Chapter 2
Estimation

In this chapter, properties of least squares estimates are examined for the model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I.$$

The chapter begins with a discussion of the concepts of identifiability and estimability in linear models. Section 2 characterizes least squares estimates. Sections 3, 4, and 5 establish that least squares estimates are best linear unbiased estimates, maximum likelihood estimates, and minimum variance unbiased estimates. The last two of these properties require the additional assumption $e \sim N(0, \sigma^2 I)$. Section 6 also assumes that the errors are normally distributed and presents the distributions of various estimates. From these distributions various tests and confidence intervals are easily obtained. Section 7 examines the model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 V,$$

where V is a known positive definite matrix. Section 7 introduces generalized least squares estimates and presents properties of those estimates. Section 8 presents the normal equations and establishes their relationship to least squares and generalized least squares estimation. Section 9 discusses Bayesian estimation.

The history of least squares estimation goes back at least to 1805, when Legendre first published the idea. Gauss made important early contributions (and claimed to have invented the method prior to 1805).

There is a huge body of literature available on estimation and testing in linear models. A few books dealing with the subject are Arnold (1981), Eaton (1983), Graybill (1976), Rao (1973), Ravishanker and Dey (2002), Rencher (2008), Scheffé (1959), Searle (1971), Seber (1966, 1977), and Wichura (2006).

R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*,
Springer Texts in Statistics, DOI 10.1007/978-1-4419-9816-3_2,

2.1 Identifiability and Estimability

A key issue in linear model theory is figuring out which parameters can be estimated and which cannot. We will see that what can be estimated are functions of the parameters that are *identifiable*. Linear functions of the parameters that are identifiable are called *estimable* and have linear unbiased estimators. These concepts also have natural applications to generalized linear models. The definitions used here are tailored to (generalized) linear models but the definition of an identifiable parameterization coincides with more common definitions of identifiability; cf. Christensen et al. (2010, Section 4.14). For other definitions, the key idea is that the distribution of Y should be either completely determined by $\mathrm{E}(Y)$ alone or completely determined by $\mathrm{E}(Y)$ along with some parameters (like σ^2) that are functionally unrelated to the parameters in $\mathrm{E}(Y)$.

Consider the general linear model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0,$$

where again Y is an $n \times 1$ vector of observations, X is an $n \times p$ matrix of known constants, β is a $p \times 1$ vector of unobservable parameters, and e is an $n \times 1$ vector of unobservable random errors whose distribution does not depend on β. We can only learn about β through $X\beta$. If x_i' is the ith row of X, $x_i'\beta$ is the ith row of $X\beta$ and we can only learn about β through the $x_i'\beta$s. $X\beta$ can be thought of as a vector of inner products between β and a spanning set for $C(X')$. Thus, we can learn about inner products between β and $C(X')$. In particular, when λ is a $p \times 1$ vector of known constants, we can learn about functions $\lambda'\beta$ where $\lambda \in C(X')$, i.e., where $\lambda = X'\rho$ for some vector ρ. These are precisely the estimable functions of β. We now give more formal arguments leading us to focus on functions $\lambda'\beta$ where $\lambda' = \rho'X$ or, more generally, vectors $\Lambda'\beta$ where $\Lambda' = P'X$.

In general, a *parameterization* for the $n \times 1$ mean vector $\mathrm{E}(Y)$ consists of writing $\mathrm{E}(Y)$ as a function of some parameters β, say,

$$\mathrm{E}(Y) = f(\beta).$$

A general linear model is a parameterization

$$\mathrm{E}(Y) = X\beta$$

because $\mathrm{E}(Y) = \mathrm{E}(X\beta + e) = X\beta + \mathrm{E}(e) = X\beta$. A parameterization is identifiable if knowing $\mathrm{E}(Y)$ tells you the parameter vector β.

Definition 2.1.1 The parameter β is *identifiable* if for any β_1 and β_2, $f(\beta_1) = f(\beta_2)$ implies $\beta_1 = \beta_2$. If β is identifiable, we say that the parameterization $f(\beta)$ is identifiable. Moreover, a vector-valued function $g(\beta)$ is identifiable if $f(\beta_1) = f(\beta_2)$ implies $g(\beta_1) = g(\beta_2)$. If the parameterization is not identifiable but nontrivial identifiable functions exist, then the parameterization is said to be *partially identifiable*.

The key point is that if β or a function $g(\beta)$ is not identifiable, it is simply impossible for one to know what it is based on knowing $\mathrm{E}(Y)$. From a statistical perspective, we are considering models for the mean vector with the idea of collecting data that will allow us to estimate $\mathrm{E}(Y)$. If actually knowing $\mathrm{E}(Y)$ is not sufficient to tell us the value of β or $g(\beta)$, no amount of data is ever going to let us estimate them.

In regression models, i.e., models for which $r(X) = p$, the parameters are identifiable. In this case, $X'X$ is nonsingular, so if $X\beta_1 = X\beta_2$, then

$$\beta_1 = (X'X)^{-1}X'X\beta_1 = (X'X)^{-1}X'X\beta_2 = \beta_2$$

and identifiability holds.

For models in which $r(X) = r < p$, there exist $\beta_1 \neq \beta_2$ but $X\beta_1 = X\beta_2$, so the parameters are not identifiable.

For general linear models, the only functions of the parameters that are identifiable are functions of $X\beta$. This follows from the next result.

Theorem 2.1.2 A function $g(\beta)$ is identifiable if and only if $g(\beta)$ is a function of $f(\beta)$.

PROOF. $g(\beta)$ being a function of $f(\beta)$ means that for some function g_*, $g(\beta) = g_*[f(\beta)]$ for all β; or, equivalently, it means that for any $\beta_1 \neq \beta_2$ such that $f(\beta_1) = f(\beta_2)$, $g(\beta_1) = g(\beta_2)$.

Clearly, if $g(\beta) = g_*[f(\beta)]$ and $f(\beta_1) = f(\beta_2)$, then $g(\beta_1) = g_*[f(\beta_1)] = g_*[f(\beta_2)] = g(\beta_2)$, so $g(\beta)$ is identifiable.

Conversely, if $g(\beta)$ is not a function of $f(\beta)$, there exists $\beta_1 \neq \beta_2$ such that $f(\beta_1) = f(\beta_2)$ but $g(\beta_1) \neq g(\beta_2)$. Hence, $g(\beta)$ is not identifiable. □

It is reasonable to estimate any identifiable function. Thus, in a linear model it is reasonable to estimate any function of $X\beta$. It is not reasonable to estimate nonidentifiable functions, because you simply do not know what you are estimating.

The traditional idea of estimability in linear models can now be presented. Estimable functions are linear functions of β that are identifiable.

Definition 2.1.3 A vector-valued linear function of β, say, $\Lambda'\beta$, is *estimable* if $\Lambda'\beta = P'X\beta$ for some matrix P.

Actually, an identifiable linear function of β is a function $g_*(X\beta)$, but since the composite function is linear and $X\beta$ is linear, the function g_* must be linear, and we can write it as a matrix P'.

Clearly, if $\Lambda'\beta$ is estimable, it is identifiable and therefore it is a reasonable thing to estimate. However, estimable functions are not the only functions of β that are reasonable to estimate. For example, the ratio of two estimable functions is not estimable, but it is identifiable, so the ratio is reasonable to estimate. You *can* estimate many functions that are not "estimable." What you cannot do is estimate nonidentifiable functions.

Unfortunately, the term "nonestimable" is often used to mean something other than "not being estimable." You can be "not estimable" by being either not linear or not identifiable. In particular, a linear function that is "not estimable" is automatically nonidentifiable. However, nonestimable is often *taken to mean* a linear function that is not identifiable. In other words, some authors (perhaps, on occasion, even this one) presume that nonestimable functions are linear, so that nonestimability and nonidentifiability become equivalent.

It should be noted that the concepts of identifiability and estimability are based entirely on the assumption that $\mathrm{E}(Y) = X\beta$. Identifiability and estimability do not depend on $\mathrm{Cov}(Y) = \mathrm{Cov}(e)$ (as long as the covariance matrix is not also a function of β).

An important property of estimable functions $\Lambda'\beta = P'X\beta$ is that although P need not be unique, its perpendicular projection (columnwise) onto $C(X)$ is unique. Let P_1 and P_2 be matrices with $\Lambda' = P_1'X = P_2'X$, then $MP_1 = X(X'X)^-X'P_1 = X(X'X)^-\Lambda = X(X'X)^-X'P_2 = MP_2$.

EXAMPLE 2.1.4. In the simple linear regression model of Example 1.0.1, β_1 is estimable because

$$\frac{1}{35}(-5,-3,-1,1,3,5)\begin{bmatrix}1&1\\1&2\\1&3\\1&4\\1&5\\1&6\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix} = (0,1)\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix} = \beta_1.$$

β_0 is also estimable. Note that

$$\frac{1}{6}(1,1,1,1,1,1)\begin{bmatrix}1&1\\1&2\\1&3\\1&4\\1&5\\1&6\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix} = \beta_0 + \frac{7}{2}\beta_1,$$

so

$$\begin{aligned}\beta_0 &= \left(\beta_0 + \frac{7}{2}\beta_1\right) - \frac{7}{2}\beta_1\\ &= \left[\frac{1}{6}\begin{pmatrix}1\\1\\1\\1\\1\\1\end{pmatrix} - \frac{7}{2}\left(\frac{1}{35}\right)\begin{pmatrix}-5\\-3\\-1\\1\\3\\5\end{pmatrix}\right]'\begin{bmatrix}1&1\\1&2\\1&3\\1&4\\1&5\\1&6\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}\end{aligned}$$

$$= \frac{1}{30}(20,14,8,2,-4,-10)\begin{bmatrix}1 & 1\\1 & 2\\1 & 3\\1 & 4\\1 & 5\\1 & 6\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}.$$

For any fixed number x, $\beta_0+\beta_1 x$ is estimable because it is a linear combination of estimable functions.

EXAMPLE 2.1.5. In the one-way ANOVA model of Example 1.0.2, we can estimate parameters like $\mu+\alpha_1$, $\alpha_1-\alpha_3$, and $\alpha_1+\alpha_2-2\alpha_3$. Observe that

$$(1,0,0,0,0,0)\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{bmatrix}\begin{bmatrix}\mu\\\alpha_1\\\alpha_2\\\alpha_3\end{bmatrix}=\mu+\alpha_1,$$

$$(1,0,0,0,-1,0)\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{bmatrix}\begin{bmatrix}\mu\\\alpha_1\\\alpha_2\\\alpha_3\end{bmatrix}=\alpha_1-\alpha_3,$$

but also

$$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3},0,\frac{-1}{2},\frac{-1}{2}\right)\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{bmatrix}\begin{bmatrix}\mu\\\alpha_1\\\alpha_2\\\alpha_3\end{bmatrix}=\alpha_1-\alpha_3,$$

and

$$(1,0,0,1,-2,0)\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{bmatrix}\begin{bmatrix}\mu\\\alpha_1\\\alpha_2\\\alpha_3\end{bmatrix}=\alpha_1+\alpha_2-2\alpha_3.$$

We have given two vectors ρ_1 and ρ_2 with $\rho_i'X\beta=\alpha_1-\alpha_3$. Using M given in Exercise 1.5.8b, the reader can verify that $M\rho_1=M\rho_2$.

