Chapter 4

General linear model: the least squares problem

4.1 Least squares (LS) problem

As observed in Chapter 1, any linear model can be expressed in the form

$$\overset{\mathbf{Y}}{\begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}} = \overset{\mathbf{X}}{\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}} \overset{\beta}{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{pmatrix}} + \overset{\epsilon}{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{pmatrix}}. \quad (4.1.1)$$

Usually $\mathbf{X}$ is a matrix of known constants representing the values of covariates, and $\mathbf{Y}$ is the vector of response and ϵ is an error vector with the assumption that $E(\epsilon|\mathbf{X}) = 0$.

The goal is to find a value of β for which $\mathbf{X}\boldsymbol{\beta}$ is a "close" approximation of $\mathbf{Y}$. In statistical terms, one would like to estimate $\boldsymbol{\beta}$ such that the "distance" between $\mathbf{Y}$ and $\mathbf{X}\boldsymbol{\beta}$ is minimum. One form of distance in real vector spaces is given by the length of the difference between two vectors $\mathbf{Y}$ and $\mathbf{X}\boldsymbol{\beta}$, namely,

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{4.1.2}$$

Note that for a given $\boldsymbol{\beta}$, both $\mathbf{Y}$ and $\mathbf{X}\boldsymbol{\beta}$ are vectors in $\mathcal{R}^n$. In addition, $\mathbf{X}\boldsymbol{\beta}$ is always a member of $\mathcal{C}(\mathbf{X})$. Thus, for given $\mathbf{Y}$ and $\mathbf{X}$, the least squares problem can be characterized as a restricted minimization problem:

Minimize $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$ over $\boldsymbol{\beta} \in \mathcal{R}^n$.

Or equivalently,

Minimize $\|\mathbf{Y} - \theta\|^2$ over $\theta \in \mathcal{C}(\mathbf{X})$.

4.2 Solution to the LS problem

Since θ belongs to the $\mathcal{C}(\mathbf{X})$, the value of θ that minimizes the distance between $\mathbf{Y}$ and θ is given by **the** orthogonal projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$ (see a formal proof below). Let

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \in \mathcal{C}(\mathbf{X}) \tag{4.2.1}$$

is the orthogonal projection of $\mathbf{Y}$ onto the $\mathcal{C}(\mathbf{X})$. Then, since $\mathcal{N}(\mathbf{X}^T) = \mathcal{C}(\mathbf{X})^{\perp}$, one can write

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}, \tag{4.2.2}$$

where $\mathbf{e} \in \mathcal{N}(\mathbf{X}^T)$. Thus,

$$\mathbf{Y} - \hat{\mathbf{Y}} \in \mathcal{N}(\mathbf{X}^T). \tag{4.2.3}$$

Lemma 4.2.1. *For any $\theta \in \mathcal{C}(\mathbf{X})$,*

$$(\mathbf{Y} - \hat{\mathbf{Y}})^T(\hat{\mathbf{Y}} - \theta) = 0. \tag{4.2.4}$$

Proof. □

Lemma 4.2.2. *$\|\mathbf{Y} - \theta\|^2$ is minimized when $\theta = \hat{\mathbf{Y}}$.*

Proof.

$$\begin{aligned}
\|\mathbf{Y} - \theta\|^2 &= (\mathbf{Y} - \theta)^T(\mathbf{Y} - \theta) \\
&= (\mathbf{Y} - \hat{\mathbf{Y}} + (\hat{\mathbf{Y}} - \theta))^T(\mathbf{Y} - \hat{\mathbf{Y}} + (\hat{\mathbf{Y}} - \theta)) \\
&= (\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \theta)^T(\hat{\mathbf{Y}} - \theta) \\
&= \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{Y}} - \theta\|^2, \qquad (4.2.5)
\end{aligned}$$

which is minimized when $\theta = \hat{\mathbf{Y}}$. □

Thus, we have figured out that $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$ is minimum when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ is such that $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$. But how do we find the orthogonal projection?

Normal equations

Notice from our discussion on the Page 110 that

$$\begin{aligned} & \mathbf{Y} - \hat{\mathbf{Y}} \in \mathcal{N}(\mathbf{X}^T) \\ \Longrightarrow \quad & \mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{Y}}) = 0 \\ \Longrightarrow \quad & \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0 \\ \Longrightarrow \quad & \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{4.2.6}$$

Equation (4.2.6) is referred to as normal equations; solution of which, if exists will lead us to the orthogonal projection.

Example 4.2.3. Example 1.1.3 (continued). The linear model in matrix form can be written as

$$\overbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}}^{\mathbf{Y}} = \overbrace{\begin{bmatrix} 1 & w_1 \\ 1 & w_2 \\ \vdots & \vdots \\ 1 & w_n \end{bmatrix}}^{\mathbf{X}} \overbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{pmatrix}}^{\epsilon}. \tag{4.2.7}$$

Here,

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum w_i \\ \sum w_i & \sum w_i^2 \end{bmatrix}, \tag{4.2.8}$$

and

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum w_i Y_i \end{pmatrix} \tag{4.2.9}$$

The normal equations are then calculated as

$$\left.\begin{aligned} \alpha n + \beta \sum w_i &= \sum Y_i \\ \alpha \sum w_i + \beta \sum w_i Y_i &= \sum w_i Y_i \end{aligned}\right\} \tag{4.2.10}$$

From the linear regression course, you know that, the solution to these normal equations is given by

$$\left.\begin{aligned} \hat{\beta} &= \frac{\sum (w_i - \bar{w})(Y_i - \bar{Y})}{\sum (w_i - \bar{w})^2} \\ \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{w}, \end{aligned}\right\} \tag{4.2.11}$$

provided $\sum (w_i - \bar{w})^2 > 0$.

Example 4.2.4. Example 1.1.7 (continued). The linear model in matrix form can be written as

$$\overbrace{\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\\ \dots\\ \mathbf{Y}_a\end{pmatrix}}^{\mathbf{Y}} = \overbrace{\begin{bmatrix}1_{n_1} & 1_{n_1} & 0_{n_1} & \dots & 0_{n_1}\\ 1_{n_2} & 0_{n_2} & 1_{n_2} & \dots & 0_{n_2}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1_{n_a} & 0_{n_a} & 0_{n_a} & \dots & 1_{n_a}\end{bmatrix}}^{\mathbf{X}} \overbrace{\begin{pmatrix}\mu\\ \alpha_1\\ \alpha_2\\ \vdots\\ \alpha_a\end{pmatrix}}^{\beta} + \overbrace{\begin{pmatrix}\epsilon_{\mathbf{1}}\\ \epsilon_{\mathbf{2}}\\ \dots\\ \epsilon_{\mathbf{a}}\end{pmatrix}}^{\epsilon}, \tag{4.2.12}$$

where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{in_i})^T$ and $\epsilon_{\mathbf{i}} = (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{in_i})^T$ for $i = 1, 2, \dots, a$. Here,

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & n_1 & n_2 & \dots & n_a\\ n_1 & n_1 & 0 & \dots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ n_a & 0 & 0 & \dots & n_a\end{bmatrix}, \tag{4.2.13}$$

and

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum_i \sum_j Y_{ij} \\ \sum_j^{n_1} Y_{1j} \\ \sum_j^{n_2} Y_{2j} \\ \vdots \\ \sum_j^{n_a} Y_{aj} \end{pmatrix} = \begin{pmatrix} Y_{\cdot\cdot} \\ Y_{1\cdot} \\ Y_{2\cdot} \\ \vdots \\ Y_{a\cdot} \end{pmatrix} = \begin{pmatrix} n\bar{Y}_{\cdot\cdot} \\ n_1\bar{Y}_{1\cdot} \\ n_2\bar{Y}_{2\cdot} \\ \vdots \\ n_a\bar{Y}_{a\cdot} \end{pmatrix} \tag{4.2.14}$$

The normal equations are then calculated as

$$\left.\begin{aligned} n\mu + \textstyle\sum_{i=1}^a n_i\alpha_i &= n\bar{Y}_{\cdot\cdot} \\ n_i\mu + n_i\alpha_i &= n_i\bar{Y}_{i\cdot}, i = 1, 2, \ldots, a. \end{aligned}\right\} \tag{4.2.15}$$

Two solutions to this set of normal equations is given by

$$\left.\begin{aligned} \hat{\mu}^{(1)} &= 0 \\ \hat{\alpha}_i^{(1)} &= \bar{Y}_{i\cdot}, i = 1, 2, \ldots, a, \end{aligned}\right\} \tag{4.2.16}$$

and

$$\left.\begin{aligned} \hat{\mu}^{(2)} &= \bar{Y}_{\cdot\cdot} \\ \hat{\alpha}_i^{(2)} &= \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}, i = 1, 2, \ldots, a. \end{aligned}\right\} \tag{4.2.17}$$

Solutions to the normal equations

In Example 4.2.3, the normal equations have a unique solutions, whereas in Example 4.2.4, there are more than one (in fact, infinitely many) solutions. Are normal equations always consistent? If we closely look at the normal equations (4.2.6)

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}, \tag{4.2.18}$$

we see that **if $\mathbf{X}^T\mathbf{X}$ is non-singular**, then there exists a unique solution to the normal equations, namely,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{4.2.19}$$

which is the case for the simple linear regression in Example 4.2.3, or more generally for any linear regression problem (multiple, polynomial).

