Chapter 4

Linear Models

4.1 Random vectors and matrices

Definition 4.1 *An $n \times p$ matrix $Z = (z_{ij})$ is a* random matrix *if its elements are random variables defined on some probability space. A* random vector *is a random matrix with one column, which we will generally denote with a lower case letter like z.*

The expectation $\mathrm{E}(Z)$ is defined element-wise:

Definition 4.2 $E(Z) = [E(z_{ij})]$.

For vectors, $p = 1$, the variance $\mathrm{Var}(z)$ is defined as

Definition 4.3 $Var(z) = (Cov(z_i, z_j)) = \Sigma = (\sigma_{ij})$*, an $n \times n$ symmetric matrix.*

The following results are stated generally without any attempt at formalism; proofs are left as exercises. Let A, B, C, a, $b, \ldots$ be fixed with dimensions that should be clear from context. Z, y, and z are random. Then:

Theorem 4.1 $E(AZB) = AE(Z)B$

Theorem 4.2 $Var(z) = \Sigma_z = E(z - Ez)(z - Ez)'$

Theorem 4.3 *If z is $n \times 1$ then* $Var(z + c) = Var(z)$.

Theorem 4.4 *If z is $n \times 1$ and B is $n \times p$, let $z = By$, so y is $p \times 1$. If E$(y) = \mu$ and Var$(y) = \Sigma$, then*

$$E(z) = BE(y) \text{ and } Var(z) = B\Sigma B'$$

Definition 4.4 *If y is an $n \times 1$ random vector, and B is an $n \times n$ symmetric matrix, then $y'By$ is called a* quadratic form.

Theorem 4.5 *If b_1 and b_2 are $m \times 1$ vectors, then Var$(b_1, y) =$ Var$({b_1}'y) = {b_1}'\Sigma b_1$ and Cov$((b_1, y), (b_2, y)) =$ Cov$({b_1}'y, {b_2}'y) = {b_1}'\Sigma b_2$.*

Definition 4.5 (Uncorrelated) *Let z_1 and z_2 be $l \times 1$ and $m \times 1$ random vectors. Then z_1 and z_2 are* uncorrelated *if for all $d_1 \in \Re^l$ and all $d_2 \in \Re^m$,*

$$Cov({d_1}'z_1, {d_2}'z_2) = 0$$

Theorem 4.6 *Suppose that $y \in \Re^n$, E$(y) = \mu$ and Var$(y) = \Sigma$. Then $z_1 = B_1 y$ and $z_2 = B_2 y$ are uncorrelated if and only if $B_1 \Sigma {B_2}' = 0$.*

Proof.

$$\begin{aligned} \text{Cov}({d_1}'z_1, {d_2}'z_2) &= \text{Cov}({d_1}'B_1 y, {d_2}'B_2 y) \\ &= ({d_1}'B_1)\Sigma({d_2}'B_2)' \\ &= {d_1}'(B_1 \Sigma {B_2}')d_2 \end{aligned}$$

which is zero for all d_1, d_2 if and only if $B_1 \Sigma {B_2}' = 0$.

Theorem 4.7 *Let P be an orthogonal projection onto some subspace of $\Re^n$, $Q = I - P$, and let y be a random n-vector with Var$(y) = \sigma^2 I$. Then:*

1. *Var$(Py) = \sigma^2 P^2 = \sigma^2 P$*
2. *Var$(Qy) = \sigma^2 Q^2 = \sigma^2 Q$*
3. *Cov$(Py, Qy) = 0$ (because $PQ = 0$).*

Theorem 4.8 *If $P_i, i = 1, \ldots, m$ are orthogonal projections such that $I = \sum P_i$, and y is a random n-vector with Var$(y) = \sigma^2 I$ and E$(y) = \mu$, then*

1. *E$(P_i y) = P_i \mu$*

2. *P_iy and P_jy are uncorrelated (by Theorem 2.31, if $P = \sum P_i$, and P and all the P_i are projections, then $P_iP_j = 0, i \neq j$).*

3. *Var$(P_iy) = \sigma^2 P_i$.*

4. *$\| y \|^2 = \| \sum P_iy \|^2 = \sum \| P_iy \|^2 = \sum y'P_iy$.*

Theorem 4.9 *Let y be a random n-vector with $E(y) = \mu,$ Var$(y) = \Sigma$. Then:*

$$\begin{aligned} E(y'My) &= E(tr(y'My)) = tr(E(yy'M)) \\ &= tr(\Sigma + \mu\mu')M = \mu'M\mu + tr(\Sigma M) \end{aligned}$$

In particular, if M is an orthogonal projection with $\mu = 0$ and $\Sigma = \sigma^2 I$, then tr$(\Sigma M) =$ tr$(M) = \rho(M)$, the dimension of $\mathcal{R}(M)$,

$$\mathrm{E}(y'My) = \sigma^2\rho(M)$$

4.2 Estimation

Let $y \in \Re^n$ be a random $n \times 1$ vector with E$(y) = \mu$ and Var$(y) = \sigma^2 I$. The standard linear model assumes that μ is a fixed vector that is in an *estimation space* $\mathcal{E} \subset \Re^n$. The standard linear model requires only that the first two moments of y be specified. Normality is more that we need. The orthogonal complement $\mathcal{E}^\perp$ will be called the *error space*.

For now, we will use the canonical inner product and norm:

$$\begin{aligned} (z_1, z_2) &= {z_1}'z_2 \\ \| y - m \|^2 &= (y - m)'(y - m) \end{aligned} \tag{4.1}$$

Definition 4.6 (Ordinary least squares) *The ordinary least squares estimator (ols) $\hat{\mu}$ of μ minimizes (4.1) over all $m \in \Re^n$.*

We have seen before that (4.1) is minimized by setting $\hat{\mu} = P_{\mathcal{E}}y$, which is a random variable because y is random. The following theorem gives the basic properties of $\hat{\mu}$.

Theorem 4.10 *If y is a random n-vector such that E$(y) = \mu \in \mathcal{E},$ Var$(y) = \sigma^2 I,$ $\dim(\mathcal{E}) = p$, and $\hat{\mu}$ is the ordinary least squares estimator of μ, then:*

1. *$E(\hat{\mu}) = \mu$ and $E(y - \hat{\mu}) = 0$,*
2. *$\| y - \mu \|^2 = \| \hat{\mu} - \mu \|^2 + \| y - \hat{\mu} \|^2$.*
3. *E$(\| y - \mu \|^2) = n\sigma^2$*
4. *E$(\| \hat{\mu} - \mu \|^2) = p\sigma^2$*
5. *E$(\| y - \hat{\mu} \|^2) = (n-p)\sigma^2$*

Proof

1. $\hat{\mu} = Py \Rightarrow \mathrm{E}(\hat{\mu}) = P\mathrm{E}(y) = P\mu = \mu.$
2. Write $y = Py + Qy$, so $y - \mu = P(y-\mu) + Qy$ since $\mu \in \mathcal{E}$. Since P and Q are orthogonal, (2) follows.
3. $\mathrm{E}(\| y - \mu \|^2) = \mathrm{E}(y-\mu)'(y-\mu) = \mathrm{E}(\mathrm{tr}(y-\mu)(y-\mu)') = \mathrm{tr}(\mathrm{E}(y-\mu)(y-\mu)') = n\sigma^2$.
4. Applying Theorem 4.9 with $\Sigma = I$, $\mathrm{E}(\| \hat{\mu} - \mu \|^2) = \mathrm{E}(\| P(y-\mu) \|^2) = \mathrm{E}(y-\mu)'P(y-\mu) = 0 + \sigma^2\mathrm{tr}(P) = p\sigma^2$.
5. $\mathrm{E}(\| y - \hat{\mu} \|^2) = (n-p)\sigma^2$ follows from 2, 3 and 4.