In the one-way analysis of covariance model,

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + e_{ij}, \quad \mathrm{E}(e_{ij}) = 0,$$

$i = 1,\dots,a$, $j = 1,\dots,N_i$, x_{ij} is a known predictor variable and γ is its unknown coefficient. γ is generally identifiable but μ and the α_is are not. The following result allows one to tell whether or not an individual parameter is identifiable.

Proposition 2.1.6 For a linear model, write $X\beta = \sum_{k=1}^{p} X_k\beta_k$ where the X_ks are the columns of X. An individual parameter β_i is not identifiable if and only if there exist scalars α_k such that $X_i = \sum_{k\neq i} X_k\alpha_k$.

PROOF. To show that the condition on X implies nonidentifiability, it is enough to show that there exist β and β_* with $X\beta = X\beta_*$ but $\beta_i \neq \beta_{*i}$. The condition $X_i = \sum_{k\neq i} X_k\alpha_k$ is equivalent to there existing a vector α with $\alpha_i \neq 0$ and $X\alpha = 0$. Let $\beta_* = \beta + \alpha$ and the proof is complete.

Rather than showing that when β_i is not identifiable, the condition on X holds, we show the contrapositive, i.e., that when the condition on X does not hold, β_i is identifiable. If there do not exist such α_ks, then whenever $X\alpha = 0$, we must have $\alpha_i = 0$. In particular, if $X\beta = X\beta_*$, then $X(\beta - \beta_*) = 0$, so $(\beta_i - \beta_{*i}) = 0$ and β_i is identifiable. □

The concepts of identifiability and estimability apply with little change to generalized linear models. In generalized linear models, the distribution of Y is either completely determined by $\mathrm{E}(Y)$ or it is determined by $\mathrm{E}(Y)$ along with another parameter ϕ that is unrelated to the parameterization of $\mathrm{E}(Y)$. A generalized linear model has $\mathrm{E}(Y) = h(X\beta)$. By Theorem 2.1.2, a function $g(\beta)$ is identifiable if and only if it is a function of $h(X\beta)$. However, the function $h(\cdot)$ is assumed to be invertible, so $g(\beta)$ is identifiable if and only if it is a function of $X\beta$. A vector-valued linear function of β, say, $\Lambda'\beta$ is identifiable if $\Lambda'\beta = P'X\beta$ for some matrix P, hence Definition 2.1.3 applies as well to define estimability for generalized linear models as it does for linear models. Proposition 2.1.6 also applies without change.

Finally, the concept of estimability in linear models can be related to the existence of linear unbiased estimators. A linear function of the parameter vector β, say $\lambda'\beta$, is estimable if and only if it admits a linear unbiased estimate.

Definition 2.1.7. An estimate $f(Y)$ of $g(\beta)$ is *unbiased* if $\mathrm{E}[f(Y)] = g(\beta)$ for any β.

Definition 2.1.8. $f(Y)$ is a *linear estimate* of $\lambda'\beta$ if $f(Y) = a_0 + a'Y$ for some scalar a_0 and vector a.

Proposition 2.1.9. A linear estimate $a_0 + a'Y$ is unbiased for $\lambda'\beta$ if and only if $a_0 = 0$ and $a'X = \lambda'$.

PROOF. $\Leftarrow$ If $a_0 = 0$ and $a'X = \lambda'$, then $\mathrm{E}(a_0 + a'Y) = 0 + a'X\beta = \lambda'\beta$.

$\Rightarrow$ If $a_0 + a'Y$ is unbiased, $\lambda'\beta = \mathrm{E}(a_0 + a'Y) = a_0 + a'X\beta$, for any β. Subtracting $a'X\beta$ from both sides gives

$$(\lambda' - a'X)\beta = a_0$$

for any β. If $\beta = 0$, then $a_0 = 0$. Thus the vector $\lambda - X'a$ is orthogonal to any vector β. This can only occur if $\lambda - X'a = 0$; so $\lambda' = a'X$. □

Corollary 2.1.10. $\lambda'\beta$ is estimable if and only if there exists ρ such that $\mathrm{E}(\rho'Y) = \lambda'\beta$ for any β.

2.2 Estimation: Least Squares

Consider the model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I.$$

Suppose we want to estimate $\mathrm{E}(Y)$. We know that $\mathrm{E}(Y) = X\beta$, but β is unknown; so all we really know is that $\mathrm{E}(Y) \in C(X)$. To estimate $\mathrm{E}(Y)$, we might take the vector in $C(X)$ that is closest to Y. By definition then, an estimate $\hat{\beta}$ is a least squares estimate if $X\hat{\beta}$ is the vector in $C(X)$ that is closest to Y. In other words, $\hat{\beta}$ is a *least squares estimate (LSE)* of β if

$$(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \min_{\beta}(Y - X\beta)'(Y - X\beta).$$

For a vector $\Lambda'\beta$, a least squares estimate is defined as $\Lambda'\hat{\beta}$ for any least squares estimate $\hat{\beta}$.

In this section, least squares estimates are characterized and uniqueness and unbiasedness properties of least squares estimates are given. An unbiased estimate of σ^2 is then presented. Finally, at the end of the section, the geometry associated with least squares estimation and unbiased estimation of σ^2 is discussed. The geometry provides good intuition for n-dimensional problems but the geometry can only be visualized in three dimensions. In other words, although it is a fine pedagogical tool, the geometry can only be spelled out for three or fewer data points. The fundamental goal of this book is to build the theory of linear models on vector space generalizations of these fundamentally geometric concepts. We now establish the *fundamental theorem of least squares estimation*, that the vector in $C(X)$ that is closest to Y is the perpendicular projection of Y onto $C(X)$.

Theorem 2.2.1. $\hat{\beta}$ is a least squares estimate of β if and only if $X\hat{\beta} = MY$, where M is the perpendicular projection operator onto $C(X)$.

PROOF. We will show that

$$(Y-X\beta)'(Y-X\beta) = (Y-MY)'(Y-MY) + (MY-X\beta)'(MY-X\beta).$$

Both terms on the righthand side are nonnegative, and the first term does not depend on β. $(Y-X\beta)'(Y-X\beta)$ is minimized by minimizing $(MY-X\beta)'(MY-X\beta)$. This is the squared distance between MY and $X\beta$. The distance is zero if and only if $MY = X\beta$, which proves the theorem. We now establish the equation.

$$\begin{aligned}(Y-X\beta)'(Y-X\beta) &= (Y-MY+MY-X\beta)'(Y-MY+MY-X\beta)\\ &= (Y-MY)'(Y-MY) + (Y-MY)'(MY-X\beta)\\ &\quad + (MY-X\beta)'(Y-MY) + (MY-X\beta)'(MY-X\beta).\end{aligned}$$

However, $(Y-MY)'(MY-X\beta) = Y'(I-M)MY - Y'(I-M)X\beta = 0$ because $(I-M)M = 0$ and $(I-M)X = 0$. Similarly, $(MY-X\beta)'(Y-MY) = 0$. □

Corollary 2.2.2. $(X'X)^{-}X'Y$ is a least squares estimate of β.

In Example 1.0.2, with M given in Exercise 1.5.8b, it is not difficult to see that

$$MY = \begin{bmatrix}\bar{y}_{1\cdot}\\ \bar{y}_{1\cdot}\\ \bar{y}_{1\cdot}\\ y_{21}\\ \bar{y}_{3\cdot}\\ \bar{y}_{3\cdot}\end{bmatrix} = \begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{bmatrix}\begin{bmatrix}0\\ \bar{y}_{1\cdot}\\ y_{21}\\ \bar{y}_{3\cdot}\end{bmatrix} = \begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\end{bmatrix}\begin{bmatrix}\bar{y}_{1\cdot}\\ 0\\ y_{21}-\bar{y}_{1\cdot}\\ \bar{y}_{3\cdot}-\bar{y}_{1\cdot}\end{bmatrix}.$$

Thus, both $\hat{\beta}_1 = (0, \bar{y}_{1\cdot}, y_{21}, \bar{y}_{3\cdot})'$ and $\hat{\beta}_2 = (\bar{y}_{1\cdot}, 0, y_{21}-\bar{y}_{1\cdot}, \bar{y}_{3\cdot}-\bar{y}_{1\cdot})'$ are least squares estimates of β. From Example 2.1.5,

$$\alpha_1 - \alpha_3 = (1,0,0,0,-1,0)X\beta = (1/3,1/3,1/3,0,-1/2,-1/2)X\beta.$$

The least squares estimates are

$$(1,0,0,0,-1,0)X\hat{\beta} = (1,0,0,0,-1,0)MY = \bar{y}_{1\cdot} - \bar{y}_{3\cdot},$$

but also

$$\begin{aligned}(1/3,1/3,1/3,0,-1/2,-1/2)X\hat{\beta} &= (1/3,1/3,1/3,0,-1/2,-1/2)MY\\ &= \bar{y}_{1\cdot} - \bar{y}_{3\cdot}.\end{aligned}$$

Moreover, these estimates do not depend on the choice of least squares estimates. Either $\hat{\beta}_1$ or $\hat{\beta}_2$ gives this result.

In Example 1.0.1, $X'X$ has a true inverse, so the unique least squares estimate of β is

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y = \begin{bmatrix}6&21\\21&91\end{bmatrix}^{-1}\begin{bmatrix}1&1&1&1&1&1\\1&2&3&4&5&6\end{bmatrix}Y.$$

An immediate result of Theorem 2.2.1, the uniqueness of perpendicular projection operators (Proposition B.34), and Theorem 2.1.2 as applied to linear models, is that the least squares estimate of any identifiable function is unique. Any least squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have $X\hat{\beta}_1 = X\hat{\beta}_2$, so if $g(\beta)$ is identifiable, $g(\hat{\beta}_1) = g(\hat{\beta}_2)$. In particular, least squares estimates of estimable functions are unique.

Corollary 2.2.3. The unique least squares estimate of $\rho' X\beta$ is $\rho' MY$.

Recall that a vector-valued linear function of the parameters, say $\Lambda'\beta$, is estimable if and only if $\Lambda' = P'X$ for some matrix P. The unique least squares estimate of $\Lambda'\beta$ is then $P'MY = \Lambda'\hat{\beta}$.

We now show that the least squares estimate of $\lambda'\beta$ is unique only if $\lambda'\beta$ is estimable.

Theorem 2.2.4. $\lambda' = \rho' X$ if $\lambda'\hat{\beta}_1 = \lambda'\hat{\beta}_2$ for any $\hat{\beta}_1$, $\hat{\beta}_2$ that satisfy $X\hat{\beta}_1 = X\hat{\beta}_2 = MY$.

PROOF. Decompose λ into vectors in $C(X')$ and its orthogonal complement. Let N be the perpendicular projection operator onto $C(X')$; then we can write $\lambda = X'\rho_1 + (I-N)\rho_2$. We want to show that $(I-N)\rho_2 = 0$. This will be done by showing that $(I-N)\rho_2$ is orthogonal to every vector in $\mathbf{R}^p$.

By assumption, $\lambda'(\hat{\beta}_1 - \hat{\beta}_2) = 0$, and we know that $\rho_1' X(\hat{\beta}_1 - \hat{\beta}_2) = 0$; so we must have $\rho_2'(I-N)(\hat{\beta}_1 - \hat{\beta}_2) = 0$, and this holds for any least squares estimates $\hat{\beta}_1$, $\hat{\beta}_2$.

