Data 100, Spring 2023

Homework #5B

Due Date: Thursday, February 23th at 11:59 PM Pacific

Total Points: 24

Submission Instructions

You must submit this assignment to Gradescope by the on-time deadline, **Thursday, February 23th 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 1A notebook.

2. Answer the below Homework 1B written questions in one of many ways:
   - 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).
4. **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 for our readers. Failure to do this may result in a score of 0 for untagged questions.

You are responsible for ensuring your submission follows our requirements. We will not be granting regrade requests nor extensions to submissions that don't follow instructions. 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 the 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 10 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 $\sum_{i=1}^{n} e_i = 0$.

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} \left(y_i - \left(\bar{y} + r\sigma_y \frac{x - \bar{x}}{\sigma_x} \right) \right) \\
&= \sum_{i=1}^{n} \left(y_i - \bar{y} - r\sigma_y \frac{x_i - \bar{x}}{\sigma_x} \right) \\
&= \left(y_1 - \bar{y} - r\sigma_y \frac{x_1 - \bar{x}}{\sigma_x} \right) + \ldots + \left(y_n - \bar{y} - r\sigma_y \frac{x_n - \bar{x}}{\sigma_x} \right) \\
&= (y_1 + \ldots + y_n) - n\bar{y} - \left(r\sigma_y \frac{x_1 - \bar{x}}{\sigma_x} + \ldots + r\sigma_y \frac{x_n - \bar{x}}{\sigma_x} \right) \\
&= \sum_{i=1}^{n} y_i - n\bar{y} - \left(r\sigma_y \frac{x_1 - \bar{x}}{\sigma_x} + \ldots + r\sigma_y \frac{x_n - \bar{x}}{\sigma_x} \right)
\end{aligned}$$

$$= -r\frac{\sigma_y}{\sigma_x}\left((x_1 - \bar{x}) + \ldots + (x_n - \bar{x})\right)$$
$$= r\frac{\sigma_y}{\sigma_x}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = r\frac{\sigma_y}{\sigma_x} \times 0 = 0. \qquad \square$$

(b) (2 points) Prove that $\bar{y} = \bar{\hat{y}}$. You may use your result from part (a).

$$\begin{aligned} 0 &= \sum_{i=1}^{n} e_i \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\ &= (y_1 - \hat{y}_1) + \ldots + (y_n - \hat{y}_n) \\ &= (y_1 + \ldots + y_n) - (\hat{y}_1 + \ldots + \hat{y}_n). \\ y_1 + \ldots + y_n &= \hat{y}_1 + \ldots + \hat{y}_n \\ \frac{y_1 + \ldots + y_n}{n} &= \frac{\hat{y}_1 + \ldots + \hat{y}_n}{n} \\ \bar{y} &= \bar{\hat{y}}. \end{aligned}$$ $\square$

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

If $(\bar{x}, \bar{y})$ is on the simple linear regression, then the residual e should be 0. We will show that is true. Recall that the simple linear regression model for a given observation x is $\hat{y} = \theta_0 + \theta_1 x$. Observe that, by plugging in $\bar{x}$ for x and $\bar{y}$ for y, we get

$$\begin{aligned} e &= \bar{y} - \hat{y} \\ &= \bar{y} - \left(\bar{y} + r\sigma_y \frac{\bar{x} - \bar{x}}{\sigma_x}\right) \\ &= \bar{y} - \left(\bar{y} + r\sigma_y \frac{0}{\sigma_x}\right) \\ &= \bar{y} - \bar{y} \\ &= 0. \end{aligned}$$

Therefore, $(\bar{x}, \bar{y})$ must be on the simple linear regression. $\square$

(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.

Observe that the correlation between residuals and the predictor variable is $r=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i-\bar{e}}{\sigma_e}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)$. We will show that $r=0$, using the errors e_i as our observation values. Recall that the two estimating equations are

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i) \tag{1}$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)x_i \tag{2}$$

and that $\bar{e}=\frac{1}{n}\sum_{i=1}^{n}e_i=0$. Subtracting $(2)-(1)\bar{x}$ gives us the equivalent equation

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)x_i-\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)\right)\bar{x}=0\iff\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)(x_i-\bar{x})=0$$

Then, observe that

$$\begin{aligned}
r&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i-\bar{e}}{\sigma_e}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{\sigma_e}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\\
&=\frac{1}{n\sigma_e\sigma_x}\sum_{i=1}^{n}\left(e_i\right)\left(x_i-\bar{x}\right)\\
&=\frac{1}{n\sigma_e\sigma_x}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(x_i-\bar{x}\right)\\
&=\frac{1}{n\sigma_e\sigma_x}*0\text{ by the estimating equations}=0.
\end{aligned}$$

□

Properties of a Linear Model With No Constant Term

2. (4 points) 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 observation.)

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.

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 10 and replace the SLR model with the model defined above.

We will take the derivative of the mean squared error function with respect to θ and set that equal to 0 (on the following page).

$$\begin{aligned} R(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)^2 \\ \frac{\partial R}{\partial \theta} = 0 &= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{\theta} x_i)x_i \\ 0 &= \sum_{i=1}^{n}(x_i y_i - \hat{\theta} x_i^2) \\ 0 &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\theta} x_i^2 \end{aligned}$$

$$\sum_{i=1}^{n} \hat{\theta} x_i^2 = \sum_{i=1}^{n} x_i y_i$$
$$\hat{\theta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$
$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Now we will show that $\hat{\theta}$ is in fact a minimizer by calculating the second derivative of $R(\theta)$.

