Basic Distributional Assumptions of the Linear Model:

1. The errors are unbiased: $E[\boldsymbol{\varepsilon}] = \mathbf{0}$.
2. The errors are uncorrelated with common variance:

$$\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}.$$

These assumptions imply that

$$E[\mathbf{Y}] = E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta},$$

$$\text{cov}(\mathbf{Y}) = \text{cov}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) =^* \text{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}.$$

*when $\mathbf{X}$ is considered fixed, not random

Results for the full rank case: Under the above assumptions, we have the following results.

1. The least squares estimate is *unbiased*:

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

2. The *covariance matrix* of the least squares estimate is

$$\begin{aligned} \text{cov}(\hat{\boldsymbol{\beta}}) &= \text{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Note that we have NOT yet assumed that errors are normally distributed.

Optimality of Least-Squares Estimates:

In general, $\hat{\boldsymbol{\beta}}$ is not unique so we consider the properties of $\hat{\boldsymbol{\theta}}$, which is unique. This is an unbiased estimate of the mean vector of $\mathbf{Y}$ ($\boldsymbol{\theta} = E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$):

$$\begin{aligned} E[\hat{\boldsymbol{\theta}}] &= E[\mathbf{P}\mathbf{Y}] \\ &= \mathbf{P}E[\mathbf{Y}] \\ &= \mathbf{P}\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}\boldsymbol{\beta} \quad \text{because} \quad \mathbf{P}\mathbf{X} = \mathbf{X} \\ &= \boldsymbol{\theta} \end{aligned}$$

The next result shows that $\hat{\boldsymbol{\theta}}$ is optimal in the sense of having minimum variance among all linear unbiased estimators. This result is the basis of the Gauss-Markov theorem on the estimation of estimable functions, which we will study in a later lecture.

Theorem: Let $\hat{\boldsymbol{\theta}}$ be the least-squares estimate of $\boldsymbol{\theta}$. For any linear combination $\mathbf{c}'\boldsymbol{\theta}$, $\mathbf{c}'\hat{\boldsymbol{\theta}}$ is (uniquely) the estimate with minimum variance among all linear unbiased estimates. We call $\mathbf{c}'\hat{\boldsymbol{\theta}}$ the BLUE (Best Linear Unbiased Estimate) of $\mathbf{c}'\boldsymbol{\theta}$.

Proof: Since $\hat{\boldsymbol{\theta}}$ is unbiased, we have that $E[\mathbf{c}'\hat{\boldsymbol{\theta}}] = \mathbf{c}'\boldsymbol{\theta}$, so $\mathbf{c}'\hat{\boldsymbol{\theta}}$ is a linear unbiased estimate of $\mathbf{c}'\boldsymbol{\theta}$.

Let $\mathbf{d}'\mathbf{Y}$ be any other linear unbiased estimate. Unbiasedness implies that $E[\mathbf{d}'\mathbf{Y}] = \mathbf{c}'\boldsymbol{\theta}$; we also know that $E[\mathbf{d}'\mathbf{Y}] = \mathbf{d}'\boldsymbol{\theta}$. Therefore $\mathbf{d}'\boldsymbol{\theta} = \mathbf{c}'\boldsymbol{\theta}$. Now, $\boldsymbol{\theta}$ is a vector in $\mathcal{R}(\mathbf{X})$ – we do not know what $\boldsymbol{\theta}$ is, but regardless of its value $\mathbf{d}'\boldsymbol{\theta} = \mathbf{c}'\boldsymbol{\theta}$. Therefore, $\mathbf{d}'\boldsymbol{\theta} = \mathbf{c}'\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \mathcal{R}(\mathbf{X})$.

It follows immediately that $(\mathbf{c} - \mathbf{d})'\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta} \in \mathcal{R}(\mathbf{X})$, so $\mathbf{c} - \mathbf{d}$ is orthogonal to $\mathcal{R}(\mathbf{X})$. Therefore, $\mathbf{P}(\mathbf{c} - \mathbf{d}) = \mathbf{0}$ and $\mathbf{Pc} = \mathbf{Pd}$. Now $\text{var}(\mathbf{c}'\hat{\boldsymbol{\theta}}) =$

$$\begin{aligned}
\text{var}(\mathbf{c}'\hat{\boldsymbol{\theta}}) &= \text{var}(\mathbf{c}'\mathbf{PY}) \\
&= \text{var}([\mathbf{Pc}]'\mathbf{Y}) \\
&= \text{var}([\mathbf{Pd}]'\mathbf{Y}) \\
&= \sigma^2(\mathbf{Pd})'\mathbf{Pd} \\
&= \sigma^2\mathbf{d}'\mathbf{Pd},
\end{aligned}$$

and $\text{var}(\mathbf{d}'\mathbf{Y}) = \sigma^2\mathbf{d}'\mathbf{d}$.