Let $\hat{\beta}_1$ be any least squares estimate and take v such that $v \perp C(X')$, then $\hat{\beta}_2 = \hat{\beta}_1 - v$ is a least squares estimate. This follows because $X\hat{\beta}_2 = X\hat{\beta}_1 - Xv = X\hat{\beta}_1 = MY$. Substituting above gives $0 = \rho_2'(I-N)(\hat{\beta}_1 - \hat{\beta}_2) = \rho_2'(I-N)v$ for any $v \perp C(X')$. Moreover, by definition of N for any $v \in C(X')$, $(I-N)v = 0$. It follows that $\rho_2'(I-N)v = 0$ for any $v \in \mathbf{R}^p$ and thus $(I-N)\rho_2 = 0$. □

When β is not identifiable, sometimes *side conditions* are arbitrarily imposed on the parameters to allow "estimation" of nonidentifiable parameters. Imposing side conditions amounts to choosing one particular least squares estimate of β. In our earlier discussion of estimation for Example 1.0.2, we presented two sets of parameter estimates. The first estimate, $\hat{\beta}_1$, arbitrarily imposed $\mu = 0$ and $\hat{\beta}_2$ arbitrarily imposed $\alpha_1 = 0$. Side conditions determine a particular least squares estimate by introducing a nonidentifiable, typically a linear nonestimable, constraint on the parameters. With $r \equiv r(X) < p$, one needs $p - r$ individual side conditions to identify the parameters and thus allow "estimation" of the otherwise nonidentifiable parameters. Initially, the model was overparameterized. A linear nonestimable constraint is chosen to remove the ambiguity. Fundamentally, one choice of side conditions is as good as any other. See the discussion near Corollary 3.3.8 for further explication of linear nonestimable constraints. The use of a side condition in one-way ANOVA is also considered in Chapter 4.

Personally, I find it silly to pretend that nonidentifiable functions of the parameters can be estimated. The one good thing about imposing arbitrary side conditions is that they allow computer programs to print out parameter estimates. But different programs use different (equally valid) side conditions, so the printed estimates may differ from program to program. Fortunately, the estimates should agree on all estimable (and, more generally, identifiable) functions of the parameters.

Least squares estimation is not a statistical procedure! Its justification as an optimal estimate is geometric, not statistical. Next we consider two statistical results on unbiased estimation related to least squares estimates. First, we note that least squares estimates of estimable functions are unbiased.

Proposition 2.2.5. If $\lambda' = \rho' X$, then $\mathrm{E}(\rho' MY) = \lambda'\beta$.

PROOF. $\mathrm{E}(\rho' MY) = \rho' M \mathrm{E}(Y) = \rho' MX\beta = \rho' X\beta = \lambda'\beta$. □

None of our results on least squares estimation involve the assumption that $\mathrm{Cov}(e) = \sigma^2 I$. Least squares provides unique estimates of identifiable functions and unbiased estimates of estimable functions, regardless of the covariance structure. The next three sections establish that least squares estimates have good statistical properties when $\mathrm{Cov}(e) = \sigma^2 I$.

We now consider unbiased estimation of the variance parameter σ^2. First write the *fitted values* (also called the *predicted values*)

$$\hat{Y} \equiv X\hat{\beta} = MY$$

and the *residuals*

$$\hat{e} \equiv Y - X\hat{\beta} = (I - M)Y.$$

The data vector Y can be decomposed as

$$Y = \hat{Y} + \hat{e} = MY + (I - M)Y.$$

The perpendicular projection of Y onto $C(X)$ (i.e., MY) provides an estimate of $X\beta$. Note that $MY = MX\beta + Me = X\beta + Me$ so that MY equals $X\beta$ plus some error where $\mathrm{E}(Me) = M\mathrm{E}(e) = 0$. Similarly, $(I - M)Y = (I - M)X\beta + (I - M)e = (I - M)e$, so $(I - M)Y$ depends only on the error vector e. Since σ^2 is a property of the error vector, it is reasonable to use $(I - M)Y$ to estimate σ^2.

Theorem 2.2.6. Let $r(X) = r$ and $\mathrm{Cov}(e) = \sigma^2 I$, then $Y'(I - M)Y/(n - r)$ is an unbiased estimate of σ^2.

PROOF. From Theorem 1.3.2 and the facts that $\mathrm{E}(Y) = X\beta$ and $\mathrm{Cov}(Y) = \sigma^2 I$, we have

$$\mathrm{E}\left[Y'(I - M)Y\right] = \mathrm{tr}\left[\sigma^2 (I - M)\right] + \beta' X'(I - M)X\beta.$$

However, $\mathrm{tr}\big(\sigma^2(I-M)\big) = \sigma^2\,\mathrm{tr}(I-M) = \sigma^2 r(I-M) = \sigma^2\,(n-r)$ and $(I-M)X = 0$, so $\beta'X'(I-M)X\beta = 0$; therefore,

$$\mathrm{E}\big[Y'(I-M)Y\big] = \sigma^2\,(n-r)$$

and

$$\mathrm{E}\big[Y'(I-M)Y/(n-r)\big] = \sigma^2. \qquad \square$$

$Y'(I-M)Y$ is called the *sum of squares for error* (SSE). It is the squared length of the residual vector $(I-M)Y$. $Y'(I-M)Y/(n-r)$ is called the *mean squared error* (MSE). It is the squared length of $(I-M)Y$ divided by the rank of $(I-M)$. In a sense, the MSE is the average squared length of $(I-M)Y$, where the average is over the number of dimensions in $C(I-M)$. The rank of $I-M$ is called the *degrees of freedom for error*, denoted dfE.

For Example 1.0.1,

$$SSE = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 2)^2 + \cdots + (y_6 - \hat{\beta}_0 - \hat{\beta}_1 6)^2$$

and $MSE = SSE/(6-2)$. For Example 1.0.2,

$$SSE = (y_{11} - \bar{y}_{1\cdot})^2 + (y_{12} - \bar{y}_{1\cdot})^2 + (y_{13} - \bar{y}_{1\cdot})^2 + (y_{21} - y_{21})^2 + (y_{31} - \bar{y}_{3\cdot})^2 + (y_{32} - \bar{y}_{3\cdot})^2$$

and $MSE = SSE/(6-3)$.

Finally, one can think about the geometry of least squares estimation in three dimensions. Consider a rectangular table. (Yes, that furniture you have in your kitchen!) Take one corner of the table to be the origin. Take $C(X)$ as the two-dimensional subspace determined by the surface of the table. Y can be any vector originating at the origin, i.e., any point in three-dimensional space. The linear model says that $\mathrm{E}(Y) = X\beta$, which just says that $\mathrm{E}(Y)$ is somewhere on the surface of the table. The least squares estimate $MY = X\hat{\beta}$ is the perpendicular projection of Y onto the table surface. The residual vector $(I-M)Y$ is the vector starting at the origin, perpendicular to the surface of the table, that reaches the same height as Y. Another way to think of the residual vector is to connect the ends of MY and Y with a line segment (that is perpendicular to the surface of the table) but then shift the line segment along the surface (keeping it perpendicular) until the line segment has one end at the origin. The residual vector is the perpendicular projection of Y onto $C(I-M)$, that is, the projection onto the orthogonal complement of the table surface. The orthogonal complement is the one-dimension space in the vertical direction that goes through the origin. Because the orthogonal complement has only one dimension, MSE is just the squared length of the residual vector.

Alternatively, one could take $C(X)$ to be a one-dimensional subspace determined by an edge of the table that includes the origin. The linear model now says that $\mathrm{E}(Y)$ is somewhere on this edge of the table. $MY = X\hat{\beta}$ is found by dropping a perpendicular from Y to the edge of the table. If you connect MY and Y, you essentially get the residual vector $(I-M)Y$, except that the line segment has to be shifted down the edge so that it has one end at the origin. The residual vector is perpendicular to the $C(X)$ edge of the table, but typically would not be perpendicular to the surface of the table. $C(I-M)$ is now the plane that contains everything (through the origin) that is perpendicular to the $C(X)$ edge of the table. In other words, $C(I-M)$ is the two-dimensional space determined by the vertical direction and the *other* edge of the table that goes through the origin. *MSE* is the squared length of the residual vector divided by 2, because $C(I-M)$ is a two-dimensional space.

2.3 Estimation: Best Linear Unbiased

Another criterion for estimation of $\lambda'\beta$ is to choose the best linear unbiased estimate (*BLUE*) of $\lambda'\beta$. We prove the Gauss–Markov theorem that least squares estimates are best linear unbiased estimates.

Definition 2.3.1. $a'Y$ is a *best linear unbiased estimate* of $\lambda'\beta$ if $a'Y$ is unbiased and if for any other linear unbiased estimate $b'Y$, $\mathrm{Var}(a'Y) \le \mathrm{Var}(b'Y)$.

Gauss–Markov Theorem 2.3.2. Consider the linear model

$$Y = X\beta + e, \qquad \mathrm{E}(e) = 0, \qquad \mathrm{Cov}(e) = \sigma^2 I.$$

If $\lambda'\beta$ is estimable, then the least squares estimate of $\lambda'\beta$ is a BLUE of $\lambda'\beta$.

PROOF. Let M be the perpendicular projection operator onto $C(X)$. Since $\lambda'\beta$ is an estimable function, let $\lambda' = \rho'X$ for some ρ. We need to show that if $a'Y$ is an unbiased estimate of $\lambda'\beta$, then $\mathrm{Var}(a'Y) \ge \mathrm{Var}(\rho'MY)$. Since $a'Y$ is unbiased for $\lambda'\beta$, $\lambda'\beta = \mathrm{E}(a'Y) = a'X\beta$ for any value of β. Therefore $\rho'X = \lambda' = a'X$. Write

$$\begin{aligned}\mathrm{Var}(a'Y) &= \mathrm{Var}(a'Y - \rho'MY + \rho'MY)\\ &= \mathrm{Var}(a'Y - \rho'MY) + \mathrm{Var}(\rho'MY) + 2\mathrm{Cov}\left[(a'Y - \rho'MY), \rho'MY\right].\end{aligned}$$

Since $\mathrm{Var}(a'Y - \rho'MY) \ge 0$, if we show that $\mathrm{Cov}[(a'Y - \rho'MY), \rho'MY] = 0$, then $\mathrm{Var}(a'Y) \ge \mathrm{Var}(\rho'MY)$ and the theorem holds.

We now show that $\mathrm{Cov}[(a'Y - \rho'MY), \rho'MY] = 0$.

$$\begin{aligned}\mathrm{Cov}\left[(a'Y - \rho'MY), \rho'MY\right] &= \mathrm{Cov}\left[(a' - \rho'M)Y, \rho'MY\right]\\ &= (a' - \rho'M)\mathrm{Cov}(Y)M\rho\\ &= \sigma^2(a' - \rho'M)M\rho\\ &= \sigma^2(a'M - \rho'M)\rho.\end{aligned}$$

As shown above, $a'X = \rho'X$, and since we can write $M = X(X'X)^{-}X'$, we have $a'M = \rho'M$. It follows that $\sigma^2(a'M - \rho'M)\rho = 0$ as required. □

Corollary 2.3.3. If $\sigma^2 > 0$, there exists a unique BLUE for any estimable function $\lambda'\beta$.

