}))^2 \ , \tag{3.401}$$

where $\mathrm{M}_{y,X}$ denotes the average over *all* values (y, X) in the population (and not only over the observed ones in the sample or over the values of an additional test data set). Then obtain the average mean squared generalization error $\overline{\mathrm{MSEP}(\eta)}$ that is an average over all training samples of the same size T which are drawn from the *same* population. Since it is usually impossible to calculate the population versions, these quantities are estimated by evaluating η on an additional test data set. The so called *poor man's* algorithms use further methods like splitting the training sample into a smaller training set and a (pseudo) test data set or cross-validation.

The bootstrap aggregating (bagging) procedure proposed by Breiman (1996) tries to improve the prediction capability by combining S approximating functions $\eta_s(X)$, $s = 1, \dots, S$, which are calculated from a set of S bootstrap samples of the training data set. The improvement is often recognized by a reduction in variance maintaining almost constant bias and thus reducing the prediction mean squared error. We refer to Efron and Tibshirani (1993) for the bootstrap methodology and its ability to estimate population quantities of the real world from the bootstrap world based on one random sample of the population. Usually, the bootstrap samples are drawn with replacement from the initial sample and with the original sample size T. The bagging predictor function is then obtained by averaging:

$$\eta(X)_{\text{bagging}} = \frac{1}{S}\sum_{s=1}^{S} \eta_s(X) \ . \tag{3.402}$$

It can be shown that when the bootstrap samples are drawn with replacement, then for reasonable sample sizes, on an average, approximately 37% cases of the training sample do not appear in a *particular* bootstrap sample. These so-called *out of bag* samples can be used as test cases to construct the improved predictor functions without the need of additional computation as in cross-validation. We conclude that bagging procedures are, at least conceptually, simple approaches to improve nonparametric regression methods. They rely on the following principles:

- Choose a (nonparametric) regression procedure (e.g., regression trees, multivariate adaptive regression splines)
- Create S bootstrap samples (with replacement in the original version) of the training data set. For each of the S samples, apply the chosen nonparametric regression procedure.
- Model averaging. Improvements are possible by using out-of-bag samples for each of the S bootstrap samples.

The boosting method proposed by Freund and Schapire (1996) is more elaborate as it tries to weight sequentially the cases of the training data set depending on the quality of their prediction. If the t^{th} observation is not well predicted by $\eta_s(X)$, then it will obtain a greater weight for learning $\eta_{s+1}(X)$. The aggregated predictor is then a linear combination of $\eta_s(X)$, which is weighted according to the quality of prediction of the training data. As with bagging, boosting was introduced in the context of machine learning and its main focus was on classification problems. Drucker (1997) considers a modification of the so-called *AdaBoost* algorithm for regression. For illustration of such algorithms, we give the algorithm presented by Drucker (1997).

Algorithm:

Consider a training sample of size T. Initially, each case in the sample receives weight $w_t^{(1)} = 1$, $t = 1, \ldots, T$.

Set $s = 1$.

Repeat the following steps while the average loss $\bar{L}$ (defined in step 5) is less than 0.5.

1. Extract T cases with replacement to form a training set. Case t is included in the set with probability $p_t^{(s)}$ proportional to its actual weight $w_t^{(s)}$, *i.e.*, $p_t^{(s)} = w_t^{(s)} / \sum_{t=1}^T w_t^{(s)}$.
2. Fit an approximating function $\eta_{(s)}(X)$ to that training set.
3. Obtain the predictions $\eta_{(s)}(x_t)$ of each case in the training set.
4. Calculate a loss L_t for each training case $t = 1, \ldots, T$. The loss may be of any functional form as long as $L \in [0, 1]$. Let $D = \max|y_t - \eta_{(s)}(x_t)|$, $t = 1, \ldots, T$. Three candidate loss functions are the linear, quadratic and exponential loss functions:
$$L_t^{(s)} = \frac{|y_t - \eta_{(s)}(x_t)|}{D} \quad (3.403)$$
$$L_t^{(s)} = \frac{|y_t - \eta_{(s)}(x_t)|^2}{D^2} \quad (3.404)$$
$$L_t^{(s)} = 1 - \exp\left\{ \frac{-|y_t - \eta_{(s)}(x_t)|}{D} \right\} . \quad (3.405)$$
5. Calculate an average loss: $\bar{L}^{(s)} = \sum_{t=1}^T p_t^{(s)} L_t^{(s)}$.
6. Form the coefficient $\beta^{(s)} = \bar{L}^{(s)}/(1 - \bar{L}^{(s)})$ where $\beta^{(s)}$ measures the confidence in predictor $\eta_{(s)}(x)$. Lower value of $\beta^{(s)}$ indicates higher confidence in the predictor.
7. Update the weight $w_t^{(s)}$ by $w_t^{(s+1)} := w_t^{(s)} {\beta^{(s)}}^{[1-\bar{L}^{(s)}]}$. The smaller the loss, the more the weight is reduced by making the probability smaller that this pattern will be picked as a member of the training set for the next predictor $\eta_{s+1}(x)$.
8. Set $s := s + 1$.

Let S predictors $\eta_{(s)}(x)$, $s = 1, \ldots, S$ are available after running the above algorithm S times. For each particular input x_t, one gets S predictions $\eta_{(s)}(x_t)$. Drucker (1997) proposed the weighted median as cumulative prediction for the case or pattern x_t:

$$h_f(x_t) = \inf\left\{ y : \sum_{s: h_{(s)}(x_t) \le y} \log(1/\beta^{(s)}) \ge \frac{1}{2} \sum_{s=1}^{S} \log(1/\beta^{(s)}) \right\} . \quad (3.406)$$

Relabel the S predictions $h_{(s)}(x_t)$, such that

$$h_{(1)}(x_t) < h_{(2)}(x_t) < \cdots < h_{(S)}(x_t) \quad (3.407)$$

and retain the associations of the $\beta^{(s)}$ with the predictions $h_{(s)}(x_t)$. Then sum the $\log(1/\beta^{(s)})$ until we reach the smallest s so that the inequality (3.406) is satisfied. The prediction from the predictor s is taken as ensemble prediction. If all $\beta^{(s)}$ are equal, then this would be the median.