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

$$\frac{d^2 R(\theta)}{d\theta^2} = -\frac{2}{n} \sum_{i=1}^{n} (-x_i^2)$$
$$= \frac{2}{n} \sum_{i=1}^{n} x_i^2 \geq 0$$

Observe that $\sum_{i=1}^{n} x_i^2 \geq 0$ because x_i^2 is always nonzero, so the sum of nonzero terms must also be nonzero. □

MSE "Minimizer"

3. (10 points) Recall from calculus that given some function $g(x)$, the x you get from solving $\frac{dg(x)}{dx} = 0$ is called a *critical point* of g – this means it could be a minimizer or a maximizer for g. In this question, we will explore some basic properties and build some intuition on why, for certain loss functions such as squared L_2 loss, the critical point of the empirical risk function (defined as an average loss on the observed data) will always be the minimizer.

   Given some linear model $f(x) = \theta x$ for some real scalar θ, we can write the empirical risk of the model f given the observed data $\{x_i, y_i\},\ for\ i \in \{1, \ldots, n\}$ as the average L_2 loss, also known as Mean Squared Error (MSE):

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

   (a) (3 points) Let's investigate one of the n functions in the summation in the MSE. Define $g_i(\theta) = \frac{1}{n}(y_i - \theta x_i)^2$ for $i \in \{1, \ldots, n\}$. In this case, note that the MSE can be written as $\sum_{i=1}^{n} g_i(\theta)$.

   Recall from calculus that we can use the 2nd derivative of a function to describe its curvature about a certain point (if it is facing concave up, down, or possibly a point of inflection). You can take the following as a fact: A function is convex if and only if the function's 2nd derivative is non-negative on its domain. Based on this property, verify that $g_i(\theta)$ is a **convex function**.

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

$$\frac{dg_i(\theta)}{d\theta} = \frac{2}{n}(y_i - \theta x_i)(-x_i) = -\frac{2}{n}(x_i y_i - \theta x_i^2)$$

$$\frac{d^2 g_i(\theta)}{d\theta^2} = -\frac{2}{n}(-x_i^2) = \frac{2x_i^2}{n} \geq 0\ \forall x_i$$

   Therefore, $g_i(\theta)$ is a convex function. □

   (b) (2 points) Briefly explain intuitively in words why given a convex function $g(\theta)$, the critical point we get by solving $\frac{dg(\theta)}{d\theta} = 0$ minimizes g. You can assume that $\frac{dg(\theta)}{d\theta}$ is a function of θ (and not a constant).

   The first derivative of a function gives us the input value that produces a relative extrema of the function g, which could either be a minimum or maximum. In

order to deduce which one it is, we need to determine the concavity of the function. Since g is a convex function, by definition, g is concave up for all x. Therefore, the extrema would be a minimum, thus making the corresponding x a minimizer of g.

(c) (3 points) Now that we have shown that each term in the summation of the MSE is a convex function, one might wonder if the entire summation is convex, given that it is a sum of convex functions.

Let's look at the formal definition of a **convex function**. Algebraically speaking, a function $g(\theta)$ is convex if for any two points $(\theta_i, g(\theta_i))$ and $(\theta_j, g(\theta_j))$ on the function,

$$g(c \times \theta_i + (1-c) \times \theta_j) \leq c \times g(\theta_i) + (1-c) \times g(\theta_j)$$

for any real constant $0 \leq c \leq 1$.

Intuitively, the above definition says that, given the plot of a convex function $g(\theta)$, if you connect 2 randomly chosen points on the function, the line segment will always lie on or above $g(\theta)$ (try this with the graph of $g(\theta) = \theta^2$).

i. (2 points) Using the definition above, show that if $g(\theta)$ and $h(\theta)$ are both convex functions, their sum $g(\theta) + h(\theta)$ will also be a convex function.

Using the formal definition of a convex function, we can observe that for two functions $g(\theta)$ and $h(\theta)$,

$$\begin{aligned}
& g(c \times \theta_i + (1-c) \times \theta_j) \leq c \times g(\theta_i) + (1-c) \times g(\theta_j) \\
+\ & h(c \times \theta_i + (1-c) \times \theta_j) \leq c \times h(\theta_i) + (1-c) \times h(\theta_j) \\
\Rightarrow\ & g(c \times \theta_i + (1-c) \times \theta_j) + h(c \times \theta_i + (1-c) \times \theta_j) \\
\leq\ & c \times g(\theta_i) + (1-c) \times g(\theta_j) + c \times h(\theta_i) + (1-c) \times h(\theta_j) \\
\leq\ & c \times (g(\theta_i) + h(\theta_i)) + (1-c) \times (g(\theta_j) + h(\theta_j))
\end{aligned}$$

ii. (1 point) Based on what you have shown in the previous part, explain intuitively why a (finite) sum of n convex functions is still a convex function when $n > 2$.

Recall that the second derivative of a sum of terms is equal to the sum of the second derivatives of each of the terms. So if we have n convex functions, each of their second derivatives will be 0 or positive, and furthermore the sum of the second derivatives must also be 0 or positive. Therefore, the finite sum will be a convex function. □

(d) (2 points) Finally, explain why in our case that, when we solve for the critical point of the MSE by taking the gradient with respect to the parameter and setting the expression to 0, it is gauranteed that the solution we find will minimize the MSE.

In our case, solving the MSE by taking the gradient with respect to the parameter and setting the expression to 0 is gauranteed to provide a solution that minimizes MSE because the second derivative, i.e the concavity of the function will always be positive. Thus, any extrema that we find from the critical values will always be a minimum of the function.

Closing note: In this question, we have discussed only the simple linear model with no constant term—a single-variable function. However, the above properties extend more generally to all multivariable linear regression models; this proof is beyond the scope of this course and is left to a future you.

Congratulations! You have finished Homework 5B!