
DSC 40A - Homework 3
Due: Friday, October 21, 2022 at 2:00PM 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 2:00PM PT on the due date. Make sure to correctly assign pages to Gradescope when submitting.

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

There are 3 required problems in this homework (Problems 1, 2, and 3). We have provided 2 additional practice problems at the very end (Problems 4 and 5) that will serve as good prep for the midterm, but you are not required to submit them (and we will not grade them).

Problem 1. We All Scream For Ice Cream!

Six days ago, you opened an ice cream shop at La Jolla Cove. Since then you've kept track of your sales per day (in dollars) and the daily high temperature (in degrees Celsius, °C). Here are all of the data points you collected:

Daily High Temperature (°C)	Sales (\$)
25	110
26	75
20	85
22	100
29	120
28	95

Throughout this question, your goal is to predict sales (y) given a daily high temperature (x).

- a) 🥑🥑🥑 Using the formulas derived in Lecture 6, determine the slope and intercept of the best linear prediction rule $H(x) = w_0 + w_1x$, according to least squares regression.

You may use a calculator (including Python), but you cannot use any tools that perform regression for you. Show all of your work.

- b) 🥑 In Lecture 8, we re-wrote the formula for the best slope, w_1^* , as

$$w_1^* = r \frac{\sigma_y}{\sigma_x}$$

where r is the correlation coefficient, σ_x is the standard deviation of the x values and σ_y is the standard deviation of the y values.

Calculate the values of r , σ_x , σ_y , using the formulae we discussed in Lecture 8. Then, use this new formula for w_1^* and verify that you get the same result as you did in part a.

- c) 🥑🥑 In parts a and b of this problem, you found the best slope w_1^* and intercept w_0^* through least squares regression. This means that your prediction rule is of the form

$$H^*(x) = w_0^* + w_1^*x$$

Determine $\sum_{i=1}^6 H^*(x_i)$ for our dataset of 6 points. The result you get back should be an integer, or very close to one (due to rounding issues).

Compare the result you get to $\sum_{i=1}^6 y_i$. What do you notice?

Hint: A good way to do this efficiently is to open a Jupyter Notebook, define a function that implements the prediction rule, and call that function 6 times, rather than writing out your formula 6 times.

- d) 🥑🥑🥑 Prove that for any dataset in general (not just the one in the question!) that

$$\sum_{i=1}^n y_i = \sum_{i=1}^n H^*(x_i)$$

Hint: This may seem daunting, but it turns out that we did most of this proof implicitly in lecture. Go back to the slides for Lecture 6 when we were minimizing $R_{sq}(w_0, w_1)$. One of the intermediate steps involved computing $\frac{\partial R_{sq}}{\partial w_0}$ and setting it equal to 0. We know that our prediction rule H^* satisfies the condition that $\frac{\partial R_{sq}}{\partial w_0} = 0$, so you can safely assume that equation still holds true here. You should take the expression for $\frac{\partial R_{sq}}{\partial w_0}$, plug in (w_0^*, w_1^*) , set it equal to 0, and arrive at the desired conclusion with just a little bit of algebraic manipulation. Don't spend too long on this question.

- e) 🥑🥑 Let's consider a more basic prediction rule. Suppose we want to instead use the prediction rule $H(x) = \alpha$, where α is some constant. To find the best such prediction rule, you minimize mean squared error $\frac{1}{n} \sum_{i=1}^n (y_i - H(x_i))^2$.

Suppose α^* is the parameter that minimizes mean squared error for this prediction rule. What is α^* – both in general for any dataset $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and for our dataset in particular?

Hint: It is *not* the result from part a or b. You should not need to use any calculus to determine the answer.

- f) 🥑🥑🥑 Let's consider another different prediction rule. Suppose we instead want to use the prediction rule $H(x) = \beta x$, where β is some constant (note that this is equivalent to finding a linear prediction rule that has no intercept). To find the best such prediction rule, you again minimize mean squared error $\frac{1}{n} \sum_{i=1}^n (y_i - H(x_i))^2$.

