BIOS 7345: Advanced Regression Analysis I

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 8: Model misspecification

Version: 09/17/2023

Table of Contents

Fixed vs. random design matrices

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 $f(\mathbf{x}, Y)$.

Fixed vs. random design matrices

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.

Fixed vs. random design matrices

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

- As a reminder:

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

- 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, $\text{Cov}[\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \succ \mathbf{0}$.

Fixed vs. random design matrices

Simulation:

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

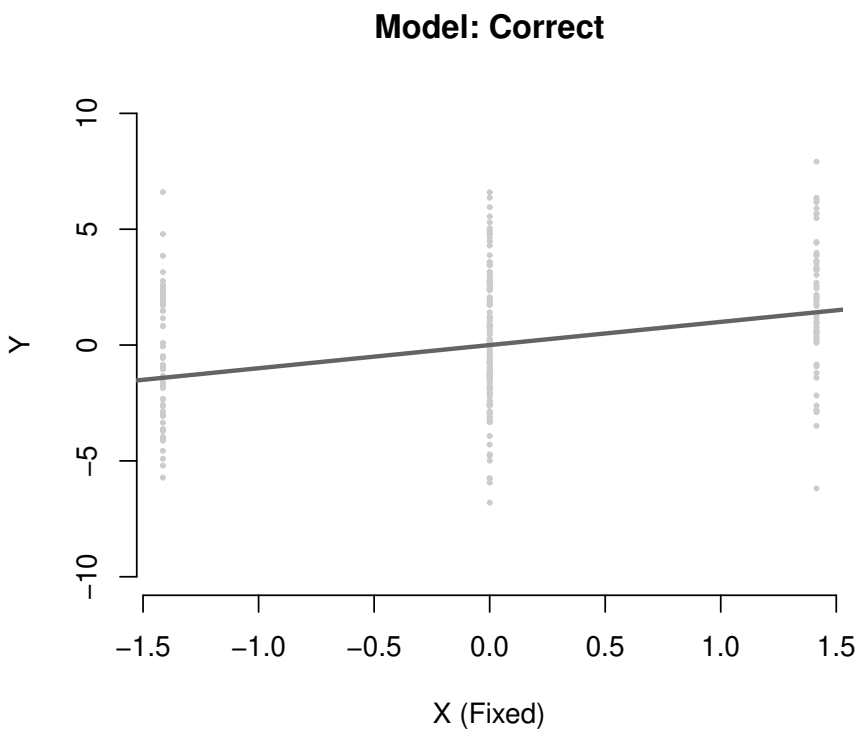
- Consider two outcome generation scenarios:
  1. $Y_i = X_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (correct).
  2. $Y_i = X_i^2 + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (incorrect).

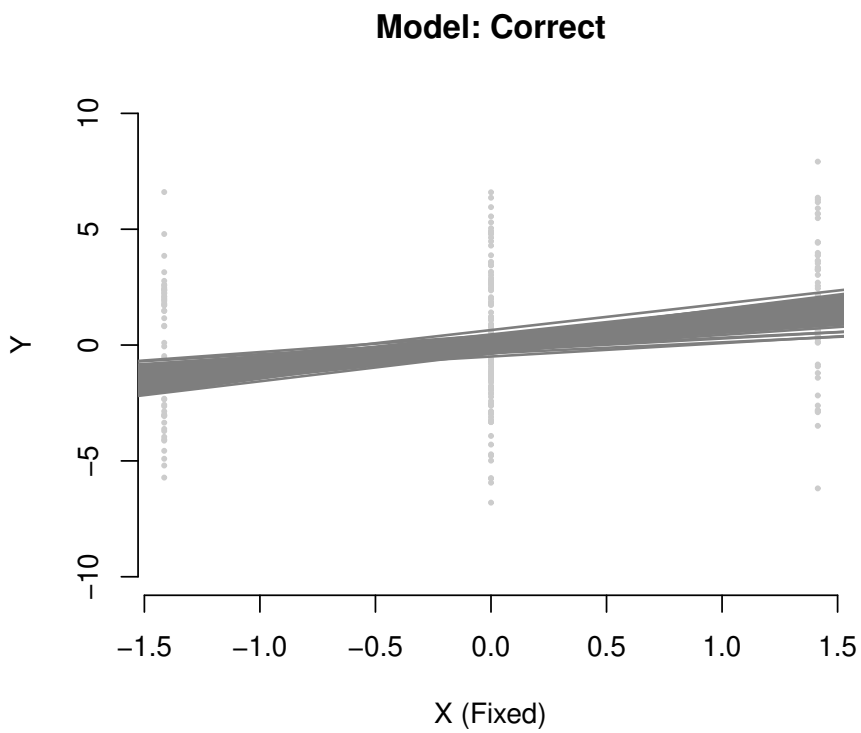
Fixed vs. random design matrices

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?

