36-401 Homework 2

Due Friday, September 12 by 5:00pm

YOUR NAME

General instructions for all assignments:

- Use this file as the template for your submission. Please write your name at the top of this page in the author section.
- Throughout, there will be placeholders for your answers/code, marked clearly in **bold**. Please put your answers in *italics* or **bold**, thereby making them easier to see. One exception is math: You don't have to write math in italics/bold, but you should write mathematical answers in LaTeX. Alternatively, you may write mathematical answers by hand; **however, if you do so, you must note in your PDF that you've submitted some answers by hand, and clearly "match" the scanned handwritten pages on your PDF in Gradescope to the corresponding question you're answering.**
- Be sure to include any code you used to arrive at your answers. Furthermore, be sure to clearly explain how you arrived at your answers (i.e., show your work for each question).
- Although it's okay to discuss homework problems with other students, all of your homework (code, written answers, etc.) should be only your own. Instances of identical, nearly identical, or copied homework will be considered cheating and plagiarism. Furthermore, you cannot copy, or nearly copy, material from someone else (including an online entity or any other resource). In other words, you must follow rules of academic integrity (as detailed in the syllabus).
- Questions will sometimes have multiple subparts, often denoted with bullets; please do your best to answer each question. That said, each subpart is denoted with the number of points allocated to that subpart. Feel free to use that as guidance for what questions you should prioritize. Partial credit will be given, so do your best to attempt each question.
- Late homeworks will not be accepted. If you're not able to complete a homework, it is better to submit a partially complete homework on time than a complete homework at a late time (because the latter will not be accepted). Also remember that there is one homework drop for the class.

Problem 1: Deriving and Leveraging Results from Class (35 points)

In class, there are several properties that we stated but did not formally prove. In the following questions, we'll see how we can derive these properties.

PART A (9pts)

In class, we considered the mean squared error (MSE) of a linear function to predict an outcome Y. In this case, the MSE is

$$\text{MSE}(\beta_0, \beta_1) = \mathbb{E}[(Y - \beta_0 - \beta_1 X)^2]$$

We stated in class that the β_0 and β_1 that minimize this MSE are:

$$\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$
$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X]$$

In this problem, prove that the above choices for β_0 and β_1 indeed minimize $\text{MSE}(\beta_0, \beta_1)$. In your answer, please also confirm that these choices *minimize* MSE (rather than maximize). As always, in your answer be sure to explain your work step-by-step. (**You should explain your work step-by-step in every homework question; this just acts as one last reminder to do so.**)

[PUT YOUR ANSWER HERE]

PART B (8pts)

For this part, consider a function of X, denoted $f(X)$. In class, I made the following claim:

$$\mathbb{E}\left[\{Y - \mathbb{E}[Y|X]\}\{\mathbb{E}[Y|X] - f(X)\}\right] = 0$$

For this problem, prove that the above result holds.

Hint: The outer expectation $\mathbb{E}[\cdot]$ in the above equation is taken over Y *and* X. Thus, both Y *and* X are random variables within the outer expectation.

[PUT YOUR ANSWER HERE]

PART C (10pts)

In class, we've stated that the least squares estimators for the slope β_1 and intercept β_0 in a simple linear regression model are

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$$
$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \overline{X}$$

Meanwhile, in class we stated that $\hat{\beta}_1$ can be written as a linear combination of the outcomes:

$$\widehat{\beta}_1 = \sum_{i=1}^n k_i Y_i, \quad \text{where } k_i = \frac{X_i - \overline{X}}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2}$$

For this part, do the following **two things**:

- (6pts) Show that, indeed, $\widehat{\beta}_1 = \sum_{i=1}^n k_i Y_i$, where the k_i are defined as above.
- (4pts) Show that the estimated intercept can also be written as a linear combination of the β_0: $\widehat{\beta}_0 = \sum_{i=1}^n \tilde{k}_i Y_i$, for some $\tilde{k}_i$ that depend on the Xs. In your answer, be explicit about what the $\tilde{k}_i$ are (i.e., define what $\tilde{k}_i$ is equal to, similar to what we did for k_i above).

Hint: For the first part, try to show that $\sum_{i=1}^n \overline{Y}(X_i - \overline{X}) = 0$. For the second part, using the result from the first part will be very helpful.