Further Readings: Random forests were introduced by Breiman (2001). They combine bagging with random subspace methods by choosing several variables randomly for each node of the tree on which the decision at that node is based upon. Each tree is fully grown and not pruned (as may be done in constructing a normal regression tree or a tree classifier). Random forests can be used for classification and regression. Recently Strobl, Boulesteix, Zeileis and Hothorn (2007) have shown that random forest variable importance measures are a sensible means for variable selection in many applications, but are not reliable in situations where potential explanatory variables vary in their scale of measurement or their number of categories. They propose to use the sampling without replacement to overcome this problem.

Bühlmann and Yu (2002) give theoretical results on bagging and a further technique which they call as subagging. Bühlmann and Yu (2003) use boosting with the squared error loss function. Tutz and Reithinger (2006) use boosting in the context of semiparametric mixed models (models with random effects), Tutz and Leitenstorfer (2007) apply boosting to monotone regression in the context of additive models and Leitenstorfer and Tutz (2007) apply it in the context of estimating smooth functions.

3.24 Projection Pursuit Regression

The term *projection pursuit* (Friedman and Tukey, 1974) describes a technique for the exploratory analysis of multivariate data. This method searches for interesting linear projections of a multivariate data set onto a linear subspace, such as, for example, a plane or a line. These low-dimensional orthogonal projections are used to reveal the structure of the high-dimensional data set.

Projection pursuit regression (PPR) constructs a model for the regression surface $y = f(X)$ using projections of the data onto planes that are spanned by the variable y and a linear projection $a'X$ of the independent variables in the direction of the vector a. Then one may define a function of merit (Friedman and Stuetzle, 1981) or a projection index (Friedman and Tukey, 1974; Jones and Sibson, 1987) $I(a)$ depending on a. Projection pursuit attempts to find directions a that give (local) optima of $I(a)$. The case $a = 0$ is excluded, and a is constrained to be of unit length (*i.e.*, any a is scaled by dividing by its length).

In linear regression the response surface is assumed to have a known functional form whose parameters have to be estimated based on a sample (y_t, x'_t). The PPR procedure models the regression surface iteratively as a sum of smooth functions of linear combinations $a'X$ of the predictors, that is, the regression surface is approximated by a sum of smooth functions

$$\phi(x) = \sum_{h=1}^{H} S_{a_h}(a_h' X) \tag{3.408}$$

(H is the counter of the runs of iteration). The algorithm is as follows (Friedman and Stuetzle, 1981):

(i) Collect a sample (y_t, x_t'), $t = 1, \ldots, T$, and assume the y_t to be centered.

(ii) Initialize residuals $r_t = y_t$, $t = 1, \ldots, T$ and set counter $H = 0$.

(iii) Choose a vector a and project the predictor variables onto one dimension $z_t = a' x_t$, $t = 1, \ldots, T$, and calculate a univariate nonparametric regression $S_a(a' x_t)$ of current residuals r_t on z_t as ordered in ascending values of z_t. These nonparametric functions are based on local averaging such as

$$S(z_t) = \text{AVE}(y_i)\,, \quad j - k \leq i \leq j + k\,, \tag{3.409}$$

where k defines the bandwidth of the smoother.

(iv) Define as a function of merit $I(a)$, for example, the fraction of unexplained variance

$$I(a) = 1 - \sum_{t=1}^{T} \frac{(r_t - S_a(a' x_t))^2}{\sum_{t=1}^{T} r_t^2}\,. \tag{3.410}$$

(v) Optimize $I(a)$ over the direction a.

(vi) Stop if $I(a) \leq \epsilon$ (a given lower bound of smoothness). Otherwise update as follows:

$$\begin{aligned} r_t &\leftarrow r_t - S_a^H(a_H' x_t)\,, \quad t = 1, \ldots, T\,, \\ H &\leftarrow H + 1\,. \end{aligned} \tag{3.411}$$

Interpretation: The PPR algorithm may be seen to be a successive refinement of smoothing the response surface by adding the optimal smoother $S_a^H(a' X)$ to the current model.

Remark: Huber (1985) and Jones and Sibson (1987) have included projection pursuit regression in a general survey of attempts at getting *interesting* projections of high-dimensional data and nonparametric fittings such as principal components, multidimensional scaling, nonparametric regression, and density estimation.

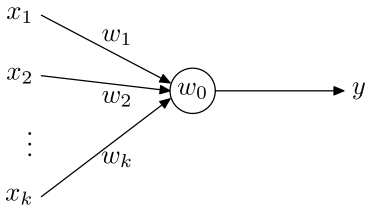

FIGURE 3.10. A single-unit perceptron

3.25 Neural Networks and Nonparametric Regression

The simplest feed-forward neural network is the so-called single-unit perceptron displayed in Figure 3.10. This perceptron consists of K input units $x_1, \dots, x_K$ and one output unit y. The input values x_i are weighted with weights w_i $(i = 1, \dots, K)$ so that the expected response y is related to the vector $x = (x_1, \dots, x_K)$ of covariates according to

$$y = w_0 + \sum_{i=1}^{K} w_i x_i \,. \tag{3.412}$$

In general, neural networks are mathematical models representing a system of interlinked computational units. Perceptrons have strong association with regression and discriminant analysis. Unsupervised networks are used for pattern classification and pattern recognition. An excellent overview on neural networks in statistics may be found in Cheng and Titterington (1994). In general, the input-output relationship at a neuron may be written as

$$y = f(x, w) \tag{3.413}$$

where $f(\cdot)$ is a known function. $f(\cdot)$ is called the activation function. Assume that we have observations $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$ of n individuals in a so-called training sample. Then the vector of weights w has to be determined such that the so-called energy or learning function

$$\mathrm{E}(w) = \sum_{j=1}^{n} \left(y^{(j)} - f(x^{(j)}, w)\right)^2 \tag{3.414}$$

is minimized with respect to w. This is just a least squares problem. To find the weight $\hat{w}$ minimizing $\mathrm{E}(w)$ we have to solve the following system of estimation equations $(k = 0, \dots, K)$

$$\frac{\partial\, \mathrm{E}(w)}{\partial w_k} = \sum_{j=1}^{n} (y^{(j)} - f(x^{(j)}, w)) \frac{\partial f(x^j, w)}{\partial w_k} = 0. \tag{3.415}$$

In practice, numerical methods are used to minimize $\mathrm{E}(w)$. Well-known techniques that have been implemented are the generalized delta rule or error back-propagation (Rumelhart, Hinton and Williams, 1986), gradient methods such as the method of steepest descent (Thisted, 1988), genetic algorithms, Newton-Raphson algorithms, and variants of them.