PROOF. Let $\lambda' = \rho'X$, and recall from Section 1 that the vector $\rho'M$ is uniquely determined by λ'. In the proof of Theorem 2.3.2, it was shown that for an arbitrary linear unbiased estimate $a'Y$,

$$\mathrm{Var}(a'Y) = \mathrm{Var}(\rho'MY) + \mathrm{Var}(a'Y - \rho'MY).$$

If $a'Y$ is a BLUE of $\lambda'\beta$, it must be true that $\mathrm{Var}(a'Y - \rho'MY) = 0$. It is easily seen that

$$0 = \mathrm{Var}(a'Y - \rho'MY) = \mathrm{Var}\big[(a' - \rho'M)Y\big] = \sigma^2(a - M\rho)'(a - M\rho).$$

For $\sigma^2 > 0$, this occurs if and only if $a - M\rho = 0$, which is equivalent to the condition $a = M\rho$. □

2.4 Estimation: Maximum Likelihood

Another criterion for choosing estimates of β and σ^2 is maximum likelihood. The likelihood function is derived from the joint density of the observations by considering the parameters as variables and the observations as fixed at their observed values. If we assume $Y \sim N\big(X\beta, \sigma^2 I\big)$, then the *maximum likelihood estimates* (*MLE*s) of β and σ^2 are obtained by maximizing

$$(2\pi)^{-n/2}[\det(\sigma^2 I)]^{-1/2} \exp\big[-(Y - X\beta)'(Y - X\beta)/2\sigma^2\big]. \tag{1}$$

Equivalently, the log of the likelihood can be maximized. The log of (1) is

$$\frac{-n}{2}\log(2\pi) - \frac{1}{2}\log[(\sigma^2)^n] - (Y - X\beta)'(Y - X\beta)/2\sigma^2.$$

For every value of σ^2, the log-likelihood is maximized by taking β to minimize $(Y - X\beta)'(Y - X\beta)$, i.e., least squares estimates are MLEs. To estimate σ^2 we can substitute $Y'(I - M)Y = (Y - X\hat{\beta})'(Y - X\hat{\beta})$ for $(Y - X\beta)'(Y - X\beta)$ and differentiate with respect to σ^2 to get $Y'(I - M)Y/n$ as the MLE of σ^2.

The MLE of σ^2 is rarely used in practice. The *MSE* is the standard estimate of σ^2. For almost any purpose except point estimation of σ^2, it is immaterial whether the *MSE* or the MLE is used. They lead to identical confidence intervals and tests for σ^2. They also lead to identical confidence regions and tests for estimable functions of β. It should be emphasized that it is not appropriate to substitute the MLE for the *MSE* and then form confidence intervals and tests as if the *MSE* were being used.

2.5 Estimation: Minimum Variance Unbiased

In Section 3, it was shown that least squares estimates give best estimates among the class of linear unbiased estimates. If the error vector is normally distributed, least squares estimates are best estimates among all unbiased estimates. In particular, with normal errors, the best estimates happen to be linear estimates. As in Section 3, a best unbiased estimate is taken to be an unbiased estimate with minimum variance.

It is not the purpose of this monograph to develop the theory of minimum variance unbiased estimation. However, we will outline the application of this theory to linear models. See Lehmann (1983, Sections 1.4, 1.5) and Lehmann (1986, Sections 2.6, 2.7, 4.3) for a detailed discussion of the definitions and theorems used here. Our model is

$$Y = X\beta + e, \qquad e \sim N\big(0, \sigma^2 I\big)\,.$$

Definition 2.5.1. A vector-valued sufficient statistic $T(Y)$ is said to be *complete* if $\mathrm{E}[h(T(Y))] = 0$ for all β and σ^2 implies that $\Pr[h(T(Y)) = 0] = 1$ for all β and σ^2.

Theorem 2.5.2. If $T(Y)$ is a complete sufficient statistic, then $f(T(Y))$ is a minimum variance unbiased estimate (MVUE) of $\mathrm{E}[f(T(Y))]$.

PROOF. Suppose $g(Y)$ is an unbiased estimate of $\mathrm{E}[f(T(Y))]$. By the Rao–Blackwell theorem (see Cox and Hinkley, 1974),

$$\mathrm{Var}(\mathrm{E}[g(Y)|T(Y)]) \le \mathrm{Var}(g(Y))\,.$$

Since $\mathrm{E}[g(Y)|T(Y)]$ is unbiased, $\mathrm{E}\{f(T(Y)) - \mathrm{E}[g(Y)|T(Y)]\} = 0$. By completeness of $T(Y)$, $\Pr\{f(T(Y)) = \mathrm{E}[g(Y)|T(Y)]\} = 1$. It follows that $\mathrm{Var}(f(T(Y))) \le \mathrm{Var}(g(Y))$. □

We wish to use the following result from Lehmann (1983, pp. 28, 46):

Theorem 2.5.3. Let $\theta = (\theta_1, \ldots, \theta_s)'$ and let Y be a random vector with probability density function

$$f(Y) = c(\theta)\exp\left[\sum_{i=1}^{s} \theta_i T_i(Y)\right] h(Y);$$

then $T(Y) = (T_1(Y), T_2(Y), \ldots, T_s(Y))'$ is a complete sufficient statistic provided that neither θ nor $T(Y)$ satisfy any linear constraints.

Suppose $r(X) = r < p$, then the theorem cannot apply to $X'Y$ because, for $b \perp C(X')$, $Xb = 0$; so $b'X'Y$ is subject to a linear constraint. We need to consider the following reparameterization. Let Z be an $n \times r$ matrix whose columns form a basis for $C(X)$. For some matrix U, we have $X = ZU$. Let $\lambda'\beta$ be an estimable function. Then for some ρ, $\lambda'\beta = \rho'X\beta = \rho'ZU\beta$. Define $\gamma = U\beta$ and consider the linear model

$$Y = Z\gamma + e, \qquad e \sim N\big(0, \sigma^2 I\big).$$

The usual estimate of $\lambda'\beta = \rho'Z\gamma$ is $\rho'MY$ regardless of the parameterization used. We will show that $\rho'MY$ is a minimum variance unbiased estimate of $\lambda'\beta$. The density of Y can be written

$$\begin{aligned} f(Y) &= (2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2} \exp\left[-(Y - Z\gamma)'(Y - Z\gamma)/2\sigma^2\right] \\ &= C_1(\sigma^2)\exp\left[-\left(Y'Y - 2\gamma'Z'Y + \gamma'Z'Z\gamma\right)/2\sigma^2\right] \\ &= C_2(\gamma, \sigma^2)\exp\left[(-1/2\sigma^2)Y'Y + (\sigma^{-2}\gamma')(Z'Y)\right]. \end{aligned}$$

This is the form of Theorem 2.5.3. There are no linear constraints on the parameters $(-1/2\sigma^2, \gamma_1/\sigma^2, \ldots, \gamma_r/\sigma^2)$ nor on $(Y'Y, Y'Z)'$, so $(Y'Y, Y'Z)'$ is a complete sufficient statistic. An unbiased estimate of $\lambda'\beta = \rho'X\beta$ is $\rho'MY = \rho'Z(Z'Z)^{-1}Z'Y$. $\rho'MY$ is a function of $Z'Y$, so it is a minimum variance unbiased estimate. Moreover, $Y'(I - M)Y/(n - r)$ is an unbiased estimate of σ^2 and $Y'(I - M)Y = Y'Y - (Y'Z)(Z'Z)^{-1}(Z'Y)$ is a function of the complete sufficient statistic $(Y'Y, Y'Z)'$, so MSE is a minimum variance unbiased estimate. We have established the following result:

Theorem 2.5.4. *MSE* is a minimum variance unbiased estimate of σ^2 and $\rho'MY$ is a minimum variance unbiased estimate of $\rho'X\beta$ whenever $e \sim N\big(0, \sigma^2 I\big)$.

2.6 Sampling Distributions of Estimates

If we continue to assume that $Y \sim N\big(X\beta, \sigma^2 I\big)$, the distributions of estimates are straightforward. The least squares estimate of $\Lambda'\beta$ where $\Lambda' = P'X$ is $\Lambda'\hat{\beta} = P'MY$. The distribution of $P'MY$ is $N\big(P'X\beta, \sigma^2 P'MIMP\big)$ or, equivalently,

$$P'MY \sim N\big(\Lambda'\beta, \sigma^2 P'MP\big).$$

Since $M = X(X'X)^- X'$, we can also write

$$\Lambda'\hat{\beta} \sim N\big(\Lambda'\beta, \sigma^2 \Lambda'(X'X)^- \Lambda\big).$$

Two special cases are of interest. First, the estimate of $X\beta$ is

$$\hat{Y} \equiv MY \sim N\big(X\beta, \sigma^2 M\big).$$

Second, if $(X'X)$ is nonsingular, β is estimable and

$$\hat{\beta} \sim N\big(\beta, \sigma^2(X'X)^{-1}\big).$$

In Section 2 it was shown that the mean square error $Y'(I-M)Y/(n-r)$ is an unbiased estimate of σ^2. We now show that $Y'(I-M)Y/\sigma^2 \sim \chi^2(n-r)$. Clearly $Y/\sigma \sim N(X\beta/\sigma, I)$, so by Theorem 1.3.3

$$Y'(I-M)Y/\sigma^2 \sim \chi^2\big(r(I-M), \beta'X'(I-M)X\beta/2\sigma^2\big).$$

We have already shown that $r(I-M) = n-r$ and $\beta'X'(I-M)X\beta/2\sigma^2 = 0$. Moreover, by Theorem 1.3.7, MY and $Y'(I-M)Y$ are independent.

Exercise 2.1 Show that for $\lambda'\beta$ estimable,

$$\frac{\lambda'\hat{\beta} - \lambda'\beta}{\sqrt{MSE\,\lambda'(X'X)^{-}\lambda}} \sim t(dfE).$$

Find the form of an α level test of $H_0 : \lambda'\beta = 0$ and the form for a $(1-\alpha)100\%$ confidence interval for $\lambda'\beta$.

Hint: The test and confidence interval can be found using the methods of Appendix E.

Exercise 2.2 Let $y_{11}, y_{12}, \ldots, y_{1r}$ be $N(\mu_1, \sigma^2)$ and $y_{21}, y_{22}, \ldots, y_{2s}$ be $N(\mu_2, \sigma^2)$ with all y_{ij}s independent. Write this as a linear model. For the rest of the problem use the results of Chapter 2. Find estimates of μ_1, μ_2, $\mu_1 - \mu_2$, and σ^2. Using Appendix E and Exercise 2.1, form an $\alpha = .01$ test for $H_0 : \mu_1 = \mu_2$. Similarly, form 95% confidence intervals for $\mu_1 - \mu_2$ and μ_1. What is the test for $H_0 : \mu_1 = \mu_2 + \Delta$, where Δ is some known fixed quantity? How do these results compare with the usual analysis for two independent samples?

Exercise 2.3 Let $y_1, y_2, \ldots, y_n$ be independent $N(\mu, \sigma^2)$. Write a linear model for these data. For the rest of the problem use the results of Chapter 2, Appendix E, and Exercise 2.1. Form an $\alpha = .01$ test for $H_0 : \mu = \mu_0$, where μ_0 is some known fixed number and form a 95% confidence interval for μ. How do these results compare with the usual analysis for one sample?

