BIOS 7345: Advanced Regression Analysis I

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 3: Ordinary least squares (full-rank case)

Version: 09/01/2023

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 $\text{rank}(\mathbf{X}) = K$.
  - We will deal with the case in which $\mathbf{X}$ is "rank-deficient" in the following set of notes.
- For the time being, assume $\mathbf{X}$ is fixed.
  - In the last section of this set, we'll justify why this simplification is usually okay.

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:

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

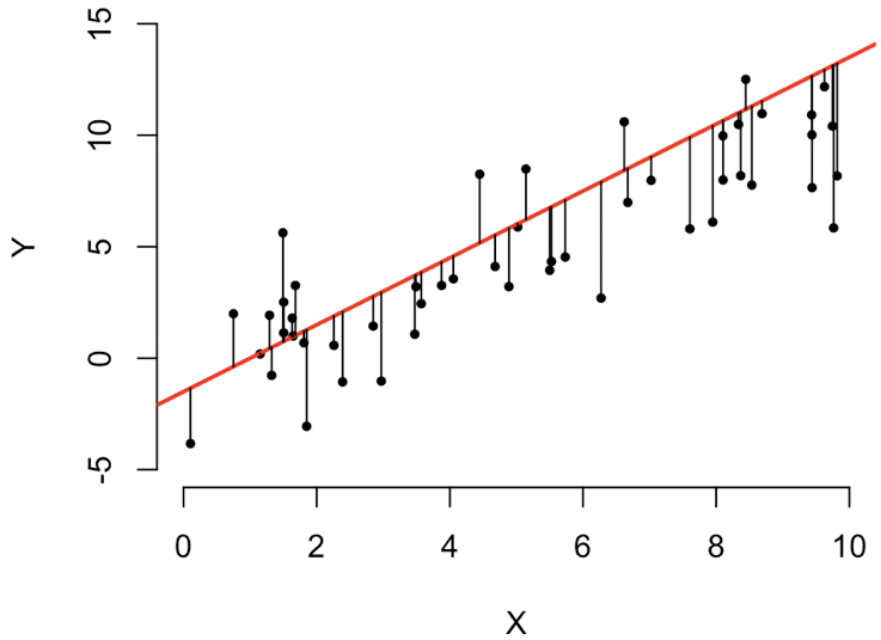
The least squares problem

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 $\boldsymbol{\beta}$:

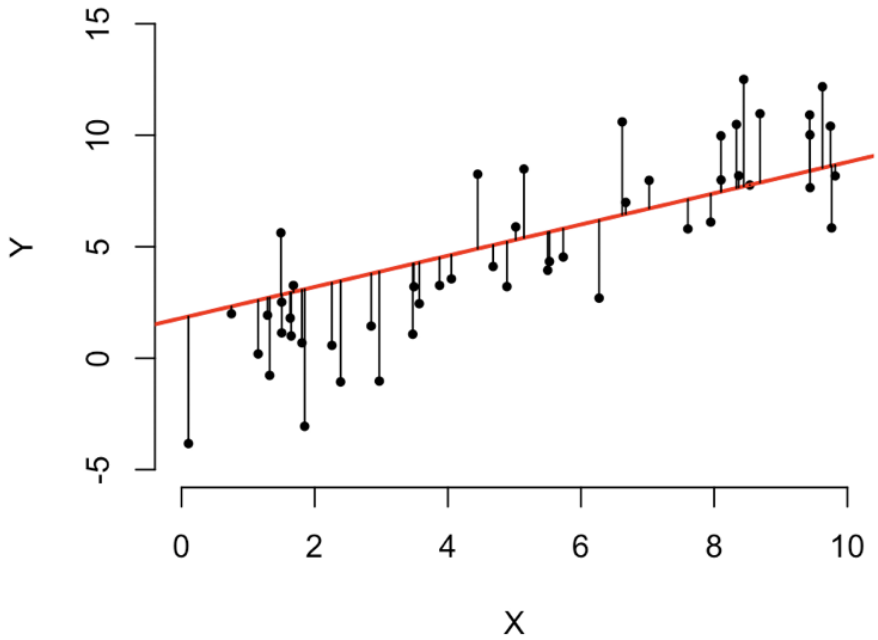
$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2.$$

Least squares in the two-dimensional case:



