Linear & Generalized Linear Models/ Advanced Regression for Independent Data

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 2: Ordinary least squares (full-rank case)

Version: 07/03/2026

Table of Contents

Setting the stage:

- We're largely assuming in this course that $N > K$.
- In this set of notes, we will also assume that $\mathbf{X}$ is of full rank (full column rank, in particular), so its columns are linearly independent. That is to say that $\text{rank}(\mathbf{X}) = K$.
    - It is possible to deal with the case in which $\mathbf{X}$ is rank-deficient— not a focus of this course, but offered in the form of appendix material that you may read independently.
- For the time being, assume $\mathbf{X}$ is fixed.
    - We will justify this simplification toward the end of this slide set. Later in the course, we'll elaborate on (and address) settings in which this simplification may not be justified.

Setting the stage:

- Let $\mathbf{X} = \begin{pmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_{K-1} \end{pmatrix}$ denote the $N \times K$ design matrix; we often have $\mathbf{x}_0$ denote a vector of ones for an intercept.
- Let $\mathbf{y} \in \mathbb{R}^N$ denote the outcome vector:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}.$$

Linear model:

- Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
- An estimate $\widehat{\boldsymbol{\beta}}$ is a least squares estimate (ordinary least squares, more specifically) if it minimizes the following objective function over all $\boldsymbol{\beta}$:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = \sum_{i=1}^{N}(Y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2.$$

The least squares problem

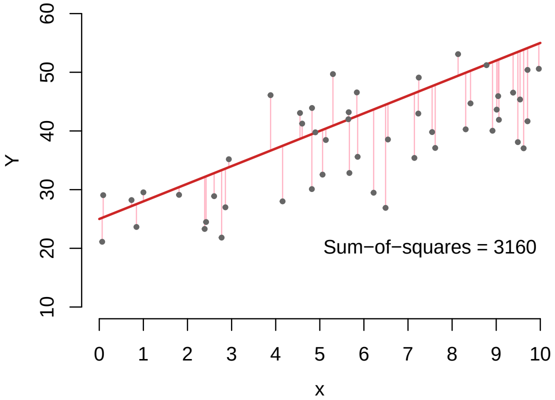

Figure 2.1: Least squares in two dimensions (line "too high").

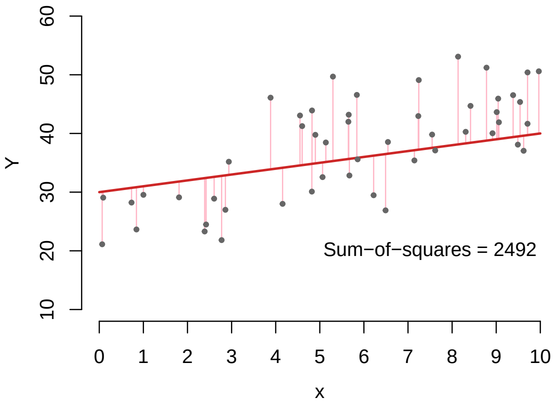

Figure 2.2: Least squares in two dimensions (line "too low").

The least squares problem

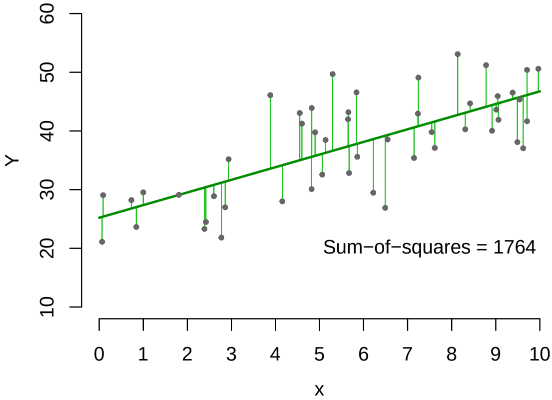

Figure 2.3: Least squares in two dimensions (line "just right").

Linear model:

- Note that $\mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X})$:

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_{K-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{K-1} \end{bmatrix}$$

$$= \beta_0 \mathbf{x}_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_{K-1} \mathbf{x}_{K-1} \in \mathcal{C}(\mathbf{X}).$$

- Therefore, we can find the OLS solution by solving the problem:

$$\underset{\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X})}{\text{minimize}} \; ||\mathbf{y} - \boldsymbol{\theta}||^2.$$

Lemma 2.1: An important decomposition

$\mathbf{y}$ can be decomposed as $\mathbf{y} = \widehat{\mathbf{y}} + \widehat{\boldsymbol{\epsilon}}$, where $\widehat{\mathbf{y}} \in \mathcal{C}(\mathbf{X})$ and $\widehat{\boldsymbol{\epsilon}} \in \mathcal{N}(\mathbf{X}^\intercal)$; this decomposition is unique.

- Note: $\mathcal{N}(\mathbf{X}^\intercal) = [\mathcal{C}(\mathbf{X})]^\perp$ is the left null space of $\mathbf{X}$, also known as the orthogonal complement of $\mathcal{C}(\mathbf{X})$.
- Note: $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the linear subspace spanned by the columns of $\mathbf{X}$.
- Note: $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ is typically referred to as the "fitted" or "predicted" value.

