

---

## DSC 40A - Homework 3

Due: Wednesday, Jan 31 at 11:59pm

---

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59pm on the due date. You can use a slip day to extend the deadline by 24 hours.


Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it.

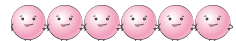
For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of 50 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

### Problem 1. Reflection and Feedback Form

 Make sure to fill out this [Reflection and Feedback Form, linked here](#) with your ucsd account for two points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

### Problem 2. Jensen Gap and Convexity

 Recall the definition of convex function, if  $f$  is convex, then  $f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2), \forall t \in [0, 1]$ . A general form of this is known as *Jensen's Inequality*: for a real convex function  $g$ , with positive weights  $a_i, \sum_i a_i = 1$ , we have

$$g\left(\sum_i a_i x_i\right) \leq \sum_i a_i g(x_i)$$

Recall that for a discrete random variable  $X$ , with possible outcome  $x_i$  and corresponding probability  $p_i$ , its expectation is defined as  $E[X] = \sum_i p_i x_i$ . The distance  $g(E[X]) - E[g(X)]$  when  $g$  is convex is known as *Jensen gap*.

Use the above facts to prove that, if  $X$  is a discrete random variable with all possible outcomes are positive, then

$$\ln E[X] - E[\ln X] \geq 0$$

### Problem 3. Combinations of Convex Functions

For each statement below, either prove the statement true using the *formal definition* of convexity from Lectures 6 and 7, or prove the statement false by finding a concrete counterexample.

- a)  The sum of two convex functions must also be convex.

- b) 🤔🤔🤔🤔 The difference of two convex functions must also be convex.

#### Problem 4. Conditions of Linear Regression

Often times in data science, one can create a decently accurate model using linear regression! It also has the added benefit of being fast, easy to interpret, and simple (compared to more complicated models like neural networks). However, there are some assumptions/conditions that need to be met in order for linear regression to work well.

In general, linear regression can be expressed as  $Y = \beta_0 + \beta_1 X + \epsilon_i$  where  $\beta_0, \beta_1$  are our unknown parameters, and  $\epsilon_i$  is a random variable that represents the error. Since this class only covers linear regression of the form  $H(\vec{x}) = w_0 + w_1 \vec{x}$ , let us explore the conditions that are unrelated to error: linearity and outliers.

Suppose you work on an avocado farm and construct the following dataset of month, average high temperature (in °F), and avocado yield (in hundreds), shown in Table 1.

| Month | Average High Temperature (°F) | Avocado Yield (hundreds) |
|-------|-------------------------------|--------------------------|
| 1     | 66                            | 18                       |
| 2     | 64                            | 22                       |
| 3     | 64                            | 19                       |
| 4     | 66                            | 17                       |
| 5     | 67                            | 14                       |
| 6     | 70                            | 10                       |
| 7     | 74                            | 6                        |
| 8     | 75                            | 2                        |
| 9     | 75                            | 3                        |
| 10    | 73                            | 6                        |
| 11    | 69                            | 10                       |
| 12    | 65                            | 18                       |

Table 1: Avocado Yield

Answer the following sub questions based on Table 1. Calculations by hand or by code are both acceptable as long as sufficient work is shown.

- a) 🤔🤔🤔🤔 Arguably, the most important condition for linear regression is linearity in the data, specifically between the  $y_i$ 's and the  $x_i$ 's. In general, it does not really make sense to fit a linear model to data that does not exhibit a linear relationship.

Perform linear regression for the linear models  $H(x^{(1)}) = a + bx^{(1)}$  and for  $H(x^{(2)}) = c + dx^{(2)}$ , where  $x^{(1)}$  is Month,  $x^{(2)}$  is Average High Temperature, and  $y$  is Avocado Yield. In the context of the data, explain the meaning of the model parameters you found for the two models.

- b) 🤔🤔🤔🤔 Additionally, you may recall the idea of a residual plot from DSC 10, where a residual is defined as  $y_i - H(x_i)$ . (Notice that least squares regression minimizes the mean squared residuals!) Draw the residual plots for the two models from part a), and use them to determine if there is linearity in the data.

- c) 🤔🤔🤔🤔 Another condition to consider checking when performing linear regression is outliers that may strongly affect the linear model. Outliers that strongly affect the model are called influential points.

Suppose that the recent storm caused your farm to flood, destroying all of your avocado trees. Alas, you have 0 avocados for January (Month = 1). Explain whether this outlier is an influential point in terms of the model parameters for the linear model  $y = c + dx_2$  where  $x_2$  is Average High Temperature and  $y$  is Avocado Yield.

