Data 100, Summer 2023

Homework #5B

Total Points: 20

Submission Instructions

You must submit this assignment to Gradescope by the on-time deadline, **Monday, July 10th at 11:59 PM Pacific**. Please read the syllabus for **the grace period policy**. No late submissions beyond the grace period will be accepted. While course staff is happy to help you if you encounter difficulties with submission, we may not be able to respond to last-minute requests for assistance (TAs need to sleep, after all!). **We strongly encourage you to plan to submit your work to Gradescope several hours before the stated deadline.** This way, you will have ample time to reach out to staff for submission support.

There are two parts to this assignment listed on Gradescope:

- **Homework 05 Coding**: Submit your Jupyter notebook zip file for Homework 5A, which can be generated and downloaded from DataHub by using the `grader.export()` cell provided.
- **Homework 05 Written**: Submit a single PDF to Gradescope that contains both (1) your answers to all manually graded questions from the Homework 5A Jupyter Notebook, and (2) your answers to all questions in this Homework 5B document.

To receive credit on this assignment, **you must submit both your coding and written portions to their respective Gradescope portals**. Your written submission (a single PDF) can be generated as follows:

1. Access your answers to manually graded Homework 5A questions in one of three ways:
    - *Automatically create PDF (recommended)*: We have provided a cell to generate your written response in the Homework 1A notebook for you. Run the cell and click to download the generated PDF. This function will extract your response to the manually-graded questions and put them on separate pages. This process may fail if your answer is not properly formatted; if this is the case, check out common errors and solutions described on Ed or follow either of the two ways described below.

- *Manually download PDF*: If there are issues with automatically generating the PDF, on DataHub, you can try downloading the PDF by clicking on `File->SaveandExportNotebookAs...->PDF`. If you choose to go this route, you must take special care to ensure all appropriate pages are chosen for each question on Gradescope.
- *Take screenshots*: If that doesn't work either, you can take screenshots of your answers (and your code if present) to manually-graded questions and include them as images in a PDF. The manually-graded questions are listed at the top of the Homework 5A notebook.

2. Answer the below Homework 5B written questions using one of the following options:
   - You can type your answers. We recommend LaTeX, the math typesetting language. Overleaf is a great tool to type in LaTeX.
   - Download this PDF, print it out and write directly on these pages. If you have a tablet, you may save this PDF and write directly on it.
   - Write your answers on a blank sheet of physical or digital paper. Note: If you write your answers on physical paper, use a scanning application (e.g., CamScanner, Apple Notes) to generate a PDF.
3. Combine these two sets of answers together into one PDF document and submit it to the appropriate Gradescope written portal. You can use PDF merging tools, e.g., Adobe Reader, Smallpdf (`https://smallpdf.com/merge-pdf`) or Apple Preview (`https://support.apple.com/en-us/HT202945`).

Important: When submitting on Gradescope, you **must tag pages to each question correctly** (it prompts you to do this after submitting your work). This significantly streamlines the grading process and allows us to release grades more quickly.

Your work will NOT be graded if you do not select pages on Gradescope. Submissions that do not follow instructions will not be regraded or extended.

If you encounter any difficulties with submission, please don't hesitate to reach out to staff prior to the deadline.

Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others, please include their names at the top of your submission.

Properties of Linear Regression Residuals

1. (10 points) In lecture, we spent a great deal of time talking about simple linear regression, which you also saw in Data 8. To briefly summarize, the simple linear regression model assumes that given a single observation x, our predicted response for this observation is $\hat{y} = \theta_0 + \theta_1 x$.

   In Lecture 9 we saw that the $\theta_0 = \hat{\theta}_0$ and $\theta_1 = \hat{\theta}_1$ that minimize the average L_2 loss for the simple linear regression model are:

   $$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$
   $$\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$$

   Or, rearranging terms, our predictions $\hat{y}$ are:

   $$\hat{y} = \bar{y} + r\sigma_y \frac{x - \bar{x}}{\sigma_x}$$

   (a) (3 points) As we saw in the lecture, a residual e_i, for data point $i \in \{1, \ldots, n\}$, is defined to be the difference between a true response y_i and predicted response $\hat{y}_i$. Specifically, $e_i = y_i - \hat{y}_i$. Note that there are n data points, and each data point is denoted by (x_i, y_i).

   Prove, using the equation for $\hat{y}$ above, that $\sum_{i=1}^{n} e_i = 0$.

   (b) (2 points) Prove that $\bar{y} = \bar{\hat{y}}$. You may use your result from part (a). Note that $\bar{\hat{y}}$ refers to the mean of all the predicted responses, that is $\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$

   (c) (2 points) Show that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

(d) (3 points) Show that the residuals are uncorrelated with the predictor variable, that is

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i-\bar{e}}{\sigma_e}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)=0,$$

where $\bar{e}=\frac{1}{n}\sum_{i=1}^{n}e_i$, $\sigma_e^2=\frac{1}{n}\sum_{i=1}^{n}(e_i-\bar{e})^2$, and $\sigma_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$. You may assume that σ_e, σ_x, and at least one residual are not exactly zero. Use the properties of estimating equations derived in lecture.

Properties of a Linear Model With No Constant Term

2. Suppose that we don't include an intercept term in our model. That is, our model is now

$$\hat{y} = \theta x,$$

where θ is the single parameter for our model that we need to optimize. (In this equation, x is a scalar, corresponding to a single feature.)

As usual, we are looking to find the value $\hat{\theta}$ that minimizes the average L_2 loss (mean squared error) across our observed data $\{(x_i, y_i)\}, for\ i \in \{1, \ldots, n\}$:

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta x_i)^2$$

The normal equations derived in the lecture no longer hold. In this problem, we'll derive a solution to this simpler model. We'll see that the least squares estimate of the slope in this model differs from the simple linear regression model, and will also explore whether or not our properties from the previous problem still hold.

(a) (4 points) Use calculus to find the minimizing $\hat{\theta}$.

That is, you may prove that:

$$\hat{\theta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Hint: You may start by following the format of SLR in lecture 9 and replace the SLR model with the model defined above.

Now let's check if the properties we proved in the last question still holds.

(b) (3 points) Do residuals still sum up to zero? In other words, is it true that $\sum_i^n e_i = 0$ for $e_i = y_i - \hat{y}_i$? If you think it's true, prove it. If you think it's false, provide a counterexample (for example, a set of observations such that this property does not hold true)

(c) (3 points) Is $(\bar{x}, \bar{y})$ still on the regression line? In other words, is it true that $\bar{y} = \hat{\theta}\bar{x}$? If you think it's true, prove it. If you think it's false, provide a counterexample (for example, a set of observations such that this property does not hold true).

Congratulations! You have finished Homework 5B!

Important: When submitting on Gradescope, you **must tag pages to each question correctly** (it prompts you to do this after submitting your work). This significantly streamlines the grading process and allows us to release grades more quickly.

Your work will NOT be graded if you do not select pages on Gradescope. Submissions that do not follow instructions will not be regraded or extended.

If you encounter any difficulties with submission, please don't hesitate to reach out to staff prior to the deadline.