Theorem 4.2.5. *Normal equations (4.2.6) are always consistent.*

Proof. From Chapter 2, Page 60, a system of equations $\mathbf{Ax} = \mathbf{b}$ is consistent iff $\mathbf{b} \in \mathcal{C}(A)$. Thus, in our case, we need to show that,

$$\mathbf{X}^T\mathbf{Y} \in \mathcal{C}(\mathbf{X}^T\mathbf{X}). \tag{4.2.20}$$

Now, $\mathbf{X}^T\mathbf{Y} \in \mathcal{C}(\mathbf{X}^T)$. If we can show that $\mathcal{C}(\mathbf{X}^T) \subseteq \mathcal{C}(\mathbf{X}^T\mathbf{X})$, then the result is established. Let us look at the following lemma first:

Lemma 4.2.6. $\mathcal{N}(\mathbf{X}^T\mathbf{X}) = \mathcal{N}(\mathbf{X})$.

Proof. . If $\mathbf{a} \in \mathcal{N}(\mathbf{X}^T\mathbf{X})$, then

$$\begin{aligned} & \mathbf{X}^T\mathbf{Xa} = 0 \implies \mathbf{a}^T\mathbf{X}^T\mathbf{Xa} = 0 \\ \implies & \|\mathbf{Xa}\| = 0 \implies \mathbf{Xa} = 0 \\ \implies & \mathbf{a} \in \mathcal{N}(\mathbf{X}). \end{aligned} \tag{4.2.21}$$

On the other hand, if $\mathbf{a} \in \mathcal{N}(\mathbf{X})$, then $\mathbf{Xa} = 0$, and hence $\mathbf{X}^T\mathbf{Xa} = 0$ which implies that $\mathbf{a} \in \mathcal{N}(\mathbf{X}^T\mathbf{X})$ which completes the proof. □

Now, from the above lemma, and from the result stated in chapter 2, page 50 and theorem 2.3.1

$$\begin{aligned} & \mathcal{N}^{\perp}(\mathbf{X}^T\mathbf{X}) = \mathcal{N}^{\perp}(\mathbf{X}) \\ \Longrightarrow \quad & \mathcal{C}(\mathbf{X}^T\mathbf{X}) = \mathcal{C}(\mathbf{X}^T), \end{aligned} \tag{4.2.22}$$

which completes the proof. □

Least squares estimator

The above theorem shows that the normal equations are always consistent. Using a g-inverse of $\mathbf{X}^T\mathbf{X}$, we can write out all possible solutions of the normal equations. Namely,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} + \left[\mathbf{I} - (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\right]\mathbf{c} \tag{4.2.23}$$

gives all possible solution to the normal equations (4.2.6) for arbitrary vector $\mathbf{c}$. The estimator $\hat{\boldsymbol{\beta}}$ is known as a least squares estimator of $\boldsymbol{\beta}$ for a given $\mathbf{c}$. Note that one could write all possible solutions using the arbitrariness of the g-inverse of $\mathbf{X}^T\mathbf{X}$.

We know that the orthogonal projection $\hat{\mathbf{Y}}$ of $\mathbf{Y}$ onto $\mathcal{C}(\mathbf{X})$ is unique. However, the solutions to the normal equations are not. Does any solution of the normal equation lead to the orthogonal projection? In fact, it does. Specifically, if $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are any two solutions to the normal equations, then

$$\mathbf{X}\hat{\boldsymbol{\beta}}_1 = \mathbf{X}\hat{\boldsymbol{\beta}}_2. \tag{4.2.24}$$

Projection and projection matrix

From the equation (4.2.23), the projection of $\mathbf{Y}$ onto the column space $\mathcal{C}(\mathbf{X})$ is given by the **prediction vector**

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} = \mathbf{P}\mathbf{Y}, \tag{4.2.25}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T$ is **the** projection matrix.

A very useful lemma:

Lemma 4.2.7. $\mathbf{X}^T\mathbf{X}\mathbf{A} = \mathbf{X}^T\mathbf{X}\mathbf{B}$ *if and only if* $\mathbf{X}\mathbf{A} = \mathbf{X}\mathbf{B}$ *for any two matrices* $\mathbf{A}$ *and* $\mathbf{B}$.

Proposition 4.2.8. *Verify (algebraically) the following results:*

1. $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T$ *is idempotent.*

2. $\mathbf{P}$ *is invariant to the choice of the g-inverse* $(\mathbf{X}^T\mathbf{X})^g$ *.*

3. $\mathbf{P}$ *is symmetric.(Note* $(\mathbf{X}^T\mathbf{X})^g$ *does not need to be symmetric).*

Proposition 4.2.9. *If* $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T$ *is the orthogonal projection onto the column space of* $\mathbf{X}^T$*, then show that*

$$\mathbf{X}^T\mathbf{P} = \mathbf{X}^T. \tag{4.2.26}$$

and

$$rank(\mathbf{P}) = rank(\mathbf{X}). \tag{4.2.27}$$

Residual vector

Definition 4.2.1. The vector $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ is known to be the residual vector.

Notice,

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}, \tag{4.2.28}$$

and $\mathbf{Y}$ can be decomposed into two orthogonal components,

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e}, \tag{4.2.29}$$

$\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$ belonging to the column space of $\mathbf{X}$ and $\mathbf{e} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}$ belonging to $\mathcal{N}(\mathbf{X}^T)$.

Example 4.2.10. Show that $\hat{\mathbf{Y}}$ and $\mathbf{e}$ are uncorrelated when the elements of $\mathbf{Y}$ are independent with equal variance.

Proof. Let $cov(\mathbf{Y}) = \sigma^2\mathbf{I}_n$. Then,

$$\begin{aligned} E(\hat{\mathbf{Y}}\mathbf{e}^T) &= E(\mathbf{P}\mathbf{Y}\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})) \\ &= \mathbf{P}E(\mathbf{Y}\mathbf{Y}^T)(\mathbf{I}_n - \mathbf{P}) \\ &= \mathbf{P}\left[\sigma^2\mathbf{I}_n + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T\right](\mathbf{I}_n - \mathbf{P}) \\ &= \sigma^2\mathbf{P}(\mathbf{I}_n - \mathbf{P}) \\ &= 0. \end{aligned} \tag{4.2.30}$$

□

Also, $E[\mathbf{e}] = 0$. Together we get, $cov(\hat{\mathbf{Y}}, \mathbf{e}) = 0$.

Example 4.2.11. For the simple linear regression problem in example (4.2.3), we find that $rank(\mathbf{X}^T\mathbf{X}) = 2$, provided $\sum(w_i - \bar{w})^2 > 0$. Then,

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum(w_i - \bar{w})^2}\begin{bmatrix} \sum w_i^2 & -\sum w_i \\ -\sum w_i & n \end{bmatrix}. \qquad (4.2.31)$$

Recall the $\mathbf{X}^T\mathbf{Y}$ matrix,

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum w_i Y_i \end{pmatrix}, \qquad (4.2.32)$$

leading to **the** least squares estimator

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \frac{1}{n\sum(w_i - \bar{w})^2}\begin{bmatrix} \sum w_i^2 & -\sum w_i \\ -\sum w_i & n \end{bmatrix}\begin{pmatrix} \sum Y_i \\ \sum w_i Y_i \end{pmatrix} \\ &= \frac{1}{n\sum(w_i - \bar{w})^2}\begin{pmatrix} \sum Y_i \sum w_i^2 - \sum w_i Y_i \sum w_i \\ n\sum w_i Y_i - \sum w_i \sum Y_i \end{pmatrix} \\ &\stackrel{?}{=} \begin{pmatrix} \bar{Y} - \hat{\beta}\bar{w} = \hat{\alpha} \\ \frac{n\sum w_i Y_i - \sum w_i \sum Y_i}{n\sum(w_i - \bar{w})^2} = \hat{\beta} \end{pmatrix}. \end{aligned} \qquad (4.2.33)$$

Example 4.2.12. For the one-way ANOVA model in Example (4.2.4),

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & n_1 & n_2 & \dots & n_a \\ n_1 & n_1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_a & 0 & 0 & \dots & n_a \end{bmatrix}, \tag{4.2.34}$$

A g-inverse is given by,

$$(\mathbf{X}^T\mathbf{X})^g = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1/n_1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1/n_a \end{bmatrix}, \tag{4.2.35}$$

The projection, $\mathbf{P}$ is obtained as,

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T = blockdiag\left\{\frac{1}{n_i}\mathbf{J}_{n_i}, i = 1, 2, \dots, a.\right\} \tag{4.2.36}$$

A solution to the normal equation is then obtained as:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 0 \\ \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{a.} \end{pmatrix}. \tag{4.2.37}$$

