36-401 Homework 3

Due Friday, September 19 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: Considering Residuals in Simple Linear Regression (60 points)

For this problem, we'll consider the simple linear regression model,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, \ldots, n$$

where $\epsilon_i | X \overset{iid}{\sim} N(0, \sigma^2)$. Specifically, we'll consider the outcomes Y_i and *true* residuals ϵ_i, and compare them to the fitted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and the estimated residuals $\hat{\epsilon}_i = Y_i - \hat{Y}_i$. Here, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the typical least squares estimators for this linear model.

PART A (8pts)

Let $\overline{Y}$ denote the sample average of $(Y_1, \ldots, Y_n)$, and let $\overline{\hat{Y}}$ denote the sample average of the $(\hat{Y}_1, \ldots, \hat{Y}_n)$. For this part, show $\overline{Y} = \overline{\hat{Y}}$, meaning that the mean outcome is equal to the mean fitted value.

[PUT YOUR ANSWER HERE]

PART B (8pts)

For this part, we'll just consider the covariate values $(X_1, \ldots, X_n)$. Show that the following equality holds:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i^2 - \overline{X}^2)$$

Thus, we can seemingly "bring the square into" the difference $(X_i - \overline{X})$.

[PUT YOUR ANSWER HERE]

PART C (10pts)

For this part, show that $\sum_{i=1}^{n} \hat{\epsilon}_i \hat{Y}_i = 0$. This means that the $\hat{\epsilon}_i$ and $\hat{Y}_i$ are uncorrelated, because the dot product of the vectors $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$ and $(\hat{Y}_1, \ldots, \hat{Y}_n)$ equals zero.

Hint: You'll likely find it helpful to use the results from Part A and Part B to prove this result. In your work for this part, just be clear when and how you are using results from Parts A and B to solve this problem. That said, it is also fine if you can solve this problem without the Part A and B results.

[PUT YOUR ANSWER HERE]

PART D (10pts)

Now we will consider the covariance of the individual outcomes Y_i and their fitted values $\hat{Y}_i$. For this part, show that

$$\text{Cov}(Y_i, \hat{Y}_i | X) = \frac{\sigma^2}{n} + \frac{\sigma^2 (X_i - \overline{X}_i)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

This means that Y_i and $\hat{Y}_i$ will always have a positive covariance, which is intuitive: We would hope that our fitted values $\hat{Y}_i$ have some positive association with the actual outcome values Y_i.

Hint: For this part, you may use previous results that we've discussed in class to derive this conditional covariance. In your answer, just be clear what results from class you're using in your derivation.

[PUT YOUR ANSWER HERE]

PART E (10pts)

For this part, derive $\mathbb{E}[\hat{\epsilon}_i | X]$ and $\text{Var}(\hat{\epsilon}_i | X)$, i.e., the conditional expectation and conditional variance of the estimated residuals $\hat{\epsilon}_i = Y_i - \hat{Y}_i$.

Hint: For this part, you may use previous results that we've discussed in class to derive this conditional expectation and variance. In your answer, just be clear what results from class you're using in your derivation.

[PUT YOUR ANSWER HERE]

PART F (6pts)

For this part, first restate the conditional variance $\text{Var}(\hat{\epsilon}_i|X)$ you found in Part E. Then, in 1-3 sentences, discuss how the true residual variance σ^2, sample size n, and empirical variance of the covariate X affect this conditional variance.

Hint: By "affect," I just mean whether the conditional variance increases, decreases, or is not affected by these quantities (σ^2, n, and empirical variance of X).

[PUT YOUR ANSWER HERE]

PART G (8pts)

In class, we've considered the residual sum of squares, which are defined as $\sum_{i=1}^n \hat{\epsilon}_i^2$. We have the following distributional result:

$$\frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sigma^2}|X \sim \chi^2_{n-2}$$

You may take this result for granted in this problem. For this part, derive the distribution of

$$\frac{\sum_{i=1}^n \epsilon_i^2}{\sigma^2}|X$$

This is the exact same quantity, but with the *true* residuals ϵ_i instead of the estimated ones, $\hat{\epsilon}_i$. Be sure to explain how you determined the distribution for $\frac{\sum_{i=1}^n \epsilon_i^2}{\sigma^2}$. After determining the distribution, answer the following: How does the variance of $\frac{\sum_{i=1}^n \epsilon_i^2}{\sigma^2}$ compare to the variance of $\frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sigma^2}$? Please answer in 1-2 sentences, and explain your reasoning.