Suppose β^* is the parameter that minimizes mean squared error for this prediction rule. What is β^* – both in general for any dataset $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and for our dataset in particular?

Hint: Unlike in the previous subpart, you will need to use calculus (similar to questions on Homework 2). Again, the result is not the result you found in part b. When determining the value of β^* for our particular dataset, please use a calculator.

- g) 🥑🥑🥑 Now, let's say that for the next 5 days, the temperature increases at a constant rate with values of 29, 30, 31, 32, 33 (In °C). However, the sales on all these days remain at a constant value of \$120. Taking these 5 new data points, what do you think will happen to the values of w_0^* and w_1^*x . Why do you think this change happens?

Problem 2.

Suppose that in 2021 we surveyed 200 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 100 and a standard deviation of 20. We then use least squares to fit a linear prediction rule $H(x) = w_0 + w_1x$, 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 200 sampled farmers, named Billy, was a very poor farmer. In 2021, his 120 avocado trees yielded only 200 avocados, the smallest total number reported by any of the survey participants.

Billy then has a conversation with a millennial, who enlightens him by pointing out that avocado yield can be increased by additional watering. Billy waters the avocado trees more frequently, and in the year 2022, his 120 trees yielded 400 avocados – wow!

Suppose we create two linear prediction rules, one using the dataset from 2021 when Billy's trees yielded 200 avocados and another using the dataset from 2022 when Billy's trees yielded 400 avocados. **Assume that all other farmers had the same number of trees and same avocado yield in both 2021 and 2022**, i.e. that only one data point is different between these two datasets.

Suppose the optimal slope and intercept fit on the first dataset are w_1^* and w_0^* , respectively, and the optimal slope and intercept fit on the second dataset 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.

Hints:

- We've seen multiple formulas for the slope of the best fit line; they could all work here, but we recommend you use the form

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

This looks slightly different than both of the forms discussed in Lecture 6, but you can prove that it's equal to both of those using one of the key facts from the end of Lecture 6.

- Note that the denominator in the calculation for both slopes does not change, since all of our x_i 's (number of avocado trees) remain untouched. In fact, only one of the n elements in the sum for the numerator changes between w_1^* and w_1' , and that is the term in the sum corresponding to Billy. As such, we recommend you break the numerator for w_1^* and w_1' into two pieces – one sum over the $n - 1$ farmers who are not Billy, plus the term $(x_i - \bar{x})y_i$ for Billy.
- The correct answer is of the form $+\frac{1}{c}$, where c is a positive integer. (The only purpose this hint serves is to give you some sort of indication if you were on the right track; it doesn't make solving the problem any easier.)

- c) 🥑🥑🥑 Let $H^*(x)$ be the linear prediction rule fit on the 2021 dataset (i.e. $H^*(x) = w_0^* + w_1^*x$) and $H'(x)$ be the linear prediction rule fit on the 2022 dataset (i.e. $H'(x) = w_0' + w_1'x$). Everything from the previous part carries over into this part as well.

Consider two new farmers, Jung and Shannon, neither of whom were part of the survey data in 2020 or 2021. Jung has 20 avocado trees and Shannon has 40 avocado trees.

Both Jung and Shannon want to try and use one of our linear prediction rules to predict their avocado yield for next year.

Suppose they both first use $H^*(x)$ to determine their predicted yields as per the first rule (when Billy only yielded 200 avocados). Just to see what happens, they both then use $H'(x)$ to determine predicted yields as per the second rule (when Billy yielded 400 avocados).

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

Hint: You should draw out both prediction rules, $H^*(x)$ and $H'(x)$. To do so, you should use the facts that

- We know the slope of $H'(x)$ is more than the slope of $H^*(x)$ (as per part b).
- As we saw in class, for any dataset, the point (\bar{x}, \bar{y}) is on the regression line. Given the information in the problem, you can determine the relative positions of \bar{y} (and hence $H(\bar{x})$) for both datasets/prediction rules.