Corresponding prediction vector $\hat{\mathbf{Y}}$ is given by,

$$\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1_{n_1}\bar{Y}_{1.} \\ 1_{n_2}\bar{Y}_{2.} \\ \vdots \\ 1_{n_a}\bar{Y}_{a.} \end{pmatrix}. \tag{4.2.38}$$

Notice that,

$$\mathbf{e} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \mathbf{Y}_1 - 1_{n_1}\bar{Y}_{1.} \\ \mathbf{Y}_2 - 1_{n_2}\bar{Y}_{2.} \\ \vdots \\ \mathbf{Y}_a - 1_{n_a}\bar{Y}_{a.} \end{pmatrix}. \tag{4.2.39}$$

$$\begin{aligned}\|\hat{\mathbf{Y}}\|^2 &= \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} = n_1\bar{Y}_{1.}^2 + n_2\bar{Y}_{2.}^2 + \ldots + n_a\bar{Y}_{a.}^2 \\ &= \sum_{i=1}^{a} n_i\bar{Y}_{i.}^2. \end{aligned} \tag{4.2.40}$$

and

$$\begin{aligned}\|\mathbf{e}\|^2 &= \mathbf{e}^T\mathbf{e} = \mathbf{Y}_1^T\mathbf{Y}_1 - n_1\bar{Y}_{1.}^2 + \mathbf{Y}_2^T\mathbf{Y}_2 - n_2\bar{Y}_{2.}^2 + \ldots + \mathbf{Y}_a^T\mathbf{Y}_a - n_a\bar{Y}_{a.}^2 \\ &= \sum_{i=1}^{a}\left\{\mathbf{Y}_i^T\mathbf{Y}_i - n_i\bar{Y}_{i.}^2\right\} \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left\{Y_{ij} - \bar{Y}_{i.}\right\}^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{a} n_i\bar{Y}_{i.}^2 \\ &= \|\mathbf{Y}\|^2 - \|\hat{\mathbf{Y}}\|^2. \end{aligned} \tag{4.2.41}$$

"Residual SS" = Total SS -"Regression SS". Or,

Total SS = "Regression SS" + "Residual SS".

Theorem 4.2.13. *If $\hat{\boldsymbol{\beta}}$ is a solution to the normal equations (4.2.6), then,*

$$\|\mathbf{Y}\|^2 = \|\hat{\mathbf{Y}}\|^2 + \|\mathbf{e}\|^2, \tag{4.2.42}$$

where, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

Proof. Left as an exercise. □

Definition 4.2.2. Regression SS, Residual SS. The quantity $\|\hat{\mathbf{Y}}\|^2$ is referred to as regression sum of squares or model sum of squares, the portion of total sum of squares explained by the linear model whereas the other part $\|\mathbf{e}\|^2$ is the error sum of squares or residual sum of squares (unexplained variation).

Coefficient of determination (R^2)

To have a general definition, let the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ contains an intercept term, meaning the first column of $\mathbf{X}$ is 1_n. Total sum of

Table 4.1: Analysis of variance

	Models with/without an intercept term	
Source	df	SS
Regression (Model)	r	$\mathbf{Y}^T\mathbf{P}\mathbf{Y}$
Residual (Error)	$n-r$	$\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}$
Total	n	$\mathbf{Y}^T\mathbf{Y}$
	Models with an intercept term	
Source	df	SS
Mean	1	$\mathbf{Y}^T 1_n 1_n^T \mathbf{Y}/n$
Regression (corrected for mean)	$r-1$	$\mathbf{Y}^T(\mathbf{P}-1_n 1_n^T/n)\mathbf{Y}$
Residual (Error)	$n-r$	$\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}$
Total	n	$\mathbf{Y}^T\mathbf{Y}$
	Models with an intercept term	
Source	df	SS
Regression (corrected for mean)	$r-1$	$\mathbf{Y}^T(\mathbf{P}-1_n 1_n^T/n)\mathbf{Y}$
Residual (Error)	$n-r$	$\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}$
Total (corrected)	$n-1$	$\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^T 1_n 1_n^T \mathbf{Y}/n$

squares corrected for the intercept term (or mean) is then written as

$$\begin{aligned} Total\ SS(corr.) &= \mathbf{Y}^T\mathbf{Y} - n\bar{Y}^2 \\ &= \mathbf{Y}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}. \end{aligned} \quad (4.2.43)$$

Similarly, the regression SS is also corrected for the intercept term and is expressed as

$$\begin{aligned} Regression\ SS(corr.) &= \mathbf{Y}^T\mathbf{P}\mathbf{Y} - n\bar{Y}^2 \\ &= \mathbf{Y}^T(\mathbf{P} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}. \end{aligned} \quad (4.2.44)$$

This is the portion of total corrected sum of squares that is purely explained by the design variables in the model. However, an equality similar to (4.4.42) applied to the corrected sums of squares still follows, and the ratio

$$R^2 = \frac{Reg.\ SS(Corr.)}{Total\ SS(Corr.)} = \frac{\mathbf{Y}^T(\mathbf{P} - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}}{\mathbf{Y}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)\mathbf{Y}} \quad (4.2.45)$$

explains the **proportion of total variation** explained by the model. This ratio is known as the coefficient of determination and is denoted by R^2.

Two important results:

Lemma 4.2.14. $\mathbf{I}_p - (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}$ *is a projection onto* $\mathcal{N}(\mathbf{X})$.

Proof. Use lemma 2.7.10. □

Lemma 4.2.15. $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g$ *is a projection onto* $\mathcal{C}(\mathbf{X}^T)$.

Proof. Use lemma 2.7.11. □

Importance:

Sometimes it is easy to obtain a basis for the null space of $\mathbf{X}$ or column space of $\mathbf{X}^T$ by careful examination of the relationship between the columns of $\mathbf{X}$. However, in some cases it is not as straightforward. In such cases, independent non-zero columns from the projection matrix $\mathbf{I}_p - (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}$ can be used as a basis for the null space of $\mathbf{X}$. Similarly, independent non-zero columns from the projection matrix $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g$ can be used as a basis for the column space of $\mathbf{X}^T$.

Example 4.2.16. Example 4.2.12 continued.

$$\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g = \begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}. \tag{4.2.46}$$

Therefore a basis for the column space of $\mathbf{X}^T$ is given by the last n columns of the above matrix. Similarly,

$$\mathbf{I}_{a+1} - (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 0 & 0 & \dots & 0 \\ -1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & \dots & 0 \end{bmatrix}. \tag{4.2.47}$$

Therefore, the only basis vector in the null space of $\mathbf{X}$ is $(1, -1_a^T)^T$.

4.3 Interpreting LS estimator

Usually, an estimator is interpreted by the quantity it estimates. Remember, a solution to the normal equation (4.2.6) is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y}$. What does $\hat{\boldsymbol{\beta}}$ really estimates?

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\beta}. \qquad (4.3.1)$$

Unless $\mathbf{X}$ has full column rank, $\hat{\boldsymbol{\beta}}$ is not an unbiased estimator of $\boldsymbol{\beta}$. It is an unbiased estimator of $\mathbf{H}\boldsymbol{\beta}$, which may not be unique (depends on g-inverse of $\mathbf{X}^T\mathbf{X}$). Therefore, when $\mathbf{X}$ is not of full column rank, the estimator $\hat{\boldsymbol{\beta}}$ is practically meaningless. Nevertheless, being a solution to the normal equations, it helps us construct useful estimators for other important functions of $\boldsymbol{\beta}$ (will discuss later).

Estimating $E(\mathbf{Y})$

Even though the normal equations (4.2.6) may not have a unique solution, it facilitates a unique LS estimator for $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ since

$$E(\hat{\mathbf{Y}}) = E(\mathbf{PY}) = \mathbf{PX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} = E(\mathbf{Y}). \qquad (4.3.2)$$

Thus $\widehat{E(\mathbf{Y})} = \hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{PY}$ is an unique unbiased estimator of $E(\mathbf{Y})$.

Introducing assumptions

So far the only assumptions we put on the response vector $\mathbf{Y}$ or equivalently on the error vector ϵ is that

$$E(\epsilon) = \mathbf{0}. \tag{4.3.3}$$

This was a defining assumption of the general linear model. This allowed us to obtain a unique unbiased estimator for the mean response $\mathbf{X}\boldsymbol{\beta}$. However, without further assumptions on the variance of the responses (or, equivalently of the random errors) it is difficult or even impossible to ascertain how efficient this estimator of the mean response is. We will introduce assumptions as we need them. Let us assume that

Assumption II. Error components are independently and identically distributed with constant variance σ^2.

Variance-covariance matrix for LS estimator

Under assumption II, $cov(\mathbf{Y}) = \sigma^2\mathbf{I}_n$. Variance-covariance matrix $cov(\hat{\boldsymbol{\beta}})$ of a LS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y}$ is given by

$$\begin{aligned} cov(\hat{\boldsymbol{\beta}}) &= cov((\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T cov(\mathbf{Y}) \left[(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\right]^T \\ &= (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X} \left[(\mathbf{X}^T\mathbf{X})^g\right]^T \sigma^2 \qquad (4.3.4) \end{aligned}$$

For full rank cases (4.3.4) reduces to the familiar form $cov(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$.