Theorem 4.11 $\| y - \hat{\mu} \|^2/(n-p)$ *is an unbiased estimate of* σ^2.

We call $\| y - \hat{\mu} \|^2$ the *residual sum of squares*.

Example. Simple random sample. Suppose that y is $n \times 1,$ $\mathrm{Var}(y) = \sigma^2 I$ and $\mathrm{E}(y) = J_n\beta$, with β an unknown parameter, and J_n is an $n \times 1$ vector of all ones. This says each coordinate of y has the same expectation and $\mathcal{E} = \mathcal{R}(J_n)$. The matrix of the projection onto $\mathcal{R}(J_n)$ is

$$P_{\mathcal{R}(1)} = \frac{J_n J_n{}'}{J_n{}' J_n} = \frac{1}{n} J_n J_n{}'$$

and $\hat{\mu} = Py = (1/n)J_n J_n{}'y = \bar{y}J_n$, the sample mean times J_n. The vector of residuals is $Qy = (I-P)y = y - \bar{y}J_n = (y_i - \bar{y})$, and $\bar{y}J_n$ and $(y_i - \bar{y})$ are uncorrelated. In addition, $\| Qy \|^2 = \sum(y_i - \bar{y})^2 = (n-1)s^2$; E$(\| Qy \|^2) = \sigma^2(\dim(\mathcal{E}^\perp)) = (n-1)\sigma^2$.

Example. General fixed effects model. The general coordinate-free fixed effect linear model is specified by

$$Y = \mu + \varepsilon,\ \text{E}(\varepsilon) = 0,\ \text{Var}(\varepsilon) = \sigma^2 I,\ \mu \in \mathcal{E} \tag{4.2}$$

where the estimation space $\mathcal{E} \subset \Re^n$. It follows immediately that $\hat{\mu} = PY$ and Var$(\hat{\mu}) = \sigma^2 P$. The residuals are given by $e = Y - \hat{\mu} = QY$, with variance Var$(e) = \sigma^2 Q$. The unbiased estimate of σ^2 is $\hat{\sigma}^2 = QY/(n - \dim(\mathcal{E}))$.

In practice, the space $\mathcal{E}$ in most problems is specified by selecting a particular matrix X whose columns span $\mathcal{E}$, $\mathcal{R}(X) = \mathcal{E}$. Thus, any $\mu \in \mathcal{E}$ can be written as $\mu = X\beta$ for some $p \times 1$ vector of coordinates β. We now have that

$$\text{E}(y) = \mu = X\beta$$

The coordinates β will be unique if the columns of X form a basis for $\mathcal{E}$; otherwise, they will not be unique; can you describe the set of all possible βs? We can use the results of the previous chapter to find explicit formulas for P and Q using one of the orthogonal decompositions from the last chapter. For example, using the QR-factorization, $X = Q_1 R$, $P = Q_1 Q_1{}'$ and $Q = I - P$.

4.3 Best Estimators

We next consider the question of best estimators of linear combinations of the elements of μ, $(b, \mu) = b'\mu$ for $b \in \Re^n$, a fixed vector. A general prescription for a "best" estimator is quite difficult since any sensible notion of the best estimator of $b'\mu$ will depend on the joint distribution of the y_is as well as on the criterion of interest. We will limit our search for a best estimator to the class of linear unbiased estimators, which of course vastly simplifies the problem, and allows a solution to the problem that only depends on the first and second moment assumptions that are part of the standard linear model.

Definition 4.7 (Linear unbiased estimators) *An estimator $\widehat{(b, \mu)}$ is linear in y if*

$$\widehat{(b, \mu)} = (c, y), y \in \Re^n$$

A linear estimator (c, y) is unbiased for (b, μ) if E$(c, y) = (b, \mu)$ for all $\mu \in \mathcal{E}$.

An unbiased estimator exists since $\mathrm{E}(b, y) = (b, \mathrm{E}(y)) = (b, \mu)$ for all $\mu \in \mathcal{E}$. We cannot expect, however, that (b, y) will be the best estimator of (b, μ). Here is an example. In the single sample case, we have $\mu \in \mathcal{E} = \mathcal{R}(J_n)$. Suppose $b' = (0, 0, 1, 0, \ldots,)$. Now $b'y = y_3$ is unbiased for μ, but it is not very efficient because it ignores all other elements of y. For example, $\bar{y}$ has smaller variance, so if the notion of "best" depends on variance, (b, y) will not be best.

Theorem 4.12 *(c, y) is unbiased for (b, μ) if and only if $Pc = Pb$. (Recall the P is the orthogonal projection on $\mathcal{E}$.)*

Proof. Assume $Pc = Pb$. Then

$$\mathrm{E}(c, y) = (c, \mu) = (c, P\mu) = (Pc, \mu) = (Pb, \mu) = (b, P\mu) = (b, \mu)$$

and so it is unbiased. Next, assume that $\mathrm{E}(c, y) = (b, \mu)$ for all $\mu \in \mathcal{E}$. Then $(c, \mu) = (b, \mu)$ implies that $(c - b, \mu) = 0$ and thus $c - b \in \mathcal{E}^{\perp}$. We then must have that $P(c - b) = 0$ and finally $Pc = Pb$.

An immediate consequence of this theorem is:

Theorem 4.13 *(c, y) is unbiased for (b, μ) if and only if*

$$c = b + Q_{\mathcal{E}} z \textit{ for some } z \in \Re^n.$$

This follows from Theorem 4.12. The set of all linear unbiased estimators forms a flat.

In the one sample case, $Q = I - J_n J_n{}'/n$, so c is of the form $b + (z - \bar{z} J_n)$ with b and any vector $z \in \Re^n$. For the special case of $n = 3$ with $b' = (1, 0, 0)$, here are some unbiased estimates:

1. If $z = (0, 0, 0)'$ then $c = b + Qz = b$.
2. If $z = (-2/3, 1/3, 1/3)'$ then $c = b + Qz = (1/3, 1/3, 1/3)'$.
3. If $z = (-4, +4, 0')$ then $c = b + Qz = (-3, 4, 0)'$
4. If $z = (z_1, z_2, z_3)'$, then $c = (1 + z_1 - \bar{z}, z_2 - \bar{z}, z_3 - \bar{z})'$.

Among the class of linear unbiased estimates, the one with the smallest variance will be considered the best estimator.

Definition 4.8 (Best linear unbiased estimates) *(c, y) is a* best linear unbiased estimate (BLUE) *of (b, μ) if*

1. *E*$(c, y) = (b, \mu)$

2. *Var*$(c, y) \leq$ *Var*(c', y) *for all* c' *such that* $Pc' = Pb$*.*

Theorem 4.14 *The unique* BLUE *of* (b, μ) *is* (Pb, y)*.*

In the single sample case we have been considering, $Pb = J_n{J_n}'b/n = \bar{b}J_n$ and $(Pb, y) = (\bar{b}J_n, y) = n\bar{b}\bar{y}$. In particular, if $b'J_n = J_n$, then the BLUE is $\bar{y}$.