[PUT YOUR ANSWER HERE]

PART D (8pts)

For this part, assume $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i | X_i \overset{iid}{\sim} N(0, \sigma^2)$. Furthermore, assume that the least squares estimator $\widehat{\beta}_0 = \sum_{i=1}^n \tilde{k}_i Y_i$, where the $\tilde{k}_i$ depend on the Xs but not the Ys (as you should show in Part

C). Given this information, what is the conditional distribution $\widehat{\beta}_0|X$? Be explicit in what the distribution is and what its parameter value(s) are, and explain your reasoning.

Hint: For this part, you may use other results about the $\widehat{\beta}_0$ least squares estimator that we've discussed in class in order to arrive at your answer. However, in your answer, just be clear about what from-class results you're using to arrive at your answer.

[PUT YOUR ANSWER HERE]

Problem 2: Fixing the Intercept (32 points)

In a typical linear regression model, we assume some intercept β_0 and slope β_1 that we must estimate. In this problem, we'll consider the consequences of using a model that fixes the intercept to a specific value, but where we still estimate the slope β_1.

Specifically, in this problem, we'll assume the following linear regression model:

$$Y_i = c + \beta_1 X_i + \epsilon_i \tag{1}$$

where c is some constant, $\mathbb{E}[\epsilon_i|X_i] = 0$, and $\text{Var}(\epsilon_i|X_i) = \sigma^2$. Thus, instead of assuming some unknown intercept β_0 that's estimated, the above model sets the intercept to a fixed value c, and only β_1 is considered a parameter that is estimated.

PART A (8pts)

Given the above model (1), derive the least squares estimator $\hat{\beta}_1$. In your answer, please also confirm that your $\hat{\beta}_1$ is indeed a *least* squares estimator rather than a "most squares" estimator. In other words, confirm that you've indeed achieved a minimizer, rather than a maximizer.

[PUT YOUR ANSWER HERE]

PART B (8pts)

Consider your least squares estimator $\hat{\beta}_1$ from Part A, and assume the above model (1) holds. For this part, please do the following two things:

- (6pts) Show $\hat{\beta}_1$ is conditionally unbiased, i.e., $\mathbb{E}[\hat{\beta}_1|X] = 0$.
- (2pts) Show $\hat{\beta}_1$ is unconditionally unbiased, i.e., $\mathbb{E}[\hat{\beta}_1] = 0$.

Hint: If you're struggling to establish unbiasedness for this problem, this may mean that your estimator $\hat{\beta}_1$ from Part A is incorrect.

[PUT YOUR ANSWER HERE]

PART C (8pts)

Consider your least squares estimator $\hat{\beta}_1$ from Part A, and assume the above model (1) holds. For this part, do the following two things:

- (4pts) Derive the conditional variance $\text{Var}(\hat{\beta}_1|X)$.
- (4pts) How does the conditional variance of your estimator compare to the conditional variance of the typical least squares estimator (if we were to assume an intercept β_0 that we had to estimate)? Please provide any mathematical reasoning you used to arrive at your answer.

Hint: In class we communicated the conditional variance of the typical least squares estimator. So, for the second part, you may want to revisit your class notes. Meanwhile, by "compare," I just mean to communicate

whether the conditional variance for your estimator is greater than, less than, or the same as the conditional variance of the typical least squares estimator (and explain your reasoning).

[PUT YOUR ANSWER HERE]

PART D (8pts)

Now let's say that the above model (1) doesn't hold. Instead, the following model holds:

$$Y_i = \tilde{c} + \beta_1 X_i + \epsilon_i$$

for some other constant $\tilde{c}$, but again $\mathbb{E}[\epsilon_i|X_i] = 0$, and $\text{Var}(\epsilon_i|X_i) = \sigma^2$. In other words, there is a potential mismatch between the constant c that we used in model (1) and the constant $\tilde{c}$ that is involved in the true model. Given this, answer the following two questions:

- (4pts) How does the difference $\tilde{c} - c$ affect the conditional bias of your estimator $\hat{\beta}_1$?
- (4pts) How does the difference $\tilde{c} - c$ affect the conditional variance of your estimator $\hat{\beta}_1$?

For each question, please provide mathematical reasoning for how you arrived at your answer.

Hint: As a reminder, the conditional bias is $\mathbb{E}[\hat{\beta}_1|X] - \beta_1$, where $\hat{\beta}_1$ is estimated using the model (1) at the beginning of the problem, but β_1 is from the *true* model that involves $\tilde{c}$. Meanwhile, when I ask "How does the difference $\tilde{c} - c$ affect" conditional bias or variance, I mean to discuss whether the conditional bias or variance increases or decreases with $\tilde{c} - c$, or does not depend on $\tilde{c} - c$ (and to explain your reasoning).