Then $\text{var}(\mathbf{d}'\mathbf{Y}) - \text{var}(\mathbf{c}'\hat{\boldsymbol{\theta}}) =$

$$\begin{aligned}
\text{var}(\mathbf{d}'\mathbf{Y}) - \text{var}(\mathbf{c}'\hat{\boldsymbol{\theta}}) &= \sigma^2\mathbf{d}'\mathbf{d} - \sigma^2\mathbf{d}'\mathbf{Pd} \\
&= \sigma^2\mathbf{d}'(\mathbf{I} - \mathbf{P})\mathbf{d} \\
&= \sigma^2\mathbf{d}'(\mathbf{I} - \mathbf{P})^2\mathbf{d} \\
&= \sigma^2[(\mathbf{I} - \mathbf{P})\mathbf{d}]'(\mathbf{I} - \mathbf{P})\mathbf{d} \\
&\geq 0, \text{ establishing minimum variance.}
\end{aligned}$$

Note $\text{var}(\mathbf{d}'\mathbf{Y}) - \text{var}(\mathbf{c}'\hat{\boldsymbol{\theta}}) = 0$ if and only if $(\mathbf{I} - \mathbf{P})\mathbf{d} = \mathbf{0}$, i.e., $\mathbf{d} = \mathbf{Pd} = \mathbf{Pc}$, i.e., $\mathbf{d}'\mathbf{Y} = (\mathbf{Pc})'\mathbf{Y} = \mathbf{c}'\mathbf{PY} = \mathbf{c}'\hat{\boldsymbol{\theta}}$. Establishing uniqueness.

Corollary: If rank$(\mathbf{X}_{n\times p}) = p$, then $\mathbf{a}'\hat{\boldsymbol{\beta}}$ is the BLUE of $\mathbf{a}'\boldsymbol{\beta}$ for any $\mathbf{a}$.

Proof: Note that rank$(\mathbf{X}_{n\times p}) = p$ implies that $\mathbf{X}'\mathbf{X}$ is invertible because rank$(\mathbf{X}'\mathbf{X}) =$ rank$(\mathbf{X})$ (Seber & Lee, A2.4). We have

$$\mathbf{a}'\boldsymbol{\beta} = \mathbf{a}'I\boldsymbol{\beta} = \mathbf{a}'\overbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}^{\text{insert}}\boldsymbol{\beta} = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\overbrace{\boldsymbol{\theta}}^{=\mathrm{X}\boldsymbol{\beta}} = \mathbf{c}'\boldsymbol{\theta}$$

where $\mathbf{c}' = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Also,

$$\mathbf{a}'\hat{\boldsymbol{\beta}} = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\theta}} = \mathbf{c}'\hat{\boldsymbol{\theta}}$$

By the theorem, $\mathbf{c}'\hat{\boldsymbol{\theta}}$ is the BLUE of $\mathbf{c}'\boldsymbol{\theta}$.

Note: The Gauss-Markov theorem generalizes the above result to the less-than-full-rank case. In the less-than-full-rank case, we will have to be more careful. Remember that $\hat{\boldsymbol{\beta}}$ is not unique, so it does not even make sense to say that $\hat{\boldsymbol{\beta}}$ is optimal in any sense.

Estimation of σ^2.

Let $\text{rank}(\mathbf{X}) = r$. Define

$$S^2 = \frac{1}{n-r}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{RSS}{n-r}.$$

This is a generalization of the sample variance.