Proof. If (c, y) is an unbiased estimator of (b, μ), then we must have that $c = b + Qz$ for some $z \in \Re^n$. We can now compute

$$\begin{aligned} c &= b + Qz \\ &= Pb + Qb + Qz \\ &= Pb + Q(b + z) \\ &= Pb + Qw \end{aligned}$$

where $w = b + z \in \Re^n$. Now for an estimator to be BLUE, it must have minimum variance, and for any w, since $(Py, Qy) = 0$,

$$\begin{aligned} \text{Var}(c, y) &= \text{Var}(Pb + Qw, y) \\ &= \text{Var}(Pb, y) + \text{Var}(Qw, y) \\ &\geq \text{Var}(Pb, y) \end{aligned}$$

with equality when $w = 0$.

The BLUE (Pb, y) of (b, μ) is often called the *Gauss-Markov estimate* and Theorem 4.14 is called the *Gauss-Markov theorem*. Since Gauss and Markov lived in different centuries, they did not collaborate.

As a special case of the Gauss-Markov theorem, suppose that $b \in \mathcal{E}$, so $Pb = b$. Then the unique BLUE is $(Pb, y) = (b, y)$. For example, in the one-sample case, we will have $b \in \mathcal{E}$ if $b = kJ_n$ for some nonzero constant k, and then the BLUE of $b'\mu$ is just $b'y = (k/n)\bar{y}$.

By the symmetry of the projection matrix P, $(Pb, y) = (b, Py) = (b, \hat{\mu})$, so we can compute the BLUE by replacing μ by $\hat{\mu}$. The variance of the Gauss-Markov estimator is Var$(Pb, y) = \sigma^2\| \; Pb \; \|^2 = \sigma^2 b'Pb =$ Var$(b, \hat{\mu})$.

4.3.1 The one-way layout

The one way layout is perhaps the simplest nontrivial example of a linear model, and it deserves careful study because most other fixed effects linear models can often be best understood relative to the one way layout. One parameterization of this model is

$$y_{ij} = \beta_i + \varepsilon_{ij},\ i = 1, \ldots, p;\ j = 1, \ldots, n_i$$

where the β_i are fixed, unknown numbers, and the ε_{ij} are random, such that $\mathrm{E}(\varepsilon_{ij}) = 0$; $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ unless $i = i'$ and $j = j'$. A matrix representation of this problem is

$$y = X\beta + \varepsilon$$

where ε is an $n \times 1$ vector, $n = \sum n_i$, and

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{pn_p} \end{pmatrix}$$

$$X = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{n_p} & 0_{n_p} & \cdots & 1_{n_p} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

It is often convenient to write $X = (X_1, \ldots, X_p)$, so that X_i is the i-th column of X. With this parameterization the columns of X are linearly independent, and in fact are orthogonal, so they form an orthogonal basis for the estimation space $\mathcal{E}$. In general linear models, or in other parameterizations of this model, the columns of the design matrix X are often linearly dependent.

Given the orthogonal basis, not an orthonormal basis because of scaling, we can easily calculate $\hat{\mu} = Py$ by projecting on each column of X separately. The result is:

$$\hat{\mu} = \sum_{i=1}^{p} \frac{(X_i, y)}{\| X_i \|^2} X_i = \sum_{i=1}^{p} \bar{y}_{i+} X_i = \begin{pmatrix} J_{n_1}\bar{y}_{1+} \\ \vdots \\ J_{n_p}\bar{y}_{p+} \end{pmatrix}$$

where we use the convention that putting a bar over a symbol implies averaging, and replacing a subscript by a “+” implies adding: thus, for example, $\bar{y}_{3+}$ is the average $(1/n_3)\sum_{j=1}^{n_3} y_{3j}$. Also,

$$QY = (I - P)Y = (y_{ij} - \bar{y}_{i+}) = \text{residuals}.$$

Since E$(\| Qy \|^2) = \dim(\mathcal{E}^\perp)\sigma^2 = (n-p)\sigma^2$,

$$\hat{\sigma}^2 = \frac{\| Qy \|^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i+})^2$$

is an unbiased estimator of σ^2.

To obtain Var$(\hat{\mu}) = \sigma^2 P$, we must obtain an expression for P. By the orthogonality of the X_i, we can write

$$P = \sum_{i=1}^{p} \frac{X_i X_i{}'}{X_i{}' X_i}$$

from which we can get an explicit expression for P as

$$P = \begin{pmatrix} P_{n_1} & \cdots & 0_{n_1} \\ \vdots & \ddots & \vdots \\ 0_{n_1} & \cdots & P_{n_p} \end{pmatrix} = \text{diag}(P_{n_i})$$

which is a block diagonal matrix with $P_{n_i} = J_{n_i} J_{n_i}{}'/n_i$. Each P_{n_i} is itself an orthogonal projection for $\mathcal{R}(J_{n_i}) \subset \Re^{n_i}$ (and tr$(P_{n_i}) = 1$). Also,

$$\| Py \|^2 = \sum \| P_{n_i} y_i \|^2 = \sum_{i=1}^{p} n_i \bar{y}_{i+}^2$$

where y_i is the $n_i \times 1$ vector $(y_{i1}, \ldots, y_{in_i})'$, and tr$(P) = \sum$tr$(P_{n_i}) = p$. From Theorem 4.1,

$$\text{E}(\| Py \|^2) = \text{tr}(\sigma^2 P) + \mu' P \mu = p\sigma^2 + \mu'\mu = p\sigma^2 + \sum n_i \beta_i^2$$

These may not be the answers you expected or find useful. Why not? We have defined $\mathcal{R}(X)$ to include the overall mean, and so the expected length of the projection onto this space is larger than a multiple of σ^2 even if all the β_i are equal. We can correct for this by projecting on the part of $\mathcal{E}$ orthogonal to the column of ones.

The space spanned by the overall mean is just $\mathcal{R}(J_n)$ with $P_1 = J_n J_n{}'/n$, and hence the projection on the part of the estimation space orthogonal to the overall mean is $P^* = (I - P_1)P = P - P_1 P$. We must have that $PP_1 = P_1$, and so by direct multiplication P^* is an orthogonal projection, and

$$P_1 P y = P_1 \begin{pmatrix} J_{n_1}\bar{y}_{1+} \\ \vdots \\ J_{n_p}\bar{y}_{p+} \end{pmatrix} = (\sum_{i=1}^{p} n_i \bar{y}_{i+}/n) J_n = \begin{pmatrix} \bar{y}_{++} \\ \vdots \\ \bar{y}_{++} \end{pmatrix}$$

and the regression sum of squares is

$$\begin{aligned}
\| P^* y \|^2 &= \| Py \|^2 + \| PP_1 y \|^2 - 2(Py, PP_1 y) \\
&= \| Py \|^2 - \| PP_1 y \|^2 \\
&= \sum_{i=1}^{p} n_i \bar{y}_{i+}^2 - n\bar{y}_{++}^2 \\
&= \sum_{i=1}^{p} n_i (\bar{y}_{i+} - \bar{y}_{++})^2
\end{aligned}$$

which is the usual answer for the projection. The expected length is

$$\mathrm{E}(\| P^* y \|^2) = (p-1)\sigma^2 + \sum n_i (\beta_i - \bar{\beta})^2$$

where $\bar{\beta} = \sum n_i \beta_i / n$ is the weighted mean of the βs. When all the β_i are equal, this last sum of squares is zero, and $\mathrm{E}(\| P^* y \|^2) = (p-1)\sigma^2$.

The quantity (c, μ) is just some linear combination of the elements of μ, or, for the one way layout, any linear combination of the group means. If we want to estimate (c, μ), then the BLUE estimator is $(Pc, y) = (c, \hat{\mu})$. For example, if $c' = (1, 0, 0, \ldots, 0)$, then $(c, \hat{\mu}) = \bar{y}_{1+}$ and $\mathrm{Var}((c, \hat{\mu})) = \sigma^2 c' P c = \sigma^2 / n_1$.