If a multilayer perceptron is considered, then it may be interpreted as a system of nonlinear regression functions that is estimated by optimizing some measure of fit. Recent developments in this field are the projection-pursuit regression (see Section 3.24) and its modifications by (Tibshirani, 1992) using so-called slide functions, and the generalized additive models (see Hastie and Tibshirani (1990)).

During the last five years a lot of publications have demonstrated the successful application of neural networks to problems of practical relevance. Among them, in the field of medicine the analysis based on a logistic regression model (see Section 10.3.1) is of special interest.

3.26 Logistic Regression and Neural Networks

Let y be a binary outcome variable and $x = (x_1, \ldots, x_K)$ a vector of covariates. As activation function $f(.)$ we choose the logistic function $l(v) = \exp(v)/(1 + \exp(v))$. Then the so-called logistic perceptron $y = l(w_0 + \sum_{i=1}^K w_i x_i)$ is modeling the relationship between y and x. The estimation equations (3.415) become (Schumacher, Roßner and Vach, 1996)

$$\frac{\partial\, \mathrm{E}(w)}{\partial w_k} = \sum_{j=1}^{n} 2 f(x^{(j)}, w)(1 - f(x^{(j)}, w)) x_k^{(j)} (y^{(j)} - f(x^{(j)}, w)) = 0. \quad (3.416)$$

For solving (3.416), the least-squares back-propagation method (Rumelhart et al., 1986) is used. It is defined by

$$\hat{w}^{(v+1)} = \hat{w}^{(v)} - \eta \partial\, \mathrm{E}(\hat{w}^{(v)})$$

for $v = 0, 1, \ldots$ and $\hat{w}^{(0)}$ a chosen starting value. The positive constant η is called the learning rate. The nonlinear model of the logistic perceptron is identical to the logistic regression model (10.61) so that the weights w can be interpreted like the regression coefficients β. (For further discussion, see Vach, Schumacher and Roßner (1996).)

3.27 Functional Data Analysis (FDA)

The occurrence and availability of high frequency data is common in a number of applied sciences like chemometrics, biometrics, econometrics and medicine. Examples of high frequency data are those variables which are measured repeatedly very often (perhaps hundreds or thousands of measurements on them). Generally the time intervals between two measurements are small and equally spaced. A reasonable assumption in such cases then is that the consecutive measurements are similar and therefore highly correlated. Thus the observed data, which are still discrete, may be assumed as observations from a (continuous) curve or a function at discrete time points. The FDA is not only used in the time series context but in other contexts also. A frequently quoted example of this type are spectrometric curves. The time axis is replaced by the wavelengths in a certain range. The measured variable is, e.g., absorbance at hundred different wavelengths. A thorough introduction to functional data analysis is given by Ramsay and Silverman (2002; 2005). An extension to nonparametric functional data analysis is discussed in Ferraty and Vieu (2006).

Let us denote one observation of a curve by a vector x_t. The whole collection of curves can be summarized in a matrix X. If, e.g., a variable is measured 512 times for 15 individuals, then x_t' is a (1×512)-vector and X is a (15×512) matrix. Ferraty and Vieu (2006) present a mathematical and statistical framework for these type of problems. The vector $x_t' = (x_t(j_1), \ldots, x_t(j_K))$ is assumed to be a discretized version of the curve $\chi_t = \{\chi_t(j) : j \in \mathcal{J}\}$ measured at K different points $j_1, \ldots, j_K$.

The functional data analysis can be used in various type of situations. Some are as follows:

1. Simple collection X of curves. Then we may be interested in estimating the mean curve or in studying the features of curves. The features of curve include the first, second and higher order derivatives.

2. A labeled collection X of curves. Additional to each curve, we observe, e.g., a binary variable y which denotes whether the curve belongs to, e.g., a healthy person ($y = 0$) or to a diseased person ($y = 1$). We may be interested in knowing whether the curves can be used to classify the persons, or in other words whether curves of healthy persons differ from the curves of diseased persons or not. In general, we observe discretized versions of curves together with a univariate or multivariate study variable y. Our interest lies in a regression of y on the curves. Marx and Eilers (1999) use the term *signal regression* for such data situations. The term *functional regression* is also used in the literature.

3. Beyond a sample of curves X, a functional study variable y is observed. The whole collection of response curves can then be summarized in a matrix Y. The objective may be in knowing how the x-curves influence the y-curves.

4. Only the study variable is a functional variable. The explanatory variables can be the usual nonfunctional variables. If the explanatory variables are factors, then we have an analog to ANOVA with a

difference that y is now a functional variable instead of a univariate random variable.

A typical feature of the collection of curves matrix X is that the number of columns (which denotes the number of repeated measurements on the functional variable) exceeds the number of rows (which denotes the number of individuals on which the curves have been observed). Therefore, e.g., a linear regression of a univariate study variable y on X is not possible without modifications. In principle, one has to look at the following problems:

- How to measure the closeness of two curves? Ferraty and Vieu (2006) propose to consider semi-metrics as a closeness measure. Such metrics can be build by using extensions of Principal Components Analysis (PCA) such as Functional Principal Components Analysis (FPCA), see Ferraty and Vieu (2006) for more details on FPCA. Another approach is Partial Least Squares (PLS) regression, see Section 3.14.4 for more details.

- Consider the situation of signal or functional regression (situation 3). Then if the error structure is assumed to be approximately normal, we can assume a linear regression model

  $$y = \beta_0 + X\beta + \epsilon \ , \tag{3.417}$$

  or, if y is, e.g., a binary or count variable, then a generalized linear model can be assumed. But since X, in general, may have more columns than rows, then OLS estimate can not be obtained. Additionally, by the construction of X, the columns may also be highly correlated. Therefore some dimension reduction is necessary in such situations.

Marx and Eilers (1999) propose a solution which is different from PCA and PLS and is based on a P-spline (penalized splines) approach. The basic idea is to constrain β in a way so that it becomes smooth. Based on B-splines, β is written as $B\alpha$ such that the model equation becomes

$$y = \beta_0 + XB\alpha + \epsilon \tag{3.418}$$

where α is another parameter with lower dimension than β. The (e.g., cubic) B-spline matrix B is constructed such that XB now has full column rank. Further smoothing is achieved by putting difference penalties on the vector α, see Marx and Eilers (1999) for a motivation of this idea. This results in a penalized least squares or penalized log-likelihood problem. Often cross validation is used to find an optimal smoothing parameter.