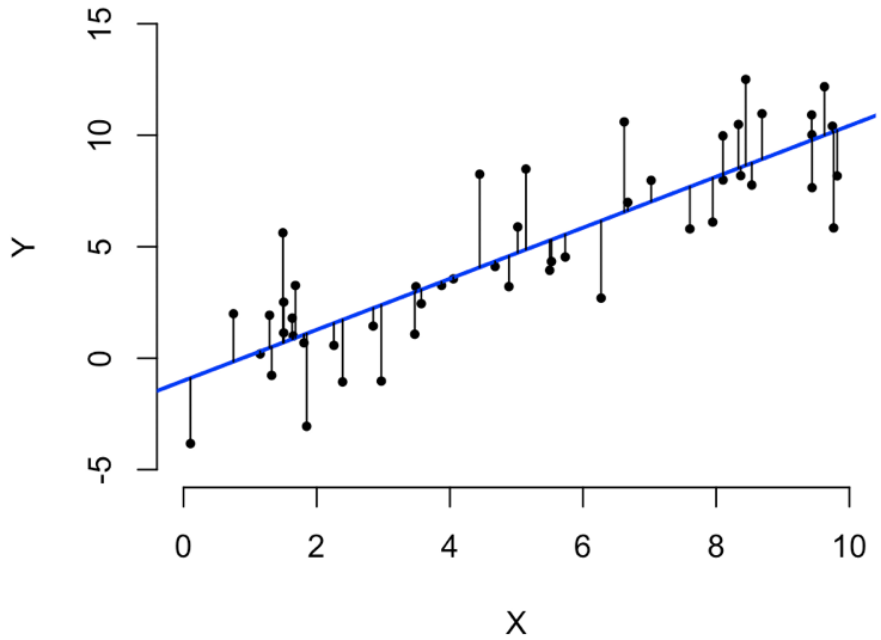
Sum of squared distances: 375.33 (Too high!)

Least squares in the two-dimensional case:



Sum of squared distances: 334.39 (Closer!)

Least squares in the two-dimensional case:



Sum of squared distances: 228.62 (Just right!)

Linear model:

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

$$\begin{aligned}\mathbf{X}\boldsymbol{\beta} &= \begin{bmatrix}\mathbf{x}_0 & \mathbf{x}_1 & \cdots & \mathbf{x}_{K-1}\end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \vdots \\ \beta_{K-1}\end{bmatrix} \\ &= \beta_0\mathbf{x}_0 + \beta_1\mathbf{x}_1 + \cdots + \beta_{K-1}\mathbf{x}_{K-1} \in \mathcal{C}(\mathbf{X}).\end{aligned}$$

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

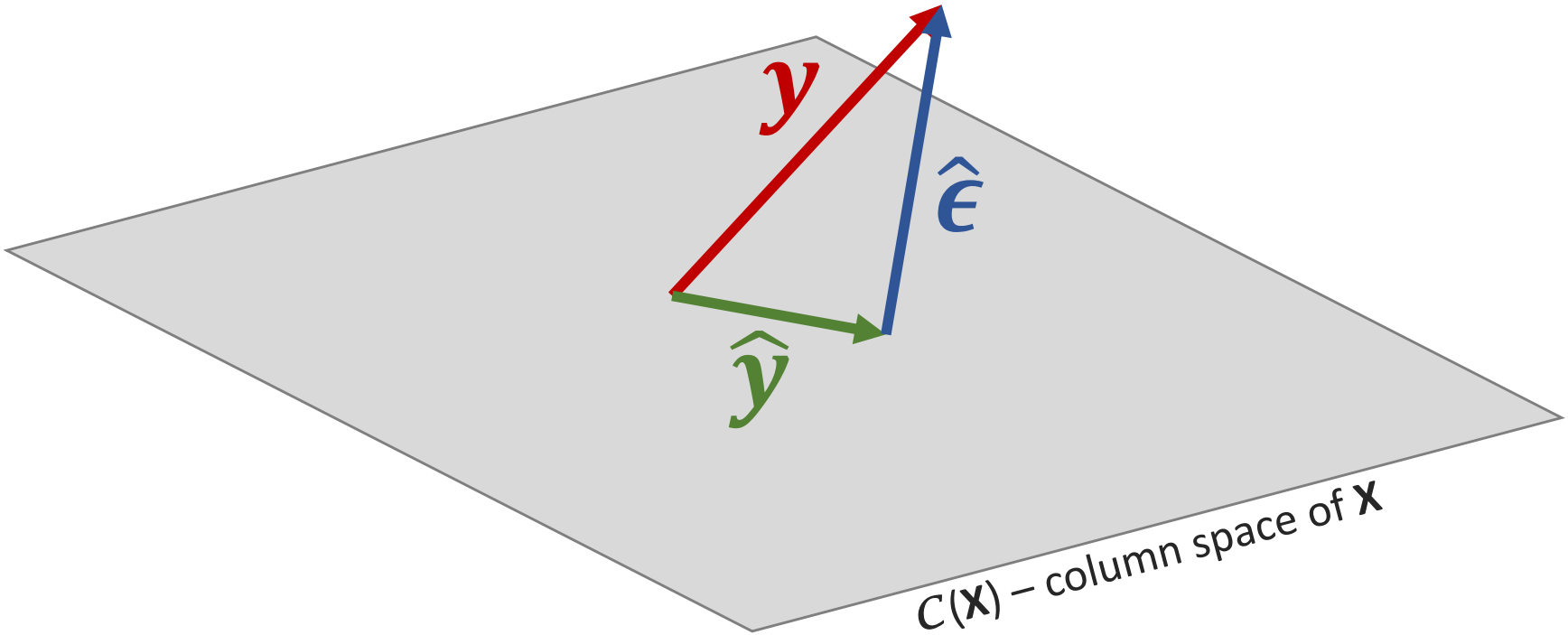
$$\underset{\boldsymbol{\theta}\in\mathcal{C}(\mathbf{X})}{\text{minimize}} \; ||\mathbf{y} - \boldsymbol{\theta}||^2.$$