4.4 Coordinates

The estimation space $\mathcal{E}$ can be viewed as the range space of an $n \times p$ matrix X, so any vector $\mu \in \mathcal{E}$ can be written as $\mu = \sum \beta_j X_j$, where the X_j are the columns of X. If X is of full rank, then the columns of X form a basis and the β_j are unique; otherwise, a subset of the columns forms a basis and the β_j are not unique. The vector $(\beta_1, \ldots, \beta_p)$ provides the *coordinates* of μ relative to X. Our goal now is to discuss $\hat{\beta}$ and its relationship to $\hat{\mu}$.

1. Since $\hat{\mu} \in \mathcal{E}$, we can write $\hat{\mu} = \sum \hat{\beta}_i X_i = X\hat{\beta}$ for some set $\hat{\beta}_i$. If the X_i are linearly dependent, the $\hat{\beta}_i$ are not unique.

2. $Qy = (I - P)y = y - Py = y - \hat{\mu} \in \mathcal{E}^{\perp}$. This is a just an expression for the residuals. *This computation does not depend in any way on coordinates, only on the definition of $\mathcal{E}$.*

3. $(y - \hat{\mu}) \perp X_i,$ for all i. Equivalently, this says that the residuals are orthogonal to all columns of X, $(y - \hat{\mu}, X_i) = 0$, even if the X_i are linearly dependent. This is also independent of coordinates.

4. Since $PX_i = X_i$, we have that $(y, X_i) = (y, PX_i) = (Py, X_i) = (\hat{\mu}, X_i)$.
5. Using the default inner product, this last result can be written as $X_i{}'y = X_i{}'\hat{\mu}$.
6. If we substitute from point 1 for $\hat{\mu}$, we find $X_i{}'y = X_i{}'X\hat{\beta}$.
7. Finally rewriting 6 for all i simultaneously,

$$X'y = (X'X)\hat{\beta} \tag{4.3}$$

Equations (4.3) are called the *normal equations*. Their solution $\hat{\beta}$ gives the coordinates of $\hat{\mu}$ relative to the columns of X. If the columns of X are linearly independent, the normal equations are consistent because $X'y \in \mathcal{R}(X') = \mathcal{R}(X'X)$, and a unique solution exists. The solution is found by multiplying both sides of the normal equations by $(X'X)^{-1}$ to get

$$\hat{\beta} = (X'X)^{-1}X'y \tag{4.4}$$

although this formula should almost never be used for computations because inverting a matrix can be highly inaccurate. If $X = Q_1R$ is the QR-factorization of X, then we get

$$\hat{\beta} = ((Q_1R)'(Q_1R))^{-1}(Q_1R)'y = (R'R)^{-1}R'Q_1'y = R^{-1}Q_1{}'Y$$

which can be solved by first computing the $p \times 1$ vector $z = Q_1{}'Y$, and then using backsolving for $\hat{\beta}$ to solve $R\hat{\beta} = z$.

If the Xs are not linearly independent but lie in a space of dimension q, then the least squares estimates of β are not unique. We first find one solution, and then will get the set of all possible solutions. Recall that, by definition, if A^- is a generalized inverse of a matrix A, and $AA^-y = y$ for all y. Hence, the vector

$$\hat{\beta}_0 = (X'X)^-X'Y$$

must be a solution since $(X'X)\hat{\beta}_0 = (X'X)(X'X)^-X'Y = X'Y$. To get the set of all solutions, we can let the generalized inverse vary over the set of all possible generalized inverses. Equivalently, consider all vectors of the form $\hat{\beta}_0 + z$. If this is to be a solution, we must have that

$$X'Y = (X'X)(\hat{\beta}_0 + z) = X'Y + (X'X)z$$

so we must have

$$(X'X)z = 0$$

and z can be any vector in the null space of $X'X$. The set of all possible solutions is a flat specified by

$$\hat{\beta}_0 + \mathrm{N}(X'X) = \hat{\beta}_0 + \mathcal{R}(X')^{\perp}$$

The set of all least squares solutions forms a flat in $\Re^p$ of dimension $p - q$, where q is the dimension of the column space of X.

Definition 4.9 *A vector $\hat{\beta}$ is an ordinary least squares (*OLS*) estimate of β if*

$$X\hat{\beta} = \hat{\mu} = Py$$

Any solution of the normal equations is an ordinary least squares estimator of β.

We now turn to moments of $\hat{\beta}$. In the full rank case, $\rho(X) = p$, and $\hat{\beta} = (X'X)^{-1}X'y$ is unique. We can then compute

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}[(X'X)^{-1}X'y] = (X'X)^{-1}X'X\beta = \beta$$

and

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= \mathrm{Var}((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'\mathrm{Var}(y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{4.5}$$

In the less than full rank case the coordinates β are not unique, and so the moments will depend on the particular way we choose to resolve the linear dependence. Using the Moore-Penrose inverse,

$$\hat{\beta} = (X'X)^{+}X'y$$

where $(X'X)^{+} = \Gamma D^{+}\Gamma'$, and $\Gamma D\Gamma'$ is the spectral decomposition of $X'X$, D is a diagonal matrix of nonnegative numbers, and D^{+} is a diagonal matrix whose nonzero elements are the inverses of the nonzero elements of D. We can find the expectation of this particular $\hat{\beta}$,

$$\begin{aligned} \mathrm{E}(\hat{\beta}) = \mathrm{E}((X'X)^{+}X'y) &= (X'X)^{+}X'X\beta \\ &= \Gamma D^{+}\Gamma'\Gamma D\Gamma'\beta \\ &= \Gamma D^{+}D\Gamma'\beta \end{aligned}$$

$$\begin{aligned}
&= \Gamma \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \Gamma'\beta \\
&= \Gamma_1{\Gamma_1}'\beta \\
&= (I - \Gamma_2{\Gamma_2}')\beta \\
&= \beta - \Gamma_2{\Gamma_2}'\beta
\end{aligned}$$

where $\Gamma = (\Gamma_1, \Gamma_2)$, and Γ_1 is the columns corresponding to the nonzero diagonals of D. In general, then, the ordinary least squares estimator is not unbiased, and the bias is given by

$$\text{Bias} = \beta - \text{E}(\hat{\beta}) = \Gamma_2{\Gamma_2}'\beta$$

Γ_2 is an orthonormal basis for $\text{N}(X'X)$.

We next turn to estimation of a linear combination $(c, \beta) = c'\beta$ of the elements of β. The natural estimator is the same linear combination of the elements of $\hat{\beta}$, so $\widehat{c'\beta} = c'\hat{\beta}$. Since the columns of Γ are a basis for $\Re^p$, any p-vector c can be written uniquely as $c = \Gamma_1 d_1 + \Gamma_2 d_2$. Then,

$$\begin{aligned}
\text{E}(c'\hat{\beta}) &= (\Gamma_1 d_1 + \Gamma_2 d_2)'\Gamma_1{\Gamma_1}'\beta \\
&= {d_1}'{\Gamma_1}'\beta \\
&= (c' - {d_2}'{\Gamma_2}')\beta \\
&= c'\beta - {d_2}'{\Gamma_2}'\beta
\end{aligned}$$

and the bias for the linear combination is

$$\text{Bias} = {d_2}'{\Gamma_2}'\beta \tag{4.6}$$

The bias will be zero when (4.6) is zero, and this can happen for all β only if d_2 = 0. This says that c must be a linear combination of only the columns of Γ_1, and these columns form an orthonormal basis for $\mathcal{R}(X'X)$, so to get unbiasedness we must have $c \in \mathcal{R}(X'X) = \mathcal{R}(X')$.

