

---

**DSC 40A - Homework 3**  
Due: Sunday, April 24, 2022 at 11:59pm PDT

---

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 PDT on Sunday.

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.

This policy also means that you **should not post or answer homework-related questions on Piazza**, which is a written medium. This includes private posts to instructors. Instead, when you need help with a homework question, talk to a classmate or an instructor in their office hours.

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. The point value of each problem or sub-problem is indicated by the number of avocados shown.

**Problem 1. Fahrenheit or Celsius?**

You wish to establish a linear relationship between the temperature in San Diego,  $x$ , and the temperature in Osaka,  $y$ .

The monthly temperature data from Homework 1 is given in the table below.

|                                      |    |    |    |    |    |    |    |    |    |    |    |    |
|--------------------------------------|----|----|----|----|----|----|----|----|----|----|----|----|
| San Diego, US ( $^{\circ}\text{F}$ ) | 66 | 66 | 67 | 69 | 69 | 72 | 76 | 77 | 77 | 74 | 70 | 66 |
| Osaka, Japan ( $^{\circ}\text{C}$ )  | 9  | 10 | 14 | 20 | 25 | 28 | 32 | 33 | 29 | 23 | 18 | 12 |

Recall that the temperatures in these places are measured in different units, Fahrenheit for San Diego and Celsius for Osaka. You'd like the relationship that you find to be in degrees Celsius. One way to do this is to convert all the San Diego temperatures to Celsius before performing least squares regression.

You friend Skip from Homework 1 is also back, and he thinks you can skip some of that work: "Why don't we perform least squares regression first, with  $x$  in Fahrenheit, and then do the Fahrenheit to Celsius conversion for both the slope and the intercept in the regression coefficients? That way we only need to do the conversion twice instead of for each data point."

- a)  Is Skip correct that you'll get the same regression coefficients either way? Show your work. Recall that if a temperature  $t$  is measured in degrees Fahrenheit, the equivalent temperature in degrees Celsius is given by  $g(t) = \frac{5}{9} \times (t - 32)$ . If Skip is not correct, can you think of a different shortcut that allows you to get the same regression coefficients without converting each data point to Celsius?
- b)  More generally, suppose we want to do least squares regression for a linear relationship:  $y = w_1x + w_0$ . How do the slope  $w_1$  and the intercept  $w_0$  of the regression line change if we replace  $x$  with a linear transformation  $f(x) = ax + b$ ?

**Problem 2. Avocado farmers**

A few years ago, a millennial decided to grow their own avocados to keep up with their rapid consumption of avocado toast. They quickly noticed that rain is fundamental to avocado production. Over the last four avocado seasons, they have recorded the number of rainy days that season,  $x$ , and the number of avocados produced that season,  $y$ .

|   |   |    |   |    |
|---|---|----|---|----|
| x | 1 | 3  | 0 | 4  |
| y | 8 | 12 | 4 | 17 |

- a) 🥑🥑 What linear relationship  $y = b_0 + b_1x$  best describes the number of avocados as a function of the number of rainy days? What is the mean squared error,  $R_{sq}((b_0, b_1); D_x)$  (where  $b_0$  and  $b_1$  are the optimal choices), for this data set?
- b) 🥑🥑 The millennial reconsiders their approach to this problem, and decides that the number inches of rain,  $z$ , may be a better predictor than the number of rainy days,  $x$ .

|   |   |    |   |    |
|---|---|----|---|----|
| z | 3 | 7  | 1 | 9  |
| y | 8 | 12 | 4 | 17 |

What linear relationship  $y = d_0 + d_1z$  best describes the number of avocados as a function of the number of inches of rainfall? What is the mean squared error,  $R_{sq}((d_0, d_1); D_z)$  (where  $d_0$  and  $d_1$  are the optimal choices), for this data set?

- c) 🥑🥑🥑🥑 In the above example, notice that  $R_{sq}((b_0, b_1); D_x) = R_{sq}((d_0, d_1); D_z)$ , so the mean squared error is the same if we use the predictor  $x$  or the predictor  $z$ . This happens because the number of inches of rainfall  $z$  is linearly related to the number of rainy days  $x$  by the following formula:  $z = 2x + 1$ .