Variance-covariance matrix for $\hat{\mathbf{Y}}$

Example 4.3.1. Show that

1. $cov(\hat{\mathbf{Y}}) = \mathbf{P}\sigma^2$.

2. $cov(\mathbf{e}) = (\mathbf{I} - \mathbf{P})\sigma^2$.

Estimating the error variance

Note that, using Theorem 3.4.7,

$$\begin{aligned} E(Residual\ SS) &= E\left[\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}\right] \\ &= trace\left\{(\mathbf{I}-\mathbf{P})\sigma^2\mathbf{I}_n\right\} + (\mathbf{X}\boldsymbol{\beta})^T(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2 trace\left\{(\mathbf{I}-\mathbf{P})\right\} + \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2(n-r), \end{aligned} \tag{4.3.5}$$

where $r = rank(\mathbf{X})$. Therefore, an unbiased estimator of the error variance σ^2 is given by

$$\hat{\sigma^2} = \frac{Residual\ SS}{n-r} = \frac{Residual\ MS}{n-r} = \frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}}{n-r}. \tag{4.3.6}$$

4.4 Estimability

Unless $\mathbf{X}$ is of full column rank, solution to the normal equations (4.2.6) is not unique. Therefore, in such cases, a solution to the normal equation does not estimate any useful population quantity. More specifically, we have shown that $E(\hat{\boldsymbol{\beta}}) = \mathbf{H}\boldsymbol{\beta}$, where $\mathbf{H} =$

$(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}$. Consider the following $\mathbf{X}^T\mathbf{X}$ matrix

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{bmatrix} \tag{4.4.1}$$

from a one-way ANOVA experiment with two treatments each replicated 3 times. Let us consider two g-inverses

$$\mathbf{G}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{bmatrix} \tag{4.4.2}$$

and

$$\mathbf{G}_2 = \begin{bmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 2/3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.4.3}$$

with

$$\mathbf{H}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{4.4.4}$$

and

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.4.5}$$

respectively. Now, if $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2)^T$, then,

$$\mathbf{H}_1\boldsymbol{\beta} = \begin{pmatrix} 0 \\ \mu + \alpha_1 \\ \mu + \alpha_2 \end{pmatrix} \tag{4.4.6}$$

whereas

$$\mathbf{H}_2\boldsymbol{\beta} = \begin{pmatrix} \mu + \alpha_1 \\ \alpha_1 - \alpha_2 \\ 0 \end{pmatrix}. \tag{4.4.7}$$

Thus two solutions to the same normal equations set estimate two different quantities. However, in practice, one would like to construct estimators that estimate the same population quantity, no matter what solution to the normal equation is used to derive that estimator. One important goal in one-way ANOVA is to estimate the difference

between two treatment effects, namely, $\delta = \alpha_1 - \alpha_2 = (0, 1, -1)\boldsymbol{\beta}$. Two different solutions based on the two g-inverses G_1 and G_2 are given by $\hat{\boldsymbol{\beta}}_1 = (0, \bar{Y}_{1.}, \bar{Y}_{2.})^T$ and $\hat{\boldsymbol{\beta}}_1 = (\bar{Y}_{2.}, \bar{Y}_{1.} - \bar{Y}_{2.}, 0)^T$. If we construct our estimator for δ based on the solution $\hat{\boldsymbol{\beta}}_1$, we obtain

$$\hat{\delta}_1 = (0, 1, -1)\hat{\boldsymbol{\beta}}_1 = \bar{Y}_{1.} - \bar{Y}_{2.}, \tag{4.4.8}$$

exactly the quantity you would expect. Now let us see if the same happens with the other solution $\hat{\boldsymbol{\beta}}_2$. For this solution,

$$\hat{\delta}_2 = (0, 1, -1)\hat{\boldsymbol{\beta}}_2 = \bar{Y}_{1.} - \bar{Y}_{2.}, \tag{4.4.9}$$

same as $\hat{\delta}_1$. Now we will show that no matter what solution you pick for the normal equation, $\hat{\delta}$ will always be the same. To see it, let us write $\hat{\delta}$ as

$$\begin{aligned} \hat{\delta} &= (0, 1, -1)(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{P}_\delta\mathbf{Y}, \end{aligned} \tag{4.4.10}$$

where $\mathbf{P}_\delta = (0, 1, -1)(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T$. If we can show that $\mathbf{P}_\delta$ does not depend on the choice of g-inverse $(\mathbf{X}^T\mathbf{X})^g$, then we are through. Let

us first look at the $\mathbf{X}^T$-matrix for this simpler version of one-way ANOVA problem:

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \tag{4.4.11}$$

Notice that, we can write $(0, 1, -1)^T$ in many ways as a linear combination of the columns of $\mathbf{X}^T$. Pick your favorite combination:

$$\begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \end{pmatrix} = \mathbf{X}^T\mathbf{c}. \tag{4.4.12}$$

Now,

$$\mathbf{P}_\delta = (0, 1, -1)(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T = \mathbf{c}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T = \mathbf{c}^T\mathbf{P}. \tag{4.4.13}$$

Since $\mathbf{P}$ does not depend on the choice of $(\mathbf{X}^T\mathbf{X})^g$, from the above equation we see that $\mathbf{P}_\delta$, and hence $\hat{\delta} = \mathbf{P}_\delta\mathbf{Y}$ is unique to the choice

of a g-inverse.

Summary

- Not all linear functions of $\boldsymbol{\beta}$ may be estimated uniquely based on the LS method.
- Linear functions $\lambda^T\boldsymbol{\beta}$ of $\boldsymbol{\beta}$, where λ is a linear combination of the columns of $\mathbf{X}^T$ allows unique estimators based on the LS estimator.

Estimable functions

Definition 4.4.1. $\hat{\theta}(\mathbf{Y})$ is an **unbiased** estimator of θ if and only if $E\left[\hat{\theta}(\mathbf{Y})\right] = \theta$, for all θ.

Definition 4.4.2. $\hat{\theta}(\mathbf{Y})$ is a **linear** estimator of θ if and only if $\hat{\theta}(\mathbf{Y}) = \mathbf{a}^T\mathbf{Y} + \mathbf{b}$, for some constant (vector) $\mathbf{b}$ and vector (matrix) $\mathbf{a}$.

Definition 4.4.3. A linear function $\theta = \lambda^T\boldsymbol{\beta}$ is linearly estimable if and only if there exists a linear function $\mathbf{c}^T\mathbf{Y}$ such that $E(\mathbf{c}^T\mathbf{Y}) =$

$\lambda^T\boldsymbol{\beta} = \theta$, for all $\boldsymbol{\beta}$.

We will drop "linearly" from "linearly estimable" for simplicity. That means "estimable" will always refer to linearly estimable unless mentioned specifically.

Example 4.4.1.

1. Components of the mean vector $\mathbf{X}\boldsymbol{\beta}$ are estimable.

2. Components of the vector $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ are estimable.

Proposition 4.4.2. *Linear combinations of estimable functions are estimable.*

Proof. Follows from the definition 4.4.3. □

Proposition 4.4.3. *A linear function $\theta = \lambda^T\boldsymbol{\beta}$ is estimable if and only if $\lambda \in \mathcal{C}(\mathbf{X}^T)$.*

Proof. Suppose $\theta = \lambda^T\boldsymbol{\beta}$ is estimable. Then, by definition, there exists a vector $\mathbf{c}$ such that

$$\begin{aligned} & E(\mathbf{c}^T\mathbf{Y}) = \lambda^T\boldsymbol{\beta}, \text{for all } \boldsymbol{\beta} \\ \Longrightarrow \quad & \mathbf{c}^T\mathbf{X}\boldsymbol{\beta} = \lambda^T\boldsymbol{\beta}, \text{for all } \boldsymbol{\beta} \\ \Longrightarrow \quad & \mathbf{c}^T\mathbf{X} = \lambda^T, \text{for all } \boldsymbol{\beta} \\ \Longrightarrow \quad & \lambda = \mathbf{X}^T\mathbf{c} \\ \Longrightarrow \quad & \lambda \in \mathcal{C}(\mathbf{X}^T). \end{aligned} \tag{4.4.14}$$

Now, suppose $\lambda \in \mathcal{C}(\mathbf{X}^T)$. This implies that $\lambda = \mathbf{X}^T\mathbf{c}$ for some $\mathbf{c}$. Then, for all $\boldsymbol{\beta}$,

$$\lambda^T\boldsymbol{\beta} = \mathbf{c}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{c}^T E(\mathbf{Y}) = E(\mathbf{c}^T\mathbf{Y}). \tag{4.4.15}$$

□

Proposition 4.4.4. *If $\theta = \lambda^T \boldsymbol{\beta}$ is estimable then there exists a* ***unique*** $\mathbf{c}^* \in \mathcal{C}(\mathbf{X})$ *such that* $\lambda = \mathbf{X}^T \mathbf{c}^*$.