2.7 Generalized Least Squares

A slightly more general linear model than the one considered so far is

$$Y = X\beta + e, \qquad \mathrm{E}(e) = 0, \qquad \mathrm{Cov}(e) = \sigma^2 V, \tag{1}$$

where V is some known positive definite matrix. By Corollary B.23, we can write $V = QQ'$ for some nonsingular matrix Q. It follows that $Q^{-1}VQ'^{-1} = I$.

Instead of analyzing model (1), we analyze the equivalent model,

$$Q^{-1}Y = Q^{-1}X\beta + Q^{-1}e. \tag{2}$$

For model (2), $\mathrm{E}(Q^{-1}e) = 0$ and $\mathrm{Cov}(Q^{-1}e) = \sigma^2 Q^{-1}VQ^{-1\prime} = \sigma^2 I$. The transformed model satisfies the assumptions made in the previously developed theory. For the transformed model, the least squares estimates minimize

$$\begin{aligned}\left(Q^{-1}Y - Q^{-1}X\beta\right)'\left(Q^{-1}Y - Q^{-1}X\beta\right) &= (Y - X\beta)'Q^{-1\prime}Q^{-1}(Y - X\beta)\\ &= (Y - X\beta)'V^{-1}(Y - X\beta).\end{aligned}$$

The estimates of β that minimize this function are called *generalized least squares estimates* because instead of minimizing the squared distance between Y and $X\beta$, a generalized squared distance determined by V^{-1} is minimized. Generalized least squares is a concept in linear model theory and should not be confused with generalized linear models. To differentiate from generalized least squares, the least squares estimation of Section 2 is sometimes called *ordinary least squares (OLS)*.

Theorem 2.7.1.

(a) $\lambda'\beta$ is estimable in model (1) if and only if $\lambda'\beta$ is estimable in model (2).

(b) $\hat{\beta}$ is a generalized least squares estimate of β if and only if

$$X\left(X'V^{-1}X\right)^{-}X'V^{-1}Y = X\hat{\beta}.$$

For any estimable function there exists a unique generalized least squares estimate.

(c) For an estimable function $\lambda'\beta$, the generalized least squares estimate is the BLUE of $\lambda'\beta$.

(d) If $e \sim N(0, \sigma^2 V)$, then for any estimable function $\lambda'\beta$, the generalized least squares estimate is the minimum variance unbiased estimate.

(e) If $e \sim N(0, \sigma^2 V)$, then any generalized least squares estimate of β is a maximum likelihood estimate of β.

PROOF.

(a) If $\lambda'\beta$ is estimable in model (1), we can write

$$\lambda' = \rho' X = (\rho' Q)Q^{-1}X;$$

so $\lambda'\beta$ is estimable in model (2). If $\lambda'\beta$ is estimable in model (2), then $\lambda' = \rho' Q^{-1}X = (\rho' Q^{-1})X$; so $\lambda'\beta$ is estimable in model (1).

(b) By Theorem 2.2.1, the generalized least squares estimates (i.e., the least squares estimates for model (2)) satisfy the equation

$$Q^{-1}X\left(X'Q^{-1\prime}Q^{-1}X\right)^{-}X'Q^{-1\prime}Q^{-1}Y = Q^{-1}X\hat{\beta}.$$

Simplifying and multiplying through on the left by Q gives the equivalent condition

$$X\left(X'V^{-1}X\right)^{-}X'V^{-1}Y = X\hat{\beta}.$$

From Theorem 2.2.3, generalized least squares estimates of estimable functions are unique.

(c) From Theorem 2.3.2 as applied to model (2), the generalized least squares estimate of $\lambda'\beta$ is the BLUE of $\lambda'\beta$ among all unbiased linear combinations of the vector $Q^{-1}Y$. However, any linear combination, in fact any function, of Y can be obtained from $Q^{-1}Y$ because Q^{-1} is invertible. Thus, the generalized least squares estimate is the BLUE.

(d) Applying Theorem 2.5.4 to model (2) establishes that the generalized least squares estimate is the MVUE from among unbiased estimates that are functions of $Q^{-1}Y$. Since Q is nonsingular, any function of Y can be written as a function of $Q^{-1}Y$; so the generalized least squares estimate is the minimum variance unbiased estimate.

(e) The likelihood functions from models (1) and (2) are identical. From model (2), a generalized least squares estimate $\hat{\beta}$ maximizes the likelihood among all functions of $Q^{-1}Y$, but since Q is nonsingular, $\hat{\beta}$ maximizes the likelihood among all functions of Y. □

Theorem 2.7.1(b) is the generalized least squares equivalent of Theorem 2.2.1. Theorem 2.2.1 relates $X\hat{\beta}$ to the perpendicular projection of Y onto $C(X)$. Theorem 2.7.1(b) also relates $X\hat{\beta}$ to a projection of Y onto $C(X)$, but in Theorem 2.7.1(b) the projection is not the perpendicular projection. If we write

$$A = X\left(X'V^{-1}X\right)^{-}X'V^{-1}, \tag{3}$$

then the condition in Theorem 2.7.1(b) is

$$AY = X\hat{\beta}.$$

We wish to show that A is a projection operator onto $C(X)$. The perpendicular projection operator onto $C(Q^{-1}X)$ is

$$Q^{-1}X\left[(Q^{-1}X)'(Q^{-1}X)\right]^{-}(Q^{-1}X)'.$$

By the definition of a projection operator,

$$Q^{-1}X\left[(Q^{-1}X)'(Q^{-1}X)\right]^{-}(Q^{-1}X)'Q^{-1}X = Q^{-1}X.$$

This can also be written as

$$Q^{-1}AX = Q^{-1}X.$$

Multiplying on the left by Q gives

$$AX = X. \tag{4}$$

From (3) and (4), we immediately have

$$AA = A,$$

so A is a projection matrix. From (3), $C(A) \subset C(X)$ and from (4), $C(X) \subset C(A)$; so $C(A) = C(X)$ and we have proven:

Proposition 2.7.2. A is a projection operator onto $C(X)$.

For an estimable function $\lambda'\beta$ with $\lambda' = \rho'X$, the generalized least squares estimate is $\lambda'\hat{\beta} = \rho'AY$. This result is analogous to the ordinary least squares result in Corollary 2.2.3. To obtain tests and confidence intervals for $\lambda'\beta$, we need to know $\text{Cov}(X\hat{\beta})$.

Proposition 2.7.3. $\text{Cov}(X\hat{\beta}) = \sigma^2 X\left(X'V^{-1}X\right)^{-}X'$.

PROOF. $\text{Cov}(X\hat{\beta}) = \text{Cov}(AY) = \sigma^2 AVA'$. From (3) and (4) it is easily seen (cf. Exercise 2.4) that

$$AVA' = AV = VA'.$$

In particular, $AV = X\left(X'V^{-1}X\right)^{-}X'$. □

Corollary 2.7.4. If $\lambda'\beta$ is estimable, then the generalized least squares estimate has $\text{Var}(\lambda'\hat{\beta}) = \sigma^2\lambda'\left(X'V^{-1}X\right)^{-}\lambda$.

It is necessary to have an estimate of σ^2. From model (2),

$$\begin{aligned}
SSE &= (Q^{-1}Y)'\left[I - Q^{-1}X\left[(Q^{-1}X)'(Q^{-1}X)\right]^{-}(Q^{-1}X)'\right](Q^{-1}Y)\\
&= Y'V^{-1}Y - Y'V^{-1}X\left(X'V^{-1}X\right)^{-}X'V^{-1}Y\\
&= Y'(I-A)'V^{-1}(I-A)Y.
\end{aligned}$$

Note that $(I-A)Y$ is the vector of residuals, so the SSE is a quadratic form in the residuals. Because Q is nonsingular, $r(Q^{-1}X) = r(X)$. It follows from model (2) that an unbiased estimate of σ^2 is obtained from

$$MSE = Y'(I-A)'V^{-1}(I-A)Y\big/[n-r(X)]\,.$$

With normal errors, this is also the minimum variance unbiased estimate of σ^2.

Suppose that e is normally distributed. From Theorem 1.3.7 applied to model (2), the MSE is independent of $Q^{-1}X\hat{\beta}$. Since $X\hat{\beta}$ is a function of $Q^{-1}X\hat{\beta}$, the MSE is independent of $X\hat{\beta}$. Moreover, $X\hat{\beta}$ is normally distributed and SSE/σ^2 has a chi-squared distribution.

A particular application of these results is that, for an estimable function $\lambda'\beta$,

$$\frac{\lambda'\hat{\beta}-\lambda'\beta}{\sqrt{MSE\,\lambda'\left(X'V^{-1}X\right)^{-}\lambda}} \sim t(n-r(X)).$$

Given this distribution, tests and confidence intervals involving $\lambda'\beta$ can be obtained as in Appendix E.

We now give a result that determines when generalized least squares estimates are (ordinary) least squares estimates. The result will be generalized in Theorem 10.4.5. The generalization changes it to an if and only if statement for arbitrary covariance matrices.

Proposition 2.7.5. If V is nonsingular and $C(VX) \subset C(X)$, then least squares estimates are BLUEs.

PROOF. The proof proceeds by showing that $A \equiv X(X'V^{-1}X)^{-}X'V^{-1}$ is the perpendicular projection operator onto $C(X)$. We already know that A is a projection operator onto $C(X)$, so all we need to establish is that if $w \perp C(X)$, then $Aw = 0$.

V being nonsingular implies that the null spaces of VX and X are identical, so $r(VX) = r(X)$. With $C(VX) \subset C(X)$, we must have $C(VX) = C(X)$. $C(VX) = C(X)$ implies that for some matrices B_1 and B_2, $VXB_1 = X$ and $VX = XB_2$. Multiplying through by V^{-1} in both equations gives $XB_1 = V^{-1}X$ and $X = V^{-1}XB_2$, so $C(X) = C(V^{-1}X)$. It follows immediately that $C(X)^{\perp} = C(V^{-1}X)^{\perp}$. Now, $w \perp C(X)$ if and only if $w \perp C(V^{-1}X)$, so

$$Aw = \left[X(X'V^{-1}X)^{-}X'V^{-1}\right]w = X(X'V^{-1}X)^{-}\left[X'V^{-1}w\right] = 0. \qquad \square$$

Frequently in regression analysis, V is a diagonal matrix, in which case generalized least squares is referred to as *weighted least squares (WLS)*. Considerable simplification results.

Exercise 2.4

(a) Show that $AVA' = AV = VA'$.

(b) Show that $A'V^{-1}A = A'V^{-1} = V^{-1}A$.

(c) Show that A is the same for any choice of $\left(X'V^{-1}X\right)^{-}$.

The following result will be useful in Section 9. It is essentially the Pythagorean theorem and can be used directly to show Theorem 2.7.1(b), that $X\hat{\beta}$ is a generalized least squares estimate if and only if $X\hat{\beta} = AY$.