Prove in general that the mean squared error does not change if we use as a predictor any linear transformation of  $x$ . For an arbitrary data set  $y_1, \dots, y_n$ , show that if  $z = c_0 + c_1x$  for some constants  $c_0, c_1 \neq 0$ , then  $R_{sq}((b_0, b_1); D_x) = R_{sq}((d_0, d_1); D_z)$ .

*Hint:* Start by expressing  $d_0, d_1$  from the relationship  $y = d_0 + d_1z$  in terms of  $c_0, c_1, b_0, b_1$ , where  $y = b_0 + b_1x$ .

### Problem 3.

🥑🥑🥑🥑 Suppose you have a data set of six data points, with two data points at each of three different  $x$  values,  $x = 5, x = 10$ , and  $x = 15$  (That is, we have  $x_i \in \{5, 5, 10, 10, 15, 15\}$  for  $i = 1, \dots, 6$ ). Show that the least squares regression line fitted to these six data points is identical to the least squares regression line fitted to the three points  $(5, \bar{y}_1), (10, \bar{y}_2), (15, \bar{y}_3)$  where  $\bar{y}_1, \bar{y}_2, \bar{y}_3$  represent the means of the two  $y$  values at each of the  $x$  values.

### Problem 4.

Suppose that we survey 100 randomly sampled avocado farmers to find out the number of avocado trees on their farm and the total number of avocados produced by those trees in a given year. In the collected survey data, we find that the number of avocado trees has a mean of 80 and a standard deviation of 30. Then we use least squares to fit a linear function  $H(x) = w_1x + w_0$ , which we will use to help other farmers predict their avocado yield based on the number of trees they have.

- a) 🥑 Is a linear function ideal here, or is there another function form for  $H(x)$  that you think would better model this scenario? Explain.
- b) 🥑🥑🥑🥑 Now suppose that one particular farmer from the 100 sampled farmers was a very poor farmer. His 15 avocado trees yielded only 150 avocados, the smallest total number reported by any of the survey participants. The farmer then has a conversation with a millennial, who enlightens him

by pointing out that avocado yield can be increased by additional watering. The farmer waters the avocado trees more frequently, and the next year, increases the yield from his 15 trees to 650 avocados. If a new linear predictor  $H'(x) = w'_1x + w'_0$  is fit using the new data (with only this one farmer's yield changed), what is the difference  $w'_1 - w_1$  between the new slope and the old?

- c) 🥑🥑🥑 Suppose some farmers who were not surveyed plan to use the data from this survey to predict how much yield to expect from their avocado farms. Who would be more affected by changing prediction rules from  $H(x)$  to the new predictor  $H'(x)$ : someone with 20 avocado trees or someone with 40 avocado trees?
- d) 🥑🥑🥑 If we had increased a different farmer's yield instead, would the original line or the new line have a steeper slope? How can you tell, based on the farmer? Is it possible that by increasing a farmer's yield, we keep the slope the same?
- e) 🥑🥑🥑🥑 In this problem, since  $x$  represents number of avocado trees and  $y$  represents yield, it is reasonable to expect that the regression line should go through the origin. In other words, if there are no avocado trees, there are no avocados. We can *force* our prediction rule to go through the origin by using a prediction rule of the form  $w_1x$  instead of the usual  $w_0 + w_1x$ . Minimize the mean square error to find the best choice of  $w_1$ , in the case that we force our prediction rule to be of the form  $w_1x$ .

**Problem 5.** 🥑🥑🥑🥑🥑🥑

|   |    |    |    |    |    |    |    |    |    |    |    |    |
|---|----|----|----|----|----|----|----|----|----|----|----|----|
| x | 66 | 66 | 67 | 69 | 69 | 72 | 76 | 77 | 77 | 74 | 70 | 66 |
| y | 9  | 10 | 14 | 20 | 25 | 28 | 32 | 33 | 29 | 23 | 18 | 12 |

For the data above, apply a suitable transformation then use linear regression to find the best fitting curve of the form:

$$x = \sqrt{ay^2 + by}.$$

|                 |     |     |     |     |     |     |     |     |     |     |     |     |
|-----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $y$             | 9   | 10  | 14  | 20  | 25  | 28  | 32  | 33  | 29  | 23  | 18  | 12  |
| $\frac{x^2}{y}$ | 484 | 435 | 320 | 238 | 190 | 185 | 180 | 179 | 204 | 238 | 272 | 363 |