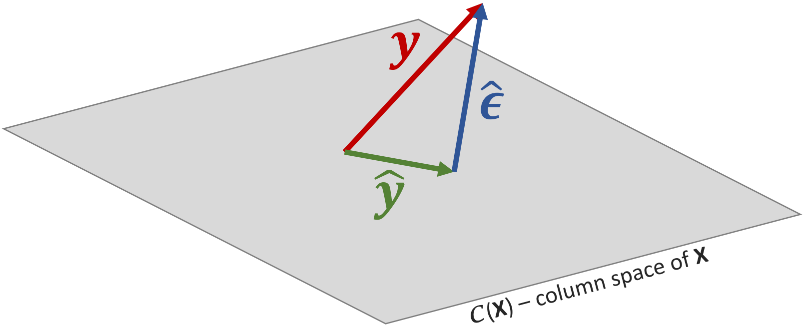

Figure 2.4: Orthogonal projection decomposition.

Lemma 2.1: Proof

- Proof of existence: $\mathcal{C}(\mathbf{X}) \oplus \mathcal{N}(\mathbf{X}^\intercal) = \mathbb{R}^N$.
- Proof of uniqueness: Suppose that $\mathbf{y} = \widehat{\mathbf{y}}_1 + \widehat{\boldsymbol{\epsilon}}_1$ and $\mathbf{y} = \widehat{\mathbf{y}}_2 + \widehat{\boldsymbol{\epsilon}}_2$. Then,

$$\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2 + \widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2 = \mathbf{0}$$

$$\begin{aligned}
\Rightarrow 0 \;&=\; (\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2 + \widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2)^\intercal(\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2 + \widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2) \\
&=\; ||\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2||^2 + ||\widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2||^2 + 2(\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2)^\intercal(\widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2) \\
&=\; ||\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2||^2 + ||\widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2||^2,
\end{aligned}$$

  where the last step follows from the fact that $(\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2) \in \mathcal{C}(\mathbf{X})$ and $(\widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2) \in \mathcal{N}(\mathbf{X}^\intercal)$ (orthogonal complements).
- Therefore, $\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2 = \mathbf{0}$ and $\widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2 = \mathbf{0}$.

Lemma 2.2: $\widehat{\mathbf{y}}$ solves the least squares problem in $\mathcal{C}(\mathbf{X})$

The orthogonal projection of $\mathbf{y}$ onto the linear subspace spanned by the columns of $\mathbf{X}$ solves the OLS minimization problem:

$$\widehat{\mathbf{y}} = \underset{\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X})}{\text{argmin}} \; ||\mathbf{y} - \boldsymbol{\theta}||^2.$$

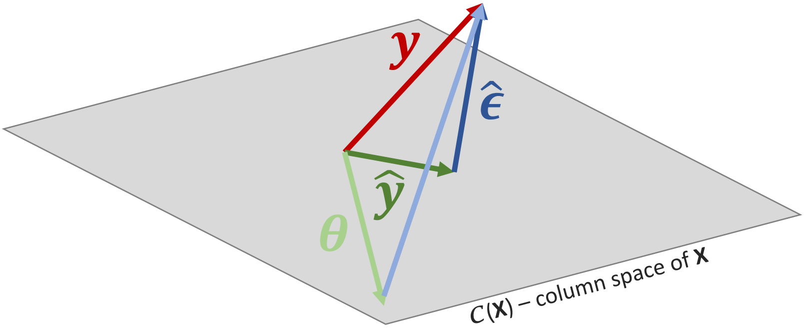

Figure 2.5: Minimizing the distance.

Lemma 2.2: Proof

- For any $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X})$, we have that $(\mathbf{y} - \widehat{\mathbf{y}})^{\intercal}(\widehat{\mathbf{y}} - \boldsymbol{\theta}) = 0$. Therefore, it follows that:

$$\begin{aligned} ||\mathbf{y} - \boldsymbol{\theta}||^2 &= ||\mathbf{y} - \widehat{\mathbf{y}} + \widehat{\mathbf{y}} - \boldsymbol{\theta}||^2 \\ &= ||\mathbf{y} - \widehat{\mathbf{y}}||^2 + ||\widehat{\mathbf{y}} - \boldsymbol{\theta}||^2, \end{aligned}$$

  which is clearly minimized by the choice $\boldsymbol{\theta} = \widehat{\mathbf{y}}$.

Reasoning through the normal equations:

- Since $\widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}} \in \mathcal{N}(\mathbf{X}^\mathsf{T})$, we know that

$$\begin{aligned} \mathbf{X}^\mathsf{T}\widehat{\boldsymbol{\epsilon}} &= \mathbf{X}^\mathsf{T}(\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{0} \\ \Rightarrow \mathbf{X}^\mathsf{T}\mathbf{y} &= \mathbf{X}^\mathsf{T}\widehat{\mathbf{y}}. \end{aligned}$$

- Since $\widehat{\mathbf{y}} \in \mathcal{C}(\mathbf{X})$, we may write $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, so that

$$\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\beta}}.$$

Lemma 2.3: $\widehat{\boldsymbol{\beta}}$ solves the least squares problem in $\mathbb{R}^K$

A least squares estimate of $\boldsymbol{\beta}$, denoted $\widehat{\boldsymbol{\beta}}$, is a solution to the normal equations, so that $\mathbf{X}^\intercal\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\intercal\mathbf{y}$.