Next, we turn to variances, still in the general parametric case with X less than full rank. For the ols estimate based on the Moore-Penrose generalized inverse, compute

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= \sigma^2(X'X)^+X'X(X'X)^+ \\
&= \sigma^2(X'X)^+ \\
&= \sigma^2\left(\Gamma \begin{pmatrix} \Delta^{-1} & 0 \\ 0 & 0 \end{pmatrix} \Gamma'\right) \\
&= \sigma^2\left((\Gamma_1, \Gamma_2) \begin{pmatrix} \Delta^{-1} & 0 \\ 0 & 0 \end{pmatrix} (\Gamma_1, \Gamma_2)'\right)
\end{aligned}$$

This variance covariance matrix is singular in general. This means that for some linear combinations $c'\hat{\beta}$, we will have Var$(c'\hat{\beta}) = 0$. Now for any linear combination, again write $c = \Gamma_1 d_1 + \Gamma_2 d_2$, and we find

$$\begin{aligned}\text{Var}(c'\hat{\beta}) &= \sigma^2 c'\Gamma_1 \Delta^{-1} {\Gamma_1}' c \\ &= \sigma^2 {d_1}' \Delta^{-1} d_1\end{aligned}$$

and this will be zero if $d_1 = 0$, or equivalently if $c = \Gamma_2 d_2$, or $c \in \mathcal{R}(\Gamma_2) = \mathrm{N}(X)$!

As a simple example, consider the two sample case, with the matrix X given by

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

so $n = 6$, $\mathcal{E} = \mathcal{R}(X)$, $\rho(\mathcal{E}) = 2$. The matrix $X'X$ is, using R,

```
> XTX
     [,1] [,2] [,3]
[1,]    6    3    3
[2,]    3    3    0
[3,]    3    0    3
```

The spectral decomposition of this matrix can be obtained using `svd` in R:

```
> S <- svd(XTX) # spectral decomposition of XTX
> print(S,digits=3)
$d
[1] 9.00e+00 3.00e+00 1.21e-32

$u
       [,1]      [,2]   [,3]
[1,] -0.816 -5.48e-17 -0.577
[2,] -0.408 -7.07e-01  0.577
[3,] -0.408  7.07e-01  0.577

$v
       [,1]      [,2]   [,3]
```

```
[1,] -0.816 -2.83e-17  0.577
[2,] -0.408 -7.07e-01 -0.577
[3,] -0.408  7.07e-01 -0.577
> Gamma1 <- S$u[,c(1,2)]
> Gamma2 <- S$u[,3]
```

We can compute the Moore Penrose g-inverse as

```
> XTXMP <- Gamma1 %*% diag( 1/S$d[c(1,2)]) %*% t(Gamma1)
> XTXMP # Moore-Penrose G-inverse, and Var(\betahat)/\sigma^2
           [,1]        [,2]        [,3]
[1,] 0.07407407  0.03703704  0.03703704
[2,] 0.03703704  0.18518519 -0.14814815
[3,] 0.03703704 -0.14814815  0.18518519
```

The Moore-Penrose inverse is singular since it has only two nonzero eigenvalues.

Let ${c_1}' = (0, 1, -1)$. Apart from σ^2, the variance of ${c_1}'\hat{\beta}$ is

```
> C <- c(0,1,-1)
> t(C) %*% XTXMP %*% C
[1,] 0.6666667
```

If ${c_2}' = (1, -1, -1)$,

```
> C <- c(1,-1,-1)
> t(C) %*% XTXMP %*% C
[1,] -2.775558e-17
```

which is zero to rounding error. The condition ${\Gamma_2}'c_1 = 0$ shows that c_1 is in the column space of Γ_1, while ${\Gamma_1}'c_2 = 0$ shows that c_2 is in the column space of Γ_2.

4.5 Estimability

The results in the last section suggest that some linear combinations of β in the less than full rank case will not be estimable.

Definition 4.10 (Estimability) *The linear parametric function $c'\beta$ is an* estimable function *if there exists a vector $a \in \Re^n$ such that*

$$E(a'y) = c'\beta \textit{ for any } \beta.$$

If X is of full column rank then all linear combinations of β are estimable, since $\hat{\beta}$ is unique; that is, take $a' = c'(X'X)^{-1}X'$. The following is the more general result:

Theorem 4.15 *$c'\beta$ is estimable if and only if $c \in \mathcal{R}(X')$. That is, we must have $c = X'\lambda$ for some $\lambda \in \Re^n$.*

Proof. Suppose $c'\beta$ is estimable. Then there exists an $a \in \Re^n$ such that

$$\mathrm{E}(a'y) = c'\beta \text{ for all } \beta$$

But

$$\mathrm{E}(a'y) = a'X\beta = c'\beta \text{ for all } \beta.$$

Thus $(c' - a'X)\beta = 0$ for all β and therefore $c = X'a$. Hence, c is a linear combination of the columns of X' (or of the rows of X), $c \in \mathcal{R}(X')$.

Now suppose $c \in \mathcal{R}(X')$. Then for some λ,

$$c'\beta = \lambda'X\beta = \lambda'\mathrm{E}(y) = \mathrm{E}(\lambda'y),$$

so $\lambda'y$ is an unbiased estimator of $c'\beta$ for all β, and thus $c'\beta$ is estimable.

The next theorem shows how to get best estimators for estimable functions.

Theorem 4.16 (Gauss-Markov Theorem, coordinate version) *If $c'\beta$ is an estimable function, then $c'\hat{\beta}$ is the unique* BLUE *of $c'\beta$.*

Proof. Since $c'\beta$ is estimable, we can find a λ such that $c = X'\lambda$ and thus $c'\beta = \lambda'X\beta = \lambda'\mu$. This shows that $c'\beta$ is estimable if this linear combination of the elements of β is equivalent to a linear combination $\lambda'\mu$ of the elements of the mean vector. This is the fundamental connection between the coordinate-free and coordinate version.

By Theorem 4.14, $\lambda'\hat{\mu} = \lambda'Py$ is the BLUE of $\lambda'\mu$. Further, $\lambda'Py = \lambda'X\hat{\beta}$ is invariant under the choice of $\hat{\beta}$ (why?). Thus we immediately have that $\lambda'\hat{\mu} = \lambda'X\hat{\beta} = c'\hat{\beta}$ is BLUE, and for each fixed λ it is unique.

Can there be more than one λ? The set of all solutions to $X'\lambda = c$ is given by:

$$\begin{aligned} \lambda &= (X')^{+}c + (I - P)z \text{ for } z \in \Re^n \\ &= (X')^{+}c + (I - X'(X')^{+})z \text{ for } z \in \Re^n \end{aligned}$$

so the set of λs forms a flat. However, since $(X')^+ = (X^+)'$,

$$\begin{aligned}\lambda' X\hat\beta = [c'X^+ + z'(I-P)]X\hat\beta &= c'X^+X\hat\beta + z'(I-P)X\hat\beta \\ &= c'\hat\beta\end{aligned}$$

since $c \in \mathcal{R}(X')$ (required for estimability) and X^+X is the orthogonal projection operator for $\mathcal{R}(X')$, and $(I-P)X = 0$. So, although λ is not unique, the resulting estimator $c'\hat\beta$ is unique.

