CHAPTER 4

Linear Statistical Models

The purpose of this chapter is to develop a theory of linear unbiased estimation that is sufficiently general to be applicable to the linear models arising in multivariate analysis. Our starting point is the classical regression model where the Gauss–Markov Theorem is formulated in vector space language. The approach taken here is to first isolate the essential aspects of a regression model and then use the vector space machinery developed thus far to derive the Gauss–Markov estimator of a mean vector.

After presenting a useful necessary and sufficient condition for the equality of the Gauss–Markov and least-squares estimators of a mean vector, we then discuss the existence of Gauss–Markov estimators for what might be called generalized linear models. This discussion leads to a version of the Gauss–Markov Theorem that is directly applicable to the general linear model of multivariate analysis.

4.1. THE CLASSICAL LINEAR MODEL

The linear regression model arises from the following considerations. Suppose we observe a random variable $Y_i \in R$ and associated with Y_i are known numbers $z_{i1},\ldots, z_{ik}$, $i = 1,\ldots, n$. The numbers $z_{i1},\ldots, z_{ik}$ might be indicator variables denoting the presence or absence of a treatment as in the case of an analysis of variance situation or they might be the numerical levels of some physical parameters that affect the observed value of Y_i. It is assumed that the mean value of Y_i is $\mathcal{E}Y_i = \sum_1^k z_{ij}\beta_j$ where the β_j are unknown parameters. It is also assumed that $\text{var}(Y_i) = \sigma^2 > 0$ and $\text{cov}(Y_i, Y_j) = 0$ if $i \neq j$. Let $Y \in R^n$ be the random vector with coordinates $Y_1,\ldots, Y_n$, let $Z = \{z_{ij}\}$ be the $n \times k$ matrix of z_{ij}'s, and let $\beta \in R^k$ be the vector with coordinates $\beta_1,\ldots, \beta_k$. In vector form, the assumptions we have made

concerning Y are that $\mathcal{E}Y = Z\beta$ and $\mathrm{Cov}(Y) = \sigma^2 I_n$. In summary, we observe the vector Y whose mean is $Z\beta$ where Z is a known $n \times k$ matrix, $\beta \in R^k$ is a vector of unknown parameters, and $\mathrm{Cov}(Y) = \sigma^2 I_n$ where σ^2 is an unknown parameter. The two essential features of this parametric model are: (i) the mean vector of Y is an unknown element of a known subspace of R^n—namely, $\mathcal{E}Y$ is an element of the range of the known linear transformation determined by Z that maps R^k to R^n; (ii) $\mathrm{Cov}(Y) = \sigma^2 I_n$—that is, the distribution of Y is weakly spherical. For a discussion of the classical statistical problems related to the above model, the reader is referred to Scheffé (1959).

Now, consider a finite dimensional inner product space $(V, (\cdot, \cdot))$. With the above regression model in mind, we define a weakly spherical linear model for a random vector with values in $(V, (\cdot, \cdot))$.

Definition 4.1. Let M be a subspace of V and let ε_0 be a random vector in V with a distribution that satisfies $\mathcal{E}\varepsilon_0 = 0$ and $\mathrm{Cov}(\varepsilon_0) = I$. For each $\mu \in M$ and $\sigma > 0$, let $Q_{\mu,\sigma}$ denote the distribution of $\mu + \sigma\varepsilon_0$. The family $\{Q_{\mu,\sigma} | \mu \in M, \sigma > 0\}$ is a *weakly spherical linear model* for $Y \in V$ if the distribution of Y is in $\{Q_{\mu,\sigma} | \mu \in M, \sigma > 0\}$.

This definition is just a very formal statement of the assumption that the mean vector of Y is an element of the subspace of M and the distribution of Y is weakly spherical so $\mathrm{Cov}(Y) = \sigma^2 I$ for some $\sigma^2 > 0$. In an abuse of notation, we often write $Y = \mu + \varepsilon$ for $\mu \in M$ where ε is a random vector with $\mathcal{E}\varepsilon = 0$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. This is to indicate the assumption that we have a weakly spherical linear parametric model for the distribution of Y. The unobserved random vector ε is often called the error vector. The subspace M is called the regression subspace (or manifold) and the subspace $M^\perp$ is called the error subspace. Further, the parameter $\mu \in M$ is assumed unknown as is the parameter σ^2. It is clear that the regression model used to motivate Definition 4.1 is a weakly spherical linear model for the observed random vector and the subspace M is just the range of Z.

Given a linear model $Y = \mu + \varepsilon$, $\mu \in M$, $\mathcal{E}\varepsilon = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I$, we now want to discuss the problem of estimating μ. The classical Gauss–Markov approach to estimating μ is to first restrict attention to linear transformations of Y that are unbiased estimators and then, within this class of estimators, find the estimator with minimum expected norm-squared deviation from μ. To make all of this precise, we proceed as follows. By a linear estimator of μ, we mean an estimator of the form AY where $A \in \mathcal{L}(V, V)$. (We could consider affine estimators $AY + v_0$, $v_0 \in V$, but the unbiasedness restriction would imply $v_0 = 0$.) A linear estimator AY of μ is unbiased iff, when $\mu \in M$ is the mean of Y, we have $\mathcal{E}(AY) = \mu$. This is equivalent

to the condition that $A\mu = \mu$ for all $\mu \in M$ since $\mathcal{E}AY = A\mathcal{E}Y = A\mu$. Thus AY is an unbiased estimator of μ iff $A\mu = \mu$ for all $\mu \in M$. Let

$$\mathcal{A} = \{A | A \in \mathcal{L}(V, V), A\mu = \mu \text{ for } \mu \in M\}.$$

The linear unbiased estimators of μ are those estimators of the form AY with $A \in \mathcal{A}$. We now want to choose the one estimator (i.e., $A \in \mathcal{A}$) that minimizes the expected norm-squared deviation of the estimator from μ. In other words, the problem is to find an element $A \in \mathcal{A}$ that minimizes $\mathcal{E}\|AY - \mu\|^2$. The justification for choosing such an A is that $\|AY - \mu\|^2$ is the squared distance between AY and μ so $\mathcal{E}\|AY - \mu\|^2$ is the average squared distance between AY and μ. Since we would like AY to be close to μ, such a criterion for choosing $A \in \mathcal{A}$ seems reasonable. The first result in this chapter, the Gauss–Markov Theorem, shows that the orthogonal projection onto M, say P, is the unique element in $\mathcal{A}$ that minimizes $\mathcal{E}\|AY - \mu\|^2$.

Theorem 4.1 (Gauss–Markov Theorem). For each $A \in \mathcal{A}$, $\mu \in M$, and $\sigma^2 > 0$,

$$\mathcal{E}\|AY - \mu\|^2 \geqslant \mathcal{E}\|PY - \mu\|^2$$

where P is the orthogonal projection onto M. There is equality in this inequality iff $A = P$.

Proof. Write $A = P + C$ so $C = A - P$. Since $A\mu = \mu$ for $\mu \in M$, $C\mu = 0$ for $\mu \in M$ and this implies that $CP = 0$. Therefore, $C(Y - \mu)$ and $P(Y - \mu)$ are uncorrelated random vectors, so $\mathcal{E}(C(Y - \mu), P(Y - \mu)) = 0$ (see Proposition 2.21). Now,

$$\begin{aligned}\mathcal{E}\|AY - \mu\|^2 &= \mathcal{E}\|A(Y - \mu)\|^2 = \mathcal{E}\|P(Y - \mu) + C(Y - \mu)\|^2 \\ &= \mathcal{E}\|P(Y - \mu)\|^2 + \mathcal{E}\|C(Y - \mu)\|^2 \\ &\geqslant \mathcal{E}\|P(Y - \mu)\|^2 = \mathcal{E}\|PY - \mu\|^2.\end{aligned}$$

The third equality results from the fact that the cross product term is zero. This establishes the desired inequality. It is clear that there is equality in this inequality iff $\mathcal{E}\|C(Y - \mu)\|^2 = 0$. However, $C(Y - \mu)$ has mean zero and covariance $\sigma^2 CC'$ so

$$\mathcal{E}\|C(Y - \mu)\|^2 = \sigma^2\langle I, CC'\rangle$$

by Proposition 2.21. Since $\sigma^2 > 0$, there is equality iff $\langle I, CC'\rangle = 0$. But $\langle I, CC'\rangle = \langle C, C\rangle$ and this is zero iff $C = A - P = 0$. □

The estimator PY of $\mu \in M$ is called the Gauss–Markov estimator of the mean vector and the notation $\hat{\mu} \equiv PY$ is used here. A moment's reflection shows that the validity of Theorem 4.1 has nothing to do with the parameter σ^2, be it known or unknown, as long as $\sigma^2 > 0$. The estimator $\hat{\mu} = PY$ is also called the least-squares estimator of μ for the following reason. Given the observation vector Y, we ask for that vector in M that is closest, in the given norm, to Y—that is, we want to minimize, over $x \in M$, the expression $\|Y - x\|^2$. But $Y = PY + QY$ where $Q = (I - P)$ so, for $x \in M$,

$$\|Y - x\|^2 = \|PY - x + QY\|^2 = \|PY - x\|^2 + \|QY\|^2.$$

The second equality is a consequence of $Qx = 0$ and $QP = 0$. Thus

$$\|Y - x\|^2 \geqslant \|QY\|^2$$

with equality iff $x = PY$. In other words, the point in M that is closest to Y is $\hat{\mu} = PY$. When the vector space V is R^n with the usual inner product, then $\|Y - x\|^2$ is just a sum of squares and $\hat{\mu} = PY \in M$ minimizes this sum of squares—hence the term least-squares estimator.