3.28 Restricted Regression

3.28.1 Problem of Selection

In plant and animal breeding we have the problem of selecting individuals for propagation on the basis of observed measurements $x_1, \ldots, x_p$ in such a way that there is improvement in a desired characteristic y_0 in the future generations. At the suggestion of R. A. Fisher that the best selection index is the regression of y_0 on $x_1, \ldots, x_p$ with individuals having a larger value preferred in selection, Fairfield Smith (1936) worked out the computational details, and Rao (1953) provided the theoretical background for the solution.

In practice, it may so happen that improvement in the main characteristic is accompanied by deterioration (side effects) in certain other desired characteristics, $y_1, \ldots, y_q$. This problem was addressed by Kempthorne and Nordskog (1959) and Tallis (1962), who modified the selection index to ensure that *no change* in $y_1, \ldots, y_q$ occurs, and subject to this condition maximum possible improvement in y_0 is achieved. Using the techniques of quadratic programming, Rao (1962; 1964) showed that a selection index can be constructed to provide maximum improvement in y_0 while ensuring that there are possible improvements in $y_1, \ldots, y_q$, but no deterioration. The theory and computations described in this section are taken from the above cited papers of Rao.

3.28.2 Theory of Restricted Regression

Let $x' = (x_1, \ldots, x_p)$ be the vector of predictors, Λ be the dispersion matrix of x, and c_i be the column vectors of the covariances $c_{i1}, \ldots, c_{ip}$, of y_i with $x_1, \ldots, x_p$, for $i = 0, 1, 2, \ldots, q$. Denote by C the partitioned matrix $(c_0, c_1, \ldots, c_q)$, and denote the dispersion matrix of $y' = (y_0, \ldots, y_q)$ by $\Sigma = (\sigma_{ij})$, $i, j = 0, 1, \ldots, q$. Let us assume that the rank of C is $q + 1$, Λ is nonsingular, and $p \geq q + 1$. If b is a p-vector, correlation of y_i and $b'x$ is

$$\frac{(b'c_i)}{\sqrt{\sigma_{ii} b' \Lambda b}} \, .$$

The problem is to choose b such that

$$\frac{(b'c_0)}{\sqrt{\sigma_{00} b' \Lambda b}} \tag{3.419}$$

is a maximum subject to the conditions

$$b'c_0 > 0, \quad b'c_i \geq 0, \quad i = 1, \ldots, q \, . \tag{3.420}$$

Note that maximizing (3.419) without any restriction leads to $b'x$, which is the linear regression of y_0 on $(x_1, \ldots, x_p)$ apart from the constant term. In such a case the selection index is $b'x$ and individuals with large values of $b'x$ are selected for future propagation.

If the constraints (3.420) are imposed to avoid side effects, then the problem is one of nonlinear programming for which the following theorems are useful.

Lemma 3.23 *Given a p-vector b satisfying the conditions (3.420), there exists a $(q+1)$-vector g such that*

(i) $m = \Lambda^{-1}Cg$ *satisfies conditions (3.420), and*

(ii) $\dfrac{m'c_0}{\sqrt{m'\Lambda m}} \geq \dfrac{b'c_1}{\sqrt{b'\Lambda b}}$.

Proof: Choose a matrix D such that $(\Lambda^{-1}C : \Lambda^{-1}D)$ is of full rank and $C'\Lambda^{-1}D = 0$, so that the spaces generated by $\Lambda^{-1}C$ and $\Lambda^{-1}D$ are orthogonal under the inner product $\alpha'\Lambda\beta$ for any two vectors α and β. Then there exist vectors g and h such that any vector b can be decomposed as

$$b = \Lambda^{-1}Cg + \Lambda^{-1}Dh = m + \Lambda^{-1}Dh\,.$$

To prove (i) observe that

$$0 \leq b'c_i = m'c_i + c_i'\Lambda^{-1}Dh = m'c_i\,, \quad i = 0, \ldots, q\,.$$

To prove (ii) we have $b'\Lambda b = m'\Lambda m + h'D'\Lambda^{-1}Dh \geq m'\Lambda m$, and since $b'c_0 = m'c_0$, we have

$$\frac{m'c_0}{\sqrt{m'\Lambda m}} \geq \frac{b'c_0}{\sqrt{b'\Lambda b}}\,.$$

Lemma 3.23 reduces the problem to that of determining m of the form $\Lambda^{-1}Cg$ where g is of a smaller order than m.

Lemma 3.24 *The problem of determining g such that with $m = \Lambda^{-1}Cg$, the conditions (3.420) are satisfied and $m'c_0/\sqrt{m'\Lambda m}$ is a maximum is equivalent to the problem of minimizing a nonnegative quadratic form $(u - \xi)'B(u-\xi)$ with u restricted to nonnegative vectors, where B and ξ are computed from the known quantities C and Λ.*

Proof: Let $v' = (v_0, v_1, \ldots, v_q)$ be a $(q+1)$-vector with all nonnegative elements and let g be a solution of

$$C'm = C'\Lambda^{-1}Cg = v$$

giving

$$g = Av\,, m = \Lambda^{-1}CAv \tag{3.421}$$

$$\frac{m'c_0}{\sqrt{m'\Lambda m}} = \frac{v_0}{\sqrt{v'Av}} \tag{3.422}$$

where $A = (C'\Lambda^{-1}C)^{-1}$. Writing $v_i/v_o = u_i, i = 1, \ldots, q$, and denoting the elements of the $(q+1)\times(q+1)$-matrix A by (a_{ij}),we can write the square of the reciprocal of (3.422) as

$$\delta + (u-\xi)'B(u-\xi) = \delta + Q(u)$$

where

$$B = (a_{ij})\,, \quad i,j = 1,\dots,q$$

and $\xi' = (\xi_1,\dots,\xi_q)$ is a solution of

$$-B\xi = \alpha_0\,, \quad \alpha_1' = (a_{01},\dots,a_{0q}) \tag{3.423}$$

and

$$\delta = a_{00} - \sum\sum\nolimits_{i,j\geq 1} a_{ij}\xi_i\xi_j\,.$$

The solution of (3.423) is

$$\xi_i = \frac{c_i'\Lambda^{-1}c_0}{c_o'\Lambda^{-1}c_0}\,, \quad i = 1,\dots,q$$

and

$$\delta = (c_0'\Lambda^{-1}c_0)^{-1}\,,$$

which are the simple functions of c_i and Λ^{-1}. Now

$$\begin{aligned} \sup_g \frac{m'c_0}{\sqrt{m'\Lambda m}} &= \sup_{u\geq 0}\{\delta + Q(u)\}^{-\frac{1}{2}} \\ &= \{\delta + \inf_{u\geq 0} Q(u)\}^{-\frac{1}{2}}\,. \end{aligned}$$

The problem is thus reduced to that of minimizing the nonnegative quadratic form $Q(u)$ with the restriction that the elements of u are nonnegative.