- The previous slide effectively proves this. Coming at it from another angle, note that this is a convex optimization problem, and we can solve for $\boldsymbol{\beta}$ directly with matrix calculus:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\text{minimize}}\ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2.$$

Lemma 2.3: Direct proof

- To see this, note that

$$\begin{aligned} ||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||^2 &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^\mathsf{T}\mathbf{y}-\mathbf{y}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta})-(\mathbf{X}\boldsymbol{\beta})^\mathsf{T}\mathbf{y}+(\mathbf{X}\boldsymbol{\beta})^\mathsf{T}\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^\mathsf{T}\mathbf{y}-(\mathbf{y}^\mathsf{T}\mathbf{X})\boldsymbol{\beta}-\boldsymbol{\beta}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{y})+\boldsymbol{\beta}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})\boldsymbol{\beta}. \end{aligned}$$

- Taking a derivative with respect to $\boldsymbol{\beta}$ and setting equal to zero,

$$\mathbf{0}=\mathbf{0}-\mathbf{X}^\mathsf{T}\mathbf{y}-\mathbf{X}^\mathsf{T}\mathbf{y}+2\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}=-2\mathbf{X}^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \quad\Rightarrow\quad \mathbf{X}^\mathsf{T}\mathbf{y}=\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta}.$$

- This is simply a restatement of the normal equations.

Closed-form expression for $\widehat{\boldsymbol{\beta}}$:

- We are currently assuming that $\mathbf{X}$ is of full (column) rank, so that $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^\intercal\mathbf{X}) = K$, and $\mathbf{X}^\intercal\mathbf{X}$ is nonsingular.
- Therefore, the normal equations have a unique solution:

$$\begin{aligned} \mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \mathbf{0} \\ \Rightarrow \mathbf{X}^\intercal\mathbf{y} &= \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} \\ \Rightarrow \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}. \end{aligned}$$

- The orthogonal projection (fitted vector) is given by:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{P}\mathbf{y}.$$

- $\mathbf{P} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ is sometimes called the "hat" matrix because $\mathbf{P}\mathbf{y} = \widehat{\mathbf{y}}$. Indeed, $\mathbf{P}$ is a projection matrix that projects $\mathbf{y}$ onto $\mathcal{C}(\mathbf{X})$.

Lemma 2.4: Some key properties

- Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$, where $\mathbf{X}$ is of full column rank. Then,
  1. $\mathbf{P}$ is a projection matrix onto $\mathcal{C}(\mathbf{X})$.
  2. $\mathbf{I} - \mathbf{P}$ is a projection matrix and onto $\mathcal{N}(\mathbf{X}^\mathsf{T}) = [\mathcal{C}(\mathbf{X})]^\perp$.
  3. $\text{rank}(\mathbf{I} - \mathbf{P}) = \text{tr}(\mathbf{I} - \mathbf{P}) = N - K$.
  4. $\mathbf{P}\mathbf{X} = \mathbf{X}$.

- Some of these we have already proven. The rest can be proven easily.

Definitions:

- The residual vector is given by:

$$\begin{aligned}
\widehat{\boldsymbol{\epsilon}} &= \mathbf{y} - \widehat{\mathbf{y}} \\
&= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{y} - \mathbf{P}\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y}.
\end{aligned}$$

- The residual sum of squares, RSS, is defined as:

$$\begin{aligned}
\text{RSS} &= \widehat{\boldsymbol{\epsilon}}^{\mathsf{T}}\widehat{\boldsymbol{\epsilon}} \\
&= \sum_{i=1}^{N} \widehat{\epsilon}_i^2 \\
&= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\
&= \mathbf{y}^{\mathsf{T}}(\mathbf{I} - \mathbf{P})\mathbf{y}.
\end{aligned}$$

Side note:

- Sometimes you will encounter the least squares equations written as:

$$\sum_{i=1}^{N} \mathbf{x}_i(y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta}) = \mathbf{0}.$$

- Similarly, you will see $\widehat{\boldsymbol{\beta}}$ written as:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\intercal\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i y_i\right).$$

- The summation notation underscores the individual contribution of each independent observation. The distinction matters much more in analysis of correlated data, in which independent *clusters* contribute to the estimating equations.

Table of Contents

Unbiasedness of OLS:

- Assuming $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}$ is unbiased:

$$
\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}] &= \mathrm{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}[\mathbf{y}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{E}[\boldsymbol{\epsilon}] \\
&= \boldsymbol{\beta} + \mathbf{0} \\
&= \boldsymbol{\beta}.
\end{aligned}
$$

Variance of OLS:

- Suppose further that $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}$. Then,

$$
\begin{aligned}
\text{Cov}[\widehat{\boldsymbol{\beta}}] &= \text{Cov}[(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}] \\
&= ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal) \text{Cov}[\mathbf{y}] ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal)^\intercal \\
&= ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal) \text{Cov}[\boldsymbol{\epsilon}] ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal)^\intercal \\
&= ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal) (\sigma^2 \mathbf{I}) ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal)^\intercal \\
&= ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal) (\sigma^2 \mathbf{I}) (\mathbf{X} (\mathbf{X}^\intercal \mathbf{X})^{-\intercal}) \\
&= \sigma^2 ((\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal) (\mathbf{X} (\mathbf{X}^\intercal \mathbf{X})^{-1}) \\
&= \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1} (\mathbf{X}^\intercal \mathbf{X}) (\mathbf{X}^\intercal \mathbf{X})^{-1} \\
&= \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1}.
\end{aligned}
$$

