Data 100, Fall 2024

Homework #5A

Due Date: Thursday, October 10th at 11:59 PM Pacific

Total Points: 26

Submission Instructions

You must submit this assignment to Gradescope by the on-time deadline, **Thursday, October 10th at 11:59 PM Pacific**. Please read the syllabus for **the Slip Day policy**. No late submissions beyond the Slip Day policy will be accepted unless additional accommodations have been arranged prior. 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 contact staff for submission support.

This assignment is entirely on paper. Your submission (a single PDF) can be generated as follows:

1. Type your answers. We recommend LaTeX, the math typesetting language. Overleaf is a great tool to type in LaTeX.
2. 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.
3. 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.

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. 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

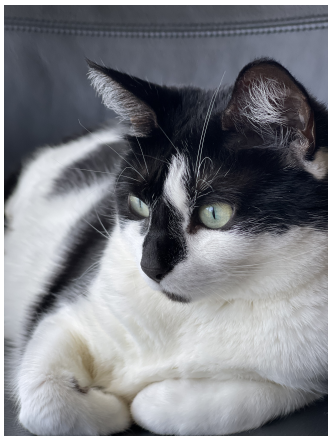
Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others, please include their names below.

Sampling

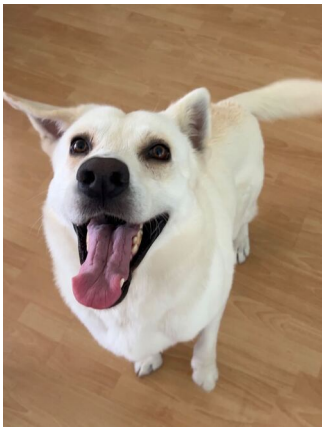
1. (7 points) Welcome to the Spring 2024 Data 100 Cutest Pets Contest! Course staff nominate their pets to participate in this contest. Students will vote on the cutest one among the nominations in the final exam.

   The nominees are:

(a) Appa (Matthew's cat)



(b) Pishi (Professor Norouzi's cat)



(c) Mimi (Shiny's dog)

   Course staff would like to predict the results for the official survey later in the semester by surveying students in the class now. This process is similar to polling that occurs before a political election.

   In this question, you are going to explore different sampling methods.

   (a) (3 points) Since her dog, Mimi, is nominated, Shiny would like to understand the class opinion before the contest. She decides to survey all Spring 2024 students enrolled in Data 100 by sending out an Ed announcement via email that asked students to choose the cutest from the three pets. You may assume no other students/users receive the survey. Shiny closes the survey 12 hours after sending it out.

   You can assume that all, and only, enrolled students are on Ed.

      i. (1 point) In Shiny's survey, which of the following is the population of interest?

      ◯ A. All UC Berkeley students

      ◯ B. All students enrolled in Data 100 across Spring 2024 and all previous semesters

      ● C. All students enrolled in Data 100 for Spring 2024

      ◯ D. All students who fill out Shiny's survey

ii. (1 point) In Shiny's survey, which of the following is the sampling frame?

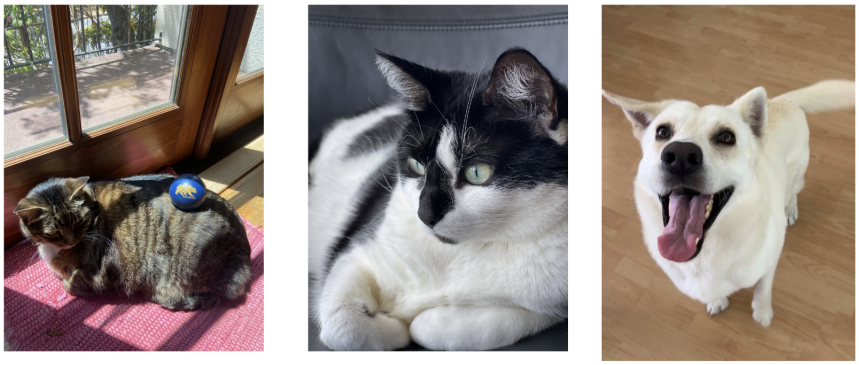
- ◯ A. All UC Berkeley students
- ◯ B. All students enrolled in Data 100 across Spring 2024 and all previous semesters
- ● C. All students enrolled in Data 100 for Spring 2024
- ◯ D. All students who fill out Shiny's survey

iii. (1 point) Which of the following is the sample?

- ◯ A. All UC Berkeley students
- ◯ B. All students enrolled in Data 100 across Spring 2024 and all previous semesters
- ◯ C. All students enrolled in Data 100 for Spring 2024
- ● D. All students who fill out Shiny's survey

(b) (4 points) In practice, we cannot get a 100% survey response rate, often because our population is too large, or because there is a time limit. In this case, very few students answered Shiny's survey before she closed it.

To get more data to predict the answer to the original question ("Which pet will win the Data 100 Cutest Pet Contest?"), Shiny decides on a different strategy: **she conducts the pre-contest survey in person in her discussion section that same week**. She then asks every student who attends the discussion that week for their opinion on the cutest of the three pets, by presenting the following slide:



i. (1 point) In this sampling scheme, which of the following is the population of interest?

- ◯ A. UC Berkeley students
- ◯ B. All students enrolled in Data 100 across Spring 2024 and all previous semesters
- ● C. All students enrolled in Data 100 for Spring 2024

◯ D. All students enrolled in Shiny's discussion section

◯ E. All students who fill out Shiny's pre-contest survey

ii. (1 point) In this sampling scheme, which of the following is the sampling frame?