If $u_0' = (u_{10},\dots,u_{q0})$ is the minimizing vector, then the optimum m is found from (3.422) as

$$m = \Lambda^{-1}CAv_0$$

and the selection index is

$$v_0'AC'\Lambda^{-1}x\,,\; v_0' = (1,u_{10},\dots,u_{q0})\,. \tag{3.424}$$

3.28.3 Efficiency of Selection

The correlation between y_0 and the best selection index (multiple regression) when there are no restrictions is

$$R_1 = \frac{1}{\sqrt{\delta\sigma_{00}}}\,.$$

With the restriction that the changes in mean values of other variables are to be in specified directions if possible, or otherwise zero, the correlation between y_0 and the best selection index is

$$R_2 = \frac{1}{\sqrt{\sigma_{00}\{\delta + \min_{u\geq 0} Q(u)\}}}\,.$$

If the restriction is such that no change in mean values of $y_1, \ldots, y_q$ is derived, then the selection index is obtained by putting $u = 0$, giving the correlation coefficient

$$R_3 = \frac{1}{\sqrt{\sigma_{00}\{\delta + \xi' B \xi\}}} .$$

It may be seen that

$$R_1 \geq R_2 \geq R_3 ,$$

which implies that selection efficiency possibly increases by generalizing the restriction of no changes to possible changes in desired directions.

The correlation coefficient between the selection index and the variables $y_i (i \neq 0)$ is

$$\frac{u_i \sqrt{\sigma_{00}}}{\sqrt{\sigma_{ii}}} R_2 , \quad i = 1, \ldots, q$$

which enables the estimation of changes in the mean value of $y_i, i = 1, \ldots, q$. When $u_i = 0$, the expected change is zero, as expected.

3.28.4 Explicit Solution in Special Cases

When $q = 1$, the solution is simple. The quadratic form $Q(u)$ reduces to

$$a_{11} \left(u_1 - \frac{c_0' \Lambda^{-1} c_1}{c_0' \Lambda^{-1} c_0} \right)^2 . \tag{3.425}$$

If $c_0' \Lambda^{-1} c_1 \geq 0$, then the minimum of (3.425) for nonnegative u_1 is zero, and the multiple regression of y_0 on $x_1, \ldots, x_p$ is $c_0' \Lambda^{-1} x$, apart from the constant term.

If $c_0' \Lambda^{-1} c_1 < 0$, then the minimum is attained when $u_1 = 0$, and using (3.425) the selection index is found to be

$$c_0' \Lambda^{-1} x - \frac{c_0' \Lambda^{-1} c_1}{c_1' \Lambda^{-1} c_1} c_1' \Lambda^{-1} x \tag{3.426}$$

which is a linear combination of the multiple regressions of y_0 and y_1 on $x_1, \ldots, x_p$. The square of the correlation between y_0 and (3.426) is

$$\sigma_{00}^{-1} \left[c_0' \Lambda^{-1} c_0 - \frac{(c_0' \Lambda^{-1} c_1)^2}{c_1' \Lambda^{-1} c_1} \right] \tag{3.427}$$

and that between y_1 and its regression on $x_1, \ldots, x_p$ is

$$\sigma_{00}^{-1} c_0' \Lambda c_0 ,$$

and the reduction in correlation due to restriction on y_1, when $c_0' \Lambda^{-1} c_1 < 0$ is given by the second term in (3.427).

The next practically important case is that of $q = 3$. The quadratic form to be minimized is

$$Q(u_1, u_2) = a_{11}(u_1 - \xi_1)^2 + 2a_{12}(u_1 - \xi_1)(u_2 - \xi_2) + a_{22}(u_2 - \xi_2)^2 .$$

A number of cases arise depending on the signs of $\xi_1, \xi_2, \ldots$

Case (i) Suppose that $\xi_1 \geq 0, \xi_2 \geq 0$. The minimum of Q is zero and the multiple regression of y_1 on $x_1, \ldots, x_p$ is the selection function.

Case (ii) Suppose that $\xi_1 < 0, \xi_2 \geq 0$. The minimum of Q is attained on the boundary $u_1 = 0$. To determine the value of u_2, we solve the equation

$$\frac{1}{2}\frac{dQ(0, u_2)}{du_2} = a_{22}(u_2 - \xi_2) - a_{12}\xi_1 = 0 ,$$

obtaining

$$u_2 = \frac{a_{12}}{a_{22}}\xi_1 + \xi_2 . \tag{3.428}$$

If $a_{12}\xi_1 + a_{22}\xi_2 \geq 0$, then the minimum value of Q is attained when $u_{10} = 0$ and u_{20} has the right-hand side value in (3.428). If $a_{12}\xi_1 + a_{22}\xi_2 < 0$, then the minimum is attained at $u_{10} = 0, u_{20} = 0$. The selection function is determined as indicated in (3.424). The case of $\xi_1 \geq 0, \xi_2 < 0$ is treated in a similar way.

Case (iii) Suppose that $\xi_1 < 0, \xi_2 < 0$. There are three possible pairs of values at which the minimum might be attained:

$$\begin{aligned} u_{10} &= 0, \quad u_{20} = \frac{a_{12}}{a_{22}}\xi_1 + \xi_2 , \\ u_{10} &= \frac{a_{12}}{a_{11}}\xi_2 + \xi_1 , \quad u_{20} = 0 , \\ u_{10} &= 0, \quad u_{20} = 0 . \end{aligned}$$

Out of these we need consider only the pairs where both coordinates are nonnegative and then choose that pair for which Q is a minimum.

