

---

## DSC 40A - Homework 2

Due: Tuesday, July 18 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.

### Notes:

- This homework involves some long calculations. You may use a calculator (Python is recommended!), but you may not use any tools that perform regression for you. Show your work by showing the mathematical expression you're evaluating with a calculator, and the numerical result; you don't need to show every intermediate step.

### Problem 1. Averaged Data Points and Their Impact on Regression Lines



Suppose you have a data set of eight data points whose coordinates are

$$(5, y_1), (5, y_2), (10, y_3), (10, y_4), (15, y_5), (15, y_6), (20, y_7), (20, y_8).$$


Define

$$\bar{y}_1 = \frac{y_1 + y_2}{2}, \quad \bar{y}_2 = \frac{y_3 + y_4}{2}, \quad \bar{y}_3 = \frac{y_5 + y_6}{2}, \quad \bar{y}_4 = \frac{y_7 + y_8}{2}.$$

Show that the least squares regression line fitted to all eight data points is identical to the least squares regression line fitted to the four points  $(5, \bar{y}_1), (10, \bar{y}_2), (15, \bar{y}_3), (20, \bar{y}_4)$ .

### Problem 2. 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.

### Problem 3. Professional Wrestling

- a) 🥑🥑🥑 In professional wrestling, wrestlers build broad and diverse move sets in order to increase fan interest and accordingly, their merchandise sales. Tony Khan, a wrestling promoter, asked one of his assistants to log wrestlers' distinct move sets for a given year and generate a lookup of wrestlers, the count of their distinct moves  $x$ , and their merchandise sales  $y$  in thousands of dollars.

Wrestler	distinct moves ( $x$ )	merchandise sales ( $y$ )
Leyla Hirsch	33	200
Adam "Hangman" Page	20	190
Dr. Britt Baker, D.M.D.	32	260
Kenny Omega	50	320
Satnam Singh	5	110

What linear relationship  $y = c_0 + c_1x$  best describes merchandise sales as a function of the number of distinct moves a wrestler has? Give exact values for  $c_0$  and  $c_1$  (do not round).

- b) 🥑🥑 Now, let's interpret the meaning of the linear function  $y = c_0 + c_1x$  that you found in part (a). What does  $c_1$  represent in terms of merchandise sales?
- c) 🥑🥑🥑 What is the mean squared error,  $MSE_x$ , for this data set, using the line you found in part (a)? Round your final answer to three decimal places.
- d) 🥑🥑🥑 Khan knows that Danny Cage, head instructor at the Monster Factory wrestling school, is one of the best trainers in the business. So he decides to quantify the value of training with Danny Cage (in years) for a given wrestler's merchandise sales. Wrestling school teaches not only distinct moves, but also other skill sets like developing a character, microphone skills, etc. For each of the five wrestlers noted, Khan recorded the number of years a wrestler has trained with Danny Cage,  $z$ , and the corresponding merchandise sales,  $y$  (in thousands of dollars).

Wrestler	years with Danny Cage ( $z$ )	merchandise sales ( $y$ )
Leyla Hirsch	12.3	200
Adam "Hangman" Page	8	190
Dr. Britt Baker, D.M.D.	12	260
Kenny Omega	18	320
Satnam Singh	3	110

What linear relationship  $y = d_0 + d_1z$  best describes merchandise sales as a function of the years training with Danny Cage? Give exact values for  $d_0$  and  $d_1$  (do not round).

- e) 🥑🥑 What is the mean squared error,  $MSE_z$ , for this data set, using the line you found in part (d)? Round your final answer to three decimal places.

### Problem 4. Vector Calculus Involving Matrices

Let  $X$  be a fixed matrix of dimension  $m \times n$ , and let  $\vec{w} \in \mathbb{R}^n$ . In this problem, you will show that the gradient of  $\vec{w}^T X^T X \vec{w}$  with respect to  $\vec{w}$  is given by

$$\frac{d}{d\vec{w}}(\vec{w}^T X^T X \vec{w}) = 2X^T X \vec{w}.$$

Let  $\vec{r}_1, \vec{r}_2, \dots, \vec{r}_m$  be the column vectors in  $\mathbb{R}^n$  that come from transposing the rows of  $X$ . For example, if

$$X = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 3 & 1 \end{bmatrix}, \text{ then } \vec{r}_1 = \begin{bmatrix} 1 \\ 4 \\ 7 \end{bmatrix} \text{ and } \vec{r}_2 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}.$$

a) 🥑🥑🥑🥑 Show that, for arbitrary  $X$  and  $\vec{w}$ , we can write

$$\vec{w}^T X^T X \vec{w} = \sum_{i=1}^m (\vec{r}_i^T \vec{w})^2.$$

*Hint:* First, show that we can write  $\vec{w}^T X^T X \vec{w}$  as a dot product of two vectors. Then, try and re-write those vectors in terms of  $\vec{r}_1, \vec{r}_2, \dots, \vec{r}_m$  and  $\vec{w}$ .