Hint: The variance of a χ^2_d distribution is $2d$.

[PUT YOUR ANSWER HERE]

Problem 2: Analyzing Rental Properties (40 points)

For this problem, we'll consider data on apartments for rent in an undisclosed city. Here is the data:

```
rentalData = read.csv("https://raw.githubusercontent.com/zjbranson/36401_fall2025/refs/heads/main/rental
```

There are only four variables in this dataset:

- `rent`: The monthly rental price of the apartment (in US dollars). This is paid by the person who rents the apartment.
- `age`: The age of the apartment (in years).
- `expense`: The monthly operating expenses and taxes for the apartment (in US dollars). This is paid by the person who owns the apartment.
- `space`: The size of the apartment (in square feet).

Let's say that, for good or bad, we have been hired by a powerful real estate broker to understand how rental prices in the city are related to aspects of apartments. To do this, we'll consider three simple linear regression models:

1) **Rent-Age Regression**: A model that regresses `rent` (the outcome) on `age` (the covariate).
2) **Rent-Expense Regression**: A model that regresses `rent` (the outcome) on `expense` (the covariate).
3) **Rent-Space Regression**: A model that regresses `rent` (the outcome) on `space` (the covariate).

The following questions ask you about these three potential regression models.

PART A (8pts)

Before implementing the regressions above, please do the following: For *each* of the three regression models, make *one* form of graphical EDA that best allows you to assess the simple linear regression model assumptions. Thus, you should make three graphs (one for each regression model). Then, after displaying your graphs, answer the following two questions:

- In 1-3 sentences, explain why your form of graphical EDA best allows you to assess the simple linear regression model assumptions.
- Based on your EDA, which of the three linear regression models would most realistically satisfy the model's assumptions? Please pick only **one** of the models, and explain your reasoning in 1-3 sentences.

```
# PUT YOUR EDA HERE
```

[PUT YOUR ANSWER HERE]

PART B (8pts)

Now run each of the three linear regression models stated at the beginning of this problem. After running your regressions, please display `summary()` output for each model. Then, for *each* model, write a one-sentence interpretation of the *point estimate for the slope parameter* of that model.

Hint: For the purposes of this question, you can pretend that the modeling assumptions are satisfied, regardless of your answer in Part A. You also don't have to consider statistical significance—or lack thereof—for this part; you only have to consider the *point estimate*.

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

PART C (8pts)

In Part B, you should have run three linear regression models. For this part, please do the following: For *each* model, compute the 95% confidence interval for the slope parameter of that model. Then, for each interval, write a 1-2 sentence practical interpretation of that confidence interval within the context of the data.

Hint: By "practical interpretation," I mean to comment on the direction and magnitude of the effect you can conclude, if anything, and the statistical inference you can conduct, if anything, from your interval (as we've discussed in class). Furthermore, for the purposes of this question, you can pretend that the modeling assumptions are satisfied, regardless of your answer in Part A.

```
# PUT YOUR CODE HERE
```

[PUT YOUR ANSWER HERE]

PART D (8pts)

For *each* of the three regression models, answer the following: Does the *intercept parameter* have a scientifically meaningful interpretation, within the context of this dataset? Answer Yes or No, and explain your reasoning in one sentence. Furthermore, for the model(s) that you state Yes (if any), interpret the point estimate and confidence interval of the intercept, based on your `summary()` output from Part B.

[PUT YOUR ANSWER HERE]

PART E (8pts)

To wrap up this homework, let's keep in mind that (as stated at the beginning of the problem) we've been hired by a powerful real estate broker. The real estate broker looks at your analyses, and states the following:

“Let’s say that I buy all of these apartments, such that I become the landlord for all of them. Your analyses suggest that if I manage to decrease the monthly expenses of each apartment by $100, then the rental price is estimated to go down by only $17.26, on average, resulting in a $82.74 increased monthly profit for each apartment, on average.”

Do you agree or disagree with the broker’s claim? State Agree or Disagree, and explain your reasoning in 1-3 sentences.

Hint: For the purposes of this question, unrealistically assume that the broker cannot directly control rental prices, and instead, rental prices are dictated by the same market forces that dictated rental prices in the data. (In real-world scenarios, it’s doubtful that a landlord would willingly let rental prices decrease.)

[PUT YOUR ANSWER HERE]