When $q > 3$, the number of different cases to be considered is large. When each $\xi_i \geq 0$, the minimum of Q is zero. But in the other cases the algorithms developed for general quadratic programming (Charnes and Cooper, 1961, pp. 682–687) may have to be adopted. It may, however, be observed that by replacing $u' = (u_1, \ldots, u_q)$ by $w' = (w_1^2, \ldots, w_q^2)$ in Q, the problem reduces to that of minimizing a quartic in $w_1, \ldots w_q$ without any restrictions. No great simplification seems to result by transforming the problem in this way. As mentioned earlier, the practically important cases correspond to $q = 2$ and 3 for which the solution is simple, as already indicated. The selective efficiency may go down rapidly with increase in the value of q.

For additional literature on selection problems with restrictions, the reader is referred to Rao (1964).

3.29 LINEX Loss Function

The squared error loss function assigns equal weight to positive and negative estimation errors of same magnitude. This may not be a reasonable proposition in many situations. For example, an under-estimation of the peak water level in the construction of water reservoir has more serious consequences than an over-estimation, see Zellner (1986); similarly under-estimation of the failure rate may result in more complaints from the customers than expected as compared to over-estimation; see also Harris (1992), Kuo and Dey (1990), Khatree (1992), Schabe (1992), Canfield (1970), Feynman (1987) for some more examples.

Relatively free from the limitation of under- and over-estimation is the LINEX (linear-exponential) loss function introduced by Varian (1975).

Let δ be the estimation error $(\hat{\beta}-\beta)$, or relative estimation error $(\hat{\beta}-\beta)/\beta$ associated with an estimator $\hat{\beta}$ for some scalar parameter β. Then the LINEX loss function is defined as

$$L(\hat{\beta};\beta) = c[\exp(\alpha\delta) - \alpha\delta - 1] \tag{3.429}$$

where $\alpha \neq 0$ and $c > 0$ are the characterizing scalars.

The value of α determines the relative losses associated with the positive and negative values of δ while the value of c specifies the factor of proportionality. The loss function (3.429) attains its minimum value as 0 for $\delta = 0$. Further, it rises exponentially on one side of zero and approximately linearly on the other side. Graphs of this function for some selected values of α have been prepared by Zellner (1986). A look at them reveals that for $\alpha > 0$, over-estimation breeds relatively larger losses in comparison to under-estimation. If $\alpha < 0$, then the reverse is true, *i.e.*, over-estimation leads to relatively smaller losses than under-estimation of the same magnitude. Thus the positive and negative values of the estimation error can be assigned possibly unequal weight in the loss function by choosing an appropriate sign for the scalar α. So far as the choice of magnitude of α is concerned, the loss function (3.429) is fairly symmetric around 0, like the squared error loss function. This is evident for small values of α from the following expansion:

$$\begin{aligned} L(\hat{\beta};\beta) &= c\left[\exp(\alpha\delta) - \alpha\delta - 1\right] \\ &= c\left[\frac{1}{2}\alpha^2\delta^2 + \frac{1}{6}\alpha^3\delta^3 + \ldots\right] . \end{aligned} \tag{3.430}$$

When the value of α is not small, the contribution of the terms of order three and more will not be negligible and asymmetry will enter. The degree of asymmetry increases for larger values of α. Thus the LINEX loss function bears a close link with the squared error loss function and offers considerable flexibility to the requirement of the given problems.

If we define the LINEX risk function as

$$\begin{aligned} R(\hat{\beta};\beta) &= \mathrm{E}[L(\hat{\beta};\beta)] \\ &= c\,[\mathrm{E}(\exp(\alpha\delta)) - \alpha\,\mathrm{E}(\delta) - 1] \end{aligned} \tag{3.431}$$

we observe that the first term on the right hand side of (3.431) contains the moment generating function of δ, provided it exists. Thus the risk function (3.431) not only depends upon the second moment of δ (*i.e.*, mean squared error or relative mean squared error) but also upon the entire set of moments. So the risk criterion takes care of all the aspects of the sampling distribution of estimator into account while the mean squared error or relative mean squared error criterion covers only one aspect of second moment.

When β is a vector of parameters $\beta_1, \ldots, \beta_K$, Zellner (1986) has provided an extension of the LINEX loss function as

$$L(\hat{\beta};\beta) = \sum_{t=1}^{K} c_t\,[\exp(\alpha_t\delta_t) - \alpha_t\delta_t - 1] \tag{3.432}$$

where $\hat{\beta}$ is an estimator of β and $\delta_t = (\hat{\beta}_t - \beta_t)$ or $(\hat{\beta}_t - \beta_t)/\beta_t$. Further, $\alpha_1, \ldots, \alpha_K$ and $c_1, \ldots, c_K$ are the scalars characterizing the loss function similar to α and c, respectively in (3.429).

The definition of the LINEX loss function is compatible with the family of loss function identified by Thompson and Basu (1996).

The usual definition of mean unbiasedness that $\mathrm{E}(\hat{\beta}) = \beta$ is inappropriate in the context of LINEX loss function because it does not distinguish between under- and over-estimation. Following Lehmann (1988), an estimator $\hat{\beta}$ of β is risk-unbiased if

$$\mathrm{E}_{\beta}[L(\hat{\beta};\beta)] \leq \mathrm{E}_{\beta}[L(\tilde{\beta};\beta)] \text{ for all } \hat{\beta} \neq \tilde{\beta}\ . \tag{3.433}$$

Under (3.429),

$$\mathrm{E}[\exp(\alpha\hat{\beta})] = \exp(\alpha\beta) \ \text{ for all } \beta\ . \tag{3.434}$$

An estimator $\hat{\beta}$ is a LINEX-unbiased estimator of β under loss function (3.429) when (3.434) holds true and β is termed as L-estimable parameter. If (3.434) does not hold true, then its bias is defined by Parsian and Sanjari (1999) as

$$\begin{aligned} L - \mathrm{Bias}(\hat{\beta}) &= \alpha^{-1}\left[\ln\{\exp(\alpha\beta)\} - \ln \mathrm{E}_{\beta}\left\{\exp(\alpha\hat{\beta})\right\}\right] \\ &= \beta - \alpha^{-1}\ln \mathrm{E}_{\beta}\left\{\exp(\alpha\hat{\beta})\right\}\ . \end{aligned}$$

More interested readers are referred to the review paper by Parsian and Kirmani (2002) for more details on the aspect of unbiased and invariant estimation of parameters under LINEX loss function.