◆ **Example 4.1.** Consider the regression model used to motivate Definition 4.1. Here, $Y \in R^n$ has a mean vector $Z\beta$ when $\beta \in R^k$ and Z is an $n \times k$ known matrix with $k \leqslant n$. Also, it is assumed that $\text{Cov}(Y) = \sigma^2 I_n$, $\sigma^2 > 0$. Therefore, we have a weakly spherical linear model for Y and $\mu \equiv Z\beta$ is the mean vector of Y. The regression manifold M is just the range of Z. To compute the Gauss–Markov estimator of μ, the orthogonal projection onto M, relative to the usual inner product on R^n, must be found. To find this projection explicitly in terms of Z, it is now assumed that the rank of Z is k. The claim is that $P \equiv Z(Z'Z)^{-1}Z'$ is the orthogonal projection onto M. Clearly, $P^2 = P$ and P is self-adjoint so P is the orthogonal projection onto its range. However, Z' maps R^n onto R^k since the rank of Z' is k. Thus $(Z'Z)^{-1}Z'$ maps R^n onto R^k. Therefore, the range of $Z(Z'Z)^{-1}Z'$ is $Z(R^k)$, which is just M, so P is the orthogonal projection onto M. Hence $\hat{\mu} = Z(Z'Z)^{-1}Z'Y$ is the Gauss–Markov and least-squares estimator of μ. Since $\mu = Z\beta$, $Z'\mu = Z'Z\beta$ and thus $\beta = (Z'Z)^{-1}Z'\mu$. There is the obvious temptation to call

$$\hat{\beta} \equiv (Z'Z)^{-1}Z'\hat{\mu} = (Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'Y = (Z'Z)^{-1}Z'Y$$

the Gauss–Markov and least-squares estimator of the parameter β.

Certainly, calling $\hat{\beta}$ the least-squares estimator of β is justified since

$$\|Y - Z\gamma\|^2 \geqslant \|Y - Z\hat{\beta}\|^2$$

for all $\gamma \in R^k$, as $Z\hat{\beta} = \hat{\mu}$ and $Z\gamma \in M$. Thus $\hat{\beta}$ minimizes the sum of squares $\|Y - Z\gamma\|^2$ as a function of γ. However, it is not clear why $\hat{\beta}$ should be called the Gauss–Markov estimator of β. The discussion below rectifies this situation. ◆

Again, consider the linear model in $(V, (\cdot, \cdot))$, $Y = \mu + \varepsilon$, where $\mu \in M$, $\mathcal{E}\varepsilon = 0$, and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. As usual, M is a linear subspace of V and ε is a random vector in V. Let $(W, [\cdot, \cdot])$ be an inner product space. Motivated by the considerations in Example 4.1, consider the problem of estimating $B\mu$, $B \in \mathcal{L}(V, W)$, by a linear unbiased estimator AY where $A \in \mathcal{L}(V, W)$. That AY is an unbiased estimator of $B\mu$ for each $\mu \in M$ is clearly equivalent to $A\mu = B\mu$ for $\mu \in M$ since $\mathcal{E}AY = A\mu$. Let

$$\mathcal{A}_1 = \{A | A \in \mathcal{L}(V, W), A\mu = B\mu \text{ for } \mu \in M\},$$

so AY is an unbiased estimator of $B\mu$, $\mu \in M$ iff $A \in \mathcal{A}_1$. The following result, which is a generalization of Theorem 4.1, shows that $B\hat{\mu}$ is the Gauss–Markov estimator for $B\mu$ in the sense that, for all $A \in \mathcal{A}_1$,

$$\mathcal{E}\|AY - B\mu\|_1^2 \geqslant \mathcal{E}\|BPY - B\mu\|_1^2.$$

Here $\|\cdot\|_1$ is the norm on the space $(W, [\cdot, \cdot])$.

Proposition 4.1. For each $A \in \mathcal{A}_1$,

$$\mathcal{E}\|AY - B\mu\|_1^2 \geqslant \mathcal{E}\|BPY - B\mu\|_1^2$$

where P is the orthogonal projection onto M. There is equality in this inequality iff $A = BP$.

Proof. The proof is very similar to the proof of Theorem 4.1. Define $C \in \mathcal{L}(V, W)$ by $C = A - BP$ and note that $C\mu = A\mu - BP\mu = B\mu - B\mu = 0$ since $A \in \mathcal{A}_1$ and $P\mu = \mu$ for $\mu \in M$. Thus $CP = 0$, and this implies that $BP(Y - \mu)$ and $C(Y - \mu)$ are uncorrelated random vectors. Since these random vectors have zero means,

$$\mathcal{E}[BP(Y - \mu), C(Y - \mu)] = 0.$$

For $A \in \mathcal{A}_1$,

$$\begin{aligned}\mathcal{E}\|AY - B\mu\|_1^2 &= \mathcal{E}\|BP(Y-\mu) + C(Y-\mu)\|_1^2 \\ &= \mathcal{E}\|BP(Y-\mu)\|_1^2 + \mathcal{E}\|C(Y-\mu)\|_1^2 \\ &\geqslant \mathcal{E}\|BP(Y-\mu)\|_1^2 = \mathcal{E}\|BPY - B\mu\|_1^2.\end{aligned}$$

This establishes the desired inequality. There is equality in this inequality iff $\mathcal{E}\|C(Y-\mu)\|_1^2 = 0$. The argument used in Theorem 4.1 applies here so there is equality iff $C = A - BP = 0$. □

Proposition 4.1 makes precise the statement that the Gauss–Markov estimator of a linear transformation of μ is just the linear transformation applied to the Gauss–Markov estimator of μ. In other words, the Gauss–Markov estimator of $B\mu$ is $B\hat{\mu}$ where $B \in \mathcal{L}(V, W)$. There is one particular case of this that is especially interesting. When $W = R$, the real line, then a linear transformation on V to W is just a linear functional on V. By Proposition 1.10, every linear functional on V has the form (x_0, x) for some $x_0 \in V$. Thus the Gauss–Markov estimator of (x_0, μ) is just $(x_0, \hat{\mu}) = (x_0, PY) = (Px_0, Y)$. Further, a linear estimator of (x_0, μ), say (z, Y), is an unbiased estimator of (x_0, μ) iff $(z, \mu) = (x_0, \mu)$ for all $\mu \in M$. For any such vector z, Proposition 4.1 shows that

$$\operatorname{var}(z, Y) \geqslant \operatorname{var}(Px_0, Y).$$

Thus the minimum of $\operatorname{var}(z, Y)$, over the class of all z's such that (z, Y) is an unbiased estimator of (x_0, μ), is achieved uniquely for $z = Px_0$. In particular, if $x_0 \in M$, $z = x_0$ achieves the minimum variance.

In the definition of a linear model, $Y = \mu + \varepsilon$, no distributional assumptions concerning ε were made, other than the first and second moment assumptions $\mathcal{E}\varepsilon = 0$ and $\operatorname{Cov}(\varepsilon) = \sigma^2 I$. One of the attractive features of Proposition 4.1 is its validity under these relatively weak assumptions. However, very little can be said concerning the distribution of $\hat{\mu} = PY$ other than $\mathcal{E}\hat{\mu} = \mu$ and $\operatorname{Cov}(\hat{\mu}) = \sigma^2 P$. In the following example, some of the implications of assuming that ε has a normal distribution are discussed.