Theorem 4.17 *Linear combinations of estimable functions are estimable. The* BLUE *of a linear combination of estimable functions is the same linear combination of the* BLUE*s of the individual functions.*

Proof Let ${c_i}'\beta$ be estimable functions, $i = 1, 2, \ldots, k$ with BLUEs ${c_i}'\hat\beta$. Set $\psi = \sum_{=1}^k a_i {c_i}'\beta$, for the a_i fixed scalars. Then:

$$\psi = a' \begin{pmatrix} {c_i}'\beta \\ \vdots \\ {c_k}'\beta \end{pmatrix} = a' \begin{pmatrix} {c_i}' \\ \vdots \\ {c_k}' \end{pmatrix} \beta = d'\beta$$

Thus $d'\beta$ is the BLUE of ψ if $d \in \mathcal{R}(X')$. But $d = \sum a_i c_i$, and each $c_i \in \mathcal{R}(X')$, so ψ is estimable.

4.5.1 One Way Anova

We return to one-way anova, now given by

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$i = 1, \ldots, p;\ j = 1, \ldots, n$, without imposing the “usual constraints” of $\sum \alpha_i = 0$, and also without dropping one of the columns of X to achieve full rank. The model is over-parameterized, since there are $p+1$ parameters, but the estimation space $\mathcal{E}$ has dimension p. Let $y = (y_{11}, y_{12}, \ldots, y_{pn})'$, $\beta' = (\mu, \alpha_1, \ldots, \alpha_p)'$, and $X = (J_n, X_1, \ldots, X_p)$, where each vector X_j has elements one for observations in group j, and 0 elsewhere. The linear model is $Y = X\beta + \varepsilon$, and since $J_n = \sum X_i$, the model is not of full rank.

First, we find the set of all ordinary least squares estimates. The first step is to find any one estimate, which we will call $\hat\beta_0$. This can be done in several ways, for example by finding the Moore-Penrose inverse of $X'X$, but for this problem there is a simpler way: simply set the first coordinate $\hat{\mu}_0$ of $\hat{\beta}_0$ to be equal to zero. This reduces us to the full rank case discussed in Section 4.3.1. It then follows immediately that $\hat{\alpha}_i = \bar{y}_{i+}$, and thus

$$\hat{\beta}_0 = \begin{pmatrix} 0 \\ \bar{y}_{1+} \\ \vdots \\ \bar{y}_{p+} \end{pmatrix}$$

is a least squares estimate. Any solution is of the form $\hat{\beta}_0 + z$, where z is in the null space of $X'X$, so z is a solution to

$$0 = X'Xz = \begin{pmatrix} np & n & \cdots & n \\ n & n & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ n & 0 & \cdots & n \end{pmatrix} z$$

Solution to these equations is any z such that $z = k(1, -1, \ldots, -1)'$ for some k, so $\hat{\beta}$ must be of the form

$$\hat{\beta} = \begin{pmatrix} 0 \\ \bar{y}_{1+} \\ \vdots \\ \bar{y}_{p+} \end{pmatrix} + k \begin{pmatrix} 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$$

Setting $k = \bar{y}_{++}$, the grand mean, gives the "usual estimates" obtained when constraining $\sum \hat{\alpha}_i = 0$.

We turn now to estimability. For $c'\beta$ to be estimable in general, we must have that $c \in \mathcal{R}(X')$, or c must be a linear combination of the rows of X, so it is of the form:

$$c = c_1 \begin{pmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \cdots + c_p \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum c_i \\ c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix}$$

So $c'\beta = (\sum c_i)\mu + c_1\alpha_1 + \cdots + c_p\alpha_p$. Thus, we can conclude that, in this particular parameterization of the one way model:

- μ is not estimable.
- α_i is not estimable.
- $\mu + \alpha_i$ is estimable.
- $\sum d_i \alpha_i$ is estimable if $\sum d_i = 0$.

What are the estimates? We can pick any solution to the normal equations and form $c'\hat{\beta}$, and the answer is always the same.

4.6 Solutions with Linear Restrictions

So far we have shown that if the model has less than full rank, the estimate of β can be biased in general, and that only certain linear combinations of β are estimable. Those linear combinations correspond to linear combinations of the elements of the mean vector μ. This suggests several sensible approaches to the problem of estimation with rank deficient models.

One possibility is to choose any $\hat{\beta}$ and proceed with the characterizations and use of estimable functions. This is potentially complex, especially in unbalanced models with many factors. The book by Searle (1971), for example, exemplifies this approach.

As an alternative, one can consider redefining the problem as follows. Given a fixed linear model $Y = X\beta + \varepsilon$ with X less than full rank, find an appropriate basis for $\mathcal{E} = \mathcal{R}(X)$. If that basis is given by $\{z_1, \ldots, z_r\}$, and the matrix whose columns are the z_i is Z, then fit the full rank model $Y = Z\gamma + \varepsilon$. All estimable functions in the original formulation are of course still estimable. This of course corresponds exactly the coordinate-free approach that is at the heart of these notes.

In the one-way anova example, we can simply delete the column of 1s to produce a full rank model. R, JMP and Arc set $\alpha_1 = 0$ to get a full rank model, but Splus and SAS use a different method, at least by default.

Occasionally, the βs may have some real meaning and we don't wish to remove columns from X. In this case we might produce a unique full rank solution by placing restrictions on β of the form

$$d_i{}'\beta = 0$$

The restricted normal equations are then

$$X'X\tilde{\beta} = X'y$$

$$d_i{}'\tilde{\beta} = 0, i = 1, 2, \ldots, t \geq p - r$$

To choose the d_i, it makes sense to require that the estimable functions in the original problem be the same as those in the constrained problem. We know that $c'\beta$ is estimable if and only if $c \in \mathcal{R}(X')$, so this is equivalent to $d_i \notin \mathcal{R}(X')$. Otherwise, we would be restricting estimable functions.

For a general statement, let $\Delta' = (d_1, \ldots, d_t), t \geq p - r$ be the matrix specifying the restrictions. Then:

Theorem 4.18 *The system:*

$$\begin{pmatrix} X'X \\ \Delta \end{pmatrix} \beta = \begin{pmatrix} X'y \\ 0 \end{pmatrix}$$

has a unique solution $\hat{\beta}$ if and only if:

1. $\rho \left[\begin{pmatrix} X \\ \Delta \end{pmatrix} \right] = p$

2. $\mathcal{R}(X') \cap \mathcal{R}(\Delta') = 0$. *This says that all functions of the form $a'\Delta\beta$ are not estimable.*

The unique solution can be computed as

$$\hat{\beta} = (X'X + \Delta'\Delta)^{-1} X'\hat{\mu} \tag{4.7}$$

where $\hat{\mu} = Py$ is the projection on the estimation space.

Proof. Only an informal justification is given. Part 1 guarantees the uniqueness of a solution. The set of solutions to the unrestricted normal equations is given by $\hat{\beta}_0 + \mathrm{N}(X'X)$ for some β_0. If we can we ensure that the solution to the restricted normal equations, which is now unique, is an element of this flat, we are done. As long as the rows of Δ lie in the space $\mathrm{N}(X'X)$, then a restriction is placed on $N(X'X)$ but not on $\mathcal{R}(X')$. Thus, Part (1) ensures uniqueness, and Part (2) ensures that the resulting estimate is an element of the original flat.

The estimable functions given restrictions are the same as those in the original problem.