S^2 is an *unbiased* estimate of σ^2. This is proved by writing

$$(n-r)S^2 = RSS = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y},$$

and applying the general result on expectation of quadratic forms (Lecture 3, p. 6), and using $\text{cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$:

$E[\mathbf{X}'\mathbf{A}\mathbf{X}] = \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

$$\begin{aligned} E[\mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}] &= \text{tr}(\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{P})) + \boldsymbol{\theta}'(\mathbf{I}-\mathbf{P})\boldsymbol{\theta} \\ &= \sigma^2\text{tr}(\mathbf{I}-\mathbf{P}) + \boldsymbol{\theta}'(\mathbf{I}-\mathbf{P})\boldsymbol{\theta} \\ &= \sigma^2(n-r) + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2(n-r) \quad \text{because} \quad (\mathbf{I}-\mathbf{P})\mathbf{X} = \mathbf{0} \end{aligned}$$

Therefore $E[S^2] = \sigma^2$.

Note: S^2 also has a minimum variance optimality property (Seber & Lee Thm 3.4). But the primary interest is in $\boldsymbol{\beta}$ and the estimate of σ^2 is used primarily to determine the standard errors for $\hat{\boldsymbol{\beta}}$.

Distributional Theory:

Normality Assumption: In addition to the assumptions $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, we now also assume that $\boldsymbol{\varepsilon}$ has a multivariate normal distribution, i.e.,

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}).$$

This immediately implies that $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Theorem: (Seber & Lee Thm 3.5).
Let $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\text{rank}(\mathbf{X}_{n\times p}) = p$. Then
(i) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.
(ii) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$.
(iii) $\hat{\boldsymbol{\beta}}$ is independent of S^2.
(iv) $RSS/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$.

Proof: (i) $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is multivariate normal by Seber & Lee Thm 2.2 (Lecture 4, p. 6). The mean and variance were derived previously.

(ii)
$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'[\text{cov}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p$, by a previous theorem (Lecture 6, p. 4).

If $\mathbf{Y} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}$ is p.d., then $(\mathbf{Y}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\theta}) \sim \chi^2_n$.

(iii)

$$\begin{aligned}
\text{cov}(\hat{\boldsymbol{\beta}}, \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= \text{cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, (\mathbf{I} - \mathbf{P})\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{Y}, \mathbf{Y})(\mathbf{I} - \mathbf{P})' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{P})' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{P}) \\
&= \mathbf{0},
\end{aligned}$$

because $\mathbf{X}'(\mathbf{I} - \mathbf{P}) = \mathbf{0}$ (Lemma, p. 2)

(iv)

We will use the previous lemma (Lecture 6, p. 8) that states:

- $\mathbf{Y} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I})$
- $\mathbf{P}_1$, $\mathbf{P}_2$ symmetric matrices
- $Q_1 = (\mathbf{Y} - \boldsymbol{\theta})'\mathbf{P}_1(\mathbf{Y} - \boldsymbol{\theta})/\sigma^2 \sim \chi^2_{r_1}$
- $Q_2 = (\mathbf{Y} - \boldsymbol{\theta})'\mathbf{P}_2(\mathbf{Y} - \boldsymbol{\theta})/\sigma^2 \sim \chi^2_{r_2}$
- $Q_1 - Q_2 \geq 0$

$\Rightarrow Q_1 - Q_2 \sim \chi^2_{r_1 - r_2}.\Box$

For the proof of (iv),

$$Q_1 \equiv ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2/\sigma^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2 \sim \chi^2_n,$$

by a general previous result.

$$Q_2 \equiv ||\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})||^2/\sigma^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p,$$

by part (ii) of this theorem.

$$Q \equiv ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2/\sigma^2 = RSS/\sigma^2.$$

Our goal to prove the distribution of Q is χ^2_{n-p}. First, we show that $Q_1 = Q + Q_2$, i.e. $Q_1 - Q = Q_2$.

Note that:

$$\sigma^2 Q = RSS = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y} = (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{P})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}),$$

because $(\mathbf{I}-\mathbf{P})\mathbf{X} = \mathbf{0}$. So

$$\begin{aligned}
\sigma^2(Q_1 - Q) &= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) - (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{P})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{I}+\mathbf{P})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{P}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{P}'\mathbf{P}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
&= (\hat{\mathbf{Y}}-\mathbf{X}\boldsymbol{\beta})'(\hat{\mathbf{Y}}-\mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) \\
&= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
&= \sigma^2 Q_2
\end{aligned}$$

Therefore, $Q_1 = Q + Q_2$ and $Q = Q_1 - Q_2$. Since Q_1 is χ^2_n and Q_2 is χ^2_p, and $Q = Q_1 - Q_2 = RSS/\sigma^2 \geq 0$, using the previous result (Lecture 6, p. 8), Q is χ^2_{n-p}.