◆ **Example 4.2.** Consider the situation treated in Example 4.1. A coordinate random vector $Y \in R^n$ has a mean vector $\mu = Z\beta$ where Z is an $n \times k$ known matrix of rank k ($k \leqslant n$) and $\beta \in R^k$ is a vector of unknown parameters. It is also assumed that $\operatorname{Cov}(Y) = \sigma^2 I_n$. The Gauss–Markov estimator of μ is $\hat{\mu} = Z(Z'Z)^{-1}Z'Y$. Since

$\beta = (Z'Z)^{-1}Z'\mu$, Proposition 4.1 shows that the Gauss–Markov estimator of β is $\hat{\beta} = (Z'Z)^{-1}Z'\hat{\mu} = (Z'Z)^{-1}Z'Y$. Now, add the assumption that Y has a normal distribution—that is, $\mathcal{L}(Y) = N(\mu, \sigma^2 I_n)$ where $\mu \in M$ and M is the range of Z. For this particular parametric model, we want to find a minimal sufficient statistic and the maximum likelihood estimators of the unknown parameters. The density function of Y, with respect to Lebesgue measure, is

$$p(y|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|y - \mu\|^2\right]$$

where $y \in R^n$, $\mu \in M$, and $\sigma^2 > 0$. Let P denote the orthogonal projection onto M, so $Q \equiv I - P$ is the orthogonal projection onto $M^{\perp}$. Since $\|y - \mu\|^2 = \|Py - \mu\|^2 + \|Qy\|^2$, the density of y can be written

$$p(y|\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|Py - \mu\|^2 - \frac{1}{2\sigma^2}\|Qy\|^2\right].$$

This shows that the pair $\{Py, \|Qy\|^2\}$ is a sufficient statistic as the density is a function of the pair $\{Py, \|Qy\|^2\}$. The normality assumption implies that PY and QY are independent random vectors as they are uncorrelated (see Proposition 3.4). Thus PY and $\|QY\|^2$ are independent. That the pair $\{Py, \|Qy\|^2\}$ is minimal sufficient and complete follows from results about exponential families (see Lehmann 1959, Chapter 2). To find the maximum likelihood estimators of $\mu \in M$ and σ^2, the density $p(y|\mu, \sigma^2)$ must be maximized over all values of $\mu \in M$ and σ^2. For each fixed $\sigma^2 > 0$,

$$\begin{aligned} p(y|\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|Py - \mu\|^2 - \frac{1}{2\sigma^2}\|Qy\|^2\right] \\ &\leqslant (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|Qy\|^2\right] \end{aligned}$$

with equality iff $\mu = Py$. Therefore, the Gauss–Markov estimator $\hat{\mu} = PY$ is the maximum likelihood estimator for μ. Of course, this also shows that $\hat{\beta} = (Z'Z)^{-1}Z'Y$ is the maximum likelihood estimator of β. To find the maximum likelihood estimator of σ^2, it remains to maximize

$$p(y|Py, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|Qy\|^2\right].$$

An easy differentiation argument shows that $p(y|Py, \sigma^2)$ is maximized for σ^2 equal to $\|Qy\|^2/n$. Thus $\tilde{\sigma}^2 \equiv \|Qy\|^2/n$ is the maximum likelihood estimator of σ^2. From our previous observation, $\hat{\mu} = PY$ and $\tilde{\sigma}^2$ are independent. Since $\mathcal{L}(Y) = N(\mu, \sigma^2 I)$,

$$\mathcal{L}(\hat{\mu}) = \mathcal{L}(PY) = N(\mu, \sigma^2 P)$$

and

$$\mathcal{L}(\hat{\beta}) = \mathcal{L}\left((Z'Z)^{-1}Z'Y\right) = N\left(\beta, \sigma^2(Z'Z)^{-1}\right).$$

Also,

$$\mathcal{L}(QY) = N(0, \sigma^2 Q)$$

since $Q\mu = 0$ and $Q^2 = Q = Q'$. Hence from Proposition 3.7,

$$\mathcal{L}\left(\frac{\|QY\|^2}{\sigma^2}\right) = \chi^2_{n-k}$$

since Q is a rank $n - k$ orthogonal projection. Therefore,

$$\mathcal{E}\tilde{\sigma}^2 = \frac{n-k}{n}\sigma^2.$$

It is common practice to replace the estimator $\tilde{\sigma}^2$ by the unbiased estimator

$$\hat{\sigma}^2 \equiv \frac{\|QY\|^2}{n-k}.$$

It is clear that $\hat{\sigma}^2$ is distributed as the constant $\sigma^2/(n-k)$ times a χ^2_{n-k} random variable. ◆

The final result of this section shows that the unbiased estimator of σ^2, derived in the example above, is in fact unbiased without the normality assumption. Let $Y = \mu + \varepsilon$ be a random vector in V where $\mu \in M \subseteq V$, $\mathcal{E}\varepsilon = 0$, and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. Given this linear model for Y, let P be the orthogonal projection onto M and set $Q = I - P$.

Proposition 4.2. Let $n = \dim V$, $k = \dim M$, and assume that $k < n$. Then the estimator

$$\hat{\sigma}^2 \equiv \frac{\|QY\|^2}{n-k}$$

is an unbiased estimator of σ^2.

Proof. The random vector QY has mean zero and $\mathrm{Cov}(QY) = \sigma^2 Q$. By Proposition 2.21,

$$\mathcal{E}\|QY\|^2 = \langle I, \sigma^2 Q\rangle = \sigma^2\langle I, Q\rangle = \sigma^2(n-k).$$

The last equality follows from the observation that for any self-adjoint operator S, $\langle I, S\rangle$ is just the sum of the eigenvalues of S. Specializing this to the projection Q yields $\langle I, Q\rangle = n - k$. □

4.2. MORE ABOUT THE GAUSS–MARKOV THEOREM

The purpose of this section is to investigate to what extent Theorem 4.1 depends on the weak sphericity assumption. In this regard, Proposition 4.1 provides some information. If we take $W = V$ and $B = I$, then Proposition 4.1 implies that

$$\mathcal{E}\|AY - \mu\|_1^2 \geqslant \mathcal{E}\|PY - \mu\|_1^2$$

where $\|\cdot\|_1$ is the norm obtained from an inner product $[\cdot,\cdot]$. Thus the orthogonal projection P minimizes $\mathcal{E}\|AY - \mu\|_1^2$ over $A \in \mathcal{A}$ no matter what inner product is used to measure deviations of AY from μ. The key to the proof of Theorem 4.1 is the relationship

$$\mathcal{E}[P(Y-\mu), (A-P)(Y-\mu)] = 0.$$

This follows from the fact that the random vectors $P(Y-\mu)$ and $(A - P)(Y-\mu)$ are uncorrelated and

$$\mathcal{E}P(Y-\mu) = \mathcal{E}(A-P)(Y-\mu) = 0 \qquad \text{for } A \in \mathcal{A}.$$

This observation is central to the presentation below. The following alternative development of linear estimation theory provides the needed generality to apply the theory to multivariate linear models.

Consider a random vector Y with values in an inner product space $(V, (\cdot,\cdot))$ and assume that the mean vector of Y, say $\mu = \mathcal{E}Y$, lies in a known regression manifold $M \subseteq V$. For the moment, we suppose that $\mathrm{Cov}(Y) = \Sigma$ where Σ is fixed and known (Σ is not necessarily nonsingular). As in the previous section, a linear estimator of μ, say AY, is unbiased iff

$$A \in \mathcal{A} \equiv \{A | A\mu = \mu, \mu \in M\}.$$

Given any inner product $[\cdot,\cdot]$ on V, the problem is to choose $A \in \mathcal{A}$ to

minimize

$$\Psi(A) = \mathcal{E}\|AY - \mu\|_1^2 = \mathcal{E}[AY - \mu, AY - \mu]$$

where the expectation is computed under the assumption that $\mathcal{E}Y = \mu$ and $\mathrm{Cov}(Y) = \Sigma$. Because of Proposition 4.1, it is reasonable to expect that the minimum of $\Psi(A)$ occurs at a point $P_0 \in \mathcal{A}$ where P_0 is a projection onto M along some subspace N such that $M \cap N = \{0\}$ and $M + N = V$. Of course, N is the null space of P_0 and the pair M, N determines P_0. To find the appropriate subspace N, write $\Psi(A)$ as

$$\begin{aligned}\Psi(A) &= \mathcal{E}\|AY - \mu\|_1^2 \\ &= \mathcal{E}\|P_0(Y - \mu) + (A - P_0)(Y - \mu)\|_1^2 \\ &= \mathcal{E}\|P_0(Y - \mu)\|_1^2 + \mathcal{E}\|(A - P_0)(Y - \mu)\|_1^2 \\ &\quad + 2\mathcal{E}[P_0(Y - \mu), (A - P_0)(Y - \mu)].\end{aligned}$$

When the third term in the final expression for $\Psi(A)$ is zero, then P_0 minimizes $\Psi(A)$. If $P_0(Y - \mu)$ and $(A - P_0)(Y - \mu)$ are uncorrelated, the third term will be zero (shown below), so the proper choice of P_0, and hence N, will be to make $P_0(Y - \mu)$ and $(A - P_0)(Y - \mu)$ uncorrelated. Setting $C = A - P_0$, it follows that $\mathcal{N}(C) \supseteq M$. The absence of correlation between $P_0(Y - \mu)$ and $C(Y - \mu)$ is equivalent to the condition

$$P_0 \Sigma C' = 0.$$

Here, C' is the adjoint of C relative to the initial inner product $(\cdot, \cdot)$ on V. Since $\mathcal{N}(C) \supseteq M$, we have

$$\mathcal{R}(C') = (\mathcal{N}(C))^{\perp} \subseteq M^{\perp}$$

and

$$\mathcal{R}(\Sigma C') \subseteq \Sigma(M^{\perp}).$$

The symbol $\perp$ refers to the inner product $(\cdot, \cdot)$. Therefore, if the null space of P_0, namely N, is chosen so that $N \supseteq \Sigma(M^{\perp})$, then $P_0 \Sigma C' = 0$ and P_0 minimizes $\Psi(A)$. Now, it remains to clean up the technical details of the above argument. Obviously, the subspace $\Sigma(M^{\perp})$ is going to play a role in what follows.

First, a couple of preliminary results.