Variance of OLS: Estimation

- We have access to our design matrix, $\mathbf{X}$, but we can't count on knowing the error variance, σ^2. How do we estimate it?
- We have to rely on a previous useful result about the expectation of quadratic forms:

$$\begin{aligned}
\mathrm{E}[\mathrm{RSS}] &= \mathrm{E}[\mathbf{y}^\intercal(\mathbf{I}-\mathbf{P})\mathbf{y}] \\
&= \mathrm{tr}((\mathbf{I}-\mathbf{P})\mathrm{Cov}[\mathbf{y}]) + (\mathrm{E}[\mathbf{y}])^\intercal(\mathbf{I}-\mathbf{P})\mathrm{E}[\mathbf{y}] \\
&= \mathrm{tr}((\mathbf{I}-\mathbf{P})\sigma^2\mathbf{I}) + (\mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta} \\
&= \sigma^2\mathrm{tr}((\mathbf{I}-\mathbf{P})) + 0 \\
&= (N-K)\sigma^2.
\end{aligned}$$

- From this, we deduce that the following estimator is unbiased for σ^2:

$$\widehat{\sigma}^2 = \frac{\mathrm{RSS}}{N-K} = \frac{1}{N-K}\sum_{i=1}^{N}\widehat{\epsilon}_i^2.$$

Independence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\sigma}}^2$:

- If we invoke the additional assumption that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, then it can be shown that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are independent.
- To see this, note that:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} = \mathbf{B}\mathbf{y}, \text{ and} \\
\widehat{\sigma}^2 &= \mathbf{y}^\intercal (\mathbf{I} - \mathbf{P}) \mathbf{y} / (N - K) = \mathbf{y}^\intercal \mathbf{A} \mathbf{y}.
\end{aligned}
$$

- By Theorem 1.4, we will have our desired result if we can verify that $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$. Indeed:

$$
\begin{aligned}
\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} &\propto (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal (\mathbf{I} - \mathbf{P}) \\
&= (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal - (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal (\mathbf{X} (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal) \\
&= (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal - (\mathbf{X}^\intercal \mathbf{X})^{-1} (\mathbf{X}^\intercal \mathbf{X}) (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \\
&= (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal - (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal = \mathbf{0}.
\end{aligned}
$$

Table of Contents

Theorem 2.1: The Gauss-Markov theorem

Suppose that each of the following conditions is satisfied:

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
- $\mathbf{X}$ has full column rank.
- $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$.

Then, $\mathbf{a}^\intercal\widehat{\boldsymbol{\beta}} = \mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ is the unique minimum-variance estimator of $\mathbf{a}^\intercal\boldsymbol{\beta}$ among all unbiased linear estimators of $\mathbf{a}^\intercal\boldsymbol{\beta}$.

- Note: $\mathbf{a}^\intercal\boldsymbol{\beta}$ represents any linear combination of the coefficients of $\boldsymbol{\beta}$ (e.g., β_1, $\beta_3 - \beta_2$, $\beta_1 + 2\beta_3$). Think about means/effects for population strata.

Optimality of OLS

Theorem 2.1: Proof

- Let $\mathbf{d}^\intercal\mathbf{y}$ denote an unbiased linear estimator of $\mathbf{a}^\intercal\boldsymbol{\beta}$.
- By unbiasedness, $\mathrm{E}[\mathbf{d}^\intercal\mathbf{y}] = \mathbf{a}^\intercal\boldsymbol{\beta}$, but because $\mathrm{E}[\mathbf{d}^\intercal\mathbf{y}] = \mathbf{d}^\intercal\mathbf{X}\boldsymbol{\beta}$, we know then that $\mathbf{d}^\intercal\mathbf{X} = \mathbf{a}^\intercal$.
- Further, note that $\mathrm{Var}(\mathbf{d}^\intercal\mathbf{y}) = \mathbf{d}^\intercal\mathrm{Cov}[\mathbf{y}]\mathbf{d} = \mathbf{d}^\intercal\mathrm{Cov}[\boldsymbol{\epsilon}]\mathbf{d} = \sigma^2\mathbf{d}^\intercal\mathbf{d}$, and $\mathrm{Var}(\mathbf{a}^\intercal\widehat{\boldsymbol{\beta}}) = \mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{a}\sigma^2 = \mathbf{d}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{d}\sigma^2$.
- So, $\mathrm{Var}(\mathbf{d}^\intercal\mathbf{y}) - \mathrm{Var}(\mathbf{a}^\intercal\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{d}^\intercal(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\mathbf{d} \geq 0$ because $\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ is positive semi-definite (a projection matrix is symmetric and has eigenvalues of only zeros and ones).
- This proves minimal variance. To prove uniqueness, note that $\mathrm{Var}(\mathbf{d}^\intercal\mathbf{y}) = \mathrm{Var}(\mathbf{a}^\intercal\widehat{\boldsymbol{\beta}}) \Leftrightarrow \mathbf{d}^\intercal(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal) = \mathbf{0}^\intercal$.
  - Put another way, $\mathbf{d}^\intercal = \mathbf{d}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal = \mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$, and hence $\mathbf{d}^\intercal\mathbf{y} = \mathbf{a}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{a}^\intercal\widehat{\boldsymbol{\beta}}$.
- Therefore, $\mathbf{a}^\intercal\widehat{\boldsymbol{\beta}}$ is the *unique* unbiased estimator having the minimum variance property.

