36-401 Homework 1

Due Friday, September 5 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: Reading a Short Regression-based News Story (9 points)

For this problem, read the short news article available on Canvas under Files/Homeworks/Homework1, called `wpArticle.pdf`. After reading the article, answer the following questions:

- (3pts) What is the main covariate/predictor variable discussed in the news story, and what is the outcome/response?

[PUT YOUR ANSWER HERE]

- (3pts) In an experiment, the covariate X is randomly assigned by investigators, but in an observational study X is not randomly assigned. (We will see later in the course that this has important implications for causal inference.) Does the article discuss an experiment or an observational study? Explain your reasoning in one sentence.

[PUT YOUR ANSWER HERE]

- (3pts) Which use of regression seems to best match the goal of the study discussed in the article: for summarizing data, prediction, or inference? Explain your reasoning in 1-2 sentences.

[PUT YOUR ANSWER HERE]

Problem 2: Expectations, Variances, and Covariances in Linear Regression (53 points)

In class, we've discussed how an outcome Y may not be linear with a covariate X, but it may be linear with a nonlinear transformation of X. In this problem, we will consider such a situation.

Specifically, assume we have $X \sim \text{Unif}(0, 1)$, and $Y = \exp(X) + \epsilon$, where $\epsilon|X \sim N(0, 1)$. Thus, the outcome Y is linear in $\exp(X)$, rather than X. Here, by $\exp(X)$, we mean $\exp(X) = e^X$.

Part A (6pts)

What is the conditional distribution $Y|X$? Please state a specific distribution, as well as the parameter value(s) for that distribution. Those parameter value(s) may or may not depend on X. Please explain in 1-3 sentences how you arrived at your answer.

[PUT YOUR ANSWER HERE]

Part B (8pts)

Derive $\mathbb{E}[Y]$. Be sure to show your work for how you arrived at your answer.

Hint: As suggested by Part A, this problem has given you information about the conditional distribution $Y|X$. How can you use that information to figure out $\mathbb{E}[Y]$?

[PUT YOUR ANSWER HERE]

Part C (9pts)

Derive $\text{Var}(Y)$. Be sure to show your work for how you arrived at your answer.

Hint: Again, this problem has given you information about the conditional distribution $Y|X$. How can you use that information to figure out $\text{Var}(Y)$?

[PUT YOUR ANSWER HERE]

Part D (9pts)

Derive $\mathbb{E}[XY]$. Be sure to show your work for how you arrived at your answer.

Hint: You'll be on the right track if you find that you have to compute the integral $\int_0^1 xe^x dx$. For that integral, it can be useful to remember integration by parts, which involves evaluating an integral that can be written as $\int_0^1 u(x)v'(x)dx$ for two functions $u(x)$ and $v(x)$, where $v'(x)$ is the derivative of $v(x)$. Thus, when computing $\int_0^1 xe^x dx$, you should consider either setting $x = u(x)$ and $e^x = v'(x)$, or vice versa (whichever you think is easier).

[PUT YOUR ANSWER HERE]

Part E (6pts)

As we'll discuss later in the course, the true value of the slope of a linear regression model that regresses Y on X is $\beta_1 = \text{Cov}(X, Y)/\text{Var}(X)$. In this problem, we know that Y is not linear in X; but in practice, we wouldn't know this, and may well fit a linear regression model. Thus, it's useful to understand the true value

that linear regression model is estimating (in this case, β_1). For this problem, derive β_1. Be sure to show your work for how you arrived at your answer.

Hint: For this problem, $\mathbb{E}[X] = 1/2$ and $\text{Var}(X) = 1/12$. Thus, this problem boils down to figuring out $\text{Cov}(X, Y)$. Your work from previous parts will likely be very helpful.

[PUT YOUR ANSWER HERE]

Part F (6pts)

As we'll discuss later in the course, the true value of the intercept of a linear regression model that regresses Y on X is $\beta_0 = \mathbb{E}[Y] - \beta_1\mathbb{E}[X]$. For this part, derive β_0.

Hint: In Part B you should have found $\mathbb{E}[Y]$, and in Part E you should have found β_1. Meanwhile, $\mathbb{E}[X] = 1/2$. So, this part is just asking you to plug in values you found in previous questions and simplify.

[PUT YOUR ANSWER HERE]

Part G (9pts)

Now let's consider the random variable $\exp(X)$. As suggested by Part E, the true value of the slope of a linear regression model that regresses Y on $\exp(X)$ is $\beta_1^* = \text{Cov}(\exp(X), Y)/\text{Var}(\exp(X))$. For this problem, derive β_1^*, and be sure to show your work for how you arrived at your answer.