Proposition 4.3. Suppose $\Sigma = \mathrm{Cov}(Y)$ in $(V, (\cdot, \cdot))$ and M is a linear subspace of V. Then:

(i) $\Sigma(M^\perp) \cap M = \{0\}$.
(ii) The subspace $\Sigma(M^\perp)$ does not depend on the inner product on V.

Proof. To prove (i), recall that the null space of Σ is

$$\{x | (x, \Sigma x) = 0\}$$

since Σ is positive semidefinite. If $u \in \Sigma(M^\perp) \cap M$, then $u = \Sigma u_1$ for some $u_1 \in M^\perp$. Since $\Sigma u_1 \in M$, $(u_1, \Sigma u_1) = 0$ so $u = \Sigma u_1 = 0$. Thus (i) holds. For (ii), let $[\cdot, \cdot]$ be any other inner product on V. Then

$$[x, y] = (x, A_0 y)$$

for some positive definite linear transformation A_0. The covariance transformation of Y with respect to the inner product $[\cdot, \cdot]$ is ΣA_0 (see Proposition 2.5). Further, the orthogonal complement of M relative to the inner product $[\cdot, \cdot]$ is

$$\{y | [x, y] = 0 \text{ for all } x \in M\} = \{y | (x, A_0 y) = 0 \text{ for all } x \in M\}$$

$$= \{A_0^{-1} u | (x, u) = 0 \text{ for all } x \in M\} = A_0^{-1}(M^\perp).$$

Thus $\Sigma(M^\perp) = (\Sigma A_0)(A_0^{-1}(M^\perp))$. Therefore, the image of the orthogonal complement of M under the covariance transformation of Y is the same no matter what inner product is used on V. □

Proposition 4.4. Suppose X_1 and X_2 are random vectors with values in $(V, (\cdot, \cdot))$. If X_1 and X_2 are uncorrelated and $\mathcal{E} X_2 = 0$, then

$$\mathcal{E} f[X_1, X_2] = 0$$

for every bilinear function f defined on $V \times V$.

Proof. Since X_1 and X_2 are uncorrelated and X_2 has mean zero, for $x_1, x_2 \in V$, we have

$$0 = \mathrm{cov}\{(x_1, X_1), (x_2, X_2)\} = \mathcal{E}(x_1, X_1)(x_2, X_2) - \mathcal{E}(x_1, X_1)\mathcal{E}(x_2, X_2)$$

$$= \mathcal{E}(x_1, X_1)(x_2, X_2).$$

However, every bilinear form f on $(V, (\cdot, \cdot))$ is given by

$$f[u_1, u_2] = (u_1, Bu_2)$$

where $B \in \mathcal{L}(V, V)$. Also, every B can be written as

$$B = \sum_i \sum_j b_{ij} y_i \square y_j$$

where $y_1, \ldots, y_n$ is a basis for V. Therefore,

$$\mathcal{E} f[X_1, X_2] = \mathcal{E} \sum \sum b_{ij}(X_1, y_i \square y_j X_2) = \sum \sum b_{ij} \mathcal{E}(y_i, X_1)(y_j, X_2) = 0.$$

□

We are now in a position to generalize Theorem 4.1. To review the assumptions, Y is a random vector in $(V, (\cdot, \cdot))$ with $\mathcal{E}Y = \mu \in M$ and $\mathrm{Cov}(Y) = \Sigma$. Here, M is a known subspace of V and Σ is the covariance of Y relative to the given inner product $(\cdot, \cdot)$. Let $[\cdot, \cdot]$ be another product on V and set

$$\Psi(A) = \mathcal{E}\|AY - \mu\|_1^2$$

for $A \in \mathcal{A}$, where $\|\cdot\|_1$ is the norm defined by $[\cdot, \cdot]$.

Theorem 4.2. Let N be any subspace of V that is complementary to M and contains the subspace $\Sigma(M^\perp)$. Here $M^\perp$ is the orthogonal complement of M relative to $(\cdot, \cdot)$. Let P_0 be the projection onto M along N. Then

$$\Psi(A) \geqslant \Psi(P_0) \quad \text{for } A \in \mathcal{A}. \tag{4.1}$$

If Σ is nonsingular, define a new inner product $(\cdot, \cdot)_\Sigma$ by

$$(x, y)_\Sigma \equiv (x, \Sigma^{-1} y).$$

Then P_0 is the unique element of $\mathcal{A}$ that minimizes $\Psi(A)$. Further, P_0 is the orthogonal projection, relative to the inner product $(\cdot, \cdot)_\Sigma$, onto M.

Proof. The existence of a subspace $N \supseteq \Sigma(M^\perp)$, which is complementary to M, is guaranteed by Proposition 4.3. Let $C \in \mathcal{L}(V, V)$ be such that $M \subseteq \mathcal{R}(C)$. Therefore,

$$\mathcal{R}(C') = (\mathcal{R}(C))^\perp \subseteq M^\perp$$

so

$$\mathcal{R}(\Sigma C') \subseteq \Sigma(M^{\perp}).$$

This implies that

$$P_0 \Sigma C' = 0$$

since $\mathcal{N}(P_0) = N \supseteq \Sigma(M^{\perp})$. However, the condition $P_0 \Sigma C' = 0$ is equivalent to the condition that $P_0(Y - \mu)$ and $C(Y - \mu)$ are uncorrelated.

With these preliminaries out of the way, consider $A \in \mathcal{A}$ and let $C = A - P_0$ so $\mathcal{N}(C) \supseteq M$. Thus

$$\begin{aligned}\Psi(A) &= \mathcal{E}\|A(Y - \mu)\|_1^2 = \mathcal{E}\|P_0(Y - \mu) + C(Y - \mu)\|_1^2 \\ &= \mathcal{E}\|P_0(Y - \mu)\|_1^2 + \mathcal{E}\|C(Y - \mu)\|_1^2 + 2\mathcal{E}\left[P_0(Y - \mu), C(Y - \mu)\right] \\ &= \mathcal{E}\|P_0(Y - \mu)\|_1^2 + \mathcal{E}\|C(Y - \mu)\|_1^2.\end{aligned}$$

The last equality follows by applying Proposition 4.4 to $P_0(Y - \mu)$ and $C(Y - \mu)$. Therefore,

$$\Psi(A) = \Psi(P_0) + \mathcal{E}\|C(Y - \mu)\|_1^2$$

so P_0 minimizes Ψ over $A \in \mathcal{A}$.

Now, assume that Σ is nonsingular. Then the subspace N is uniquely defined ($N = \Sigma(M^{\perp})$) since $\dim(\Sigma(M^{\perp})) = \dim(M^{\perp})$ and $M + \Sigma(M^{\perp}) = V$. Therefore, P_0 is uniquely defined as its range and null space have been specified. To show that P_0 uniquely minimizes Ψ, for $A \in \mathcal{A}$, we have

$$\Psi(A) = \Psi(P_0) + \mathcal{E}\|C(Y - \mu)\|_1^2$$

where $C = A - P_0$. Thus $\Psi(A) > \Psi(P_0)$ with equality iff

$$\mathcal{E}\|C(Y - \mu)\|_1^2 = 0.$$

This expectation can be zero iff $C(Y - \mu) = 0$ (a.e.) and this happens iff the covariance transformation of $C(Y - \mu)$ is zero in some (and hence every) inner product. But in the inner product $(\cdot, \cdot)$,

$$\mathrm{Cov}(C(Y - \mu)) = C\Sigma C'$$

and this is zero iff $C = 0$ as Σ is nonsingular. Therefore, P_0 is the unique

minimizer of Ψ. For the last assertion, let N_1 be the orthogonal complement of M relative to the inner product $(\cdot, \cdot)_\Sigma$. Then,

$$N_1 = \{y | (x, y)_\Sigma = 0 \text{ for all } x \in M\} = \{y | (x, \Sigma^{-1} y) = 0 \text{ for all } x \in M\}$$

$$= \{\Sigma y | (x, y) = 0 \text{ for all } x \in M\} = \Sigma(M^\perp).$$

Since $\mathcal{R}(P_0) = \Sigma(M^\perp)$, it follows that P_0 is the orthogonal projection onto M relative to $(\cdot, \cdot)_\Sigma$. □

In all of the applications of Theorem 4.2 in this book, the covariance of Y is nonsingular. Thus the projection P_0 is unique and $\hat{\mu} = P_0 Y$ is called the Gauss–Markov estimator of $\mu \in M$. In the context of Theorem 4.2, if $\mathrm{Cov}(Y) = \sigma^2 \Sigma$ where Σ is known and nonsingular and $\sigma^2 > 0$ is unknown, then $P_0 Y$ is still the Gauss–Markov estimator for $\mu \in M$ since $(\sigma^2 \Sigma)(M^\perp) = \Sigma(M^\perp)$ for each $\sigma^2 > 0$. That is, the presence of an unknown scale parameter σ^2 does not affect the projection P_0. Thus P_0 still minimizes Ψ for each fixed $\sigma^2 > 0$.