Unpacking the meaning of BLUE

- BLUE: **B**est **L**inear **U**nbiased **E**stimator
- More specifically:
  - Best: “Lowest variance.”
  - Linear: “Linear in $\mathbf{y}$.”
  - Unbiased: “Unbiased for $\mathbf{a}^\intercal \boldsymbol{\beta}$.”
- No other linear (in $\mathbf{y}$) and unbiased (for $\mathbf{a}^\intercal \boldsymbol{\beta}$) estimator has smaller variance than $\mathbf{a}^\intercal \widehat{\boldsymbol{\beta}}$.
- Though this section is only three sides, it contains information about one of the most crucial themes of this course—one which we will revisit (repeatedly).

Table of Contents

More on the sampling distribution

Central limit theorem for linear regression:

- Still assuming the case of full rank, recall that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.
- Because $\mathbf{X}$ has an additional row added with each independent observation, we cannot directly apply the classical central limit theorem to determine an asymptotic distribution for $\widehat{\boldsymbol{\beta}}$.
- The Lindeberg-Feller central limit theorem gives us a regularity condition under which $\widehat{\boldsymbol{\beta}}$ achieves asymptotic normality.
- Note that we are still assuming that the model is correctly specified and that $\mathbf{X}$ is fixed.

Theorem 2.2: A central limit theorem for OLS

Suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$, with $N^{-1}\mathbf{X}^\intercal\mathbf{X} \longrightarrow \mathbf{S} \succ 0$. Let $\gamma_N = \max_i ||\mathbf{x}_i||$, where $\mathbf{x}_i$ denotes the (length-K) vector of covariates for subject i. If γ_N is bounded and $N^{-1}\gamma_N^2 \longrightarrow 0$, then $(\mathbf{X}^\intercal\mathbf{X})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

- The technical details of this are beyond the scope of this course, but these conditions can be interpreted to say that no observation can exert undue influence on the sample.

Table of Contents

Semi-parametric model:

- Suppose that instead of $\mathbf{X}$ being fixed by design, N independent observations are drawn from the joint distribution of $(\mathbf{x}, Y)$.
- The semi-parametric formulation of the regression model is given by:

$$\mathrm{E}[Y|\mathbf{x}] = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}.$$

- If the model is correctly specified, then the normal equations are still a perfectly reasonable way to estimate $\boldsymbol{\beta}$.
- To see this, note that:

$$\mathrm{E}_{\mathbf{X},\mathbf{y}}[\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]] = \mathrm{E}_{\mathbf{X}}[\mathrm{E}_{\mathbf{y}|\mathbf{X}}[\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]] = \mathbf{0}.$$

Estimating equations:

- The normal equations that we derived for OLS form what we call *unbiased estimating equations*.
  - In the world of likelihood, the equation $\dot{\ell}(\theta; X) = 0$ is an example of an unbiased estimating equation for θ given that the score function has expectation zero. Estimating equations needn't arise from a likelihood.
- First order of business: develop notation that will help us understand the asymptotic behavior of the OLS estimate even when $\mathbf{X}$ is random.
- For the time being:
  - Let $\widehat{\boldsymbol{\beta}}_N$ denote the solution to the normal equations based on N independent observations.
  - Let $\boldsymbol{\beta}_0$ denote the true, unknown value of the parameter to be estimated.

Estimating equations: Setting up notation

- $\mathbf{G}(\boldsymbol{\beta}; \mathbf{x}, Y) = \mathbf{x}(Y - \mathbf{x}^\intercal \boldsymbol{\beta})$.
  - This is the *estimating function*.
  - Represents contribution of one observation to the estimating equations.
  - In this course, we will always choose $\mathbf{G}$ to be analytic (it will have a Taylor series about $\boldsymbol{\beta}_0$).
- $\mathbb{G}_N(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) = \mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^N \mathbf{G}(\boldsymbol{\beta}; \mathbf{x}_i, Y_i) = \sum_{i=1}^N \mathbf{x}_i(Y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta})$.
  - Note that $\mathbb{G}_N(\boldsymbol{\beta}) = \mathbf{0}$ represents the *estimating equation*.
  - It is a restatement of the normal equations that makes the contribution of each observation more explicit. This will be amazingly useful, particularly later in the course when our solutions to these kinds of equation cannot be written in closed form.

Estimating equations: Setting up notation

- $\mathbf{A}(\boldsymbol{\beta}) = \mathrm{E}\Big[-\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}(\boldsymbol{\theta};\mathbf{x},Y)\big|_{\boldsymbol{\theta}=\boldsymbol{\beta}}\Big]$.
  - The purpose of this has not been made apparent yet.
  - In the case of the normal equations for OLS, it is straightforward to show that $\mathbf{A}(\boldsymbol{\beta}) = \mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]$.
- $\mathbf{B}(\boldsymbol{\beta}) = \mathrm{E}[\mathbf{G}(\boldsymbol{\beta};\mathbf{x},Y)\mathbf{G}(\boldsymbol{\beta};\mathbf{x},Y)^\intercal]$.
  - The purpose of this has not been made apparent yet.
  - In the case of the normal equations for OLS, it is straightforward to show that under homoscedasticity, $\mathbf{B}(\boldsymbol{\beta}) = \sigma^2\mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]$.
- Later in the course, we won't be so lucky as to have such beautiful expressions pop out!
  - ...But don't worry – they still won't be too bad! :)
- Does the similarity of $\mathbf{A}(\boldsymbol{\beta})$ and $\mathbf{B}(\boldsymbol{\beta})$ remind you of anything you might have learned in a previous course? It's no accident!

