Let $X_{np}$ denote the design matrix of the linear model. When $\text{rank}(X) = r < p$, $X^T X$ is singular; in this situation $\hat{\beta} = (X^T X)^{-1} X^T y$ is an LSE, where $(X^T X)^{-1}$ is a generalized inverse of $X^T X$.

In Lemma 2.3 we showed that $X(X^T X)^{-1} X^T Y = X$.

In the proof, we assumed that $[X(X^T X)^{-1} X^T Y]_T = X^T X (X^T X)^{-1} X^T Y$ without proof. Although this is true, it needs to be proved. Here is a complete proof of Lemma 2.3.

Proof: By the definition of g-inverse,

$$X^T X (X^T X)^{-1} X^T Y = X^T Y$$

Take transpose of both sides, we have

$$X^T X [(X^T X)^{-1} X^T Y]_T = X^T Y,$$

indicating that $[(X^T X)^{-1} X^T Y]_T$ is also a g-inverse of $X^T X$.

Let $D = X(X^T X)^{-1} X^T Y - X$.

Then $D^T D = [X^T X (X^T X)^{-1} X^T Y]_T - [X^T X (X^T X)^{-1} X^T Y]_T X^T X (X^T X)^{-1} X^T Y - X^T X [X^T X]_T X^T Y - X^T X (X^T X)^{-1} X^T Y + X^T X$

$$= X^T X (X^T X)^{-1} X^T Y - X^T X + X^T X$$

$$= X^T X$$

Therefore, $D = 0$, which further implies that

$$X(X^T X)^{-1} X^T Y = X$$.

Because $[(X^T X)^{-1} X^T Y]_T$ is also a g-inverse, we have

$$X = X(X^T X)^{-1} X^T Y = X[(X^T X)^{-1} X^T Y]_T X^T Y$$

Take transpose we have

$$X^T = X^T X (X^T X)^{-1} X^T = X^T X [((X^T X)^{-1} X^T Y]_T X^T Y$$
2.3 The fitted values and residuals

Fitted value: \( \hat{\theta} = X\hat{\beta} = \hat{Y} = (Y_1, \ldots, Y_n)^T \)

The uniqueness of the prediction

When \( X \) is less than full rank, there are infinitely many LSEs. Do different LSEs lead to different predictions?

**Theorem 2.4** The fitted values, i.e., \( \hat{Y} = X\hat{\beta} \), are unique.

**Proof.** According to Urquhart (1969), all the LSEs of \( \hat{\beta} \) can be written into \( \hat{\beta} = (X^TX)^{-1}X^TY \).

- Define \( P_X = X(X^TX)^{-1}X^T \). Then \( P_X \) is invariant to the choice of \( (X^TX)^{-1} \). The proof is in Appendix B of Seber & Lee.

**Proposition 2.6** \( P_X \) is a projection matrix.

**Proof.**

1. \( P_X^2 = P_XP_X = X(X^TX)^{-1}X^T \cdot X(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T \) \( \stackrel{\text{Lemma 2.3}}{=} P_X \)

This shows that \( P_X \) is idempotent.

2. \( P_X^T = (P_X^2)^T = P_X^T P_X \)

\[
= X[(X^TX)^{-1}X^T \cdot X:]^T \stackrel{\text{Lemma 2.3}}{=} X[X^TX]^T X^T
\]

\[
= X[(X^TX)^{-1}X^T \cdot X]X^T \hspace{0.5cm} (X^TX)^{-1}X^T
\]

\[
= X[(X^TX)^{-1}X^T \cdot X]X^T \hspace{0.5cm} (X^TX)^{-1}X^T
\]

\[
= P_X^2 \equiv P_X
\]

This shows that \( P_X \) is symmetric.

By 1) and 2) and the definition of projection matrix, \( P_X \) is a projection matrix.

**Note.**

1. \( P_X \) is also called the hat matrix: \( P_X \hat{Y} = \hat{Y} \)

2. Another commonly used notation for the hat matrix is \( H \).
Because $P_x$ is a projection matrix, we have

1. $\text{rank}(P_x) = \text{trace}(P_x)$
2. $I - P_x$ is also a projection matrix.
3. The eigenvalues of $P_x$ are equal to either 0 or 1, and the number of 1's is the same as $\text{rank}(P_x)$.