Proof. Proposition 4.4.3 indicates that there exists a $\mathbf{c}$ such that $\lambda = \mathbf{X}^T \mathbf{c}$. But any vector $\mathbf{c}$ can be written as a direct sum of two unique components belonging to two orthogonal complements. Thus, we can find $\mathbf{c}^* \in \mathcal{C}(\mathbf{X})$ and $\mathbf{c}^{**} \in \mathcal{N}(\mathbf{X}^T)$ such that

$$\mathbf{c} = \mathbf{c}^* + \mathbf{c}^{**}. \tag{4.4.16}$$

Now

$$\lambda = \mathbf{X}^T \mathbf{c} = \mathbf{X}^T \mathbf{c}^* + \mathbf{X}^T \mathbf{c}^{**} = \mathbf{X}^T \mathbf{c}^*. \tag{4.4.17}$$

Hence, the proof. □

Proposition 4.4.5. *Collection of all possible estimable functions constitutes a vector space of dimension $r = rank(\mathbf{X})$.*

Proof. **Hint:** (i) Show that linear combinations of estimable functions are also estimable, and (ii) Use proposition 4.4.3. □

Methods to determine estimability

Method 1. $\lambda^T\boldsymbol{\beta}$ is estimable if and only if it can be expressed as a linear combinations of the rows of $\mathbf{X}\boldsymbol{\beta}$.

Method 2. $\lambda^T\boldsymbol{\beta}$ is estimable if and only if $\lambda^T\mathbf{e} = 0$ for all bassis vectors $\mathbf{e}$ of the null space of $\mathbf{X}$.

Method 3. $\lambda^T\boldsymbol{\beta}$ is estimable if and only if λ is a linear combination of the basis vectors of $\mathcal{C}(\mathbf{X}^T)$.

Example 4.4.6. Multiple linear regression (Example 1.1.5 continued..) In the case of multiple regression with p independent variables (which may include the intercept term) and n observations $(n > p)$, columns of $\mathbf{X}$ are all independent. Therefore, $\mathcal{N}(\mathbf{X}) = \{0\}$. By method 2, all linear functions of $\boldsymbol{\beta}$ are estimable. In particular,

1. Individual coefficients β_j are estimable.

2. Differences between two coefficients are estimable.

Example 4.4.7. Example 4.2.12 continued.

1. Treatment-specific means $\mu + \alpha_i, i = 1, 2, \ldots, a$ are estimable (using Method 1).

2. Difference between two treatment effects $(\alpha_i - \alpha_{i'})$ is estimable. (Follows from the above, or can be inferred by Method 2).

3. In general, any linear combination $\lambda^T \boldsymbol{\beta} = \lambda_0 \mu + \sum_{i=1}^{a} \lambda_i \alpha_i$ is estimable if and only if $\lambda_0 = \sum_{i=1}^{a} \lambda_i$. (Use Method 2).

Example 4.4.8. Two-way nested design. Suppose n_i patients are randomized to the ith level of treatment A, $i = 1, 2, \ldots, a$ and within the ith treatment group a second randomization is done to b_i levels of treatment B which are unique to each level of treatment A. The linear model for this problem can be written as

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}, \\ & i = 1, 2, \ldots, a; j = 1, 2, \ldots, b_i; k = 1, 2, \ldots, n_{ij}. \end{aligned} \tag{4.4.18}$$

Then the $\mathbf{X}-$matrix for this problem is given by

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{4.4.19}$$

where we have simplified the problem by taking $a = 2, b_1 = b_2 = 2$, and $n_{11} = n_{12} = n_{21} = n_{22} = 2$. Clearly $rank(\mathbf{X}) = 4$. Dimension of the null space of $\mathbf{X}$ is 7 - 4 = 3. A set of basis vectors for the null space of $\mathbf{X}$ can be written as:

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ -1 \\ -1 \end{pmatrix} \tag{4.4.20}$$

Thus, using Method 2, $\lambda^T\boldsymbol{\beta}$ is estimable if

$$\lambda^T\mathbf{e}_j = 0, \ j = 1, 2, 3. \tag{4.4.21}$$

Specifically, if $\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})^T$, then $\lambda^T\boldsymbol{\beta}$ is es-

timable if the following three conditions are satisfied:

$$\begin{aligned}
&(1)\ \ \lambda_0 = \sum_{i=1}^{2}\sum_{j=1}^{2}\lambda_{ij},\\
&(2)\ \ \lambda_1 = \sum_{j=1}^{2}\lambda_{1j},\\
&(3)\ \ \lambda_2 = \sum_{j=1}^{2}\lambda_{2j}.
\end{aligned} \tag{4.4.22}$$

Let us consider some special cases:

1. Is α_1 estimable?

2. Is $\mu + \alpha_1$ estimable?

3. Is $\alpha_1 - \alpha_2$ estimable?

4. Is $\alpha_1 - \alpha_2 + (\beta_{11} + \beta_{12})/2 - (\beta_{21} + \beta_{22})/2$ estimable?

Definition 4.4.4. Least squares estimator of an estimable function $\lambda^T\boldsymbol{\beta}$ is given by $\lambda^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a solution to the normal equations (4.2.6).

Properties of least squares estimator

Proposition 4.4.9. Uniqueness. *Least squares estimator (of an estimable function) is invariant to the choice of a solution to the normal equations.*

Proof. Let us consider the class of solutions from the normal equations

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y}.$$

Least squares estimator of a an estimable function $\lambda^T\boldsymbol{\beta}$ is then given by

$$\lambda^T\hat{\boldsymbol{\beta}} = \lambda^T(\mathbf{X}^T\mathbf{X})^T\mathbf{X}^T\mathbf{Y}. \tag{4.4.23}$$

From proposition 4.4.4, since $\lambda^T\boldsymbol{\beta}$ is estimable, there exists a unique $\mathbf{c} \in \mathcal{C}(\mathbf{X})$ such that

$$\lambda = \mathbf{X}^T\mathbf{c}. \tag{4.4.24}$$

Therefore, Equation (4.4.23) combined with (4.4.24) leads to

$$\lambda^T\hat{\boldsymbol{\beta}} = \mathbf{c}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} = \mathbf{c}^T\mathbf{P}\mathbf{Y}. \tag{4.4.25}$$

Since both $\mathbf{c}$ and $\mathbf{P}$ are unique (does not depend on the choice of g-inverse), the result follows. □

Proposition 4.4.10. Linearity and Unbiasedness. *LS estimator is linear and unbiased.*

Proof. Left as an exercise. □

Proposition 4.4.11. Variance. *Under Assumption II,*

$$Var(\lambda^T\hat{\boldsymbol{\beta}}) = \sigma^2\lambda^T(\mathbf{X}^T\mathbf{X})^g\lambda. \tag{4.4.26}$$

Proof.

$$\begin{aligned} Var(\lambda^T\hat{\boldsymbol{\beta}}) &= Var\left[\lambda^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y}\right] \\ &= \lambda^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T cov(\mathbf{Y})\left\{\lambda^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\right\}^T \\ &= \sigma^2\lambda^T(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\left\{(\mathbf{X}^T\mathbf{X})^g\right\}^T\lambda \\ &\stackrel{?}{=} \sigma^2\lambda^T(\mathbf{X}^T\mathbf{X})^g\lambda. \end{aligned} \tag{4.4.27}$$

□

Proposition 4.4.12. Characterization. *If an estimator $\lambda^T\hat{\boldsymbol{\beta}}$ of a linear function $\lambda^T\boldsymbol{\beta}$ is invariant to the choice of the solutions $\hat{\boldsymbol{\beta}}$ to the normal equations, then $\lambda^T\boldsymbol{\beta}$ is estimable.*

Proof. For a given g-inverse $\mathbf{G}$ of $\mathbf{X}^T\mathbf{X}$, consider the general form of the solutions to the normal equations:

$$\hat{\boldsymbol{\beta}} = \mathbf{G}\mathbf{X}^T\mathbf{Y} + (\mathbf{I} - \mathbf{G}\mathbf{X}^T\mathbf{X})\mathbf{c} \tag{4.4.28}$$

for any vector $\mathbf{c} \in \mathcal{R}^p$. Then,

$$\begin{aligned} \lambda^T\hat{\boldsymbol{\beta}} &= \lambda^T\left\{\mathbf{G}\mathbf{X}^T\mathbf{Y} + (\mathbf{I} - \mathbf{G}\mathbf{X}^T\mathbf{X})\mathbf{c}\right\} \\ &= \lambda^T\mathbf{G}\mathbf{X}^T\mathbf{Y} + \lambda^T(\mathbf{I} - \mathbf{G}\mathbf{X}^T\mathbf{X})\mathbf{c}. \end{aligned} \tag{4.4.29}$$

Since $\mathbf{G}$ is given, in order for the above to be equal for all $\mathbf{c}$, we must have

$$\lambda^T(\mathbf{I} - \mathbf{G}\mathbf{X}^T\mathbf{X}) = 0. \tag{4.4.30}$$

Or, equivalently,

$$\lambda^T = \lambda^T\mathbf{G}\mathbf{X}^T\mathbf{X}. \tag{4.4.31}$$

This last equation implies that $\lambda \in \mathcal{C}(\mathbf{X}^T)$. This completes the proof. □

Theorem 4.4.13. Gauss-Markov Theorem. *Under Assumptions I and II, if $\lambda^T\boldsymbol{\beta}$ is estimable, then the least squares estimator $\lambda^T\hat{\boldsymbol{\beta}}$ is the unique minimum variance linear unbiased estimator.*

In the econometric literature, minimum variance is referred to as best and along with the linearity and unbiasedness the least squares estimator becomes best linear unbiased estimator (BLUE).