- d) 🥑🥑🥑 As we saw, when Billy’s yield changed from from 200 in 2021 to 400 in 2022, the slope of the new prediction rule $H'(x)$ became more than the slope of the old prediction rule $H^*(x)$.

In each of the following cases, suppose Billy’s avocado yield still increases from 2021 to 2022.

- If Billy instead had 80 avocado trees, which slope would be larger: $H^*(x)$ or $H'(x)$?
- If Billy instead had 100 avocado trees, 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.

Hint: It might help to look at the work you did in part b for guidance. Alternatively, draw a picture.

Problem 3. Least Absolute Deviation Regression

In this class, we explored least squares regression and defined it as the problem of finding the values of w_1 (slope) and w_0 (intercept) that minimize the function

$$R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$$

The resulting values are called w_0^* and w_1^* , and the resulting prediction function is $H^*(x) = w_0^* + w_1^*x$.

What is important to notice here is that we used the squared loss function, $(y_i - (w_0 + w_1 x_i))^2$ as our metric for deviation. What if we used a different loss function?

We are going to introduce another type of linear regression: least absolute deviation (LAD) regression. We will define least absolute deviation regression in terms of the absolute loss function rather than the squared loss function to measure how far away our predictions are from the data. That is, we will try to instead minimize

$$R_{abs}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n |y_i - (w_0 + w_1 x_i)|$$

Unlike with least squares regression, with LAD regression we cannot just take the gradient of R_{abs} , set it equal to zero, and solve for the values of w_0 and w_1 . In order to generate the optimal LAD regression line we are going to leverage a very useful theorem:

If you have a data set with n data points in \mathbb{R}^k , where $k \leq n$, then one of the optimal LAD regression lines must pass through k data points.

Notice that unlike with least squares regression, the LAD regression line may not be unique!

This theorem is useful to us because it allows us to adopt a very conceptually simple, albeit not very efficient, strategy to compute an optimal LAD regression line.

Our data will be in \mathbb{R}^2 , because we have two parameters to tweak (w_0 and w_1). Since our data is in \mathbb{R}^2 , we will generate all possible unique pairs of points (since one of the optimal LAD regression lines must pass through 2 data points) and calculate the slope w_1 and intercept w_0 of the line between each pair. Then we'll just select which w_0, w_1 pair among these finite options has the smallest value of $R_{abs}(w_0, w_1)$. This is guaranteed by the theorem to be an optimal LAD regression line.

- a) 🥑🥑 If you are given n data points, how many pairs of points are there? Give your answer in terms of n .

Hint: Try it out on some small values of n and look for a pattern. Note that if you have two data points (x_1, y_1) and (x_2, y_2) , this counts as only one pair of points because the line from (x_1, y_1) and (x_2, y_2) is the same as the line from (x_2, y_2) to (x_1, y_1) .

- b) 🥑🥑🥑🥑 Let's find the LAD line. Recall from the problem description the procedure outlined to generate an optimal LAD regression line. In the supplemental Jupyter Notebook, [which can be accessed by clicking this link](#), functions to generate all possible unique pairs of points and the respective lines for these unique pairs are already implemented for you. You will need to implement two more functions.
- The first, `mean_absolute_error`, should calculate the mean absolute error given the data and the values of w_0 and w_1 that define the line between a given pair of points.
 - The second, `find_best_line`, should pick the best (w_0, w_1) pair based on whichever has the lowest mean absolute error. If multiple (w_0, w_1) pairs have the same lowest mean absolute error, you can select any one of them.

Turn in screenshots of both of these functions as well as the best (w_0, w_1) pair calculated for the data.