Hint: Remember, we are given at the beginning of the problem that, truly, Y is linearly related to $\exp(X)$. Thus, you should be able to guess what β_1^* is, and this question is asking you to derive β_1^*. In your derivation, you will likely come across quantities that you derived in previous questions. You may refer to your past work in your answer to this question; just be sure to clearly communicate what past work you're using (i.e., what work from what part) to arrive at your answer.

[PUT YOUR ANSWER HERE]

Problem 3: Simulating a Data-Generating Process (38 points)

In this problem, we'll again consider the data-generating process in Problem 2, where we have a covariate $X \sim \text{Unif}(0, 1)$ and an outcome $Y = \exp(X) + \epsilon$, where $\epsilon|X \sim N(0, 1)$. **However, you do not need to complete Problem 2 in order to complete this problem.** Instead, we'll focus on using R to simulate the data-generating process from Problem 2, and consider estimators using simulated data.

Part A (8pts)

For this part, complete the following tasks:

- Write code below that simulates the above data-generating process involving X and Y, for `n = 500` observations. Thus, your code should generate 500 random draws for X and 500 random draws for Y, based on the above data-generating process. Furthermore, please write your code in such a way that your draws stay the same every time you Knit your .Rmd file.

Hint: If you're not sure how to do the last part—where you make sure your draws stay the same every time you Knit—look back at the R tutorial I posted on Canvas (under Files).

```
# PUT YOUR CODE HERE
```

- Consider the population-level quantities $\mathbb{E}[X]$ and $\mathbb{E}[Y]$. What are natural estimators for $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, based on your random draws? Please first write a mathematical equation for each of your estimators in terms of the 500 random draws $(X_1, Y_1), \ldots, (X_{500}, Y_{500})$. Then, write code below that computes and prints each of your estimators, based on the 500 random draws. Finally, state in words what your estimates are for $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, based on your random draws. (**In all assignments, if you are**

asked to answer a question, please be sure to do outside of code chunks, even if you think your code and output is self-explanatory.)

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

Part B (8pts)

Consider the population-level quantities:

- $\text{Var}(X)$
- $\text{Cov}(X, Y)$
- $\text{Var}(\exp(X))$
- $\text{Cov}(\exp(X), Y)$

For each of these quantities, write a mathematical equation of a natural estimator for that quantity, in terms of the 500 random draws $(X_1, Y_1), \ldots, (X_{500}, Y_{500})$. Then, write code below that computes and prints each of your estimators, based on the 500 random draws. Finally, state in words what your estimates are for each of these quantities, based on your random draws.

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

Part C (6pts)

As stated in Problem 2, the true slope of the linear regression model that regresses Y on X is

$$\beta_1 = \text{Cov}(X, Y)/\text{Var}(X)$$

whereas the true slope when regressing Y on $\exp(X)$ is

$$\beta_1^* = \text{Cov}(\exp(X), Y)/\text{Var}(\exp(X))$$

Notice that β_1 and β_1^* are in terms of variances and covariances. Meanwhile, in Part B, you should have computed estimates for these variances and covariances. Thus, one idea for estimating β_1 and β_1^* is to simply plug in estimates for the variances and covariances. Given this idea, write code below that computes and prints out estimates for β_1 and β_1^*, using your estimates from Part B. Then, state in words what your estimates are for each of these quantities, based on your output.

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

Part D (8pts)

We'll again consider estimating β_1 and β_1^* from Part C, but now with linear regression models. For this part, use the `lm()` function to appropriately estimate β_1 and β_1^* via linear regression. After defining your linear regression models, use the `coefficients()` function to print out your estimates for β_1 and β_1^*. Then, state in words what your estimates are for each of these quantities, based on your output. Finally, answer the following in 1-2 sentences: How do your estimates for β_1 and β_1^* here compare to your estimates from Part C?

Hint: If you're unsure how to use the `lm()` function, you should look at the R tutorial on Canvas (under Files). Meanwhile, by "compare" your estimates, I simply mean state whether the estimates here appear similar or different to the estimates in Part C.

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

Part E (8pts)

To wrap up this homework, use your 500 random draws from Part A to create a scatterplot with X on the x-axis and Y on the y-axis. Then, add the following two lines to your scatterplot:

- The estimated regression line for the $Y \sim X$ linear model (i.e., the linear regression model that regresses Y on X). This should correspond to $\hat{\beta}_0 + \hat{\beta}_1 X$, where these estimates are computed from `lm()`. Please make this line black on your plot.
- The true regression function line, which is $\exp(X)$. Please make this line red on your plot.

For this part, you just have to produce the desired plot.

Hint: If you're unsure how to make a scatterplot or add lines to it, you should review the R tutorial on Canvas (under Files). When plotting your lines in R, you may have to sort your observations (from decreasing to increasing) in order for your lines to display properly. To do this in R, use the `sort()` function. For example, `sort(x)` would sort the vector `x` from least to greatest.

```
# PUT YOUR CODE HERE
```