- d) 🍌🍌🍌🍌 Outliers in the x-direction are called high leverage points. Suppose that February (Month = 2) turned out to be especially cold this winter, with an average high temperature of 55. Miraculously, your avocado trees managed to produce 3,400 avocados (Avocado Yield = 34)! Explain whether this high leverage point is an influential point.

### Problem 5. Six Data Points

🍌🍌🍌🍌🍌🍌 Suppose you have a data set of six data points whose coordinates are

$$(5, y_1), (5, y_2), (10, y_3), (10, y_4), (15, y_5), (15, y_6).$$

Define

$$\bar{y}_1 = \frac{y_1 + y_2}{2}, \quad \bar{y}_2 = \frac{y_3 + y_4}{2}, \quad \bar{y}_3 = \frac{y_5 + y_6}{2}.$$

Show that the least squares regression line fitted to all six data points is identical to the least squares regression line fitted to the three points  $(5, \bar{y}_1)$ ,  $(10, \bar{y}_2)$ ,  $(15, \bar{y}_3)$ .

### Problem 6. Holler for Haaland

Suppose that in 2018 we collected data about 200 randomly sampled professional soccer players to find out how many goals they scored that year and their corresponding market value, which is the amount of money they would be sold for if another team wanted them. In the collected survey data, we find that the goals scored had a mean of 31 and a standard deviation of 6. We then use least squares to fit a linear prediction rule  $H(x) = w_0 + w_1x$ , which we will use to help other players predict their market value in millions of dollars ( $y$ ) based on how many goals they scored ( $x$ ).

- a) 🍌🍌🍌🍌 Erling Haaland was one of the professional players in our sample. Suppose that in 2018, he scored 16 goals and his market value was only 20 million, the smallest market value in our sample.

In 2019, Haaland moved to the Bundesliga, a much more competitive league. In 2019, he again scored 16 goals, but his market value shot up to 80 million!

Suppose we create two linear prediction rules, one using the dataset from 2018 when Haaland had a market value of 20 million and another using the dataset from 2019 when Haaland had a market value of 80 million. Assume that all other players scored the same amount of goals and had the same market value in both datasets. That is, only this one data point is different between these two datasets.

Suppose the optimal slope and intercept fit on the first dataset (2018) are  $w_1^*$  and  $w_0^*$ , respectively, and the optimal slope and intercept fit on the second dataset (2019) are  $w_1'$  and  $w_0'$ , respectively.

What is the difference between the new slope and the old slope? That is, what is  $w_1' - w_1^*$ ? The answer you get should be a number with no variables.

**Note:** Since we want to predict market value in millions of dollars, use 20 instead of 20,000,000 for Haaland's market value in 2018.

**Hint:** There are many equivalent formulas for the slope of the regression line. We recommend using this one for this problem:

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- b) 🍌🍌🍌🍌 Let  $H^*(x)$  be the linear prediction rule fit on the 2018 dataset (i.e.  $H^*(x) = w_0^* + w_1^*x$ ) and  $H'(x)$  be the linear prediction rule fit on the 2019 dataset (i.e.  $H'(x) = w_0' + w_1'x$ ).

Consider two other players, Lozano and Messi, neither of whom were part of our original sample in 2018. Suppose that in 2022, Lozano had 18 goals and Messi had 25 goals.

Both Lozano and Messi want to try and use one of our linear prediction rules to predict their market value for next year.

Suppose they both first use  $H^*(x)$  to determine their predicted yields as per the first rule (when Haaland was only worth 20 million). Then, they both then use  $H'(x)$  to determine predicted yields as per the second rule (when Haaland was worth 80 million).

Whose prediction changed more by switching from  $H^*(x)$  to  $H'(x)$  – Lozano’s or Messi’s?

**Hint:** You should draw a picture of both prediction rules,  $H^*(x)$  and  $H'(x)$ . You already know how the slope of these lines differs from part (b). Can you identify a point that each line must go through?

c) 🤔🤔🤔🤔 In this problem, we’ll consider how our answer to part (b) might have been different if Haaland had more goals in both 2018 and 2019.

- Suppose Haaland instead had 31 goals in both 2018 and 2019. If his market value increased from 2018 to 2019, and everyone else’s data stayed the same, which slope would be larger:  $H^*(x)$  or  $H'(x)$ ?
- Suppose Haaland instead had 45 goals in both 2018 and 2019. If his market value increased from 2018 to 2019, and everyone else’s data stayed the same, which slope would be larger:  $H^*(x)$  or  $H'(x)$ ?

You don’t have to actually calculate the new slopes, but given the information in the problem and the work you’ve already done, you should be able to answer the question and give brief justification.