4.6.1 More one way anova

For the one way anova model with p levels and n_i observations in level i, the "usual constraint" is of the form $\sum a_i\alpha_i = 0$. Most typically, one takes all the $a_i = 1$, which comes from writing:

$$\begin{aligned} y_{ij} &= \mu_i + \varepsilon_{ij} \\ &= \bar{\mu} + (\mu_i - \bar{\mu}) + \varepsilon_{ij} \\ &= \mu + \alpha_i + \varepsilon_{ij} \end{aligned}$$

Now $\hat{\mu} = Py = (\bar{y}_{1+}J_{n_1}, \ldots, \bar{y}_{p+}J_{n_p})'$, and since $X = (J, X_1, \ldots, X_p)$,

$$X'\hat{\mu} = \begin{pmatrix} y_{++} \\ y_{1+} \\ \vdots \\ y_{p+} \end{pmatrix}$$

$$X'X = \begin{pmatrix} \sum n_i & n_1 & \cdots & n_p \\ n_1 & n_1 & \cdots & 0 \\ \vdots & & \ddots & \\ n_p & 0 & \cdots & n_p \end{pmatrix}$$

If we impose the usual constraint $\sum \alpha_i = 0$, we get

$$\Delta'\Delta = \begin{pmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (0, 1, \cdots, 1)$$

and

$$X'X + \Delta'\Delta = \begin{pmatrix} \sum n_i & n_1 & \cdots & n_p \\ n_1 & n_1 + 1 & \cdots & 1 \\ \vdots & & \ddots & \\ n_p & 1 & \cdots & n_p + 1 \end{pmatrix}$$

In the *balanced* case, $n_1 = \cdots = n_p = n$, this matrix can we written in partitioned form as

$$X'X + \Delta'\Delta = \begin{pmatrix} np & nJ_p{}' \\ nJ_p & nI + J_pJ_p{}' \end{pmatrix}$$

The inverse of this matrix in the balanced case can be computed for general n and p using two results concerning patterned matrices. First, if

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

B_{11} and B_{22} are all full rank square matrices, then

$$B^{-1} = \begin{pmatrix} (B_{11} - B_{12}B_{22}^{-1}B_{21})^{-1} & -B_{11}^{-1}B_{12}(B_{22} - B_{21}B_{11}^{-1}B_{12})^{-1} \\ -B_{22}^{-1}B_{21}(B_{11} - B_{12}B_{22}^{-1}B_{21})^{-1} & (B_{22} - B_{21}B_{11}^{-1}B_{12})^{-1} \end{pmatrix}$$

Also, provided that $a + (p-1)b \neq 0$,

$$((a-b)I + bJ_p{J_p}')^{-1} = \frac{1}{a-b}\left(I - \frac{b}{a+(p-1)b}J_p{J_p}'\right)$$

These two results can be used to show that, in the balanced case,

$$(X'X + \Delta'\Delta)^{-1} = \frac{1}{np^2}\begin{pmatrix} n+p & -n{J_p}' \\ -nJ_p & p^2I + (n-p)J_p{J_p}' \end{pmatrix}$$

from which we can compute the restricted least squares solution can be found by substituting $\hat{\mu}_0$ for $X\hat{\beta}$ in (4.7)

$$\hat{\tilde{\beta}}' = \begin{pmatrix} \bar{y}_{++} \\ \bar{y}_{1+} - \bar{y}_{++} \\ \vdots \\ \bar{y}_{p+} - \bar{y}_{++} \end{pmatrix}$$

the “usual” estimates that are presented in elementary textbooks.

In the general n_i case the algebra is less pleasant, and we find

$$\hat{\tilde{\beta}}' = \begin{pmatrix} \sum n_i\bar{y}_{i+}/\sum n_i \\ \bar{y}_{1+} - \sum n_i\bar{y}_{i+}/\sum n_i \\ \vdots \\ \bar{y}_{p+} - \sum n_i\bar{y}_{i+}/\sum n_i \end{pmatrix}$$

This may not be the answer you expected. The restricted estimate of the parameter $\bar{\mu}$ is the average of the group averages, weighted by sample size. The definition of the *population characteristic* $\bar{\mu}$ thus depends on the *sampling design*, namely on the n_i, and this seems rather undesirable.

The alternative is to use a different set of constraints, namely that $\sum n_i \alpha_i = 0$. Given these constraints, one can show that

$$\hat{\tilde{\beta}}' = \begin{pmatrix} \bar{y}_{++} \\ \bar{y}_{1+} - \bar{y}_{++} \\ \vdots \\ \bar{y}_{p+} - \bar{y}_{++} \end{pmatrix}$$

Now the estimates are more appealing, but the *constraints* depend on the sampling design. This is also unattractive.

What is the solution to this problem? *Only consider summaries that are estimable functions*, that is, only consider linear combinations of $\mathrm{E}(y) = \mu$, and give up using parameters on expecting parameters to be interpretable. In the one-way design, for example, the group means $\mathrm{E}(y_{ij})$ are always estimable, as are contrasts among them. These are the quantities that should be used to summarize the analysis.

4.7 Generalized Least Squares

We now consider estimation in the expanded class:

$$\mathrm{E}(Y) = \mu \in \mathcal{E}; \mathrm{Var}(e) = \sigma^2 \Sigma \tag{4.8}$$

where Σ is known and positive definite. Perhaps the easiest way to handle this problem is to transform it to the $\mathrm{Var}(y) = \sigma^2 I$ case. Using the spectral decomposition:

$$\begin{array}{rcl} \Sigma & = & \Gamma D \Gamma' \\ & & \Gamma D^{1/2} D^{1/2} \Gamma' \\ & & \Gamma D^{1/2} \Gamma' \Gamma D^{1/2} \Gamma' \\ & & \Sigma^{1/2} \Sigma^{1/2} \end{array} \tag{4.9}$$

So $\Sigma^{1/2}$ is a symmetric square root of Σ. Define $z = (\Sigma^{1/2})^{-1} y = \Sigma^{-1/2} y$. Then $\mathrm{E}(z) = \Sigma^{-1/2} \mathrm{E}(y) = \Sigma^{-1/2} \mu$ and $\mathrm{Var}(z) = \sigma^2 \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = \sigma^2 I$. We can then use ordinary least squares on z by projecting on the space in which $\Sigma^{-1/2}\mu$ lives, and get an estimate of $\Sigma^{-1/2}\mu$. We can then back-transform (multiply by $\Sigma^{1/2}$) to get an estimate of μ.

Let's look first at a parametric version. If we have a full rank parameterization, $Y = X\beta + \varepsilon$, with Var$(\varepsilon) = \sigma^2\Sigma$, then, if $z = \Sigma^{-1/2}y$ and

$$\Sigma^{-1/2}y = z = \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\varepsilon = W\beta + \varepsilon^*$$

and

$$\hat{\beta} = (W'W)^{-1}W'z = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

$$\text{E}(\hat{\beta}) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}X\beta = \beta$$

$$\text{Var}(\hat{\beta}) = \sigma^2(X'\Sigma^{-1}X)^{-1}$$

$$\hat{\sigma}^2 = \frac{\parallel z - \hat{\mu}_z \parallel^2}{n-p} = \frac{(y-\hat{\mu})'\Sigma^{-1}(y-\hat{\mu})}{n-p}$$