Proof. Uniqueness follows from the proposition 4.4.9. Linearity and unbiasedness follows from the proposition 4.4.10. The only thing remains to be shown is that no other linear unbiased estimator of $\lambda^T\boldsymbol{\beta}$ can have smaller variance than $\lambda^T\hat{\boldsymbol{\beta}}$.

Since $\lambda^T\boldsymbol{\beta}$ is estimable, there exists a $\mathbf{c}$ such that $\lambda = \mathbf{X}^T\mathbf{c}$. Let $\mathbf{a} + \mathbf{d}^T\mathbf{Y}$ be any other linear unbiased estimator of $\lambda^T\boldsymbol{\beta}$. Then, we

must have $\mathbf{a} = 0$ and $\lambda^T = \mathbf{d}^T\mathbf{X}$. Then,

$$\begin{aligned} \mathbf{X}^T\mathbf{d} &= \mathbf{X}^T\mathbf{c} \\ &\Longrightarrow \mathbf{X}^T(\mathbf{c}-\mathbf{d}) = 0 \\ &\Longrightarrow (\mathbf{c}-\mathbf{d}) \in \mathcal{N}(\mathbf{X}^T) \\ &\Longrightarrow \mathbf{P}(\mathbf{c}-\mathbf{d}) = 0 \\ &\Longrightarrow \mathbf{Pc} = \mathbf{Pd}. \end{aligned} \tag{4.4.32}$$

Now, by proposition 4.4.11,

$$var(\lambda^T\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{c}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{c} = \sigma^2\mathbf{c}^T\mathbf{Pc}. \tag{4.4.33}$$

and

$$var(\mathbf{d}^T\mathbf{Y}) = \sigma^2\mathbf{d}^T\mathbf{d}. \tag{4.4.34}$$

Thus,

$$\begin{aligned} var(\mathbf{d}^T\mathbf{Y}) - var(\lambda^T\hat{\boldsymbol{\beta}}) &= \sigma^2\left\{\mathbf{d}^T\mathbf{d} - \mathbf{c}^T\mathbf{P}\mathbf{c}\right\} \\ &= \sigma^2\left\{\mathbf{d}^T\mathbf{d} - \mathbf{c}^T\mathbf{P}^2\mathbf{c}\right\} \\ &= \sigma^2\left\{\mathbf{d}^T\mathbf{d} - \mathbf{d}^T\mathbf{P}^2\mathbf{d}\right\} \\ &= \sigma^2\mathbf{d}^T(\mathbf{I}-\mathbf{P})\mathbf{d} \qquad (4.4.35) \\ &\geq 0. \end{aligned}$$

Therefore the LS estimator has the minimum variance among all linear unbiased estimators. Equation (4.4.35) shows that $var(\mathbf{d}^T\mathbf{Y}) = var(\lambda^T\hat{\boldsymbol{\beta}})$ if and only if $(\mathbf{I}-\mathbf{P})\mathbf{d} = 0$, or equivalently, $\mathbf{d} = \mathbf{P}\mathbf{d} = \mathbf{P}\mathbf{c}$, leading to $\mathbf{d}^T\mathbf{Y} = \mathbf{c}^T\mathbf{P}\mathbf{Y} = \mathbf{c}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} = \lambda^T\hat{\boldsymbol{\beta}}$. □

Example 4.4.14. Example 4.4.8 continued.

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 8 & 4 & 4 & 2 & 2 & 2 & 2 \\ 4 & 4 & 0 & 2 & 2 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 2 & 2 \\ 2 & 2 & 0 & 2 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 & 0 & 2 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 & 2 \end{bmatrix}, \tag{4.4.36}$$

a g-inverse of which is given by,

$$(\mathbf{X}^T\mathbf{X})^g = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \end{bmatrix}. \tag{4.4.37}$$

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}^T \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{pmatrix} = \begin{pmatrix} Y_{...} \\ Y_{1..} \\ Y_{2..} \\ Y_{11.} \\ Y_{12.} \\ Y_{21.} \\ Y_{22.} \end{pmatrix} \tag{4.4.38}$$

Thus, a solution to the normal equations is given by

$$\begin{pmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_{11} \\ \hat{\beta}_{12} \\ \hat{\beta}_{21} \\ \hat{\beta}_{22} \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \bar{Y}_{11.} \\ \bar{Y}_{12.} \\ \bar{Y}_{21.} \\ \bar{Y}_{22.} \end{pmatrix}. \tag{4.4.39}$$

Therefore **the** linear MVUE (or BLUE) of the estimable function $\alpha_1 - \alpha_2 + (\beta_{11} + \beta_{12})/2 - (\beta_{21} + \beta_{22})/2$ is given by $(\bar{Y}_{11.} + \bar{Y}_{12.})/2 - (\bar{Y}_{21.} + \bar{Y}_{22.})/2$.

4.4.1 A comment on estimability and Missing data

The concept of estimability is very important in drawing statistical inference from a linear model. What effects can be estimated from an experiment totally depends on how the experiment was designed. For instance, in a two-way nested model, difference between two main effects is not estimable, whereas difference between two nested effects within the same main effect is. In an over-parameterized one-way ANOVA model (One-way ANOVA with an intercept term), the treatment effects are not estimable while the difference between any two pair of treatments is estimated by the difference in corresponding cell means.

When observations in some cells are missing, the problem of estimability becomes more acute. We illustrate the concept by using an example. Consider the two-way nested design considered in Example 4.4.8. Suppose after planning the experiment, the observation corresponding to the last two rows of $\mathbf{X}$ matrix could not be observed.

Thus the observed design matrix is given by

$$\mathbf{X}_M = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{4.4.40}$$

How does this effect the estimability of certain functions? Note that $rank(\mathbf{X}_M) = 3$. A basis for the null space of $\mathbf{X}^T$ is given by

$$\left\{ \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \\ 1 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ -1 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \mathbf{e}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}. \tag{4.4.41}$$

1. Is α_1 estimable?

   $\alpha_1 = (0, 1, 0, 0, 0, 0, 0)\boldsymbol{\beta} = \lambda_1^T \boldsymbol{\beta}$.

   $\lambda_1^T \mathbf{e}_1 = 0 \neq \lambda_1^T \mathbf{e}_2 \ \rightarrow$ Not estimable.

2. Is $\mu + \alpha_1$ estimable?

3. Is $\alpha_1 - \alpha_2$ estimable?

4. Is $\alpha_1 - \alpha_2 + (\beta_{11} + \beta_{12})/2 - (\beta_{21} + \beta_{22})/2$ estimable?

   Here, $\lambda_4^T = (0, 1, -1, 1/2, 1/2, -1/2, -1/2)$, and

   $\lambda_4^T \mathbf{e}_1 \neq 0 \ \rightarrow$ Not estimable.

5. Is $\alpha_1 - \alpha_2 + (\beta_{11} + \beta_{12})/2 - \beta_{21}$ estimable?

   Here, $\lambda_5^T = (0, 1, -1, 1/2, 1/2, -1, 0)$, and you can check that

   $\lambda_5^T \mathbf{e}_j = 0, \ j = 1, 2, 3, 4 \ \rightarrow$ Estimable.

4.5 Least squares estimation under linear constraints

Often it is desirable to estimate the parameters from a linear model under certain linear constraints. Two possible scenarios where such constrained minimization of the error sum of squares ($\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$) becomes handy are as follows.

1. **Converting a non-full rank model to a full rank model.**

   A model of non-full rank can be transformed into a full rank model by imposing a linear constraint on the model. Let us take a simple example of a balanced one-way ANOVA with two treatments. The over-parameterized version of this model can be written as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},\ i = 1, 2;\ j = 1, 2, \ldots, n. \qquad (4.5.1)$$

   We know from our discussion that α_i is not estimable in this model. We also know that the $\mathbf{X}$-matrix is not of full rank,

leading to more than one solutions for the normal equations

$$\begin{aligned} 2\mu + \alpha_1 + \alpha_2 &= 2\bar{Y}_{..} \\ \mu + \alpha_1 &= \bar{Y}_{1.} \\ \mu + \alpha_2 &= \bar{Y}_{2.} \end{aligned} \tag{4.5.2}$$

One traditional way of obtaining a unique solution is to impose some restrictions on the parameters. A popular one is to treat one of the treatment effect as a reference by setting it equal to zero. Treating $\alpha_2 = 0$ leads to the solution $\hat{\alpha}_1 = \bar{Y}_{1.} - \bar{Y}_{2.}$ and $\hat{\mu} = \bar{Y}_{2.}$. Another restriction that is commonly applied is that the treatment effects are centered to zero. That is, $\alpha_1 + \alpha_2 = 0$. If we apply this last restriction to the above normal equations, we obtain a unique solution: $\hat{\mu} = \bar{Y}_{..}$, $\hat{\alpha}_1 = \bar{Y}_{1.}$, and $\hat{\alpha}_2 = \bar{Y}_{2.}$.