Lemma 2.7.6 Let $A = X(X'V^{-1}X)^{-}X'V^{-1}$, then

$$\begin{aligned}(Y - X\beta)V^{-1}(Y - X\beta) &= (Y - AY)'V^{-1}(Y - AY) + (AY - X\beta)V^{-1}(AY - X\beta)\\ &= (Y - AY)'V^{-1}(Y - AY) + (\hat{\beta} - \beta)'(X'V^{-1}X)(\hat{\beta} - \beta)\end{aligned}$$

where $\hat{\beta} = (X'V^{-1}X)^{-}X'V^{-1}Y$.

PROOF. Following the proof of Theorem 2.2.2, write $(Y - X\beta) = (Y - AY) + (AY - X\beta)$ and eliminate cross product terms using Exercise 2.4 and

$$(I - A)'V^{-1}(AY - X\beta) = V^{-1}(I - A)(AY - X\beta) = 0. \qquad \square$$

Exercise 2.5 Show that A is the perpendicular projection operator onto $C(X)$ when the inner product between two vectors x and y is defined as $x'V^{-1}y$.

Hint: Recall the discussion after Definition B.50.

2.8 Normal Equations

An alternative method for finding least squares estimates of the parameter β in the model

$$Y = X\beta + e, \qquad \mathrm{E}(e) = 0, \qquad \mathrm{Cov}(e) = \sigma^2 I$$

is to find solutions of what are called the *normal equations*. The normal equations are defined as

$$X'X\beta = X'Y.$$

They are usually arrived at by setting equal to zero the partial derivatives of $(Y - X\beta)'(Y - X\beta)$ with respect to β.

Corollary 2.8.2 shows that solutions of the normal equations are least squares estimates of β. Recall that, by Theorem 2.2.1, least squares estimates are solutions of $X\beta = MY$.

Theorem 2.8.1. $\hat{\beta}$ is a least squares estimate of β if and only if $(Y - X\hat{\beta}) \perp C(X)$.

PROOF. Since M is the perpendicular projection operator onto $C(X)$, $(Y - X\hat{\beta}) \perp C(X)$ if and only if $M(Y - X\hat{\beta}) = 0$, i.e, if and only if $MY = X\hat{\beta}$. $\square$

Corollary 2.8.2. $\hat{\beta}$ is a least squares estimate of β if and only if $X'X\hat{\beta} = X'Y$.

PROOF. $X'X\hat{\beta} = X'Y$ if and only if $X'(Y - X\hat{\beta}) = 0$, which occurs if and only if $(Y - X\hat{\beta}) \perp C(X)$. □

For generalized least squares problems, the normal equations are found from model (2.7.2). The normal equations simplify to

$$X'V^{-1}X\beta = X'V^{-1}Y.$$

2.9 Bayesian Estimation

Bayesian estimation incorporates the analyst's subjective information about a problem into the analysis. It appears to be the only logically consistent method of analysis, but not the only useful one. Some people object to the loss of objectivity that results from using the analyst's subjective information, but either the data are strong enough for reasonable people to agree on their interpretation or, if not, analysts should be using their subjective (prior) information for making appropriate decisions related to the data.

There is a vast literature on Bayesian statistics. Three fundamental works are de Finetti (1974, 1975), Jeffreys (1961), and Savage (1954). Good elementary introductions to the subject are Lindley (1971) and Berry (1996). A few of the well-known books on the subject are Berger (1993), Box and Tiao (1973), DeGroot (1970), Geisser (1993), Raiffa and Schlaifer (1961), and Zellner (1971). My favorite is now Christensen, Johnson, Branscum, and Hanson (2010), which also includes many more references to many more excellent books. Christensen et al. contains a far more extensive discussion of Bayesian linear models than this relatively short section.

Consider the linear model

$$Y = X\beta + e, \quad e \sim N(0, \sigma^2 I),$$

where $r(X) = r$. It will be convenient to consider a full rank reparameterization of this model,

$$Y = Z\gamma + e, \quad e \sim N(0, \sigma^2 I),$$

where $C(X) = C(Z)$. As in Sections 1.2 and 2.4, this determines a density for Y given γ and σ^2, say, $f(Y|\gamma, \sigma^2)$. For a Bayesian analysis, we must have a joint density for γ and σ^2, say $p(\gamma, \sigma^2)$. This distribution reflects the analyst's beliefs, prior to collecting data, about the process of generating the data. We will actually specify this distribution conditionally as $p(\gamma, \sigma^2) = p(\gamma|\sigma^2)p(\sigma^2)$. In practice, it is difficult to specify these distributions for γ given σ^2 and σ^2. Convenient choices that have minimal impact on (most aspects of) the analysis are the (improper) reference priors $p(\gamma|\sigma^2) = 1$ and $p(\sigma^2) = 1/\sigma^2$. These are improper in that neither prior density integrates to 1. Although these priors are convenient, a true Bayesian analysis requires the specification of proper prior distributions.

Specifying prior information is difficult, particularly about such abstract quantities as regression coefficients. A useful tool in specifying prior information is to think in terms of the mean of potential observations. For example, we could specify a vector of predictor variables, say $\tilde{z}_i$, and specify the distribution for the mean of observations having those predictor variables. With covariates $\tilde{z}_i$, the mean of potential observables is $\tilde{z}_i'\gamma$. Typically, we assume that $\tilde{z}_i'\gamma$ has a $N(\tilde{y}_i, \sigma^2\tilde{w}_i)$ distribution. One way to think about $\tilde{y}_i$ and $\tilde{w}_i$ is that $\tilde{y}_i$ is a prior guess for what one would see with covariates $\tilde{z}_i$, and $1/\tilde{w}_i$ is the number of observations this guess is worth. To specify a proper prior distribution for γ given σ^2, we specify independent priors at vectors $\tilde{z}_i$, $i = 1,\ldots,r$, where r is the dimension of γ. As will be seen later, under a mild condition this prior specification leads easily to a proper prior distribution on γ. Although choosing the $\tilde{z}_i$s and specifying the priors may be difficult to do, it is much easier to do than trying to specify an intelligent joint prior directly on γ. If one wishes to specify only partial prior information, one can simply choose fewer than r vectors $\tilde{z}_i$ and the analysis follows as outlined below. In fact, using the reference prior for γ amounts to not choosing any $\tilde{z}_i$s. Again, the analysis follows as outlined below. Bedrick, Christensen, and Johnson (1996) and Christensen et al. (2010) discuss these techniques in more detail.

I believe that the most reasonable way to specify a proper prior on σ^2 is to think in terms of the variability of potential observables around some fixed mean. Unfortunately, the implications of this idea are not currently as well understood as they are for the related technique of specifying the prior for γ indirectly through priors on means of potential observables. For now, we will simply consider priors for σ^2 that are inverse gamma distributions, i.e., distributions in which $1/\sigma^2$ has a gamma distribution. An inverse gamma distribution has two parameters, a and b. One can think of the prior as being the equivalent of $2a$ (prior) observations with a prior guess for σ^2 of b/a.

It is convenient to write $\tilde{Y} = (\tilde{y}_1,\ldots,\tilde{y}_r)'$ and $\tilde{Z}$ as the $r \times r$ matrix with ith row $\tilde{z}_i'$. In summary, our distributional assumptions are

$$\begin{aligned} Y|\gamma,\sigma^2 &\sim N(Z\gamma,\sigma^2 I), \\ \tilde{Z}\gamma|\sigma^2 &\sim N(\tilde{Y},\sigma^2 D(\tilde{w})), \\ \sigma^2 &\sim InvGa(a,b). \end{aligned}$$

We assume that $\tilde{Z}$ is nonsingular, so that the second of these induces the distribution

$$\gamma|\sigma^2 \sim N(\tilde{Z}^{-1}\tilde{Y}, \sigma^2\tilde{Z}^{-1}D(\tilde{w})\tilde{Z}^{-1\prime}).$$

Actually, any multivariate normal distribution for γ given σ^2 will lead to essentially the same analysis as given here.

A Bayesian analysis is based on finding the distribution of the parameters given the data, i.e., $p(\gamma,\sigma^2|Y)$. This is accomplished by using Bayes's theorem, which states that

$$p(\gamma,\sigma^2|Y) = \frac{f(Y|\gamma,\sigma^2)p(\gamma,\sigma^2)}{f(Y)}.$$

If we know the numerator of the fraction, we can obtain the denominator by

$$f(Y) = \int f(Y|\gamma,\sigma^2)p(\gamma,\sigma^2)\,d\gamma d\sigma^2.$$

In fact, because of this relationship, we only need to know the numerator up to a constant multiple, because any multiple will cancel in the fraction.

Later in this section we will show that

$$\begin{aligned} p(\gamma,\sigma^2|Y) \propto & \left(\sigma^2\right)^{-(n+r)/2} p(\sigma^2) \\ & \times \exp\left\{\frac{-1}{2\sigma^2}\left[(\gamma-\hat{\gamma})'(Z'Z+\tilde{Z}'D^{-1}(\tilde{w})\tilde{Z})(\gamma-\hat{\gamma})\right]\right\} \\ & \times \exp\left\{\frac{-1}{2\sigma^2}\left[(Y-Z\hat{\gamma})'(Y-Z\hat{\gamma})+(\tilde{Y}-\tilde{Z}\hat{\gamma})'D^{-1}(\tilde{w})(\tilde{Y}-\tilde{Z}\hat{\gamma})\right]\right\}, \end{aligned} \tag{1}$$

where

$$\hat{\gamma} = \left(Z'Z+\tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)^{-1}\left[Z'Y+\tilde{Z}'D^{-1}(\tilde{w})\tilde{Y}\right]. \tag{2}$$

The joint posterior (post data) density is the righthand side of (1) divided by its integral with respect to γ and σ^2.

The form (1) for the joint distribution given the data is not particularly useful. What we really need are the marginal distributions of σ^2 and functions $\rho'Z\gamma \equiv \rho'X\beta$, and predictive distributions for new observations.

As will be shown, the Bayesian analysis turns out to be quite consistent with the frequentist analysis. For the time being, we use the reference prior $p(\sigma^2) = 1/\sigma^2$. In our model, we have $\tilde{Z}\gamma$ random, but it is convenient to think of $\tilde{Y}$ as being r independent prior observations with mean $\tilde{Z}\gamma$ and weights $\tilde{w}$. Now consider the generalized least squares model

$$\begin{bmatrix} Y \\ \tilde{Y} \end{bmatrix} = \begin{bmatrix} Z \\ \tilde{Z} \end{bmatrix}\gamma + \begin{bmatrix} e \\ \tilde{e} \end{bmatrix}, \quad \begin{bmatrix} e \\ \tilde{e} \end{bmatrix} \sim N\left(\begin{bmatrix} 0_{n\times 1} \\ 0_{r\times 1} \end{bmatrix}, \sigma^2\begin{bmatrix} I_n & 0 \\ 0 & D(\tilde{w}) \end{bmatrix}\right). \tag{3}$$

This generalized least squares model can also be written as, say,

$$Y_* = Z_*\gamma + e_*, \quad e_* \sim N(0,\sigma^2 V_*).$$

The generalized least squares estimate from this model is $\hat{\gamma}$ as given in (2). In the Bayesian analysis, $\hat{\gamma}$ is the expected value of γ given Y. Let *BMSE* denote the mean squared error from the (Bayesian) generalized least squares model with *BdfE* degrees of freedom for error.