Consider a random vector Y taking values in $(V, (\cdot, \cdot))$ with $\mathcal{E} Y = \mu \in M$ and

$$\mathrm{Cov}(Y) = \sigma^2 \Sigma_1, \qquad \sigma^2 > 0.$$

Here, Σ_1 is assumed known and positive definite while $\sigma^2 > 0$ is unknown. Theorem 4.2 implies that the Gauss–Markov estimator of μ is $\hat{\mu} = P_0 Y$ where P_0 is the projection onto M along $\Sigma_1(M^\perp)$. Recall that the least-squares estimator of μ is PY where P is the orthogonal projection onto M in the given inner product, that is, P is the projection onto M along $M^\perp$.

Proposition 4.5. The Gauss–Markov and least-squares estimators of μ are the same iff $\Sigma_1(M) \subseteq M$.

Proof. Since P_0 and P are both projections onto M, $P_0 Y = PY$ iff both P_0 and P have the same null spaces—that is, the Gauss–Markov and least-squares estimators are the same iff

$$\Sigma_1(M^\perp) = M^\perp.$$

Since Σ_1 is nonsingular and self-adjoint, this condition is equivalent to the condition $\Sigma_1(M) \subseteq M$. □

The above result shows that if $\Sigma_1(M) \subseteq M$, we are free to compute either P or P_0 to find $\hat{\mu}$. The implications of this observation become clearer in the next section.

4.3. GENERALIZED LINEAR MODELS

First, consider the linear model introduced in Section 4.2. The random vector Y in $(V, (\cdot, \cdot))$ has a mean vector $\mu \in M$ where M is a subspace of V and $\mathrm{Cov}(Y) = \sigma^2\Sigma_1$. Here, Σ_1 is a fixed positive definite linear transformation and $\sigma^2 > 0$. The essential features of this linear model are: (i) the mean vector of Y is assumed to be an element of a known subspace M and (ii) the covariance of Y is an element of the set $\{\sigma^2\Sigma_1 | \sigma^2 > 0\}$. The assumption concerning the mean vector of Y is not especially restrictive since no special assumptions have been made about the subspace M. However, the covariance structure of Y is quite restricted. The set $\{\sigma^2\Sigma_1 | \sigma^2 > 0\}$ is an open half line from $0 \in \mathcal{L}(V, V)$ through the point $\Sigma_1 \in \mathcal{L}(V, V)$ so the set of the possible covariances for Y is a one-dimensional set. It is this assumption concerning the covariance of Y that we want to modify so that linear models become general enough to include certain models in multivariate analysis. In particular, we would like to discuss Example 3.2 within the framework of linear models.

Now, let M be a fixed subspace of $(V, (\cdot, \cdot))$ and let γ be an arbitrary set of positive definite linear transformations on V to V. We say that $\{M, \gamma\}$ is the *parameter set* of a linear model for Y if $\mathcal{E}Y = \mu \in M$ and $\mathrm{Cov}(Y) \in \gamma$. For a general parameter set $\{M, \gamma\}$, not much can be said about a linear model for Y. In order to restrict the class of parameter sets under consideration, we now turn to the question of existence of Gauss–Markov estimators (to be defined below) for μ. As in Section 4.1, let

$$\mathcal{A} = \{A | A \in \mathcal{L}(V, V), A\mu = \mu \text{ for } \mu \in M\}.$$

Thus a linear transformation of Y is an unbiased estimator of $\mu \in M$ iff it has the form AY for $A \in \mathcal{A}$. The following definition is motivated by Theorem 4.2.

Definition 4.2. Let $\{M, \gamma\}$ be the parameter set of a linear model for Y. For $A_0 \in \mathcal{A}$, A_0Y is a Gauss–Markov estimator of μ iff

$$\mathcal{E}_\Sigma \|AY - \mu\|^2 \geqslant \mathcal{E}_\Sigma \|A_0Y - \mu\|^2$$

for all $A \in \mathcal{A}$ and $\Sigma \in \gamma$. The subscript Σ on the expectation means that the expectation is computed when $\mathrm{Cov}(Y) = \Sigma$.

When $\gamma = \{\sigma^2 I | \sigma^2 > 0\}$, Theorem 4.1 establishes the existence and uniqueness of a Gauss–Markov estimator for μ. More generally, when $\gamma = \{\sigma^2 \Sigma_1 | \sigma^2 > 0\}$, Theorem 4.2 shows that the Gauss–Markov estimator for μ is $P_1 Y$ where P_1 is the orthogonal projection onto M relative to the inner product $(\cdot, \cdot)_1$ given by

$$(x, y)_1 \equiv \left(x, \Sigma_1^{-1} y\right), \qquad x, y \in V.$$

The problem of the existence of a Gauss–Markov estimator for general γ is taken up in the next paragraph.

Suppose that $\{M, \gamma\}$ is the parameter set for a linear model for Y. Consider a fixed element $\Sigma_1 \in \gamma$, and let $(\cdot, \cdot)_1$ be the inner product on V defined by

$$(x, y)_1 \equiv \left(x, \Sigma_1^{-1} y\right), \qquad x, y \in V.$$

As asserted in Theorem 4.2, the unique element in $\mathcal{A}$ that minimizes $\mathcal{E}_{\Sigma_1} \|AY - \mu\|^2$ is P_1—the orthogonal projection onto M relative to $(\cdot, \cdot)_1$. Thus if a Gauss–Markov estimator $A_0 Y$ exists according to Definition 4.2, A_0 must be P_1. However, exactly the same argument applies for $\Sigma_2 \in \gamma$, so A_0 must be P_2—the orthogonal projection onto M relative to the inner product defined by Σ_2. These two projections are the same iff $\Sigma_1(M^\perp) = \Sigma_2(M^\perp)$—see Theorem 4.2. Since Σ_1 and Σ_2 were arbitrary elements of γ, the conclusion is that a Gauss–Markov estimator can exist iff $\Sigma_1(M^\perp) = \Sigma_2(M^\perp)$ for all $\Sigma_1, \Sigma_2 \in \gamma$. Summarizing this leads to the following.

Proposition 4.6. Suppose that $\{M, \gamma\}$ is the parameter set of a linear model for Y in $(V, (\cdot, \cdot))$. Let Σ_1 be a fixed element of γ. A Gauss–Markov estimator of μ exists iff

$$\Sigma(M^\perp) = \Sigma_1(M^\perp) \quad \text{for all } \Sigma \in \gamma.$$

When a Gauss–Markov estimator of μ exists, it is $\hat{\mu} = PY$ where P is the orthogonal projection onto M relative to any inner product $[\cdot, \cdot]$ given by $[x, y] = (x, \Sigma^{-1} y)$ for some $\Sigma \in \gamma$.

Proof. It has been argued that a Gauss–Markov estimator for μ can exist iff $\Sigma_1(M^\perp) = \Sigma_2(M^\perp)$ for all $\Sigma_1, \Sigma_2 \in \gamma$. This is clearly equivalent to $\Sigma(M^\perp) = \Sigma_1(M^\perp)$ for all $\Sigma \in M$. The second assertion follows from the observation that when $\Sigma(M^\perp) = \Sigma_1(M^\perp)$, then all the projections onto M, relative to the inner products determined by elements of γ, are the same. That $\hat{\mu} = PY$ is a consequence of Theorem 4.2. □

An interesting special case of Proposition 4.6 occurs when $I \in \gamma$. In this case, choose $\Sigma_1 = I$ so a Gauss–Markov estimator exists iff $\Sigma(M^\perp) = M^\perp$ for all $\Sigma \in \gamma$. This is clearly equivalent to $\Sigma(M) = M$ for all $\Sigma \in \gamma$, which is equivalent to the condition

$$\Sigma(M) \subseteq M \quad \text{for all } \Sigma \in \gamma$$

since each $\Sigma \in \gamma$ is nonsingular. It is this condition that is verified in the examples that follow.

◆ **Example 4.3.** As motivation for the discussion of the general multivariate linear model, we first consider the multivariate version of the k-sample situation. Suppose X_{ij}'s, $j = 1,\ldots, n_i$ and $i = 1,\ldots, k$, are random vectors in R^p. It is assumed that $\mathcal{E}X_{ij} = \mu_i$, $\mathrm{Cov}(X_{ij}) = \Sigma$, and different random vectors are uncorrelated. Form the random matrix X whose first n_1 rows are $X'_{1j}, j = 1,\ldots, n_1$, the next n_2 rows of X are $X'_{2j}, j = 1,\ldots, n_2$, and so on. Then X is a random vector in $(\mathcal{L}_{p,n}, \langle \cdot, \cdot \rangle)$ where $n = \sum_1^k n_i$. It was argued in the discussion following Proposition 2.18 that

$$\mathrm{Cov}(X) = I_n \otimes \Sigma$$

relative to the inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{L}_{p,n}$. The mean of X, say $\mu = \mathcal{E}X$, is an $n \times p$ matrix whose first n_1 rows are all μ'_1, whose next n_2 rows are all μ'_2, and so on. Let B be the $k \times p$ matrix with rows $\mu'_1,\ldots, \mu'_k$. Thus the mean of X can be written $\mu = ZB$ where Z is an $n \times k$ matrix with the following structure: the first column of Z consists of n_1 ones followed by $n - n_1$ zeroes, the second column of Z consists of n_1 zeroes followed by n_2 ones followed by $n - n_1 - n_2$ zeroes, and so on. Define the linear subspace M of $\mathcal{L}_{p,n}$ by