The matrix of the projection is $\Sigma^{-1/2}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1/2}$, which is symmetric, and hence is an orthogonal projection. Now all computations have been done in the z coordinates, so in particular $X\hat{\beta}$ estimates $\mu_z = \Sigma^{-1/2}\mu$. Since linear combinations of Gauss-Markov estimates are Gauss-Markov, it follows immediately that

$$\hat{\mu} = \Sigma^{1/2}\hat{\mu}_z$$

4.7.1 A direct solution via inner products

An alternative approach to the generalized least squares problem is to *change the inner product*. Suppose we have a random vector y with mean μ and covariance matrix Σ. Then for any fixed vectors a and b, using the standard inner product $(a, b) = a'b$ we find

$$\begin{aligned} \text{Cov}((a,y),(b,y)) &= \text{Cov}(\sum a_i x_i, \sum b_i x_i) \\ &= \sum\sum a_i b_j \text{Cov}(y_i, y_j) \\ &= \Sigma a_i b_i \sigma_{ij} \\ &= (a, \Sigma b) \end{aligned}$$

Suppose A is some $n \times n$ symmetric full rank matrix (or linear transformation. We can define a new inner product $(\bullet, \bullet)_A$ by

$$(a, b)_A = (a, Ab) = a'Ab$$

In this new inner product, we have

$$\begin{aligned} Cov((a,y)_A,(b,y)_A) &= \text{Cov}((a,Ax),(b,Ax)) \\ &= \text{Cov}((Aa,y),(Ab,y)) \\ &= (Aa,\Sigma Ab) \\ &= (a,A\Sigma Ab) \\ &= (a,\Sigma Ab)_A \end{aligned}$$

We are free to choose any positive definite symmetric A we like, in particular if we set $A = \Sigma^{-1}$, then

$$Cov((a,y)_{\Sigma^{-1}},(b,y)_{\Sigma^{-1}}) = (a,b)_{\Sigma^{-1}}$$

so virtually all of the results we have obtained for linear models assuming the identity covariance matrix (so $(a,b) = a'b$) hold when Var$(y) = \Sigma$ if we change the inner product to $(a,b)_{\Sigma^{-1}}$.

Consider the inner product space given by $(\Re^n, (\bullet,\bullet)_{\Sigma^{-1}})$, and E$(Y) = \mu \in \mathcal{E}$ and Var$(Y) = \sigma^2\Sigma$. Let P_Σ be the projection on $\mathcal{E}$ in this inner product space, and let Q_Σ be the projection on the orthogonal complement of this space, so $y = P_\Sigma y + Q_\Sigma y$.

Theorem 4.19 $P_\Sigma = X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}$.

Proof. We will prove that P_Σ is an orthogonal projection (it is symmetric and idempotent) and that it projects on the range space of X.

Idempotency: $P_\Sigma P_\Sigma = X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} = P_\Sigma$.

Symmetry: $(P_\Sigma x,y)_{\Sigma^{-1}} = x'P_\Sigma{}'\Sigma^{-1}y = x'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y = (x,P_\Sigma y)_{\Sigma^{-1}}$.

Range: $\mathcal{R}(P_\Sigma) \subset \mathcal{R}(X)$ since $P_\Sigma = X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} = XC$ for some matrix C, and the column space of XC must be contained in the column space of X. But $\dim(\mathcal{E}) = p$ and $\dim(\mathcal{R}(P_\Sigma)) =$ tr$(P_\Sigma) =$ tr$(X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}) =$ tr$((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}X) =$ tr$(I_p) = p$. Since the dimensions match, we must have $\mathcal{R}(P_\Sigma) = \mathcal{E}$.

We have the usual relationships:

$$\begin{gathered} y = P_\Sigma y + Q_\Sigma Y = \hat{\mu} + (y - \hat{\mu}) \\ \parallel y \parallel_\Sigma^2 = \parallel P_\Sigma y \parallel^2 + \parallel Q_\Sigma y \parallel_{\Sigma^{-1}}^2 \\ \parallel Q_\Sigma y \parallel_{\Sigma^{-1}}^2 = [Q_\Sigma y, Q_\Sigma y] = (y-\hat{\mu})'\Sigma^{-1}(y-\hat{\mu}) \\ \hat{\sigma}^2 = \frac{(y-\hat{\mu})'\Sigma^{-1}(y-\hat{\mu})}{n-p} \end{gathered}$$

4.8 Equivalence of OLS and Generalized Least Squares

The ordinary least squares and generalized least squares estimators are, in general, different. Are there circumstances (other than the trivial $\Sigma = I$) when they are the same?

Theorem 4.20 *The ordinary least squares estimate $\hat{\beta} = \text{OLS}$ and the Generalized least Squares estimate $\tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ are the same if and only if:*

$$\mathcal{R}(\Sigma^{-1}X) = \mathcal{R}(X)$$

Proof. Assume $\hat{\beta} = \tilde{\beta}$. Then for all $y \in \Re^n$:

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

implies

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

Taking transposes, we find

$$X(X'X)^{-1} = \Sigma^{-1/2}X(X'\Sigma^{-1}X)^{-1}$$

and thus $\mathcal{R}(\Sigma^{-1/2}X) = \mathcal{R}(X)$ because $(X'X)$ and $(X'\Sigma^{-1}X)$ are nonsingular and hence serve only to transform from one basis to another.

Next, suppose that $\mathcal{R}(X) = \mathcal{R}(\Sigma^{-1/2}X)$. The columns of X form a basis for $\mathcal{R}(X)$ and the columns of $\Sigma^{-1}X$ form a basis for $\mathcal{R}(X)$. We know that there exists a nonsingular matrix A that takes us from one basis to another basis, so $\Sigma^{-1}X = XA$ for some $p \times p$ matrix $A > 0$. Thus:

$$\begin{aligned} (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y &= (A'X'X)^{-1}A'X'y \\ &= (X'X)^{-1}A^{-T}A'X'y \\ &= (X'X)^{-1}X'y \end{aligned}$$

Corollary 4.21 $\mathcal{R}(\Sigma^{-1}X) = \mathcal{R}(X) = \mathcal{R}(\Sigma X)$*, so* Σ *need not be inverted to apply the theory.*

Proof.

$$\begin{aligned} \mathcal{R}(X) &= \{w|\Sigma\Sigma^{-1}Xz = w, z \in \Re^p\} \\ &= \{w|\Sigma z_1 = w, z_1 \in \mathcal{R}(\Sigma^{-1}X)\} \\ &= \{w|\Sigma z_1, z_1 \in \mathcal{R}(X)\} \\ &= \mathcal{R}(\Sigma X) \end{aligned}$$

To use this equivalence theorem (due to W. Kruskal), we usually characterize the Σs for a given X for which $\hat{\beta} = \tilde{\beta}$. If X is completely arbitrary, then only $\Sigma = \sigma^2 I$ works.

For example, if $J_n \in \mathcal{R}(X)$, then any Σ of the form:

$$\Sigma = \sigma^2(1-\rho)I - \sigma^2 \rho J_n {J_n}'$$

with $-1/(n-1) < \rho < 1$ will work. This is the model for intra-class correlation. To apply the theorem, we write,

$$\Sigma X = \sigma^2(1-\rho)X + \sigma^2 \rho J_n {J_n}' X$$

so for $i > 1$, the i-th column of ΣX is

$$(\Sigma X)_i = \sigma^2(1-\rho)X_i + \sigma^2 \rho J_n a_i$$

with $a_i = {J_n}' X_i$. Thus, the i-th column of ΣX is a linear combination of the i-th column of x and the column of 1s. For the first column of ΣX, we compute $a_1 = n$ and

$$(\Sigma X)_1 = \sigma^2(1-\rho)X_1 + n\sigma^2\rho 1 = \sigma^2(1+\rho(n-1))1$$

so $\mathcal{R}(\Sigma X) = \mathcal{R}(X)$ as required, provided $1 + \rho(n-1) \neq 0$ or $\rho > -1/(n-1)$.