In the frequentist generalized least squares analysis, for fixed γ with random $\hat{\gamma}$ and *BMSE*,

$$\frac{\lambda'\hat{\gamma}-\lambda'\gamma}{\sqrt{BMSE\,\lambda'\left(Z'Z+\tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)^{-1}\lambda}} \sim t(BdfE).$$

In the Bayesian analysis the same distribution holds, but for fixed $\hat{\gamma}$ and *BMSE* with random γ. Frequentist confidence intervals for $\lambda'\gamma$ are identical to Bayesian posterior probability intervals for $\lambda'\gamma$. Note that for estimating a function $\rho' X\beta$, simply write it as $\rho' X\beta = \rho' Z\gamma$ and take $\lambda' = \rho' Z$.

In the frequentist generalized least squares analysis, for fixed σ^2 and random $BMSE$,

$$\frac{(BdfE)BMSE}{\sigma^2} \sim \chi^2(BdfE).$$

In the Bayesian analysis, the same distribution holds, but for fixed $BMSE$ and random σ^2. Confidence intervals for σ^2 are identical to Bayesian posterior probability intervals for σ^2.

In the frequentist generalized least squares analysis, a prediction interval for a future independent observation y_0 with predictor vector z_0 and weight 1 is based on the distribution

$$\frac{y_0 - z_0'\hat{\gamma}}{\sqrt{BMSE\left[1 + z_0'\left(Z'Z + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)^{-1} z_0\right]}} \sim t(BdfE), \qquad (4)$$

where $\hat{\gamma}$ and $BMSE$ are random and y_0 is independent of Y for given γ and σ^2, see Exercise 2.10.1. In the Bayesian analysis, the same distribution holds, but for fixed $\hat{\gamma}$ and $BMSE$. Standard prediction intervals for y_0 are identical to Bayesian prediction intervals.

If we specify an improper prior on γ using fewer than r vectors $\tilde{z}_i$, these relationships between generalized least squares and the Bayesian analysis remain valid. In fact, for the reference prior on γ, i.e., choosing no $\tilde{z}_i$s, the generalized least squares model reduces to the usual model $Y = X\beta + e$, $e \sim N(0, \sigma^2 I)$, and the Bayesian analysis becomes analogous to the usual ordinary least squares analysis.

In the generalized least squares model (3), $BdfE = n$. If we take $\sigma^2 \sim InvGa(a,b)$, the only changes in the Bayesian analysis are that $BMSE$ changes to $[(BdfE)BMSE + 2b]/(BdfE + 2a)$ and the degrees of freedom for the t and χ^2 distributions change from $BdfE$ to $BdfE + 2a$. With reference priors for both γ and σ^2, $BdfE = n - r$, as in ordinary least squares.

EXAMPLE 2.9.1. Schafer (1987) presented data on 93 individuals at the Harris Bank of Chicago in 1977. The response is beginning salary in hundreds of dollars. There are four covariates: sex, years of education, denoted EDUC, years of experience, denoted EXP, and time at hiring as measured in months after 1-1-69, denoted TIME. This is a regression, so we can take $Z \equiv X$, $\gamma \equiv \beta$, and write $\tilde{X} \equiv \tilde{Z}$. With an intercept in the model, Johnson, Bedrick, and I began by specifying five covariate vectors $\tilde{x}_i' = (1, SEX_i, EDUC_i, EXP_i, TIME_i)$, say, $(1,0,8,0,0)$, $(1,1,8,0,0)$, $(1,1,16,0,0)$, $(1,1,16,30,0)$, and $(1,1,8,0,36)$, where a SEX value of 1 indicates a male. For example, the vector (1,0,8,0,0) corresponds to a male with 8 years of education, no previous experience, and starting work on 1-1-69. Thinking about the mean salary for each set of covariates, we chose $\tilde{y}' = (40, 40, 60, 70, 50)$, which reflects a prior belief that starting salaries are the same for equally qualified men and women and a belief that salary is increasing as a function of EDUC, EXP, and TIME. The weights $\tilde{w}_i$ are all chosen to be 0.4, so that in total the prior carries the same weight as two sampled observations. The induced prior on β given σ^2 has mean vector $(20, 0, 2.50, 0.33, 0.28)'$ and standard deviation vector $\sigma \equiv (2.74, 2.24, 0.28, 0.07, 0.06)'$.

To illustrate partial prior information, we consider the same example, only with the fifth "prior observation" deleted. In this instance, the prior does not reflect any information about the response at TIMEs other than 0. The prior is informative about the mean responses at the first four covariate vectors but is noninformative (the prior is constant) for the mean response at the fifth covariate vector. Moreover, the prior is constant for the mean response with any other choice of fifth vector, provided this vector is linearly independent of the other four. (All such vectors must have a nonzero value for the time component.) In this example, the induced improper distribution on β has the same means and standard deviations for β_1, β_2, β_3, and β_4, but is flat for β_5.

We specify a prior on σ^2 worth $2a = 2$ observations with a prior guess for σ^2 of $b/a = 25$. The prior guess reflects our belief that a typical salary has a standard deviation of 5.

Using our informative prior, the posterior mean of β is

$$\hat{\beta} = (33.68, 6.96, 1.02, 0.18, 0.23)'$$

with $BdfE = 95$ and $BMSE = 2404/95 = 25.3053$. The standard deviations for β are $\sqrt{95/93}(310, 114, 23.43, 6.76, 5.01)'/100$. In the partially informative case discussed above, the posterior mean is

$$\hat{\beta} = (34.04, 7.11, 0.99, 0.17, 0.23),$$

$BdfE = 94$, $BMSE = 2383/94 = 25.3511$, and the standard deviations for β are $\sqrt{94/92}(313, 116, 23.78, 6.80, 5.06)'/100$. Using the standard reference prior, i.e., using ordinary least squares without prior data, $\hat{\beta} = (35.26, 7.22, 0.90, 0.15, 0.23)'$, $BdfE = 88$, $BMSE = 2266/88 = 25.75$, and the standard deviations for β are $\sqrt{88/86}(328, 118, 24.70, 7.05, 5.20)'/100$. The 95% prediction interval for $x_0 = (1, 0, 10, 3.67, 7)'$ with a weight of 1 is $(35.97, 56.34)$ for the informative prior, $(35.99, 56.38)$ with partial prior information, and $(36.27, 56.76)$ for the reference prior.

2.9.1 Distribution Theory

For the time being, we will assume relation (1) for the joint posterior and use it to arrive at marginal distributions. Afterwards, we will establish relation (1).

The distribution theory for the Bayesian analysis involves computations unlike anything else done in this book. It requires knowledge of some basic facts.

The density of a gamma distribution with parameters $a > 0$ and $b > 0$ is

$$g(\tau) = \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp[-b\tau]$$

for $\tau > 0$. A $Gamma(n/2, 1/2)$ distribution is the same as a $\chi^2(n)$ distribution. We will not need the density of an inverse gamma, only the fact that y has a $Gamma(a,b)$ distribution if and only if $1/y$ has an $InvGa(a,b)$ distribution. The improper reference distribution corresponds to $a = 0, b = 0$.

The t distribution is defined in Appendix C. The density of a $t(n)$ distribution is

$$g(w) = \left[1 + \frac{w^2}{n}\right]^{-(n+1)/2} \Gamma\left(\frac{n+1}{2}\right) \Big/ \left[\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}\right].$$

Eliminating the constants required to make the density integrate to 1 gives

$$g(w) \propto \left[1 + \frac{w^2}{n}\right]^{-(n+1)/2}.$$

Bayesian linear models involve multivariate t distributions, cf. DeGroot (1970, Section 5.6). Let $W \sim N(\mu, V)$, $Q \sim \chi^2(n)$, with W and Q independent. Then if

$$Y \sim (W - \mu)\frac{1}{\sqrt{Q/n}} + \mu,$$

by definition

$$Y \sim t(n, \mu, V).$$

For an r vector Y, a multivariate $t(n)$ distribution with center μ and dispersion matrix V has density

$$g(y) = \left[1 + \frac{1}{n}(y-\mu)'V^{-1}(y-\mu)\right]^{-(n+r)/2} \times \Gamma\left(\frac{n+r}{2}\right) \Big/ \left\{\Gamma\left(\frac{n}{2}\right)(n\pi)^{r/2}[\det(V)]^{1/2}\right\}.$$

This distribution has mean μ for $n > 1$ and covariance matrix $[n/(n-2)]V$ for $n > 2$. To get noninteger degrees of freedom a, just replace Q/n in the definition with bT/a, where $T \sim Gamma(a/2, b/2)$ independent of W.

Note that from the definition of a multivariate t,

$$\frac{\lambda'Y - \lambda'\mu}{\sqrt{\lambda'V\lambda}} \sim \frac{(\lambda'W - \lambda'\mu)/\sqrt{\lambda'V\lambda}}{\sqrt{Q/n}} \sim t(n). \tag{5}$$

Proceeding with the Bayesian analysis, to find the marginal posterior of γ let

$$Q = \left[(Y - Z\hat{\gamma})'(Y - Z\hat{\gamma}) + (\tilde{Y} - \tilde{Z}\hat{\gamma})'D^{-1}(\tilde{w})(\tilde{Y} - \tilde{Z}\hat{\gamma})\right] + \left[(\gamma - \hat{\gamma})'(Z'Z + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Z})(\gamma - \hat{\gamma})\right].$$

From (1),

$$p(\gamma|Y) \propto \int \left(\sigma^2\right)^{-(n+r)/2} p(\sigma^2) \exp\left\{\frac{-1}{2\sigma^2} Q\right\} d\sigma^2.$$

Transforming σ^2 into $\tau = 1/\sigma^2$ gives $\sigma^2 = 1/\tau$ and $d\sigma^2 = |-\tau^{-2}|d\tau$. Thus

$$p(\gamma|Y) \propto \int (\tau)^{(n+r)/2} p(1/\tau) \exp\{-\tau Q/2\}\, \tau^{-2} d\tau.$$

Note that if σ^2 has an inverse gamma distribution with parameters a and b, then τ has a gamma distribution with parameters a and b; so $p(1/\tau)\tau^{-2}$ is a gamma density and

$$p(\gamma|Y) \propto \int (\tau)^{(n+r+2a-2)/2} \exp\{-\tau(Q+2b)/2\}\, d\tau.$$

The integral is a gamma integral, e.g., the gamma density given earlier integrates to 1, so

$$p(\gamma|Y) \propto \Gamma[(n+r+2a)/2] \Big/ [(Q+2b)/2]^{(n+r+2a)/2}$$

or

$$p(\gamma|Y) \propto [Q+2b]^{-(n+r+2a)/2}.$$

We can rewrite this as

$$\begin{aligned} p(\gamma|Y) &\propto \left[(BdfE)(BMSE) + 2b + (\gamma-\hat{\gamma})' \left(Z'Z + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)(\gamma-\hat{\gamma})\right]^{-(n+r+2a)/2} \\ &\propto \left[1 + \frac{1}{n+2a} \frac{(\gamma-\hat{\gamma})' \left(Z'Z + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)(\gamma-\hat{\gamma})}{[(BdfE)(BMSE)+2b]/(n+2a)}\right]^{-(n+2a+r)/2}, \end{aligned}$$

so

$$\gamma|Y \sim t\left(n+2a, \hat{\gamma}, \frac{(BdfE)(BMSE)+2b}{n+2a} \left(Z'Z + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)^{-1}\right).$$

Together with (5), this provides the posterior distribution of $\lambda'\gamma$.