Lemma 3.1: 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 $\widehat{\boldsymbol{\epsilon}} \in \mathcal{N}(\mathbf{X}^T)$; this decomposition is unique.

- Note: $\mathcal{N}(\mathbf{X}^T) = [\mathcal{C}(\mathbf{X})]^{\perp}$ 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.

The least squares problem



Lemma 3.1: Proof

- Proof of existence: $\mathcal{C}(\mathbf{X}) \oplus \mathcal{N}(\mathbf{X}^T) = \mathbb{R}^N$.
- 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,

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

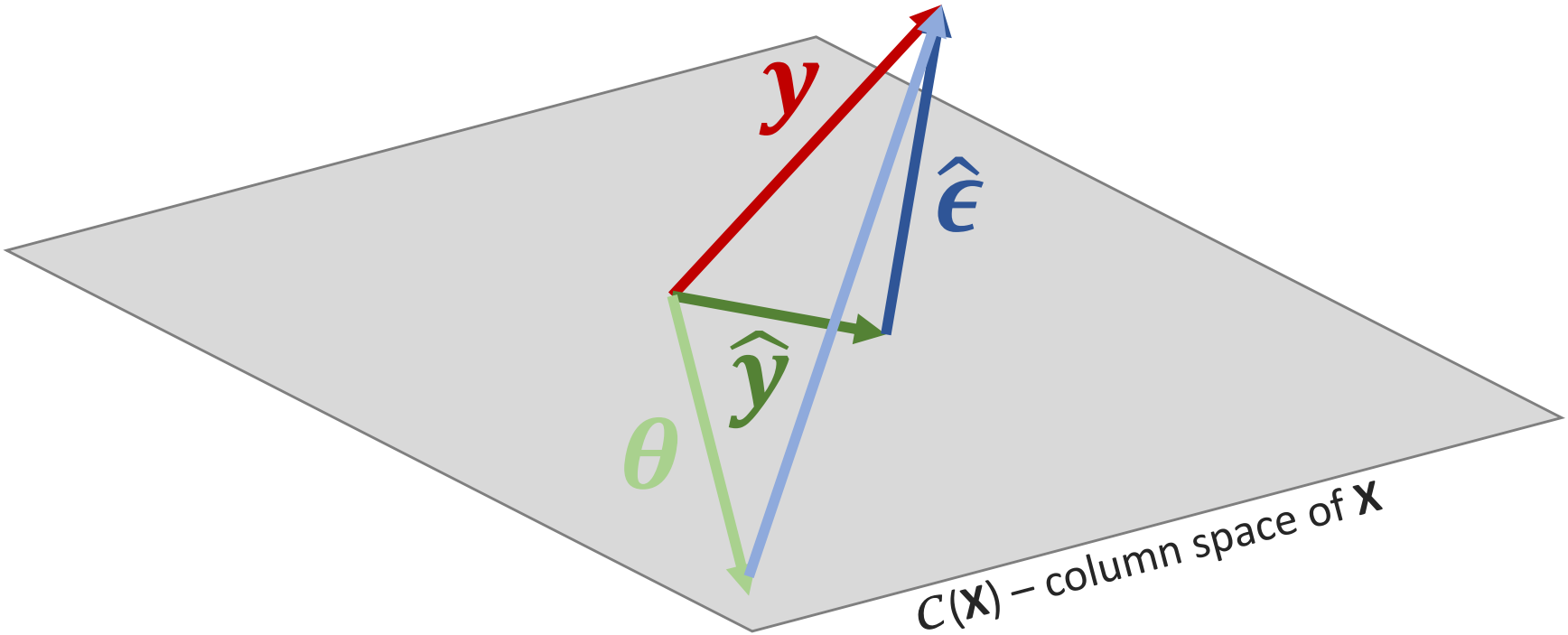
where the last step follows from the fact that $(\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2) \in \mathcal{C}(\mathbf{X})$ and $(\widehat{\boldsymbol{\epsilon}}_1 - \widehat{\boldsymbol{\epsilon}}_2) \in \mathcal{N}(\mathbf{X}^T)$. Therefore, $\widehat{\mathbf{y}}_1 - \widehat{\mathbf{y}}_2 = \mathbf{0}$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