- c) 🥑🥑 At the top of the notebook, which can be found [at this link](#), we already defined a function that calculates the least squares regression line for you. Now that you've calculated the least absolute deviation regression line, let's plot both of these lines together to see the difference! Generate a scatter plot with the data in black, the least squares line in blue, and the LAD line red. Turn in a picture of your plot.
- d) 🥑🥑 Given your knowledge of the loss functions behind least absolute deviation and least squares regression, provide one advantage and one disadvantage of using LAD over least squares for regression.

Optional Problems

The problems in this section are entirely optional. You don't have to do them, and we will not grade them. Instead, they're here for extra practice with linear regression.

Problem 4. Feet or Meters?

You wish to establish a linear relationship between the average height of girls age 2 to 12, x , and the average height of boys in the same age range, y .

The average height data by age is given in the table below.

Age	2	3	4	5	6	7	8	9	10	11	12
Avg Height of Girls in Feet, x	2.83	3.08	3.33	3.5	3.92	4	4.17	4.42	4.58	4.75	5
Avg Height of Boys in Meters, y	0.89	0.96	1.02	1.09	1.16	1.22	1.28	1.34	1.4	1.48	1.54

The heights are measured in different units, feet for girls and meters for boys. You'd like the relationship that you find to be in meters only. One way to do this is to convert all the girls' heights to meters before performing least squares regression.

Your friend Skip thinks you can skip some of that work: "Why don't we perform least squares regression first, with x in feet, and then do the feet to meters 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. If Skip is not correct, is there a different shortcut that allows you to get the same regression coefficients without converting each data point to meters? Recall that if a height h is measured in feet, the equivalent height in meters is given by $g(h) = 0.3048 \cdot h$.
- b) 🥑🥑🥑 More generally, suppose we want to do least squares regression for a linear relationship: $y = w_0 + w_1x$. 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 5. Restaurants

There are a lot of factors that help determine whether a restaurant will be successful or not. In this problem, let's consider how restaurants may determine how busy they will be.

- a) 🥑🥑🥑 Below is a list of a few restaurants' average Yelp rating versus their "busyness score". Naturally, we might expect that restaurants with a higher average Yelp rating would be busier.

Average Yelp Rating (x)	1.33	2.17	3.22	3.72	4.03	4.37	4.91
Busyness Score (y)	25	35	60	65	70	82	90

What linear relationship $y = c_0 + c_1x$ best describes the busyness score as a function of the average Yelp rating? What is the mean squared error, MSE_x , for this data set?

- b) 🥑🥑🥑 You reconsider the problem and think it's likely that there should be a strong connection between the busyness of a restaurant and their annual revenue. For the same set of restaurants, you collect the following data, which shows each restaurant's average yearly revenue in millions of dollars and their "busyness score."

Annual Revenue in Millions (z)	1.96	3.64	5.74	6.74	7.36	8.04	9.12
Busyness Score (y)	27	42	55	65	70	83	91

What linear relationship $y = d_0 + d_1z$ best describes the busyness score as a function of the annual revenue in millions? What is the mean squared error, MSE_z , for this data set?

- c) 🥑 Both regression lines you found in parts (a) and (b) have positive slope. Based on this, if a restaurant wants to increase how busy they will be, should they focus on increasing their average Yelp rating or increasing their yearly revenue? Explain.
- d) 🥑🥑🥑🥑 In the above example, notice that $MSE_x = MSE_z$, which means the mean squared error is the same if we use the predictor x or the predictor z . This happens because the average revenue in millions z is linearly related to the average Yelp rating x by the following formula: $z = 2x - 0.7$.

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 $z = ax + b$ for some constants $a, b \neq 0$, then $MSE_x = MSE_z$.

Hint: Start by using the result in Problem 4 (b) and expressing the best intercept and slope d_0^*, d_1^* from the relationship $y = d_0 + d_1z$ in terms of c_0^*, c_1^*, a, b , where c_0^*, c_1^* are coefficients inferred for the linear relationship: $y = c_0 + c_1x$.