Now consider the marginal (posterior) distribution of σ^2.

$$\begin{aligned} p(\sigma^2|Y) \propto{}& \left(\sigma^2\right)^{-n/2} p(\sigma^2) \\ &\times \exp\left\{\frac{-1}{2\sigma^2}\left[(Y-Z\hat{\gamma})'(Y-Z\hat{\gamma}) + (\tilde{Y}-\tilde{Z}\hat{\gamma})'D^{-1}(\tilde{w})(\tilde{Y}-\tilde{Z}\hat{\gamma})\right]\right\} \\ &\times \int \left(\sigma^2\right)^{-r/2} \exp\left\{\frac{-1}{2\sigma^2}\left[(\gamma-\hat{\gamma})'(Z'Z+\tilde{Z}'D^{-1}(\tilde{w})\tilde{Z})(\gamma-\hat{\gamma})\right]\right\} d\gamma. \end{aligned}$$

The term being integrated is proportional to a normal density, so the integral is a constant that does not depend on σ^2. Hence,

$$\begin{aligned} p(\sigma^2|Y) \propto{}& \left(\sigma^2\right)^{-n/2} p(\sigma^2) \\ &\times \exp\left\{\frac{-1}{2\sigma^2}\left[(Y-Z\hat{\gamma})'(Y-Z\hat{\gamma}) + (\tilde{Y}-\tilde{Z}\hat{\gamma})'D^{-1}(\tilde{w})(\tilde{Y}-\tilde{Z}\hat{\gamma})\right]\right\} \end{aligned}$$

or, using the generalized least squares notation,

$$p(\sigma^2|Y) \propto \left(\sigma^2\right)^{-n/2} p(\sigma^2)\exp\left[\frac{-1}{2\sigma^2}(BdfE)(BMSE)\right].$$

We transform to the *precision*, $\tau \equiv 1/\sigma^2$. The $InvGa(a,b)$ distribution for σ^2 yields

$$\begin{aligned} p(\tau|Y) &\propto (\tau)^{n/2}(\tau)^{a-1}\exp[-b\tau]\exp\left[\frac{-\tau}{2}(BdfE)(BMSE)\right] \\ &= (\tau)^{[(n+2a)/2]-1}\exp\left[-\frac{2b+(BdfE)(BMSE)}{2}\tau\right]; \end{aligned}$$

so

$$\tau|Y \sim Gamma\left(\frac{n+2a}{2}, \frac{2b+(BdfE)(BMSE)}{2}\right).$$

It is not difficult to show that

$$[2b+(BdfE)(BMSE)]\tau|Y \sim Gamma\left(\frac{n+2a}{2}, \frac{1}{2}\right),$$

i.e.,

$$\frac{2b+(BdfE)(BMSE)}{\sigma^2}\Big|Y \sim Gamma\left(\frac{n+2a}{2}, \frac{1}{2}\right).$$

As mentioned earlier, for the reference distribution with $a=0, b=0$,

$$\frac{(BdfE)(BMSE)}{\sigma^2}\Big|Y \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right) = \chi^2(n).$$

Finally, we establish relation (1).

$$\begin{aligned} p(\gamma,\sigma^2|Y) &\propto f(Y|\gamma,\sigma^2)p(\gamma|\sigma^2)p(\sigma^2) \\ &\propto \left\{\left(\sigma^2\right)^{-n/2}\exp\left[-(Y-Z\gamma)'(Y-Z\gamma)/2\sigma^2\right]\right\} \\ &\quad \times\left\{\left(\sigma^2\right)^{-r/2}\exp\left[-\left(\gamma-\tilde{Z}^{-1}\tilde{Y}\right)'\left(\tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)\left(\gamma-\tilde{Z}^{-1}\tilde{Y}\right)\Big/2\sigma^2\right]\right\} \\ &\quad \times p(\sigma^2) \end{aligned}$$

Most of the work involves simplifying the terms in the exponents. We isolate those terms, deleting the multiple $-1/2\sigma^2$.

$$\begin{aligned} &(Y-Z\gamma)'(Y-Z\gamma)+\left(\gamma-\tilde{Z}^{-1}\tilde{Y}\right)'\left(\tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)\left(\gamma-\tilde{Z}^{-1}\tilde{Y}\right) \\ &= (Y-Z\gamma)'(Y-Z\gamma)+(\tilde{Y}-\tilde{Z}\gamma)'D(\tilde{w})^{-1}(\tilde{Y}-\tilde{Z}\gamma) \\ &= \left(\begin{bmatrix} Y \\ \tilde{Y}\end{bmatrix}-\begin{bmatrix} Z \\ \tilde{Z}\end{bmatrix}\gamma\right)'\begin{bmatrix} I & 0 \\ 0 & D(\tilde{w})^{-1}\end{bmatrix}\left(\begin{bmatrix} Y \\ \tilde{Y}\end{bmatrix}-\begin{bmatrix} Z \\ \tilde{Z}\end{bmatrix}\gamma\right). \end{aligned}$$

Write

$$Y_* = \begin{bmatrix} Y \\ \tilde{Y} \end{bmatrix}, \quad Z_* = \begin{bmatrix} Z \\ \tilde{Z} \end{bmatrix}, \quad V_* = \begin{bmatrix} I & 0 \\ 0 & D(\tilde{w}) \end{bmatrix}$$

and apply Lemma 2.7.6 to get

$$\begin{aligned}
&(Y - Z\gamma)'(Y - Z\gamma) + \left(\gamma - \tilde{Z}^{-1}\tilde{Y}\right)' \left(\tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right) \left(\gamma - \tilde{Z}^{-1}\tilde{Y}\right) \\
&= (Y_* - Z_*\gamma)'V_*^{-1}(Y_* - Z_*\gamma) \\
&= (Y_* - A_*Y_*)'V_*^{-1}(Y_* - A_*Y_*) + (\hat{\gamma} - \gamma)'(Z_*'V_*^{-1}Z_*)(\hat{\gamma} - \gamma),
\end{aligned}$$

where $A_* = Z_*(Z_*'V_*^{-1}Z_*)^{-1}Z_*'V_*^{-1}$ and

$$\begin{aligned}
\hat{\gamma} &= (Z_*'V_*^{-1}Z_*)^{-1}Z_*'V_*^{-1}Y_* \\
&= \left(Z'Z + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Z}\right)^{-1} \left[Z'Y + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Y}\right].
\end{aligned}$$

Finally, observe that

$$\begin{aligned}
&(Y_* - A_*Y_*)'V_*^{-1}(Y_* - A_*Y_*) + (\hat{\gamma} - \gamma)'(Z_*'V_*^{-1}Z_*)(\hat{\gamma} - \gamma) \\
&= \left[(Y - Z\hat{\gamma})'(Y - Z\hat{\gamma}) + (\tilde{Y} - \tilde{Z}\hat{\gamma})'D^{-1}(\tilde{w})(\tilde{Y} - \tilde{Z}\hat{\gamma})\right] \\
&\quad + \left[(\gamma - \hat{\gamma})'(Z'Z + \tilde{Z}'D^{-1}(\tilde{w})\tilde{Z})(\gamma - \hat{\gamma})\right].
\end{aligned}$$

Substitution gives (1).

Exercise 2.6 Prove relation (4).

2.10 Additional Exercises

Exercise 2.10.1 Consider a regression model $Y = X\beta + e$, $e \sim N(0, \sigma^2 I)$ and suppose that we want to predict the value of a future observation, say y_0, that will be independent of Y and be distributed $N(x_0'\beta, \sigma^2)$.

(a) Find the distribution of

$$\frac{y_0 - x_0'\hat{\beta}}{\sqrt{MSE\left[1 + x_0'(X'X)^{-1}x_0\right]}}.$$

(b) Find a 95% prediction interval for y_0.

Hint: A prediction interval is similar to a confidence interval except that, rather than finding parameter values that are consistent with the data and the model, one finds new observations y_0 that are consistent with the data and the model as determined by an α level test.

(c) Let $\eta \in (0, 0.5]$. The 100ηth percentile of the distribution of y_0 is, say, $\gamma(\eta) = x_0'\beta + z(\eta)\sigma$. (Note that $z(\eta)$ is a negative number.) Find a $(1-\alpha)100\%$

lower confidence bound for $\gamma(\eta)$. In reference to the distribution of y_0, this lower confidence bound is referred to as a lower η tolerance point with confidence coefficient $(1-\alpha)100\%$. For example, if $\eta = 0.1$, $\alpha = 0.05$, and y_0 is the octane value of a batch of gasoline manufactured under conditions x_0, then we are 95% confident that no more than 10% of all batches produced under x_0 will have an octane value below the tolerance point.

Hint: Use a noncentral t distribution based on $x_0'\hat{\beta} - \gamma(\eta)$.

Comment: For more detailed discussions of prediction and tolerance (and we all know that tolerance is a great virtue), see Geisser (1993), Aitchison and Dunsmore (1975), and Guttman (1970).

Exercise 2.10.2 Consider the model

$$Y = X\beta + b + e, \qquad \mathrm{E}(e) = 0, \qquad \mathrm{Cov}(e) = \sigma^2 I,$$

where b is a known vector. Show that Proposition 2.1.3 is not valid for this model by producing a linear unbiased estimate of $\rho' X\beta$, say $a_0 + a'Y$, for which $a_0 \neq 0$.

Hint: Modify $\rho' MY$.

Exercise 2.10.3 Consider the model $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, e_is i.i.d. $N(0,\sigma^2)$. Use the data given below to answer (a) through (d). Show your work, i.e., do not use a regression or general linear models computer program.

(a) Estimate β_1, β_2, and σ^2.
(b) Give 95% confidence intervals for β_1 and $\beta_1 + \beta_2$.
(c) Perform an $\alpha = 0.01$ test for $H_0 : \beta_2 = 3$.
(d) Find an appropriate P value for the test of $H_0 : \beta_1 - \beta_2 = 0$.

obs.	1	2	3	4	5	6	7	8
y	82	79	74	83	80	81	84	81
x_1	10	9	9	11	11	10	10	12
x_2	15	14	13	15	14	14	16	13

Exercise 2.10.4 Consider the model $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, e_is i.i.d. $N(0,\sigma^2)$. There are 15 observations and the sum of the squared observations is $Y'Y = 3.03$. Use the normal equations given below to answer parts (a) through (c).

(a) Estimate β_1, β_2, and σ^2.
(b) Give 98% confidence intervals for β_2 and $\beta_2 - \beta_1$.
(c) Perform an $\alpha = 0.05$ test for $H_0 : \beta_1 = 0.5$.

The normal equations are

$$\begin{bmatrix} 15.00 & 374.50 \\ 374.50 & 9482.75 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 6.03 \\ 158.25 \end{bmatrix}.$$

Exercise 2.10.5 Consider the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + e_i,$$

where the predictor variables take on the following values.

i	1	2	3	4	5	6	7
x_{i1}	1	1	-1	-1	0	0	0
x_{i2}	1	-1	1	-1	0	0	0

Show that β_0, β_1, β_2, $\beta_{11} + \beta_{22}$, β_{12} are estimable and find (nonmatrix) algebraic forms for the estimates of these parameters. Find the *MSE* and the standard errors of the estimates.