Lemma 3.2: $\widehat{\mathbf{y}}$ solves the least squares problem in $\mathcal{C}(\mathbf{X})$

The orthogonal projection of $\mathbf{y}$ onto the linear subspace spanned by the column of $\mathbf{X}$ solves the OLS minimization problem. In other words:

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

The least squares problem



Lemma 3.2: Proof

- For any $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X})$, we have that $(\mathbf{y} - \widehat{\mathbf{y}})^T(\widehat{\mathbf{y}} - \boldsymbol{\theta}) = 0$. 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 $\widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}} \in \mathcal{N}(\mathbf{X}^T)$, we know that

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

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

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

Lemma 3.3: $\widehat{\boldsymbol{\beta}}$ 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 $\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$.

- 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 3.3: Direct proof

- To see this, note that

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

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

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

- 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 $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^T\mathbf{X}) = K$, and $\mathbf{X}^T\mathbf{X}$ is nonsingular.
- Therefore, the normal equations have a unique solution:

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

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

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

- $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ 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 3.4: Some key properties

- Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, 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 $\mathcal{N}(\mathbf{X}^T) = [\mathcal{C}(\mathbf{X})]^{\perp}$.
  3. $\text{rank}(\mathbf{I} - \mathbf{P}) = \text{trace}(\mathbf{I} - \mathbf{P}) = N - K$.
  4. $\mathbf{PX} = \mathbf{X}$.

- 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:

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

Side note:

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

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

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

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

- The summation notation underscores the individual contributions of each independent observation. The distinction matters much more in analysis of correlated data, in which independent *clusters* contribute to the estimating equations (BIOS 7346).

Table of Contents

Bias and variance of the OLS estimator

Unbiasedness of OLS:

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

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

Bias and variance of the OLS estimator

Variance of OLS:

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

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

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:

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

- 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 $\widehat{\sigma}^2$:

- 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:

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

- By Theorem 2.4, we will have our desired result if we can verify that $\mathbf{B\Sigma A} = \mathbf{0}$. Indeed:

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

Table of Contents

Optimality of OLS

Theorem 3.1: 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, $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}} = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the unique minimum-variance estimator of $\mathbf{a}^T\boldsymbol{\beta}$ among all unbiased linear estimators of $\boldsymbol{a}^T\boldsymbol{\beta}$.

- Note: $\boldsymbol{a}^T\boldsymbol{\beta}$ represents any linear combination of the coefficients of $\boldsymbol{\beta}$ (e.g., β_1, $\beta_3 - \beta_2$, $\beta_1 + 2\beta_3$).