Estimating equations: Invoking a Taylor expansion

- Because $\widehat{\boldsymbol{\beta}}_N$ solves the estimating equations, it follows that:

$$\mathbf{0} = \frac{1}{N}\mathbb{G}_N(\widehat{\boldsymbol{\beta}}_N; \mathbf{X}, \mathbf{y}).$$

- Because $\mathbf{G}$ is analytic, we can push this further under suitable regularity conditions by doing a Taylor approximation about $\boldsymbol{\beta}_0$:

$$\begin{aligned}
\mathbf{0} &\approx \frac{1}{N}\mathbb{G}_N(\boldsymbol{\beta}_0; \mathbf{X}, \mathbf{y}) + \frac{\partial}{\partial \boldsymbol{\beta}}\left[\frac{1}{N}\mathbb{G}_N(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y})\right]\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{G}(\boldsymbol{\beta}_0; \mathbf{x}_i, Y_i) + \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{G}(\boldsymbol{\beta}; \mathbf{x}_i, Y_i)\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right](\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0).
\end{aligned}$$

Estimating equations: Rearrangement

- Assume that $\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{G}(\boldsymbol{\beta}_0; \mathbf{x}_i, Y_i)$ is invertible.
- Rearranging the equation on the prior slide (and leaving the details surrounding the regularity conditions on the remainder term of the Taylor expansion to a more theory-oriented course):

$$(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \approx \left[-\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{G}(\boldsymbol{\beta}; \mathbf{x}_i, Y_i) \bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right]^{-1} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{G}(\boldsymbol{\beta}_0; \mathbf{x}_i, Y_i) \right].$$

Estimating equations: Invoking asymptotics

- Multiplying both sides by $\sqrt{N}$, we then have:

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \approx \left[-\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{G}(\boldsymbol{\beta}; \mathbf{x}_i, Y_i) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right]^{-1} \left[\frac{\sqrt{N}}{N} \sum_{i=1}^{N} \mathbf{G}(\boldsymbol{\beta}_0; \mathbf{x}_i, Y_i) \right].$$

- By the weak law of large numbers,

$$\left[-\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{G}(\boldsymbol{\beta}; \mathbf{x}_i, Y_i) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right] \xrightarrow{p} \mathbf{A}(\boldsymbol{\beta}_0).$$

- By the central limit theorem,

$$\frac{\sqrt{N}}{N} \sum_{i=1}^{N} \mathbf{G}(\boldsymbol{\beta}_0; \mathbf{x}_i, Y_i) = \left[\sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{G}(\boldsymbol{\beta}_0; \mathbf{x}_i, Y_i) - \mathbf{0} \right) \right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{B}(\boldsymbol{\beta}_0)).$$

Estimating equations: Invoking more asymptotics

- By Slutsky's theorem, it then follows that

$$\widehat{\boldsymbol{\beta}}_N \dot{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{N}[\mathbf{A}(\boldsymbol{\beta}_0)]^{-1}\mathbf{B}(\beta_0)[\mathbf{A}(\boldsymbol{\beta}_0)]^{-\intercal}\right).$$

- Invoking what we found previously for the normal equations,

$$\begin{aligned}\frac{1}{N}[\mathbf{A}(\boldsymbol{\beta}_0)]^{-1}\mathbf{B}(\beta_0)[\mathbf{A}(\boldsymbol{\beta}_0)]^{-\intercal} &= \frac{1}{N}[\mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]]^{-1}\left(\sigma^2[\mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]]\right)[\mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]]^{-\intercal} \\ &= \frac{1}{N}\sigma^2[\mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]]^{-1}.\end{aligned}$$

- Note: $N^{-1}\mathbf{X}^\intercal\mathbf{X} = N^{-1}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^\intercal \xrightarrow{p} \mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]$, lending itself to the following estimator:

$$\widehat{\mathrm{Cov}}[\widehat{\boldsymbol{\beta}}_N] = \widehat{\sigma}^2(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

Unbiasedness:

- When $\mathbf{X}$ is random, the closed-form expression for $\widehat{\boldsymbol{\beta}}$ can be used to justify unbiasedness without the estimating equations approach.
- Assuming $\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$ is unbiased:

$$
\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}] &= \mathrm{E}_{\mathbf{X}}[\mathrm{E}_{\mathbf{y}|\mathbf{X}}[\widehat{\boldsymbol{\beta}}]] \\
&= \mathrm{E}_{\mathbf{X}}[\mathrm{E}_{\mathbf{y}|\mathbf{X}}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}]] \\
&= \mathrm{E}_{\mathbf{X}}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathrm{E}_{\mathbf{y}|\mathbf{X}}[\mathbf{y}]] \\
&= \mathrm{E}_{\mathbf{X}}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\boldsymbol{\beta})] \\
&= \mathrm{E}_{\mathbf{X}}[(\mathbf{X}^\intercal\mathbf{X})^{-1}(\mathbf{X}^\intercal\mathbf{X})\boldsymbol{\beta}] \\
&= \mathrm{E}_{\mathbf{X}}[\boldsymbol{\beta}] = \boldsymbol{\beta}.
\end{aligned}
$$

Variance of $\widehat{\boldsymbol{\beta}}$ under homoscedasticity:

- When $\mathbf{X}$ is random, the closed-form expression for $\widehat{\boldsymbol{\beta}}$ can be used to approximate the variance without the estimating equations approach.