$$M = \{\mu | \mu = ZB, B \in \mathcal{L}_{p,k}\}$$

so M is the range of $Z \otimes I_p$ as a linear transformation on $\mathcal{L}_{p,k}$ to $\mathcal{L}_{p,n}$. Further, set

$$\gamma = \{I_n \otimes \Sigma | \Sigma \in \mathcal{L}_{p,p}, \Sigma \text{ positive definite}\}$$

and note that γ is a set of positive definite linear transformations on $\mathcal{L}_{p,n}$ to $\mathcal{L}_{p,n}$. Therefore, $\mathcal{E}X \in M$ and $\mathrm{Cov}(X) \in \gamma$, and $\{M, \gamma\}$ is a parameter set for a linear model for X. Since $I_n \otimes I_p$ is the identity

linear transformation on $\mathcal{L}_{p,n}$ and $I_n \otimes I_p \in \gamma$, to show that a Gauss–Markov estimator for $\mu \in M$ exists, it is sufficient to verify that, if $x \in M$, then $(I_n \otimes \Sigma)x \in M$. For $x \in M$, $x = ZB$ for some $B \in \mathcal{L}_{p,k}$. Therefore,

$$(I_n \otimes \Sigma)(ZB) = ZB\Sigma = (Z \otimes I_p)(B\Sigma),$$

which is an element of M. Thus M is invariant under each element of γ so a Gauss–Markov estimator for μ exists. Since the identity is an element of γ, the Gauss–Markov estimator is just the orthogonal projection of X on M relative to the given inner product $\langle \cdot\,, \cdot \rangle$. To find this projection, we argue as in Example 4.1. The regression subspace M is the range of $Z \otimes I_p$ and, clearly, Z has rank k. Let

$$P = (Z \otimes I_p)\left[(Z \otimes I_p)'(Z \otimes I_p)\right]^{-1}(Z \otimes I_p)'$$

$$= (Z \otimes I_p)\left[(Z'Z) \otimes I_p\right]^{-1}(Z' \otimes I_p) = Z(Z'Z)^{-1}Z' \otimes I_p,$$

which is an orthogonal projection; see Proposition 1.28. To verify that P is the orthogonal projection onto M, it suffices to show that the range of P is M. For any $x \in \mathcal{L}_{p,n}$,

$$Px = \left(Z(Z'Z)^{-1}Z' \otimes I_p\right)x = (Z \otimes I_p)\left[(Z'Z)^{-1}Z'x\right],$$

which is an element of M since $(Z'Z)^{-1}Z'x \in \mathcal{L}_{p,k}$. However, if $x \in M$, then $x = ZB$ and $Px = P(ZB) = ZB$—that is, P is the identity on M. Hence, the range of P is M and the Gauss–Markov estimator of μ is

$$\hat{\mu} = PX = Z(Z'Z)^{-1}Z'X.$$

Since $\mu = ZB$,

$$B = (Z'Z)^{-1}Z'\mu = \left((Z'Z)^{-1}Z' \otimes I_p\right)\mu$$

and, by Proposition 4.1,

$$\hat{B} = \left((Z'Z)^{-1}Z' \otimes I_p\right)\hat{\mu} = (Z'Z)^{-1}Z'X$$

is the Gauss–Markov estimator of the matrix B. Further, $\mathcal{E}(\hat{B}) = B$

and

$$\begin{aligned}\operatorname{Cov}(\hat{B}) &= \operatorname{Cov}\left[\left((Z'Z)^{-1}Z' \otimes I_p\right)X\right] \\ &= \left((Z'Z)^{-1}Z' \otimes I_p\right)(I_n \otimes \Sigma)\left(Z(Z'Z)^{-1} \otimes I_p\right) \\ &= (Z'Z)^{-1} \otimes \Sigma.\end{aligned}$$

For the particular matrix Z, $Z'Z$ is a $k \times k$ diagonal matrix with diagonal entries $n_1, \ldots, n_k$ so $(Z'Z)^{-1}$ is diagonal with diagonal elements $n_1^{-1}, \ldots, n_k^{-1}$. A bit of calculation shows that the matrix $\hat{B} = (Z'Z)^{-1}Z'X$ has rows $\bar{X}_1', \ldots, \bar{X}_k'$ where

$$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

is the sample mean in the ith sample. Thus the Gauss–Markov estimator of the ith mean μ_i is $\bar{X}_i$, $i = 1, \ldots, k$. ◆

It is fairly clear that the explicit form of the matrix Z in the previous example did not play a role in proving that a Gauss–Markov estimator for the mean vector exists. This observation leads quite naturally to what is usually called the general linear model of multivariate analysis. After introducing this model in the next example, we then discuss the implications of adding the assumption of normality.

◆ **Example 4.4 (Multivariate General Linear Model).** As in Example 4.3, consider a random matrix X in $(\mathcal{L}_{p,n}, \langle \cdot, \cdot \rangle)$ and assume that (i) $\mathcal{E}X = ZB$ where Z is a known $n \times k$ matrix of rank k and B is a $k \times p$ matrix of parameters, (ii) $\operatorname{Cov}(X) = I_n \otimes \Sigma$ where Σ is a $p \times p$ positive definite matrix—that is, the rows of X are uncorrelated and each row of X has covariance matrix Σ. It is clear we have simply abstracted the essential features of the linear model in Example 4.3 into assumptions for the linear model of this example. The similarity between the current example and Example 4.1 should also be noted. Each component of the observation vector in Example 4.1 has become a vector, the parameter vector has become a matrix, and the rows of the observation matrix are still uncorrelated. Of course, the rows of the observation vector in Example 4.1 are just scalars. For the example at hand, it is clear that

$$M \equiv \{\mu | \mu = ZB, B \in \mathcal{L}_{p,k}\}$$

is a subspace of $\mathfrak{L}_{p,n}$ and is the range of $Z \otimes I_p$. Setting

$$\gamma = \{I_n \otimes \Sigma | \Sigma \text{ is a } p \times p \text{ positive definite matrix}\},$$

$\{M, \gamma\}$ is the parameter set of a linear model for X. More specifically, the linear model for X is that $\mathcal{E}X = \mu \in M$ and $\mathrm{Cov}(X) \in \gamma$. Just as in Example 4.3, M is invariant under each element of γ so a Gauss–Markov estimator of $\mu = \mathcal{E}X$ exists and is PX where

$$P \equiv Z(Z'Z)^{-1}Z' \otimes I_p$$

is the orthogonal projection onto M relative to $\langle \cdot, \cdot \rangle$. Mimicking the argument given in Example 4.3 yields

$$\hat{B} = (Z'Z)^{-1}Z'X = \left((Z'Z)^{-1}Z' \otimes I_p\right)X$$

and

$$\mathrm{Cov}(\hat{B}) = (Z'Z)^{-1} \otimes \Sigma.$$

In addition to the linear model assumptions for X, we now assume that $\mathfrak{L}(X) = N(ZB, I_n \otimes \Sigma)$ so X has a normal distribution in $(\mathfrak{L}_{p,n}, \langle \cdot, \cdot \rangle)$. As in Example 4.2, a discussion of sufficient statistics and maximum likelihood estimators follows. The density function of X with respect to Lebesgue measure is

$$p(x|\mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2} \times \exp\left[-\tfrac{1}{2}\langle x - \mu, (I_n \otimes \Sigma^{-1})(x - \mu)\rangle\right],$$

as discussed in Chapter 3. Let $P_0 = Z(Z'Z)^{-1}Z'$ and $Q_0 = I - P_0$ so $P = P_0 \otimes I_p$ is the orthogonal projection onto M and $Q \equiv Q_0 \otimes I_p$ is the orthogonal projection onto $M^{\perp}$. Note that both P and Q commute with $I_n \otimes \Sigma$ for any Σ. Since $\mu \in M$, we have

$$\begin{aligned}
&\langle x - \mu, (I_n \otimes \Sigma^{-1})(x - \mu)\rangle \\
&\quad = \langle P(x - \mu) + Qx, (I_n \otimes \Sigma^{-1})(P(x - \mu) + Qx)\rangle \\
&\quad = \langle P(x - \mu), (I_n \otimes \Sigma^{-1})P(x - \mu)\rangle + \langle Qx, (I_n \otimes \Sigma^{-1})Qx\rangle
\end{aligned}$$

because $\langle Qx, (I_n \otimes \Sigma^{-1})P(x - \mu)\rangle = \langle x, Q(I_n \otimes \Sigma^{-1})P(x - \mu)\rangle$

$= 0$ since $Q(I_n \otimes \Sigma^{-1})P = QP(I_n \otimes \Sigma^{-1}) = 0$. However,

$$\begin{aligned}
&\langle Qx, (I_n \otimes \Sigma^{-1})Qx\rangle \\
&\quad = \langle x, Q(I_n \otimes \Sigma^{-1})Qx\rangle = \langle x, Q(I_n \otimes \Sigma^{-1})x\rangle \\
&\quad = \langle x, (Q_0 \otimes \Sigma^{-1})x\rangle = \langle x, Q_0 x\Sigma^{-1}\rangle \\
&\quad = \operatorname{tr}(x\Sigma^{-1}x'Q_0) = \operatorname{tr}(x'Q_0x\Sigma^{-1}).
\end{aligned}$$