2. **Testing a linear hypothesis.** One major goal in statistical analysis involving linear models is to test certain hypothesis regarding the parameters. A certain linear hypothesis can be tested by comparing the residual sum of squares from the model

under the null hypothesis to the same from unrestricted model (no hypothesis). Details will follow in Chapter 6.

4.5.1 Restricted Least Squares

Suppose the linear model is of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{4.5.3}$$

where a set of linear restrictions

$$\mathbf{A}^T\boldsymbol{\beta} = \mathbf{b} \tag{4.5.4}$$

has been imposed on the parameters for given matrices $\mathbf{A}$ and $\mathbf{b}$. We want to minimize the residual sum of squares

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.5.5}$$

for $\boldsymbol{\beta}$ to obtain the LS estimators under the constraints (4.5.4).The problem can easily be written as a Lagrangian optimization problem by constructing the objective function

$$E = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda^T(\mathbf{A}^T\boldsymbol{\beta} - \mathbf{b}), \tag{4.5.6}$$

which needs to be minimized unconditionally with respect to $\boldsymbol{\beta}$ and λ. Taking the derivatives of (4.5.6) with respect to $\boldsymbol{\beta}$ and λ and setting them equal to zero, we obtain,

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\lambda = \mathbf{X}^T\mathbf{Y} \tag{4.5.7}$$

$$\mathbf{A}^T\boldsymbol{\beta} = \mathbf{b}. \tag{4.5.8}$$

The above equations will be referred to as restricted normal equations (RNE). We will consider two different scenarios.

CASE I. $\mathbf{A}^T\boldsymbol{\beta}$ is estimable.

A set of q linear constraints $\mathbf{A}^T\boldsymbol{\beta}$ is estimable if and only if each constraint is estimable. If we write $\mathbf{A}$ as $(\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_q)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_q)$, then $\mathbf{A}^T\boldsymbol{\beta}$ is estimable iff each component $\mathbf{a}_i^T\boldsymbol{\beta}$ is estimable. Although, the q constraints need not be independently estimable, but we assume that they are so that is $rank(\mathbf{A}) = q$. If they are not, one can easily reduce them into a set of linearly independent constraints.

Now, if $(\hat{\boldsymbol{\beta}}_r, \hat{\lambda}_r)$ is a solution to the restricted normal equations,

then from (4.5.7) we obtain,

$$\hat{\boldsymbol{\beta}}_r = (\mathbf{X}^T\mathbf{X})^g(\mathbf{X}^T\mathbf{Y} - \mathbf{A}\hat{\lambda}_r) = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^g\mathbf{A}\hat{\lambda}_r. \tag{4.5.9}$$

From (4.5.8), using (4.5.9), assuming the required inverse exists,

$$\hat{\lambda}_r = \left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}(\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b}). \tag{4.5.10}$$

But we have not yet shown that $\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}$ is invertible. The following proposition takes care of that.

Proposition 4.5.1. *In terms of the notations of this section, when $\mathbf{A}^T\boldsymbol{\beta}$ is estimable,*

$$rank(\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}) = rank(\mathbf{A}) = q. \tag{4.5.11}$$

Proof. □

Using (4.5.9) and (4.5.10), it is possible to express the restricted least square estimator $\hat{\boldsymbol{\beta}}_r$ in terms of an unrestricted LS estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^g\mathbf{A}\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}(\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b}). \qquad (4.5.12)$$

Example 4.5.2. Take the simple example of one-way balanced ANOVA from the beginning of this section. Consider the restriction $\alpha_1 - \alpha_2 = 0$, which can be written as $\mathbf{A}^T\boldsymbol{\beta} = 0$, where

$$\mathbf{A} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \qquad (4.5.13)$$

A g-inverse of the $\mathbf{X}^T\mathbf{X}$-matrix is given by

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1/n & 0 \\ 0 & 0 & 1/n \end{pmatrix} \qquad (4.5.14)$$

with corresponding unrestricted solution,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0 \\ \bar{Y}_{1\cdot} \\ \bar{Y}_{2\cdot} \end{pmatrix}. \qquad (4.5.15)$$

$$\mathbf{A}^T\hat{\boldsymbol{\beta}} = \bar{Y}_{1.} - \bar{Y}_{2.}. \tag{4.5.16}$$

$$\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A} = 2/n. \tag{4.5.17}$$

$$(\mathbf{X}^T\mathbf{X})^g\mathbf{A} = \begin{pmatrix} 0 \\ 1/n \\ -1/n \end{pmatrix}. \tag{4.5.18}$$

Using these in equation 4.5.12, we obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}}_r &= \begin{pmatrix} 0 \\ \bar{Y}_{1.} \\ \bar{Y}_{2.} \end{pmatrix} - \begin{pmatrix} 0 \\ 1/n \\ -1/n \end{pmatrix} \left(\frac{n}{2}\right)(\bar{Y}_{1.} - \bar{Y}_{2.}). \\ &= \begin{pmatrix} 0 \\ (\bar{Y}_{1.} + \bar{Y}_{2.})/2 \\ (\bar{Y}_{1.} + \bar{Y}_{2.})/2. \end{pmatrix} \end{aligned} \tag{4.5.19}$$

Is this restricted solution unique? Try with a different g-inverse. (Note you do not have to compute $\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}$, as it is invariant to the choice of a g-inverse.)

Properties of restricted LS estimator

Proposition 4.5.3. *1.* $E\left[\hat{\boldsymbol{\beta}}_r\right] = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{H}\boldsymbol{\beta} = E\left[\hat{\boldsymbol{\beta}}\right]$.

2. $cov\left(\hat{\boldsymbol{\beta}}_r\right) = \sigma^2\left\{(\mathbf{X}^T\mathbf{X})^g\mathbf{D}\left[(\mathbf{X}^T\mathbf{X})^g\right]^T\right\}$, *where* $\mathbf{D} = \mathbf{I} - \mathbf{A}\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}\mathbf{A}^T$.

3. $E(RSS_r) = E\left[(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_r)^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_r)\right] = (n - r + q)\sigma^2$.

Proof. We will leave the first two as exercises. For the third one,

$$\begin{aligned}
RSS_r &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_r)^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_r) \\
&= (\underbrace{\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}}_{\in\mathcal{N}(\mathbf{X}^T)} + \underbrace{\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_r}_{\in\mathcal{C}(\mathbf{X})})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_r) \\
&= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_r)^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_r) \\
&= RSS + (\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b})^T\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}\mathbf{A}^T\left\{(\mathbf{X}^T\mathbf{X})^g\right\}^T \\
&\qquad \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}(\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b}) \\
&= RSS + (\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b})^T\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}(\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b}). \\
E(RSS_r) &= E(RSS) + E\left\{(\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b})^T\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}(\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b})\right\} \\
&= (n - r)\sigma^2 + trace\left\{\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}cov(\mathbf{A}^T\hat{\boldsymbol{\beta}} - \mathbf{b})\right\} \\
&= (n - r)\sigma^2 + trace\left\{\left[\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right]^{-1}\sigma^2\mathbf{A}^T(\mathbf{X}^T\mathbf{X})^g\mathbf{A}\right\} \\
&= (n - r + q)\sigma^2. \qquad (4.5.20)
\end{aligned}$$

□

CASE II. $\mathbf{A}^T\boldsymbol{\beta}$ is not estimable.

A set of q linear constraints $\mathbf{A}^T\boldsymbol{\beta}$ is non-estimable if and only if each constraint is non-estimable and no linear combination of the linear constraints is estimable. Assume as before that columns of $\mathbf{A}$ are independent. That is, $rank(\mathbf{A}) = q$. This means $\mathbf{Ac} \notin \mathcal{C}(\mathbf{X}^T)$, for all $p \times 1$ vectors $\mathbf{c}$ (why?). This in turn implies that

$$\mathcal{C}(\mathbf{A}) \cap \mathcal{C}(\mathbf{X}^T) = \{0\}. \tag{4.5.21}$$

On the other hand, from the RNEs,

$$\mathbf{A}\hat{\lambda}_r = \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_r) \in \mathcal{C}(\mathbf{X}^T). \tag{4.5.22}$$

But by definition,

$$\mathbf{A}\hat{\lambda}_r \in \mathcal{C}(\mathbf{A}). \tag{4.5.23}$$

Together we get,

$$\mathbf{A}\hat{\lambda}_r = 0. \tag{4.5.24}$$

Since the columns of $\mathbf{A}$ are independent, this last equation implies that $\hat{\lambda}_r = 0$. The normal equation (4.5.7) then reduces to

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}, \tag{4.5.25}$$

which is the normal equation for the unrestricted LS problem. Thus RNEs in this case have a solution

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y}, \text{ and} \tag{4.5.26}$$

$$\hat{\lambda}_r = 0. \tag{4.5.27}$$

Therefore, in this case the residual sums of squares from restricted and unrestricted model are identical. *i.e.* $RSS_r = RSS$.