[PUT YOUR ANSWER HERE]

Problem 3: Analyzing Ecological Data (33 points)

For this problem, we'll practice appropriate exploratory data analysis (EDA) and linear regression models using real data. Specifically, we'll look at data measuring different bird species. Here is the code to read in the data:

```
birds = read.csv("https://raw.githubusercontent.com/zjbranson/36401_fall2025/refs/heads/main/bird-divers
```

There are several variables in this dataset, and for this homework we'll only focus on two: `allelicRichness` and `breedingRangeSize`. Allelic richness is a measure of genetic diversity of a particular species; meanwhile, the breeding range size measures the size of a species' range (i.e., how large of an area a species lives), in 10,000 square kilometers. Some ecological theories suggest that as the area of a species' habitat increases, its genetic diversity will increase (because the species has more room to evolve and differentiate). Genetic diversity is a way to measure the overall resilience of a species; thus, understanding the relationship between allelic richness and breeding range size can suggest how a species' resilience is related to the amount of space it occupies.

In this problem, we'll consider using the **square root** of `breedingRangeSize` as a covariate, and `allelicRichness` as the outcome. By using the square-root transformation, the resulting covariate will be on the kilometer scale, rather than the squared kilometer scale.

PART A (8pts)

First we'll conduct EDA. For this part, create a histogram of the covariate, a histogram of the outcome, and a scatterplot of the outcome and covariate. You may use base R or `ggplot` to make your visualizations, but please make sure your visualizations are appropriately labeled and readable. After making your visualizations, describe each variable's distribution (based on your histograms) in 1-2 sentences, and describe the relationship between the two variables (based on your scatterplot) in 1-2 sentences.

Hint: Remember, the **square root** of breeding range size is our covariate!

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

PART B (9pts)

Consider the **normal simple linear regression model** that we've discussed in class. For this part, please refer to your three visualizations in Part A to answer the questions below. Each question involves a Yes or No choice. **For each question, you must take a clear Yes or No stance. If your answer does not clearly take a Yes or No stance (e.g., "Yes I think so, but on the other hand. . . "), we will deduct points.**

- (3pts) First, look at the histogram of the covariate. Does this histogram suggest any potential violations of assumptions involved in the normal simple linear regression model? State Yes or No, and explain your reasoning in 1-2 sentences.

[PUT YOUR ANSWER HERE]

- (3pts) Now look at the histogram of the outcome. Does this histogram suggest any potential violations of assumptions involved in the normal simple linear regression model? State Yes or No, and explain your reasoning in 1-2 sentences.

[PUT YOUR ANSWER HERE]

- (3pts) Finally, look at the scatterplot of the outcome and covariate. Does this scatterplot suggest any potential violations of assumptions involved in the normal simple linear regression model? State Yes or No, and explain your reasoning in 1-2 sentences.

[PUT YOUR ANSWER HERE]

PART C (10pts)

Now fit a simple linear regression model, where the **square root** of breeding range size is used as the covariate and allelic richness is used as the outcome. Please display your estimates for the intercept and slope from this linear regression model. Then, answer the following questions:

- (5pts) In one sentence, write an interpretation for the estimated slope **within the context of this dataset and application**. Furthermore, does the linear regression model's slope have a scientifically meaningful interpretation within the context of this dataset? State Yes or No, and explain your reasoning in 1-2 sentences.
- (5pts) In one sentence, write an interpretation for the estimated intercept **within the context of this dataset and application**. Furthermore, does the linear regression model's intercept have a scientifically meaningful interpretation within the context of this dataset? State Yes or No, and explain your reasoning in 1-2 sentences.

Hint: By "scientifically meaningful interpretation," I mean that interpreting the parameter is relevant for studying the ecological theories of interest for this dataset (as stated at the beginning of this problem). Furthermore, your answers must make a clear Yes or No stance—otherwise, we will deduct points.

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

PART D (6pts)

To wrap up this homework, make a histogram of the **residuals** for the linear regression model you ran in Part C. Then, answer the following: Does your histogram suggest any potential violations of assumptions involved in the normal simple linear regression model? State Yes or No, and explain your reasoning in 1-2 sentences.

Hint: To obtain residuals in R, you can use the `residuals()` function. Furthermore, as in previous questions, you must make a clear Yes or No stance; otherwise, we will deduct points.

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]