Thus

$$\begin{aligned}
&\langle x - \mu, (I_n \otimes \Sigma^{-1})(x - \mu)\rangle \\
&\quad = \langle Px - \mu, (I_n \otimes \Sigma^{-1})(Px - \mu)\rangle + \operatorname{tr}(x'Q_0x\Sigma^{-1}).
\end{aligned}$$

Therefore, the density $p(x|\mu, \Sigma)$ is a function of the pair $\{Px, x'Q_0x\}$ so the pair $\{Px, x'Q_0x\}$ is sufficient. That this pair is minimal sufficient and complete for the parametric family $\{p(\cdot|\mu, \Sigma); \mu \in M, \Sigma \text{ positive definite}\}$ follows from exponential family theory. Since $P(I_n \otimes \Sigma)Q = PQ(I_n \otimes \Sigma) = 0$, the random vectors PX and QX are independent. Also, $X'Q_0X = (QX)'(QX)$ so the random vectors PX and $X'Q_0X$ are independent. In other words, $\{PX, X'Q_0X\}$ is a sufficient statistic and PX and $X'Q_0X$ are independent. To derive the maximum likelihood estimator of $\mu \in M$, fix Σ. Then

$$\begin{aligned}
p(x|\mu, \Sigma) &= (2\pi)^{-np/2}|\Sigma|^{-n/2} \\
&\quad \times \exp\left[-\tfrac{1}{2}\langle Px - \mu, (I_n \otimes \Sigma^{-1})(Px - \mu)\rangle - \tfrac{1}{2}\operatorname{tr} x'Q_0x\Sigma^{-1}\right] \\
&\leqslant (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left[-\tfrac{1}{2}\operatorname{tr} x'Q_0x\Sigma^{-1}\right]
\end{aligned}$$

with equality iff $\mu = Px$. Thus the maximum likelihood estimator of μ is $\hat{\mu} = PX$, which is also the Gauss–Markov and least-squares estimator of μ. It follows immediately that

$$\hat{B} = (Z'Z)Z'X$$

is the maximum likelihood estimator of B, and

$$\mathcal{L}(\hat{B}) = N(B, (Z'Z)^{-1} \otimes \Sigma).$$

To find the maximum likelihood estimator of Σ, the function

$$p(x|\hat{\mu}, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left[-\tfrac{1}{2}\operatorname{tr} x'Q_0x\Sigma^{-1}\right]$$

must be maximized over all $p \times p$ positive definite matrices Σ. When $x'Q_0x$ is positive definite, this maximum occurs uniquely at

$$\hat{\Sigma} \equiv \frac{1}{n}x'Q_0x$$

so the maximum likelihood estimator of Σ is stochastically independent of $\hat{\mu}$. A proof that $\hat{\Sigma}$ is the maximum likelihood estimator of Σ and a derivation of the distribution of $\hat{\Sigma}$ is deferred until later. ◆

The principal result of this chapter, Proposition 4.6, gives necessary and sufficient conditions on the parameter set $\{M, \gamma\}$ of a linear model in order that the Gauss–Markov estimator of $\mu \in M$ exists. Many of the classical parametric models in multivariate analysis are in fact linear models with a parameter set $\{M, \gamma\}$ so that there is a Gauss–Markov estimator for $\mu \in M$. For such models, the additional assumption of normality implies that $\hat{\mu}$ is also the maximum likelihood estimator of μ, and the estimation of μ is relatively easy if we are satisfied with the maximum likelihood estimator. For the time being, let us agree that the problem of estimating μ has been solved in these models. However, very little has been said about the estimation of the covariance other than in Example 4.4. To be specific, assume $\mathcal{L}(X) = N(\mu, \Sigma)$ where $\mu \in M \subseteq (V, (\cdot, \cdot))$ and $\{M, \gamma\}$ is the parameter set of this linear model for x. Assume that $I \in \gamma$ and $\hat{\mu} = PX$ is the Gauss–Markov estimator for μ so $\Sigma M = M$ for all $\Sigma \in \gamma$. Here, P is the orthogonal projection onto M in the given inner product on V. It follows immediately from Proposition 4.6 that $\hat{\mu} = PX$ is also the maximum likelihood estimator of $\mu \in M$. Substituting $\hat{\mu}$ into the density of X yields

$$p(x|\hat{\mu}, \Sigma) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left[-\tfrac{1}{2}(Qx, \Sigma^{-1}Qx)\right]$$

where $n = \dim V$ and $Q = I - P$ is the orthogonal projection onto $M^{\perp}$. Thus to find the maximum likelihood estimator of $\Sigma \in \gamma$, we must compute

$$\sup_{\Sigma \in \gamma} p(x|\hat{\mu}, \Sigma) \equiv p(x|\hat{\mu}, \hat{\Sigma});$$

assuming that the supremum is attained at a point $\hat{\Sigma} \in \gamma$. Although many

examples of explicit sets γ are known where $\hat{\Sigma}$ is not too difficult to find, general conditions on γ that yield an explicit $\hat{\Sigma}$ are not available. This overview of the maximum likelihood estimation problem in linear models where Gauss–Markov estimators exists has been given to provide the reader with a general framework in which to view many of the estimation and testing problems to be discussed in later chapters.

PROBLEMS

1. Let Z be an $n \times k$ matrix (not necessarily of full rank) so Z defines a linear transformation on R^k to R^n. Let M be the range of Z and let $z_1, \ldots, z_k$ be the columns of Z.

(i) Show that $M = \text{span}\{z_1, \ldots, z_k\}$.

(ii) Show that $Z(Z'Z)^- Z'$ is the orthogonal projection onto M where $(Z'Z)^-$ is the generalized inverse of $Z'Z$.

2. Suppose $X_1, \ldots, X_n$ are i.i.d. from a density $p(x|\beta) = f(x - \beta)$ where f is a symmetric density on R^1 and $\int x^2 f(x)\, dx = 1$. Here, β is an unknown translation parameter. Let $X \in R^n$ have coordinates $X_1, \ldots, X_n$.

(i) Show that $\mathcal{L}(X) = \mathcal{L}(\beta e + \varepsilon)$ where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with density f. Show that $\mathcal{E}X = \beta e$ and $\text{Cov}(X) = I_n$.

(ii) Based on (i), find the Gauss–Markov estimator of β.

(iii) Let U be the vector of order statistics for X ($U_1 < U_2 < \cdots < U_n$) so $\mathcal{L}(U) = \mathcal{L}(\beta e + \nu)$ where ν is the vector of order statistics of ε. Show that $\mathcal{E}(U) = \beta e + a_0$ where $a_0 = \mathcal{E}\nu$ is a known vector (f is assumed known), and $\text{Cov}(U) = \Sigma_0 \equiv \text{Cov}(\nu)$ where Σ_0 is also known. Thus $\mathcal{L}(U - a_0) = \mathcal{L}(\beta e + (\nu - a_0))$ where $\mathcal{E}(\nu - a_0) = 0$ and $\text{Cov}(\nu - a_0) = \Sigma_0$. Based on this linear model, find the Gauss–Markov estimator for β.

(iv) How do these two estimators of β compare?

3. Consider the linear model $Y = \mu + \varepsilon$ where $\mu \in M$, $\mathcal{E}\varepsilon = 0$, and $\text{Cov}(\varepsilon) = \sigma^2 I_n$. At times, a submodel of this model is of interest. In particular, assume $\mu \in \omega$ where ω is a linear subspace of M.

(i) Let $M - \omega = \{x | x \in M, x \perp \omega\}$. Show that $M - \omega = M \cap \omega^\perp$.

(ii) Show that $P_M - P_\omega$ is the orthogonal projection onto $M - \omega$ and verify that $\|(P_M - P_\omega)x\|^2 = \|P_M x\|^2 - \|P_\omega x\|^2$.

4. For this problem, we use the notation of Problem 1.15. Consider subspaces of R^{IJ} given by

$$M_0 = \{y | y_{ij} = y_{..} \quad \text{for all } i, j\}$$

$$M_1 = \{y | y_{ij} = y_{ik} \quad \text{for all } j, k;\ i = 1, \ldots, I\}$$

$$M_2 = \{y | y_{ij} = y_{kj} \quad \text{for all } i, k;\ j = 1, \ldots, J\}$$

(i) Show that $\mathcal{R}(A) = M_0$, $\mathcal{R}(B_1) = M_1 - M_0$, and $\mathcal{R}(B_2) = M_2 - M_0$.

Let M_3 be the range of B_3.

(ii) Show that $R^{IJ} = M_0 \oplus (M_1 - M_0) \oplus (M_2 - M_0) \oplus M_3$.