Now that we have written

$$\vec{w}^T X^T X \vec{w} = \sum_{i=1}^m (\vec{r}_i^T \vec{w})^2$$

we can apply the chain rule, along with the result of part (a) above, to conclude that

$$\begin{aligned} \frac{d}{d\vec{w}} (\vec{w}^T X^T X \vec{w}) &= \sum_{i=1}^m 2(\vec{r}_i^T \vec{w}) \frac{d}{d\vec{w}} (\vec{r}_i^T \vec{w}) \\ &= \sum_{i=1}^m 2(\vec{r}_i^T \vec{w}) \vec{r}_i \end{aligned}$$

b) 🥑🥑🥑🥑 Next, show that, for arbitrary  $X$  and  $\vec{w}$ , we can write

$$2X^T X \vec{w} = \sum_{i=1}^m 2(\vec{r}_i^T \vec{w}) \vec{r}_i$$

*Hint 1:* Use the column-mixing interpretation of matrix-vector multiplication from Module 10.

*Hint 2:* It is likely that you'll need to use one of your intermediate results from part (a).

Since you've shown that  $\frac{d}{d\vec{w}} (\vec{w}^T X^T X \vec{w})$  and  $2X^T X \vec{w}$  are both equal to the same expression,  $\sum_{i=1}^m 2(\vec{r}_i^T \vec{w}) \vec{r}_i$ , you have proven that they are equal to one another, i.e. that

$$\frac{d}{d\vec{w}} (\vec{w}^T X^T X \vec{w}) = 2X^T X \vec{w}$$

as desired.

## Problem 5. Sums of Residuals

Let's start by recalling the idea of orthogonality from linear algebra. This will allow us to prove a powerful result regarding linear regression.

Two vectors are **orthogonal** if their dot product is 0, i.e. for  $\vec{a}, \vec{b} \in \mathbb{R}^n$ :

$$\vec{a}^T \vec{b} = 0 \implies \vec{a}, \vec{b} \text{ are orthogonal}$$

Orthogonality is a generalization of perpendicularity to multiple dimensions. (Two orthogonal vectors in 2D meet at a right angle.)

Suppose we want to represent the fact that some vector  $\vec{b}$  is orthogonal to many vectors  $\vec{a}_1, \vec{a}_2, \dots, \vec{a}_d$  all at once. It turns out that we can do this by creating a new  $n \times d$  matrix  $A$  whose columns are the vectors  $\vec{a}_1, \vec{a}_2, \dots, \vec{a}_d$ , and writing  $A^T \vec{b} = 0$ .

For instance, suppose  $\vec{a}_1 = \begin{bmatrix} 8 \\ 4 \\ -2 \end{bmatrix}$ ,  $\vec{a}_2 = \begin{bmatrix} 3 \\ 5 \\ 1 \end{bmatrix}$ , and  $\vec{b} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$ . Then,

$$A = \begin{bmatrix} 8 & 3 \\ 4 & 5 \\ -2 & 1 \end{bmatrix} \implies A^T = \begin{bmatrix} 8 & 4 & -2 \\ 3 & 5 & 1 \end{bmatrix}$$

Note that the product  $A^T \vec{b}$  involves taking the dot product of each row in  $A^T$  with  $\vec{b}$ .

$$A^T \vec{b} = \begin{bmatrix} 8 & 4 & -2 \\ 3 & 5 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 8(1) + 4(-1) + (-2)(2) \\ 3(1) + 5(-1) + 2(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Since  $A^T \vec{b} = \vec{0}$ , then it is the case that  $\vec{b}$  is orthogonal to each row of  $A^T$ , and hence orthogonal to each column of  $A$ .

(We will not use this fact in this class, but if  $A^T \vec{b} = 0$ , it also means that  $\vec{b}$  is orthogonal to the **column space** of  $A$ , which is the space of all linear combinations of the columns of  $A$ . As a good exercise in linear algebra, see if you can prove this result!)

- a) 🥝🥝🥝 Suppose  $\vec{1}$  is a vector in  $\mathbb{R}^n$  containing the value 1 for each element, i.e.  $\vec{1} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$ .

For any other vector  $\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}$ , what is the value of  $\vec{1}^T \vec{b}$ , i.e. what is the dot product of  $\vec{1}$  and  $\vec{b}$ ?

- b) 🥝🥝🥝🥝 Now, consider the typical multiple regression scenario where our prediction rule has an intercept term ( $w_0$ ):

$$H(\vec{x}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \dots + w_d x^{(d)}.$$

For this scenario,  $X$  is a  $n \times (d+1)$  design matrix,  $\vec{y} \in \mathbb{R}^n$  is an observation vector, and  $\vec{w} \in \mathbb{R}^{(d+1)}$  is the parameter vector. We'll use  $\vec{w}^*$  to denote the optimal parameter vector, or the one that satisfies the normal equations.

Show that the error vector,  $\vec{e} = \vec{y} - X\vec{w}^*$ , is orthogonal to the columns of  $X$ .

*Hint:* Use the normal equations and the definition of orthogonality to the columns of a matrix given in the problem description.

- c) 🥝🥝🥝🥝🥝 We define the  $i$ th **residual** to be the difference between the actual and predicted values for individual  $i$  in our data set. In other words, the  $i$ th residual  $e_i$  is

$$e_i = (\vec{y} - X\vec{w}^*)_i$$

Here,  $(\vec{y} - X\vec{w}^*)_i$  is referring to element  $i$  of the vector  $\vec{y} - X\vec{w}^*$ . We use the letter  $e$  for residuals because residuals are also known as errors.

Using what you learned in parts (a) and (b), show that for multiple linear regression with an intercept term, the sum of the residuals is zero, that is

$$\sum_{i=1}^n e_i = 0.$$