◯ A. UC Berkeley students

◯ B. All students enrolled in Data 100 across Spring 2024 and all previous semesters

◯ C. All students enrolled in Data 100 for Spring 2024

● D. All students enrolled in Shiny's discussion section

◯ E. All students who fill out Shiny's pre-contest survey

iii. (1 point) Which of the following is the sample?

◯ A. UC Berkeley students

◯ B. All students enrolled in Data 100 across Spring 2024 and all previous semesters

◯ C. All students enrolled in Data 100 for Spring 2024

◯ D. All students enrolled in Shiny's discussion section

● E. All students who fill out Shiny's pre-contest survey

iv. (1 point) Which of the following best characterizes the sample?

◯ A. Simple Random Sample

● B. Convenience Sample

◯ C. Probability Sample

Properties of Simple Linear Regression

2. (7 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 $\hat{y} = \theta_0 + \theta_1 x$. (Note: In this problem we write (θ_0, θ_1) instead of (a, b) to more closely mirror the multiple linear regression model notation.)

   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 lecture, a residual e_i 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 $\sum_{i=1}^{n} e_i = 0$.

$e_i = y_i - (\bar{y} + r\sigma_y \frac{x_i - \bar{x}}{\sigma_x})$

$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \bar{y} - r\sigma_y \frac{x_i - \bar{x}}{\sigma_x})$

$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} - r\sigma_y \frac{1}{\sigma_x} \sum_{i=1}^{n} (x_i - \bar{x})$

$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - n\bar{y}$

$\sum_{i=1}^{n} e_i = n\bar{y} - n\bar{y} = 0$

   (b) (2 points) Using your result from part (a), prove that $\bar{y} = \bar{\hat{y}}$.

$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$

$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \rightarrow \bar{y} = \bar{\hat{y}}$

$\frac{\sum_{i=1}^{n} y_i}{n} \quad \frac{\sum_{i=1}^{n} \hat{y}_i}{n}$

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

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

$\hat{y} = \bar{y} + r\sigma_y \underbrace{\frac{\bar{x} - \bar{x}}{\sigma_x}}_{0}$

$\hat{y} = \bar{y}$

Therefore, the point $(\bar{x}, \bar{y})$ because when $x = \bar{x}$, $\hat{y} = \bar{y}$.

Properties of a Linear Model With No Constant Term

3. (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 (MSE) 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 estimating 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.

Use calculus to find the minimizing $\hat{\theta}$.

That is, simply prove that:

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

Hint: You can start by following the format of SLR in lecture 10 and replace the SLR model with the model defined above.

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

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

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

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

$$0 = \sum_{i=1}^{n} x_i y_i - \theta\sum_{i=1}^{n} x_i^2$$

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

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

MSE "Minimizer"

4. (8 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 (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**.

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

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

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

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

   $(Y_i - \theta x_i)(0) + (0 - x_i)(x_i)$

   $-x_i^2$

   $g_i(\theta) = \frac{2x_i^2}{n}$ non-negative, so convex function

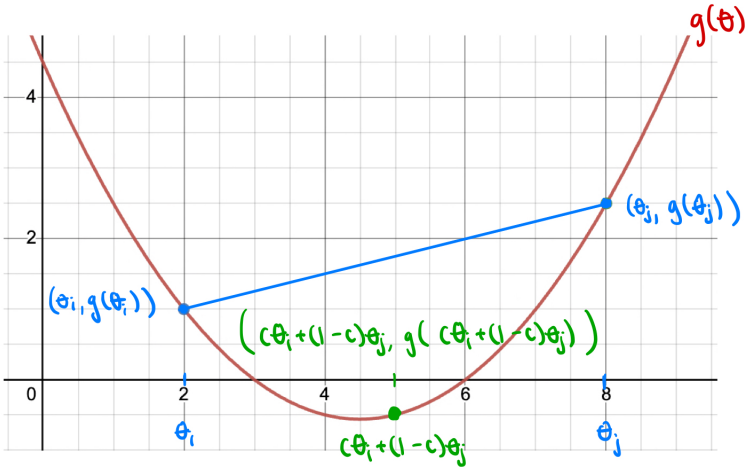
   (b) (3 points) Now that we have shown that each term in the summation of the MSE is a convex function, one might wonder if the entire summation is convex, given that it is a sum of convex functions.

   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 $0 \leq c \leq 1$.

The function g evaluated on any point between θ_i and θ_j will always lie at or below the secant line connecting $g(\theta_i)$ and $(g(\theta_j)$



See a graph in this Wikipedia article https://en.wikipedia.org/wiki/Convex_function.

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.

since you are adding two positive function $g(\theta)+h(\theta)$ will always be positive, therefore it will be a convex function

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 $n > 2$.

A sum of convex functions remains convex because each function satisfies the convexity property, and adding doesn't have any effect on the property. The combined function still ensures the weighted average of values lies below or on the secant line.

(c) (2 points) Remember from part (a) that the MSE can be written as:

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

We solve for its critical point by taking the gradient with respect to parameter θ and setting that expression to 0. Explain why this solution is guaranteed to minimize the MSE.

The solution is gurantee to minimize the MSE because convexity gurantees any local minimum is a global minimum. Due to the derivation being a convex function shows that $g_i(\theta)$ is convex. Therefore, when it's set to 0 you can find the global minimum of the MSE

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.