3.30 Balanced Loss Function

Performance of any estimation procedure for the parameters in a model is generally evaluated by either the goodness of fitted model or the concentration of the estimates around the true parameter values. In the linear model $y = X\beta + \epsilon$, if $\tilde{\beta}$ denotes any estimator of β, then the goodness of the fitted model is reflected in the residual vector $(X\tilde{\beta} - y)$. Similarly, the pivotal quantity for measuring the concentration of estimates around the true parameter values is the estimation error $(\tilde{\beta} - \beta)$. Accordingly, the quadratic loss function for the goodness of fit of the model is

$$(X\tilde{\beta} - y)'(X\tilde{\beta} - y) \tag{3.435}$$

while the commonly employed loss function for the precision of estimation are squared error loss function

$$(\tilde{\beta} - \beta)'(\tilde{\beta} - \beta) \tag{3.436}$$

or weighted squared error loss function

$$(\tilde{\beta} - \beta)'X'X(\tilde{\beta} - \beta) \ . \tag{3.437}$$

Both the criterion are important and it may often be desirable to employ both the criteria simultaneously in practice; see, for instance, Zellner (1994), Shalabh (1995) and Toutenburg and Shalabh (1996; 2000) for more details and some illustrative examples. Accordingly, considering both the criteria of the goodness of fit and precision of estimation together, Zellner (1994) has proposed the following balanced loss function:

$$BL(\tilde{\beta}) = \omega(X\tilde{\beta} - y)'(X\tilde{\beta} - y) + (1 - \omega)(\tilde{\beta} - \beta)'X'X(\tilde{\beta} - \beta) \tag{3.438}$$

where ω is a scalar between 0 and 1. When $\omega = 1$, the loss function (3.438) reflects the goodness of fitted model and when $\omega = 0$, the loss function (3.438) reflects the precision of estimation. Any other value of ω between 0 and 1 provides the weight to the goodness of fit.

From the viewpoint of the prediction of values of study variable within the sample, the loss functions (3.435) and (3.437) can be regarded as arising from the prediction of actual values y by $X\tilde{\beta}$ and the prediction of average values $\mathrm{E}(y) = X\beta$ by $X\tilde{\beta}$, respectively. Such details are discussed later in Chapter 6. Further, using the idea of simultaneous prediction of actual and average values of study variable (cf., Section 6.8), Shalabh (1995) has presented the following predictive loss function

$$\begin{aligned} PL(\tilde{\beta}) &= \omega^2(X\tilde{\beta} - y)'(X\tilde{\beta} - y) + (1 - \omega)^2(\tilde{\beta} - \beta)'X'X(\tilde{\beta} - \beta) \\ &\quad +2\omega(1 - \omega)(X\tilde{\beta} - y)'X(\tilde{\beta} - \beta) \end{aligned} \tag{3.439}$$

where ω is a scalar between 0 and 1. Such loss function is an extension of the balanced loss function (3.438) and also takes care of the covariability between the goodness of fit and precision of estimation.

Looking at the functional forms of the balanced loss function and the predictive loss function, Shalabh, Toutenburg and Heumann (2006) proposed the following extended balanced loss function:

$$\begin{aligned} WL(\tilde{\beta}) &= \lambda_1(X\tilde{\beta} - y)'(X\tilde{\beta} - y) + \lambda_2(\tilde{\beta} - \beta)'X'X(\tilde{\beta} - \beta) \\ &\quad +(1 - \lambda_1 - \lambda_2)(X\tilde{\beta} - y)'X(\tilde{\beta} - \beta) \end{aligned} \tag{3.440}$$

where λ_1 and λ_2 are scalars between 0 and 1 which characterize the loss functions.

Clearly, the function (3.440) encompasses the loss functions (3.435), (3.437), (3.438) and (3.439) as particular cases. Thus it is fairly general and sufficiently flexible.

For illustration, we consider the OLS estimator of β as $b = (X'X)^{-1}X'y$. The risk function of b is under (3.440) is

$$\begin{aligned} R(b) &= \mathrm{E}\,[WL(b)] \\ &= \sigma^2\lambda_1 n - \sigma^2 p(\lambda_1 - \lambda_2)\ . \end{aligned} \tag{3.441}$$

It is clear from (3.441) that the performance of OLSE is affected by the goodness of fit as well as the precision of estimation.

3.31 Complements

3.31.1 Linear Models without Moments: Exercise

In the discussion of linear models in the preceding sections of this chapter, it is assumed that the error variables have second-order moments. What properties does the OLSE, $\hat{\beta} = (X'X)^{-1}X'y$, have if the first- and second-order moments do not exist? The question is answered by Jensen (1979) when ϵ has a spherical distribution with the density

$$\mathcal{L}(y) = \sigma^{-T}\Psi_T\{(y - X\beta)'(y - X\beta)/\sigma^2\}. \tag{3.442}$$

We represent this class by $S_k(X\beta, \sigma^2 I)$, where k represents the integral order of moments that ϵ admits. If $k = 0$, no moments exist. Jensen (1979) proved among other results the following.

Theorem 3.25 (Jensen, 1979) *Consider $\hat{\beta} = (X'X)^{-1}X'y$ as an estimator β in the model $y = X\beta + \epsilon$. Then*

(i) *If $\mathcal{L}(y) \in S_0(X\beta, \sigma^2 I)$, then $\hat{\beta}$ is median unbiased for β and $\hat{\beta}$ is at least as concentrated about β as any other median unbiased estimator of β.*
[Note that an s-vector $t \in \mathbb{R}^s$ is said to be modal unbiased for $\theta \in \mathbb{R}^s$ if $a't$ is modal unbiased for $a'\theta$ for all a.]

(ii) *If $\mathcal{L}(y) \in S_1(X\beta, \sigma^2 I)$, then $\hat{\beta}$ is unbiased for β and is at least as concentrated around β as any other unbiased linear estimator.*

(iii) *If $\mathcal{L}(y) \in S_0(X\beta, \sigma^2 I)$ and in addition unimodal, then $\hat{\beta}$ is modal unbiased for β.*

3.31.2 Nonlinear Improvement of OLSE for Nonnormal Disturbances

Consider the linear regression model (3.23). The Gauss-Markov Theorem states that $b = (X'X)^{-1}X'y$ is the best linear unbiased estimator for β, that is, $\mathrm{Var}(\tilde{b}) - \mathrm{Var}(b)$ is nonnegative definite for any other linear unbiased estimator $\tilde{b}$. If ϵ is multinormally distributed, then b is even the best unbiased estimator.