Optimality of OLS

Gauss-Markov theorem: Proof

- Let $\mathbf{d}^T\mathbf{y}$ denote an unbiased linear estimator of $\mathbf{a}^T\boldsymbol{\beta}$.
- By unbiasedness, $\mathrm{E}[\mathbf{d}^T\mathbf{y}] = \mathbf{a}^T\boldsymbol{\beta}$, but because $\mathrm{E}[\mathbf{d}^T\mathbf{y}] = \mathbf{d}^T\mathbf{X}\boldsymbol{\beta}$, we know then that $\mathbf{d}^T\mathbf{X} = \mathbf{a}^T$.
- Further, note that $\mathrm{Var}(\mathbf{d}^T\mathbf{y}) = \mathbf{d}^T\mathrm{Cov}[\mathbf{y}]\mathbf{d} = \mathbf{d}^T\mathrm{Cov}[\boldsymbol{\epsilon}]\mathbf{d} = \sigma^2\mathbf{d}^T\mathbf{d}$, and $\mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\beta}}) = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\sigma^2 = \mathbf{d}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}\sigma^2$.
- So, $\mathrm{Var}(\mathbf{d}^T\mathbf{y}) - \mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{d}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{d} \geq 0$ because $\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ 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 $\mathrm{Var}(\mathbf{d}^T\mathbf{y}) = \mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\beta}}) \Leftrightarrow \mathbf{d}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \mathbf{0}^T$.
  - Put another way, $\mathbf{d}^T = \mathbf{d}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, and hence $\mathbf{d}^T\mathbf{y} = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{a}^T\widehat{\boldsymbol{\beta}}$.
- Therefore, $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$ is the *unique* unbiased estimator having the minimum variance property.

Optimality of OLS

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 $\mathbf{a}^T\boldsymbol{\beta}$."
- No other linear (in $\mathbf{y}$) and unbiased (for $\mathbf{a}^T\boldsymbol{\beta}$) estimator has smaller variance than $\mathbf{a}^T\widehat{\boldsymbol{\beta}}$.
- Recall the bias-variance trade-off. Is it really necessary only to consider linear unbiased estimators? There may be reasons to introduce a small amount of bias for the purpose of greatly reducing the variance.

Table of Contents

More on the sampling distribution

Central limit theorem for linear regression:

- Still assuming the case of full rank, recall that $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- 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 3.2: A central limit theorem for OLS

Let $\mathbf{A} := (\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{X}^T$ (which is $K \times N$), with columns given by $\mathbf{a}_{N1}, \ldots, \mathbf{a}_{NN}$. Let $\mathbf{X}$ be fixed and of full rank, and suppose the following conditions (Lindeberg conditions) hold:

1. $\sum_{i=1}^{N} ||\mathbf{a}_{Ni}||^k \longrightarrow 0$ for $k > 2$.
2. $\mathrm{E}[|\epsilon|^k] < \infty$ for $k > 2$.

Then, $(\mathbf{X}^T\mathbf{X})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \longrightarrow_d \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

- The technical details of this are beyond the scope of this course, but these conditions can be heuristically 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:

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

- 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:

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

Random design matrices

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

- $\mathbf{G}(\boldsymbol{\beta}; \mathbf{x}, Y) = \mathbf{x}(Y - \mathbf{x}^T\boldsymbol{\beta})$
  - 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 $\boldsymbol{\beta}_0$).
- $\mathbb{G}_N(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^N \mathbf{G}(\boldsymbol{\beta}; \mathbf{x}_i, Y_i) = \sum_{i=1}^N \mathbf{x}_i(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})$
  - 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

- $\mathbf{A}(\boldsymbol{\beta}) = \mathrm{E}\left[-\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{G}(\boldsymbol{\theta}; \mathbf{x}, Y)\big|_{\boldsymbol{\theta}=\boldsymbol{\beta}}\right]$
  - 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 $\mathbf{A}(\boldsymbol{\beta}) = \mathrm{E}[\mathbf{x}\mathbf{x}^T]$.
- $\mathbf{B}(\boldsymbol{\beta}) = \mathrm{E}[\mathbf{G}(\boldsymbol{\beta}; \mathbf{x}, Y)\mathbf{G}(\boldsymbol{\beta}; \mathbf{x}, Y)^T]$
  - 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, $\mathbf{B}(\boldsymbol{\beta}) = \sigma^2\mathrm{E}[\mathbf{x}\mathbf{x}^T]$.
- Later in the course, we won't be so lucky as to have such beautiful expressions pop out! :)

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):

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

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,

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

- By the central limit theorem,

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

Estimating equations: Invoking more asymptotics

- By Slutsky's theorem, it then follows that

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

- Invoking what we found previously for the normal equations,

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

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

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

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:

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

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.

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

- This derivation relies on the fact that $(\mathbf{X}^T\mathbf{X})^{-1}$ is positive definite, a reality without which this approximation might not hold.

Summary

So far:

- Ordinary least squares in the case of a full-rank design matrix.

Summary

Up next:

- Ordinary least squares in the case of a rank-deficient design matrix.