**Residuals**

**Def. Residuals** $e = y - \hat{y}$

**Note:** $e = y - \hat{y} = y - X\hat{\beta} = y - Px y = (I - P_x)y$

**Def. Residual Sum of Squares:** $e^T e$

**Note:** $e^T e = y^T (I - P_x)^T (I - P_x) y = y^T (I - P_x) (I - P_x) y$

$= y^T (I - P_x) y$ (because $I - P_x$ is a projection matrix)

2.4 The Generalized Least Squares Estimates

**Expectations and Covariances**

**Proposition 2.7** Let $Z$ be a random matrix and $A$, $B$, and $C$ be known matrices. Then

$E[AZB + C] = AE[Z]B + C$

**Proof.** Let $W = AZB + C$, then

$W_{ij} = \sum_k a_{ik} Z_{kl} b_{lj} + c_{ij}$

$\Rightarrow E[W_{ij}] = \sum_k a_{ik} E[Z_{kl}] b_{lj} + c_{ij}$

$= \sum_k a_{ik} (E[Z])_{kl} b_{lj} + c_{ij}$

**Proposition 2.8** For random vectors $Y$ and $Z$, known matrices $A$ and $B$, and constants $c$ and $d$

2. $\text{cov}(AY + BZ + c) = A \text{cov}(Y, Z) B^T$
Proof of 2. Let $U = Ay + c$ and $W = Bz + d$. Then
\[ E[U] = AE[y] + c \]
\[ U - E[U] = A(y - E[y]) \]
\[ E[W] = BE[z] + d \]
\[ W - E[W] = B(z - E[z]) \]

By the definition of Cov,
\[ \text{Cov}(Ay + c, Bz + d) = E[(U - E[U])(W - E[W])^T] \]
\[ = E[A(y - E[y])(z - E[z])^T B] \]
\[ = AE[(y - E[y])(z - E[z])^T] B \]
\[ = A \text{Cov}(y, z) B \]

the generalized LSE

Consider the linear model $y = \beta x + \varepsilon$ with $E[\varepsilon] = 0$ and $\text{Var}[\varepsilon] = \Sigma$, where $\Sigma$ is positive definite. Let $\Sigma^{1/2}$ be the square root matrix of $\Sigma$. Consider $z = \Sigma^{-1/2} y$, we have
\[ E[z] = \Sigma^{-1/2} E[y] = \Sigma^{-1/2} \beta \]
\[ \text{Var}[z] = \Sigma^{-1/2} \text{Var}[y] \Sigma^{-1/2} = \Sigma \]

Consider a new design matrix $W = \Sigma^{1/2} x$ and the response vector $Z = \Sigma^{1/2} y$. The LSE for $\beta$ under the new model is
\[ (W^T W) \beta = W^T Z \]

Use the original notations, we have the generalized least squares estimate
\[ \hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y \]
Moments (mean and variance) of LSE

When $X$ is less than full rank, there are infinitely many LSEs of $\beta$. Since $\hat{\beta}$ is not unique, $\beta$ is not estimable.

Is any of the LSEs unbiased for $\beta$? NO.

Consider an arbitrary $\hat{\beta}$. If it is unbiased, we have $\beta = E[\hat{\beta}] = E[(X^T)^{-1}X^TY] = (X^T)^{-1}X^T \beta$ for all $\beta$. Thus $(X^T)^{-1}X^T = I_p$, contradicting with the fact that $X$ is less than full rank.

Definition. $a^T \beta$ is said to be estimable iff $\exists$ $C$ s.t. $E[C^T \gamma] = a^T \beta$

for any $\beta$.

e.g. $X \beta$ is always estimable, as $E[Y\beta] = X \beta$

e.g. $Y_{ij} = U_{ij} + \epsilon_{ij}$, $i=1,2; j=1,2$, $E[\epsilon_{ij}] = 0$

$X \beta_1 = E[Y_{11}]$, so $X \beta_1$ is estimable

$X \beta_1 - X \beta_2 = E[Y_{11} - Y_{12}]$, so $X \beta_1 - X \beta_2$ is estimable

How about $U$?

Theorem 2.9 $a^T \beta$ is estimable iff any one of the following conditions is true:

1. $a \in C(X^T)$ [Note: $C(X^T) = C(X^T X)$]
2. $a^T (X^T X)^{-1} X = a^T$
3. $E[a^T \hat{\beta}] = a^T \beta$ for all $\beta$, where $\hat{\beta}$ is an LSE.

Proof 1. $a^T \beta$ is estimable iff $\exists C$ s.t. $E[C^T \gamma] = a^T \beta$ for all $\beta$

But $E[C^T \gamma] = C^T E[\gamma] = C^T \beta$

So $a^T \beta$ is estimable iff $\exists C$ s.t. $C^T \beta = a^T \beta$ for all $\beta$, which is true iff $a = X^T C$ $\iff$ $a \in C(X^T)$. 