4.6 Problems

1. The least squares estimator of β can be obtained by minimizing $\|\mathbf{Y} - \mathbf{X}\beta\|^2$. Use the derivative approach to derive the normal equations for estimating β.

2. For the linear model

$$y_i = \mu + \alpha x_i + \epsilon_i, i = 1, 2, 3,$$

where $x_i = (i - 1)$.

   (a) Find $\mathbf{P}$ and $I - \mathbf{P}$.

   (b) Find a solution to the equation $\mathbf{X}\beta = \mathbf{P}\mathbf{Y}$.

   (c) Find a solution to the equation $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$. Is this solution same as the solution you found for the previous equation?

   (d) What is the null space of $\mathbf{X}^T$ for this problem?

3. Show that, for any general linear model, the solutions to the system of linear equations $\mathbf{X}\beta = \mathbf{P}\mathbf{Y}$ are the same as the solutions to the normal equations $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$.

4. Show that

   (a) $I - \mathbf{P}$ is a projection matrix onto the null space of $\mathbf{X}^T$, and

   (b) $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^g$ is a projection onto the column space of $\mathbf{X}^T$.

5. (a) If A^g is a generalized inverse of A, then show that $A^- = A^gAA^g + (I - A^gA)B + C(I - AA^g)$ is also a g-inverse of A for any conformable matrices B nd C.

   (b) In class, we have shown that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y}$ is a solution to the normal equations $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$ for a given g-inverse$(\mathbf{X}^T\mathbf{X})^g$ of $\mathbf{X}^T\mathbf{X}$. Show that $\tilde{\beta}$ is a solution to the normal equations if and only if there exists a vector z such that

$\tilde{\beta} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} + (I - (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X})z$. (Thus, by varying z, one can swipe out all possible solutions to the normal equations.)

(c) In fact, $\tilde{\beta} = \mathbf{G}\mathbf{X}^TY$ generates all solutions to the normal equations, for all possible generalized inverses $\mathbf{G}$ of $\mathbf{X}^T\mathbf{X}$. To show this, start with the general solution $\tilde{\beta} = (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{Y} + (I - (\mathbf{X}^T\mathbf{X})^g\mathbf{X}^T\mathbf{X})z$ (from part (b)). Also take it as a fact that for a given non-zero vector $\mathbf{Y}$ and an arbitrary vector z, there exists an arbitrary matrix $\mathbf{M}$ such that $z = \mathbf{M}\mathbf{Y}$. Use this fact, along with the result from part (a) to write $\tilde{\beta}$ as $\mathbf{G}\mathbf{X}^T\mathbf{Y}$ where $\mathbf{G}$ is a g-inverse of $\mathbf{X}^T\mathbf{X}$.

6. For the general one-way ANOVA model,

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, n_i,$$

(a) What is the $\mathbf{X}$ matrix?

(b) Find $r(\mathbf{X})$.

(c) Find a basis for the null space of $\mathbf{X}$.

(d) Give a basis for the set of all possible linearly independent estimable functions.

(e) Give conditions under which $c_0\mu + \sum_{i=1}^a c_i\alpha_i$ is estimable. In particular, is μ estimable? Is $\alpha_1 - \alpha_2$ estimable?

(f) Obtain a solution to the normal equation for this problem and find the least square estimator of $\alpha_a - \alpha_1$.

7. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,\ \ E(\epsilon) = 0,\ \ cov(\epsilon) = \sigma^2 I_n. \tag{4.6.1}$$

Follow the following steps to show that if $\lambda^T\beta$ is estimable, then $\lambda^T\hat{\beta}$ is the BLUE of $\lambda^T\beta$, where $\hat{\beta}$ is a solution to the normal equations $(\mathbf{X}^T\mathbf{X})\beta = \mathbf{X}^T\mathbf{Y}$.

(a) Consider another linear unbiased estimator $c + d^t\mathbf{Y}$ of $\lambda^T\beta$. Show that c must be equal to zero and $d^T\mathbf{X} = \lambda^T$.

(b) Now we will show that $var(c + d^T\mathbf{Y})$ can be written as the $var(\lambda^T\hat{\beta})$ plus some non-negative quantity. To do this, write

$$var(c + d^t\mathbf{Y}) = var(d^T\mathbf{Y}) = var(\lambda^T\hat{\beta} + \underbrace{d^T\mathbf{Y} - \lambda^T\hat{\beta}}_{g(\mathbf{Y})}).$$

Show that $g(\mathbf{Y})$ defined in this manner is a linear function of $\mathbf{Y}$.

(c) Show that $\lambda^T\hat{\beta}$ and $g(\mathbf{Y})$ are uncorrelated. **Hint: Use (i)** $cov(A\mathbf{Y}, B\mathbf{Y}) = Acov(\mathbf{Y})B^T$ **(ii) Result from part (b).**

(d) Hence

$$var(c + d^T\mathbf{Y}) = var(d^T\mathbf{Y}) = var(\lambda^T\hat{\beta}) + \ldots.$$

In other words, variance of any other linear unbiased estimator is greater than or equal to the variance of the least square estimator.

(e) Show that $var(c + d^T\mathbf{Y}) = var(\lambda^T\hat{\beta})$ only if $c + d^T\mathbf{Y} = \lambda^T\hat{\beta}$.

8. One example of a simple two-way nested model is as follows. Suppose two instructors taught two classes using Teaching Method I, and three instructors taught two classes with Teaching Method II. Let Y_{ijk} is the average score for the kth class taught by jth instructor with ith teaching method. The model can be written as:

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}.$$

Assume $E(\epsilon_{ijk}) = 0$, and $cov(\epsilon_{ijk}, \epsilon_{i_1j_1k_1}) = \sigma^2$, if $i = i_1, j = j_1, k = k_1$; 0, otherwise.

(a) Write this model as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, explicitly describing the X matrix and β.

(b) Find r, the rank of $\mathbf{X}$. Give a basis for the null space of $\mathbf{X}$.

(c) Write out the normal equations and give a solution to the normal equations.

(d) How many linearly independent estimable functions can you have in this problem? Provide a list of such estimable functions and give the least squares estimators for each one.

(e) Show that the difference in the effect of two teaching methods is not estimable.

9. Consider the linear model

$$Y_{ij} = \sum_{k=0}^{i-1} \beta_k + \epsilon_{ij},\ i = 1, 2, 3; j = 1, 2; \tag{4.6.2}$$

with $E(\epsilon_{ij}) = 0$; $Var(\epsilon_{ij}) = \sigma^2$; $cov(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ whenever $i' \neq i$ or $j' \neq j$.

9(a) Write the above model in the form of a general linear model. Find $rank(\mathbf{X})$.

9(b) Find $\beta = (\beta_0, \beta_1, \beta_2)^T$ such that the quantity

$$E = \sum_{i=1}^{3} \sum_{j=1}^{2} \left(Y_{ij} - \sum_{k=0}^{i-1} \beta_k \right)^2 \tag{4.6.3}$$

is minimized. Call it $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$.

9(c) Find the mean and variance of $\hat{\beta}$.

For the rest of the parts of this question, assume that ϵ_{ij}'s are normally distributed.

9(d) What is the distribution of $\hat{\beta}$?

9(e) What is the distribution of $\hat{\beta}_1$?

9(f) What is the distribution of $D = \hat{\beta}_1 - \hat{\beta}_2$?

9(g) Find the distribution of

$$\hat{E} = \sum_{i=1}^{3} \sum_{j=1}^{2} \left(Y_{ij} - \sum_{k=0}^{i-1} \hat{\beta}_k \right)^2. \tag{4.6.4}$$

9(h) Are D and $\hat{E}$ independent?

9(i) Find the distribution of $D/\sqrt{\hat{E}}$.

10. Consider the analysis of covariance model

$$Y_{ij} = \mu + \alpha_i + \gamma X_{ij} + \epsilon_{ij}, i = 1, 2;\ j = 1, 2, \ldots, n,$$

where X_{ij} represents the value of a continuous explanatory variable.

(a) Write this model as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, explicitly describing the $\mathbf{X}$ matrix and $\boldsymbol{\beta}$.

(b) Find r, the rank of $\mathbf{X}$. Give a basis for the null space of $\mathbf{X}$.

(c) Give a basis for the null space of $\mathbf{X}$.

(d) Is the regression coefficient γ estimable?

(e) Give conditions under which a linear function $a\mu+b\alpha_1+c\alpha_2+d\gamma$ will be estimable.

For the rest of the problem, assume $n = 5$**, and** $X_{i1} = -2, X_{i2} = -1, X_{i3} = 0, X_{i4} = 1,$ **and** $X_{i5} = 2, i = 1, 2$**.**

(f) Give an expression for the LS estimator of γ and $\alpha_1 - \alpha_2$, if exists.

(g) Obtain the LS estimator of γ under the restriction that $\alpha_1 = \alpha_2$.

(h) Obtain the LS estimator of $\alpha_1 - \alpha_2$ under the restriction that $\gamma = 0$.

(i) Obtain the LS estimator of γ under the restriction that $\alpha_1 + \alpha_2 = 0$.