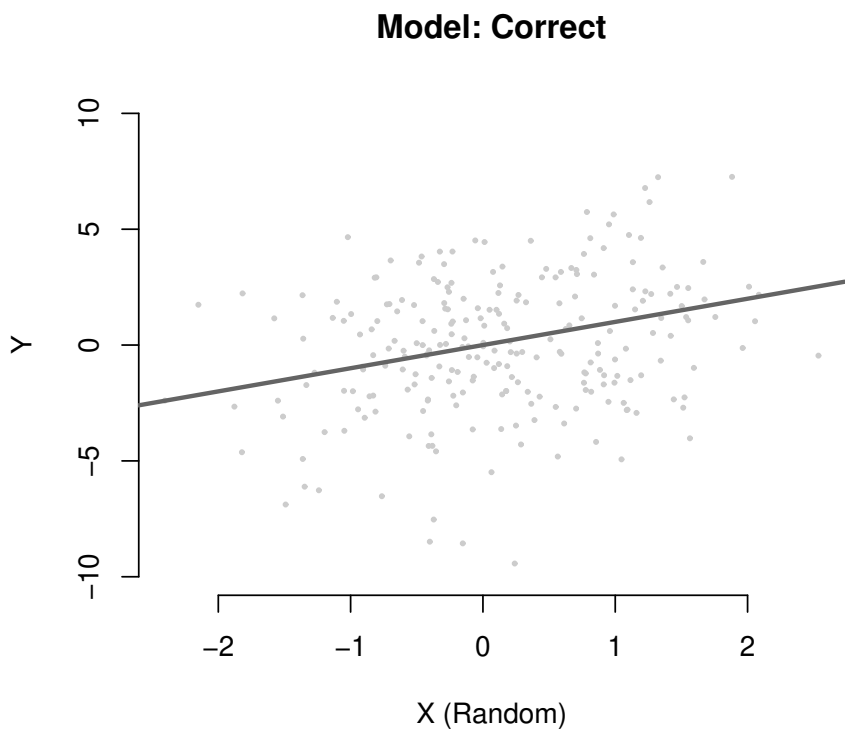
Simulation: Results

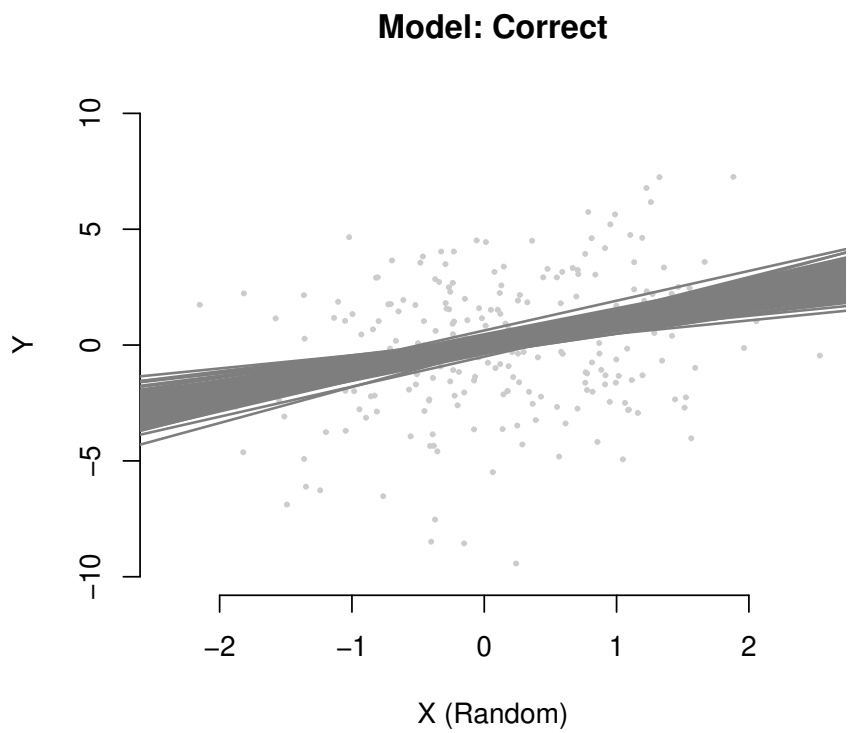




Empirical standard error of $\widehat{\beta}_1$ across simulations: 0.20

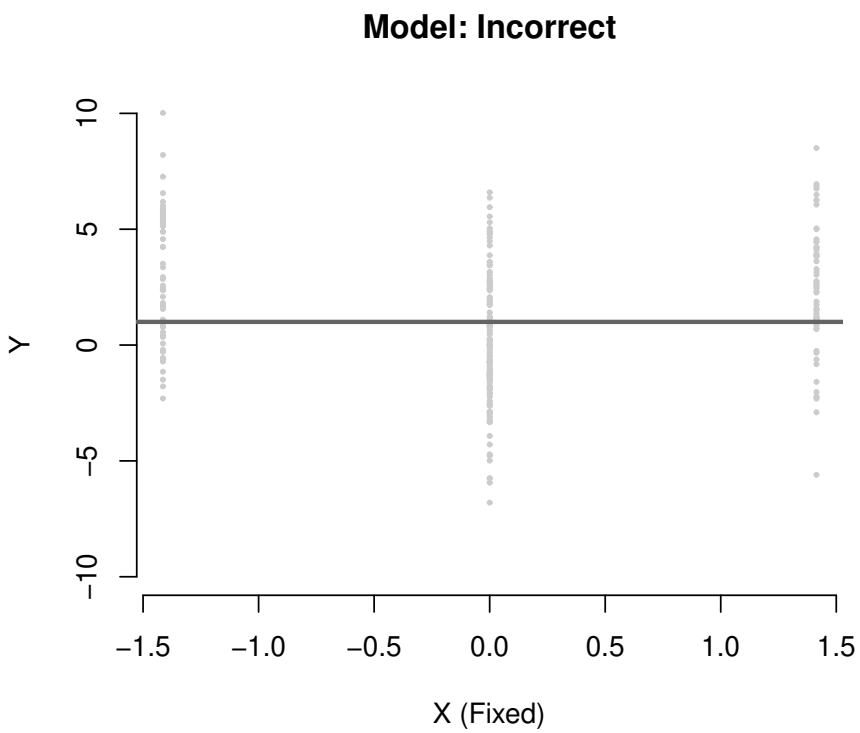
Simulation: Results

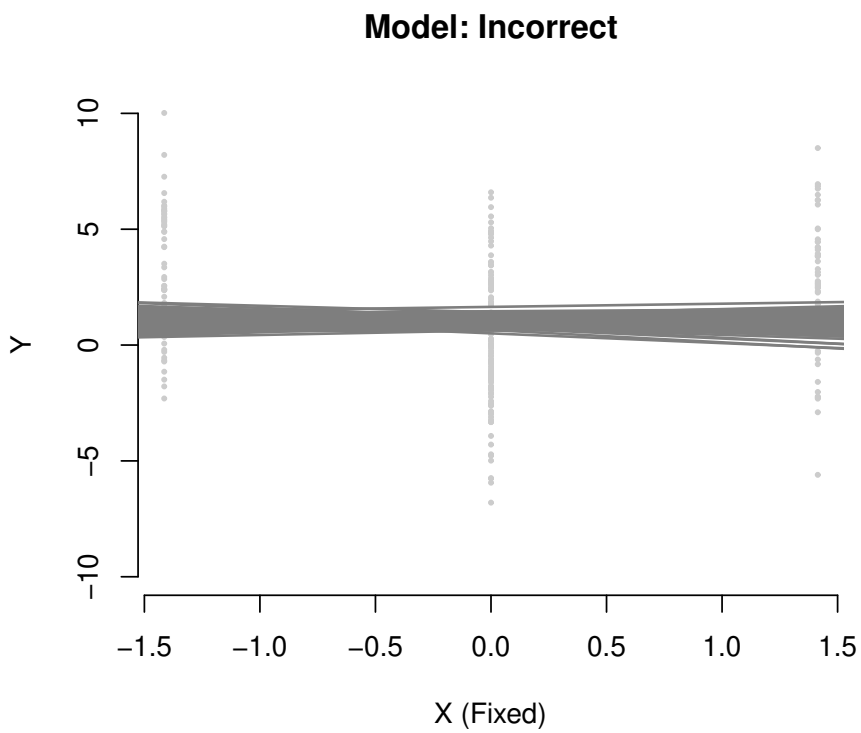