$$
\begin{aligned}
\mathrm{Cov}[\widehat{\boldsymbol{\beta}}] &= \mathrm{Cov}[\mathrm{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \mathrm{E}[\mathrm{Cov}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \mathrm{Cov}[\boldsymbol{\beta}] + \mathrm{E}[\mathrm{Cov}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \mathbf{0} + \mathrm{E}[\mathrm{Cov}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \mathrm{E}[\mathrm{Cov}[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}|\mathbf{X}]] \\
&= \vdots \\
&= \mathrm{E}[\sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}] \approx \sigma^2\mathrm{E}[\mathbf{X}^\intercal\mathbf{X}]^{-1} \approx \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}.
\end{aligned}
$$

- To see this, note that $\mathrm{E}[\mathbf{X}^\intercal\mathbf{X}] = N \cdot \mathrm{E}[\mathbf{x}\mathbf{x}^\intercal]$ so that $\mathbf{X}^\intercal\mathbf{X} \approx \mathrm{E}[\mathbf{X}^\intercal\mathbf{X}]$ for large N, and use a first-order Taylor approximation (details omitted).
- This derivation relies on the fact that $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ is positive definite, a reality without which this approximation might not hold.

Table of Contents

Recall: Assumptions about the nature of $\mathbf{X}$

- The design matrix, $\mathbf{X}$, is considered to be fixed when it would not vary across study replicates.
  - Example: two-arm (1:1) clinical trial of $N = 500$ individuals with blocked randomization, I know in advance that $N_0 = N_1 = 250$.
  - Example: observational study in which I sample $N_0 = 100$ smokers and $N_1 = 100$ non-smokers.
  - There is no randomness in the structure of the design matrix. Data form an empirical estimate of the *conditional* distribution of $f(Y|X)$.
- The design matrix, $\mathbf{X}$, is considered to be random when its exact value would vary across study replicates.
  - Example: two-arm (1:1) clinical trial of $N = 500$ individuals with pure randomization (e.g., coin flip), N_0 and N_1 are not known in advance (they are random variables).
  - Example: observational study with samples $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_N, Y_N)$
  - Data are an empirical estimate of the *joint* distribution $f(\mathbf{x}, Y)$.

Recall: Assumptions about the nature of $\mathbf{X}$

- In much of our discussion so far, it has been convenient to presume the design matrix, $\mathbf{X}$, to be fixed in order to develop the theory.
- However, we made a couple of technically sophisticated arguments that the conclusions from the fixed-$\mathbf{X}$ theory wouldn't be wildly different than what happens in the random-$\mathbf{X}$ case.
- I'd like to highlight a point of nuance regarding the implications of mean model misspecification.

Misspecification of the mean model

Recall: Assumptions about the nature of $\mathbf{X}$

- As a reminder (when the specifications of the model are fully correct):

$$
\begin{aligned}
\text{Cov}[\widehat{\boldsymbol{\beta}}] &= \text{Cov}[\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \text{E}[\text{Cov}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \vdots \\
&= \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1} \text{ if } \mathbf{X} \text{ is fixed, or} \\
&\approx \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1} \text{ if } \mathbf{X} \text{ is random.}
\end{aligned}
$$

- In the "fixed" case: $\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ is *always* constant—so $\text{Cov}[\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] = \mathbf{0}$, regardless of whether the model is correctly specified.
- In the "random" case: constancy of $\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ can only be met if the mean model is correctly specified; otherwise, $\text{Cov}[\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \succ \mathbf{0}$.
- Natural question: How does the distinction between "fixed" and "random" play out when the mean model is *not* correct?

Simulation:

- Consider two study design scenarios:
  1. $X_i \sim \mathcal{N}(\mu = 0, \sigma^2 = 1),\ i = 1, \ldots, 4n \qquad$ (random).
  2. $X_i = \begin{cases} -\sqrt{2} & \text{for } i = 1, \ldots, n \\ 0 & \text{for } i = n+1, \ldots, 3n \\ \sqrt{2} & \text{for } i = 3n+1, \ldots, 4n \end{cases} \qquad$ (fixed).
- Consider two outcome generation scenarios:
  1. $Y_i = X_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad$ (correct).
  2. $Y_i = X_i^2 + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad$ (incorrect).

Simulation: What will happen?

- Because X has variance one in both study design scenarios, and the error variance is shared between the outcome generating scenarios, comparing the four overall scenarios is "apples-to-apples."
- Also note that $\beta_1 = 1$ in both scenarios, although this is less important to the matter at hand.
- We anticipate that the variance of $\widehat{\beta}_1$ from OLS will be *higher* than the others in a specific combination of study design/outcome scenarios. Which combination, specifically?

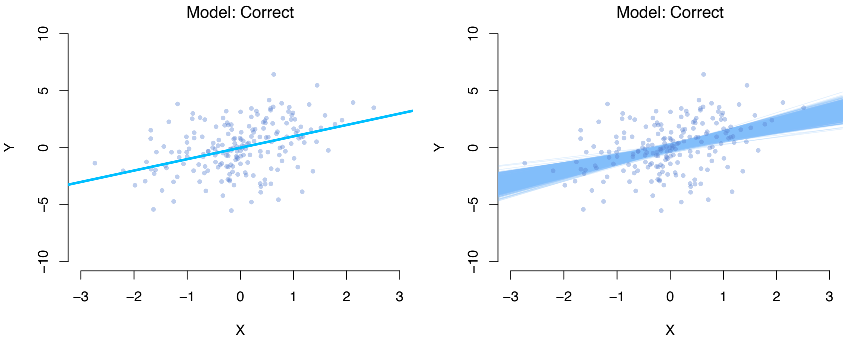

Figure 2.6: Random design/correct model ($SE[\widehat{\beta}_1] = 0.14$).