(iii) Show that a vector μ is in $M = M_0 \oplus (M_1 - M_0) \oplus (M_2 - M_0)$ iff μ can be written as $\mu_{ij} = \alpha + \beta_i + \gamma_j$, $i = 1, \ldots, I, j = 1, \ldots, J$, where α, β_i, and γ_j are scalars that satisfy $\Sigma\beta_i = \Sigma\gamma_j = 0$.

5. (The $\mathcal{F}$-test.) Most of the classical hypothesis testing problems in regression analysis or ANOVA can be described as follows. A linear model $Y = \mu + \varepsilon$, $\mu \in M$, $\mathcal{E}\varepsilon = 0$, and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$ is given in $(V, (\cdot, \cdot))$. A subspace ω of M ($\omega \neq M$) is given and the problem is to test $H_0 : \mu \in \omega$ versus $H_1 : \mu \notin \omega$, $\mu \in M$. Assume that $\mathcal{L}(Y) = N(\mu, \sigma^2 I)$ in $(V, (\cdot, \cdot))$.

(i) Show that the likelihood ratio test of H_0 versus H_1 rejects for large values of $F = \|P_{M-\omega} Y\|^2 / \|Q_M Y\|^2$ where $Q_M = I - P_M$.

(ii) Under H_0, show that F is distributed as the ratio of two independent chi-squared variables.

6. In the notation of Problem 4, consider $Y \in R^{IJ}$ with $\mathcal{E}Y = \mu \in M$ (M is given in (iii) of Problem 4). Under the assumption of normality, use the results of Problem 5 to show that the $\mathcal{F}$-test for testing $H_0 : \beta_1 = \beta_2 = \cdots = \beta_I$ rejects for large values of

$$\frac{J\Sigma_i (\bar{y}_{i.} - \bar{y}_{..})^2}{\Sigma_i \Sigma_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2}.$$

Identify ω for this problem.

7. (The normal equations.) Suppose the elements of the regression subspace $M \subseteq R^n$ are given by $\mu = X\beta$ where X is $n \times k$ and $\beta \in R^k$. Given an observation vector y, the problem is to find $\hat{\mu} = P_M y$. The

equations (in β)

$$X'y = X'X\beta, \qquad \beta \in R^k \tag{4.2}$$

are often called the normal equations.

(i) Show that (4.2) always has a solution $b \in R^k$.

(ii) If b is any solution to (4.2), show that $Xb = P_M y$.

8. For $Y \in R^n$, assume $\mu = \mathcal{E}Y \in M$ and $\mathrm{Cov}(Y) \in \gamma$ where $\gamma = \{\Sigma | \Sigma = \alpha P_e + \beta Q_e, \alpha > 0, \beta > 0\}$. As usual, e is the vector of ones, P_e is the orthogonal projection onto $\mathrm{span}\{e\}$, and $Q_e = I - P_e$.

(i) If $e \in M$ or $e \in M^\perp$, show that the Gauss–Markov and least-squares estimators for μ are the same for each α and β.

(ii) If $e \notin M$ and $e \notin M^\perp$, show that there are values of α and β so that the least-squares and Gauss–Markov estimators of μ differ.

(iii) If $\mathcal{L}(Y) = N(\mu, \Sigma)$ with $\Sigma \in \gamma$ and $M \subseteq (\mathrm{span}\{e\})^\perp$ $(M \neq (\mathrm{span}\{e\})^\perp)$, find the maximum likelihood estimates for μ, α, and β. What happens when $M = \mathrm{span}\{e\}$?

9. In the linear model $Y = X\beta + \varepsilon$ on R^n with $X: n \times k$ of full rank, $\mathcal{E}\varepsilon = 0$, and $\mathrm{Cov}(\varepsilon) = \sigma^2\Sigma_1$ (Σ_1 is positive definite and known), show that $\hat{\mu} = X(X'\Sigma_1^{-1}X)^{-1}X'\Sigma_1^{-1}Y$ and $\hat{\beta} = (X'\Sigma_1^{-1}X)^{-1}X'\Sigma_1^{-1}Y$.

10. (Invariance in the simple linear model.) In $(V, (\cdot,\cdot))$, suppose that $\{M, \gamma\}$ is the parameter set for a linear model for Y where $\gamma = \{\Sigma | \Sigma = \sigma^2 I, \sigma > 0\}$. Thus $\mathcal{E}Y = \mu \in M$ and $\mathrm{Cov}(Y) \in \gamma$. This problem has to do with the invariance of this linear model under affine transformations:

(i) If $\Gamma \in \mathcal{O}(V)$ satisfies $\Gamma(M) \subseteq M$, show that $\Gamma'(M) \subseteq M$.

Let $\mathcal{O}_M(V)$ be those $\Gamma \in \mathcal{O}(V)$ that satisfy $\Gamma(M) \subseteq M$.

(ii) For $x_0 \in M$, $c > 0$, and $\Gamma \in \mathcal{O}_M(V)$, define the function (c, Γ, x_0) on V to V by $(c, \Gamma, x_0)y = c\Gamma y + x_0$. Show that this function is one-to-one and onto and find the inverse of this function. Show that this function maps M onto M.

(iii) Let $\tilde{Y} = (c, \Gamma, x_0)Y$. Show that $\mathcal{E}\tilde{Y} \in M$ and $\mathrm{Cov}(\tilde{Y}) \in \gamma$. Thus (M, γ) is the parameter set for $\tilde{Y}$ and we say that the linear model for Y is invariant under the transformation (c, Γ, x_0).

Since $\mathcal{E}Y = \mu$, it follows that $\mathcal{E}\tilde{Y} = (c, \Gamma, x_0)\mu$ for $\mu \in M$. If $t(Y)$ (t maps V into M) is *any* point estimator for μ, then it seems plausible to use $t(\tilde{Y})$ as a point estimator for $(c, \Gamma, x_0)\mu = c\Gamma\mu + x_0$. Solving for μ, it then seems plausible to use $c^{-1}\Gamma'(t(\tilde{Y}) - x_0)$ as a point estimator for μ. Equating these estimators of μ leads to $t(Y) = c^{-1}\Gamma'(t(c\Gamma Y +$

$x_0) - x_0)$ or

$$t(c\Gamma Y + x_0) = c\Gamma t(Y) + x_0. \tag{4.3}$$

An estimator that satisfies (4.3) for all $c > 0$, $\Gamma \in \mathcal{O}_M(V)$, and $x_0 \in M$ is called *equivariant*.

(iv) Show that $t_0(Y) = P_M Y$ is equivariant.

(v) Show that if t maps V into M and satisfies the equation $t(\Gamma Y + x_0) = \Gamma t(Y) + x_0$ for all $\Gamma \in \mathcal{O}_M(V)$ and $x_0 \in M$, then $t(Y) = P_M Y$.

11. Consider $U \in R^n$ and $V \in R^n$ and assume $\mathcal{L}(U) = N(Z_1\beta_1, \sigma_{11} I_n)$ and $\mathcal{L}(V) = N(Z_2\beta_2, \sigma_{22} I_n)$. Here, Z_i is $n \times k$ of rank k and $\beta_i \in R^k$ is an unknown vector of parameters, $i = 1, 2$. Also, $\sigma_{ii} > 0$ is unknown, $i = 1, 2$. Now, let $X = (UV) : n \times 2$ so $\mu = \mathcal{E}X$ has first column $Z_1\beta_1$ and second column $Z_2\beta_2$.

(i) When U and V are independent, then $\mathrm{Cov}(X) = I_n \otimes A$ where

$$A = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}.$$

In this case, show that the Gauss–Markov and least-squares estimates for μ are the same. Further, show that the Gauss–Markov estimates for β_1 and β_2 are the same as what we obtain by treating the two regression problems separately.

(ii) Now, suppose $\mathrm{Cov}(X) = I_n \otimes \Sigma$ where

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

is positive definite and unknown. For general Z_1 and Z_2, show that the regression subspace of X is not invariant under all $I_n \otimes \Sigma$ so the Gauss–Markov and least-squares estimators are not the same in general. However, if $Z_1 = Z_2$, show that the results given in Example 4.4 apply directly.

(iii) If the column space of Z_1 equals the column space of Z_2, show that the Gauss–Markov and least-squares estimators of μ are the same for each $I_n \otimes \Sigma$.

NOTES AND REFERENCES

1. Scheffé (1959) contains a coordinate account of what might be called univariate linear model theory. The material in the first section here follows Kruskal (1961) most closely.

2. The result of Proposition 4.5 is due to Kruskal (1968).

3. Proposition 4.3 suggests that a theory of best linear unbiased estimation can be developed in vector spaces without inner products (i.e., dual spaces are not identified with the vector space via the inner product). For a version of such a theory, see Eaton (1978).

4. The arguments used in Section 4.3 were used in Eaton (1970) to help answer the following question. Given $X \in \mathcal{L}_{p,n}$ with $\mathrm{Cov}(X) = I_n \otimes \Sigma$ where Σ is unknown but positive definite, for what subspaces M does there exist a Gauss–Markov estimator for $\mu \in M$? In other words, with γ as in Example 4.4, for what M's can the parameter set $\{M, \gamma\}$ admit a Gauss–Markov estimator? The answer to this question is that M must have the form of the subspaces considered in Example 4.4. Further details and other examples can be found in Eaton (1970).