Empirical standard error of $\widehat{\beta}_1$ across simulations: 0.20

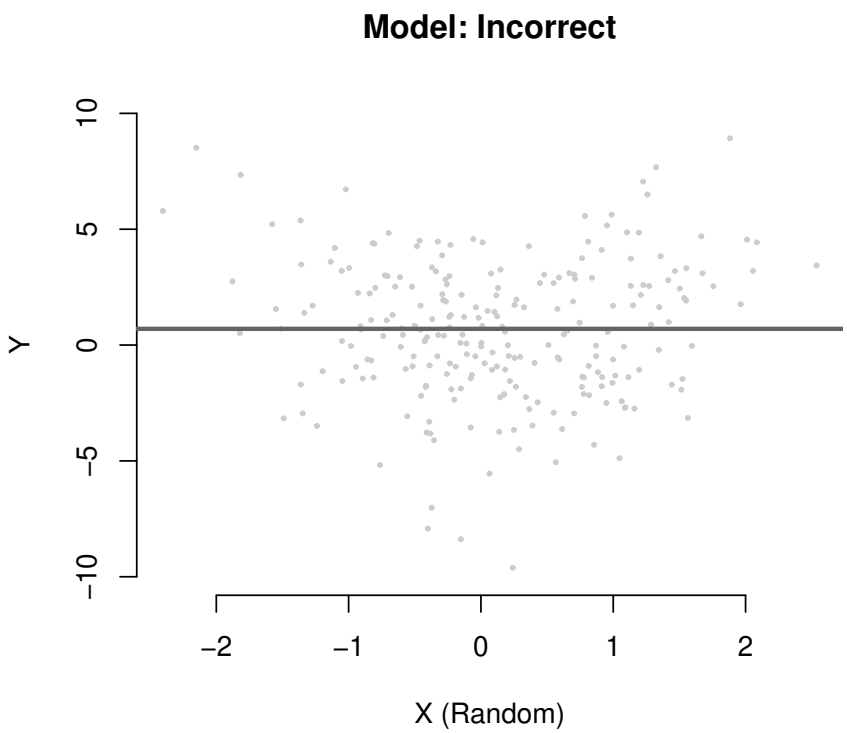
Simulation: Results

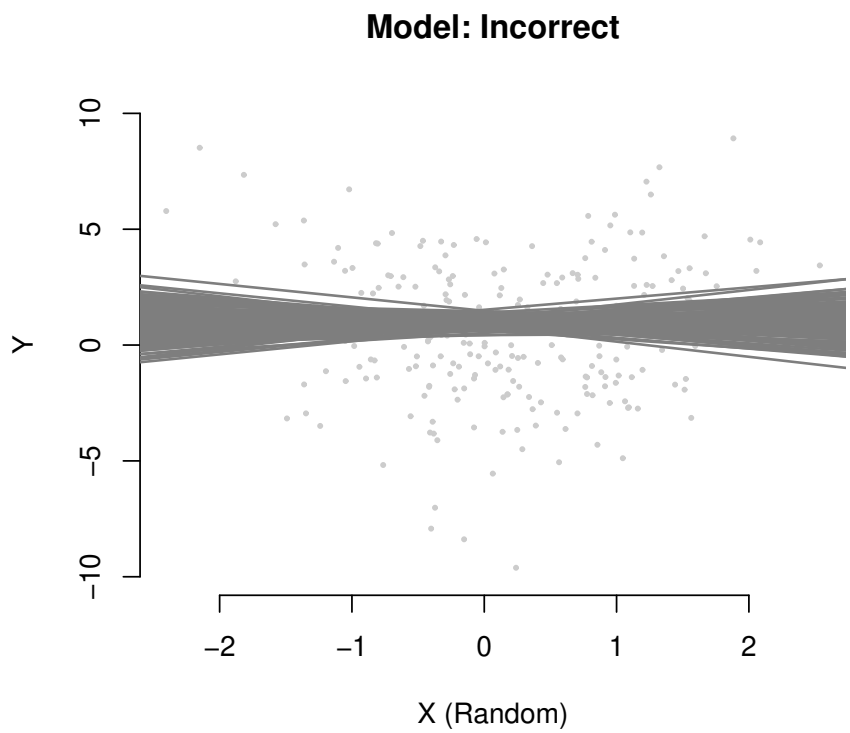




Empirical standard error of $\widehat{\beta}_1$ across simulations: 0.20

Simulation: Results





Empirical standard error of $\widehat{\beta}_1$ across simulations: 0.29

Fixed vs. random design matrices

Simulation: Why did it happen?

- Scenario 4 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 $\mathrm{E}[\widehat{\beta}_1|\mathbf{X}] > 0$.