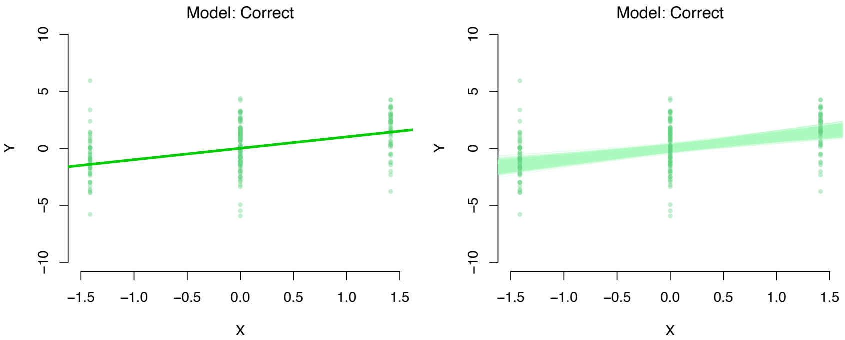

Figure 2.7: Fixed design/correct model ($\mathrm{SE}[\widehat{\beta}_1] = 0.14$).

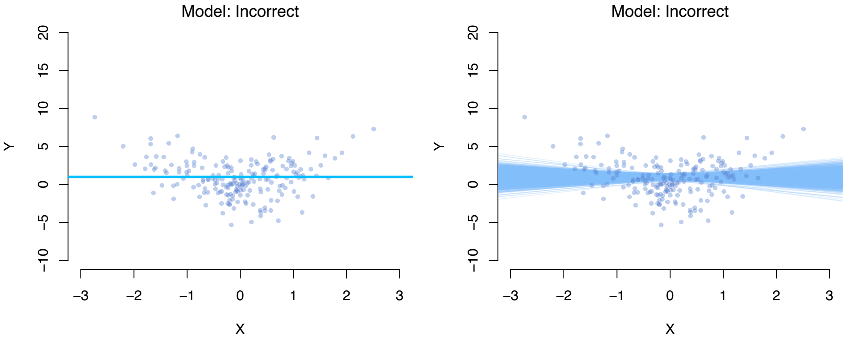

Figure 2.8: Random design/incorrect model ($\text{SE}[\widehat{\beta}_1] = 0.26$).

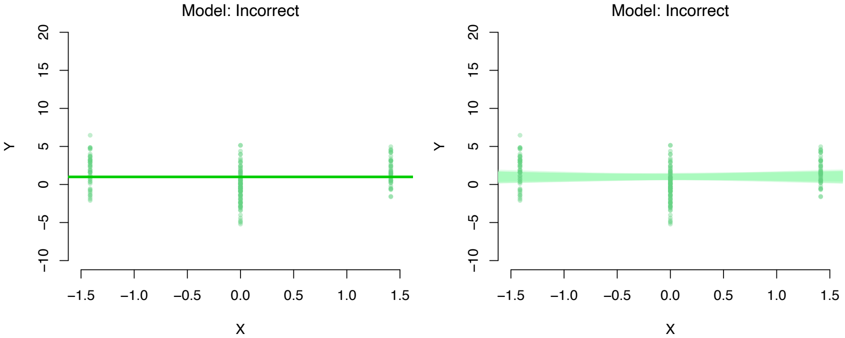

Figure 2.9: Fixed design/incorrect model (SE$[\widehat{\beta}_1] = 0.14$).

Simulation: Why did it happen?

- Scenario 3 was clearly not like the others.
- When the mean model is not correctly specified, the quantity $\mathrm{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ depends upon the value of $\mathbf{X}$.
- Because $\mathbf{X}$ is random, this introduces a source of variation that would not be present if $\mathbf{X}$ were fixed.
- Over *all* samples of $\mathbf{X}$, it is clear that $\mathrm{E}[\widehat{\beta}_1] = 0$. But because the model is not correct, $\mathrm{E}[\widehat{\beta}_1|\mathbf{X}]$ is not zero for all $\mathbf{X}$.
- Consider hypothetical simulation replicates in which the values of $\mathbf{X}$ "tend to be lower" by chance. In such replicates, we will tend to see that $\mathrm{E}[\widehat{\beta}_1|\mathbf{X}] < 0$. On the other hand (and symmetrically), simulation replicates in which the values of $\mathbf{X}$ "tend to be higher," will tend to be such that $\mathrm{E}[\widehat{\beta}_1|\mathbf{X}] > 0$.

Further thoughts:

- Even though Scenario 3 looks like the "problem case," that distinct honor is actually held by Scenario 4 (for reasons we have not yet discussed, but will later in the semester).
- In a nutshell, we will need a special method to properly *estimate* the variance in this scenario (stick around—we'll deal with this sometime in October).

Summary

This unit:

- Ordinary least squares.
- Bias and variance.
- The Gauss-Markov theorem.
- Nuances regarding misspecification of the mean model and its relation to fixed/random design.

Summary: So far

- Random vectors and matrices; multivariate normal theory.
- Ordinary least squares.

Summary: Coming up

- Hypothesis testing and ANOVA.
- Weighted least squares.
- Confidence regions and prediction.
- Diagnostics.
- Exponential families.
- Generalized linear models.
- Sandwich and bootstrap.
- Quasi-likelihood.
- Hypothesis testing for GLMs.
- Diagnostics for GLMs.
- Further considerations for binary outcomes.
- Nonlinear least squares.