Hence, if ϵ is not multinormally distributed, there is a potential of nonlinear unbiased estimators for β that improve upon b.

- What is the most general description of such estimators?
- What is the best estimator within this class?

Remark. This problem was proposed by G. Trenkler. Related work may be found in Kariya (1985) and Koopmann (1982).

3.31.3 A Characterization of the Least Squares Estimator

Consider the model $y = X\beta + \epsilon$ with $\mathrm{Cov}(\epsilon) = \sigma^2 I$, $\mathrm{rank}(X) = K$, the size of vector β, and a submodel $y_{(i)} = X_{(i)}\beta + \epsilon_{(i)}$ obtained by choosing $k \leq T$ rows of the original model. Further, let

$$\hat{\beta} = (X'X)^{-1}X'y\,, \quad \hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-}X'_{(i)}y_{(i)} \tag{3.443}$$

be the LSEs from the original and the submodel respectively. Subramanyam (1972) and Rao and Precht (1985) proved the following result.

Theorem 3.26 *Denoting $d_{(i)} = |X'_{(i)}X_{(i)}|$, we have*

$$\hat{\beta} = \frac{\sum_{i=1}^{c} d_{(i)}\hat{\beta}_{(i)}}{\sum_{i=1}^{c} d_{(i)}} \tag{3.444}$$

where c is the number of all possible subsets of size k from $\{1, \ldots, T\}$.

The result (3.444), which expresses $\hat{\beta}$ as a weighted average of $\hat{\beta}_{(i)}$, is useful in regression diagnostics. We may calculate all possible $\hat{\beta}_{(i)}$ and look for consistency among them. If some appear to be much different from others, then we may examine the data for outliers or existence of clusters and consider the possibility of combining them with a different set of weights (some may be zero) than those in (3.444). Further results of interest in this direction are contained in Wu (1986).

3.31.4 A Characterization of the Least Squares Estimator: A Lemma

Consider the model $\epsilon_i = y_i - x_i'\beta$, $i = 1, 2, \ldots$, in which $\epsilon_1, \epsilon_2, \ldots$, are independently and identically distributed with mean 0 and variance σ^2, and the x_i''s are K-vectors of constants. Let the $K \times n$-matrix $X' = (x_1, \ldots, x_n)$ of constants be of rank K. Define $h_{ii}(n) = x_i'(X'X)^{-1}x_i$ and $b = (X'X)^{-1}X'y$ where $y = (y_1, \ldots, y_n)'$. Then for any $r \times K$-matrix C of constants and of rank $r \leq K$,

$$\sigma^{-2}(Cb - C\beta)'[C(X'X)^{-1}C']^{-1}(Cb - C\beta) \to \chi_r^2$$

if (and only if) $\max_{1 \leq i \leq n} h_{ii}(n) \to \infty$.

This result and the condition on $h_{ii}(n)$ were obtained by Srivastava (1971; 1972) using a lemma of Chow (1966).

3.32 Exercises

Exercise 1. Define the principle of least squares. What is the main reason to use $e'e$ from (3.6) instead of other objective functions such as $\max_t |e_t|$ or $\sum_{t=1}^{T} |e_t|$?

Exercise 2. Discuss the statement: In well-designed experiments with quantitative x-variables it is not necessary to use procedures for reducing the number of included x-variables after the data have been obtained.

Exercise 3. Find the least squares estimators of β in $y = X\beta + \epsilon$ and $y = \alpha 1 + X\beta + \epsilon^*$, where 1 denotes a column vector with all elements unity. Compare the dispersion matrices as well as the residual sums of squares.

Exercise 4. Consider the two models $y_1 = \alpha_1 1 + X\beta + \epsilon_1$ and $y_2 = \alpha_2 1 + X\beta + \epsilon_2$ (with 1 as above). Assuming ϵ_1 and ϵ_2 to be independent with same distributional properties, find the least squares estimators of α_1, α_2, and β.

Exercise 5. In a bivariate linear model, the OLSE's are given by $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 = \sum(x_t - \bar{x})(y_t - \bar{y}) / \sum(x_t - \bar{x})^2$. Calculate the covariance matrix $\mathrm{V}\binom{b_0}{b_1}$. When are b_0 and b_1 uncorrelated?

Exercise 6. Show that the estimator minimizing the generalized variance (determinant of variance-covariance matrix) in the class of linear and unbiased estimators of β in the model $y = X\beta + \epsilon$ is nothing but the least squares estimator.

Exercise 7. Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the least squares estimators of β from $y_1 = X\beta + \epsilon_1$ and $y_2 = X\beta + \epsilon_2$. If β is estimated by $\hat{\beta} = w\hat{\beta}_1 + (1-w)\hat{\beta}_2$ with $0 < w < 1$, determine the value of w that minimizes the trace of the dispersion matrix of $\hat{\beta}$. Does this value change if we minimize $\mathrm{E}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)$?

Exercise 8. Demonstrate that the best quadratic estimator of σ^2 is $(T - K + 2)^{-1}y'(I - P)y$, where P is the projection matrix on $\mathcal{R}(X)$.

Exercise 9. Let the following model be given:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

(i) Formulate the hypothesis H_0: $\beta_2 = 0$ as a linear restriction $r = R\beta$ on β.

(ii) Write down the test statistic for testing H_0: $\beta_2 = 0$.

Exercise 10. Describe a procedure for testing the equality of first p elements of β_1 and β_2 in the model $y_1 = X_1\beta_1 + \epsilon_1$ and $y_2 = X_2\beta_2 + \epsilon_2$. Assume that $\epsilon_1 \sim N(0, \sigma^2 I_{n_1})$ and $\epsilon_2 \sim N(0, \sigma^2 I_{n_2})$ are stochastically independent.

Exercise 11. If $\hat{\theta}_i$ is a MVUE (minimum variance unbiased estimator) of $\theta_i, i = 1, \ldots, k$, then $a_1\hat{\theta}_1 + \ldots + a_k\hat{\theta}_k$ is a MVUE of $a_1\theta_1 + \ldots + a_k\theta_k$ for any $a_1, \ldots, a_k$.