Fixed vs. random design matrices

Further thoughts:

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

Table of Contents

Underfitting and overfitting

Underfitting:

- Let's continue with the theme of mean model misspecification—but let's work with the case in which the design matrix is fixed.
- Suppose the true data generating mechanism is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

  with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$.
- Without loss of generality, assume the columns of $\mathbf{Z}$ are linearly independent of the columns of $\mathbf{X}$.
- Suppose we fit the model $\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, omitting $\mathbf{Z}$.
- What will the impact be on:
  - Bias of $\widehat{\boldsymbol{\beta}}$?
  - Covariance of $\widehat{\boldsymbol{\beta}}$?
  - Estimated error variance?

Underfitting: Bias of $\widehat{\boldsymbol{\beta}}$

- We can determine a form for the bias:

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

- Estimate of $\boldsymbol{\beta}$ biased *unless* $\mathbf{X}^T\mathbf{Z} = \mathbf{0}$ (should remind you of precision variables).

Underfitting: Covariance of $\widehat{\boldsymbol{\beta}}$

- We can determine a form for the covariance of $\widehat{\boldsymbol{\beta}}$:

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

Underfitting: Estimated error variance

- We can determine a form for the estimated error variance (but let's start with RSS):

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

- Therefore,

$$\mathrm{E}[S^2] = \frac{\mathrm{E}[\mathrm{RSS}]}{N-K} = \sigma^2 + \frac{(\mathbf{Z}\boldsymbol{\alpha})^T(\mathbf{I}-\mathbf{P_X})(\mathbf{Z}\boldsymbol{\alpha})}{N-K}.$$

- This will exceed σ^2 unless $\mathbf{Z}^T\mathbf{X} = \mathbf{0}$.

Overfitting:

- Now, suppose the true data generating mechanism is given by:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon},$$

  with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$.
- Suppose we fit the model $\mathrm{E}[\mathbf{y}|\mathbf{X}_1, \mathbf{X}_2] = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$.
- What will the impact be on:
  - Bias of $\widehat{\boldsymbol{\beta}}$?
  - Covariance of $\widehat{\boldsymbol{\beta}}$?
  - Estimated error variance?

Overfitting: Bias of $\widehat{\boldsymbol{\beta}}$

- We can determine a form for the bias:

$$\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}] - \boldsymbol{\beta} &= \mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] - \boldsymbol{\beta} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}[\mathbf{y}] - \boldsymbol{\beta} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}_1\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix} \begin{bmatrix}\boldsymbol{\beta}_1 \\ \mathbf{0}\end{bmatrix} - \boldsymbol{\beta} \\
&= \begin{bmatrix}\boldsymbol{\beta}_1 \\ \mathbf{0}\end{bmatrix} - \boldsymbol{\beta} = \begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}.
\end{aligned}$$

- Estimate of $\boldsymbol{\beta}$ unbiased.

Underfitting and overfitting

Overfitting: Covariance of $\widehat{\boldsymbol{\beta}}$

- We can determine a form for the covariance of $\widehat{\boldsymbol{\beta}}$:
  - Homework! :)

Overfitting: Estimated error variance

- We can determine a form for the estimated error variance (but let's start with RSS):

$$\begin{aligned} \mathrm{E}[\mathrm{RSS}] &= \mathrm{E}[\mathbf{y}^T(\mathbf{I}-\mathbf{P}_\mathbf{X})\mathbf{y}] \\ &= \sigma^2\mathrm{trace}(\mathbf{I}-\mathbf{P}_\mathbf{X}) + (\mathbf{X}\boldsymbol{\beta}_1)^T(\mathbf{I}-\mathbf{P}_\mathbf{X})(\mathbf{X}\boldsymbol{\beta}_1) \\ &= (N-K)\sigma^2. \end{aligned}$$

- Therefore,

$$\mathrm{E}[S^2] = \frac{\mathrm{E}[\mathrm{RSS}]}{N-K} = \sigma^2.$$

- The error variance is unbiased.

Underfitting and overfitting

Careful:

- The term "bias" has a mathematical definition, but it is often used to describe different things.
- In these examples, we are specifically making a comparison of estimators to the natural corresponding parameters in the data generating mechanism.
- If I fit an unadjusted model, my estimate can still be unbiased for the unadjusted parameter—it's just not unbiased for the parameter of the data generating mechanism.

Summary

So far:

- Various forms of model misspecification.
  - There are others, some of which are left to homework and/or lab!

Summary

Up next:

- Prediction.