DSC 40A - Homework 3

Due: Friday, January 28, 2022 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 has 55 avocados but it will still be graded out of 50 points. This means you can get over 100 percent on this assignment!
- 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 all of your work.
- For Problem 5, parts (b), (c), and (d), you'll need to code your answers in Python. We've provided a supplementary Jupyter notebook (linked). You'll need to turn in your completed Python file to Gradescope separately from the rest of this homework, in a file called `hw3code.py` . We'll grade parts (b) and (c) using an autograder, so explanations are not necessary for those parts. Part (d) requires a plot, and no explanation is needed there either.

Problem 1. Equivalent Formulas

In class, we showed that the least squares solutions for the slope and intercept are given by

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

$$w_0 = \overline{y} - w_1 \cdot \overline{x}$$

In this problem, you will show the equivalence of two other common forms of these solutions. It can be useful to have multiple equivalent formulas because some properties can be easier to prove when we start with the solutions in a certain form. After doing this problem, feel free to start at any of these equivalent forms when solving other problems in this class.

a) 🥑🥑 Show that the following form is equivalent to the form above from class.

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

$$w_0 = \overline{y} - w_1 \cdot \overline{x}$$

Hint: We did something similar in Lecture 10.

b) 🥑🥑🥑 In DSC 10, we wrote the slope and intercept of the regression line as

$$w_1 = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$w_0 = \overline{y} - w_1 \cdot \overline{x}$$

where r is the correlation coefficient, defined in your DSC 10 textbook (linked) as "the average of the products of the two variables, when both variables are measured in standard units."

Show that these formulas are equivalent to the ones above from class.

Problem 2. Nonlinear Transformation

🥑🥑🥑🥑🥑

x	1	2	2.5	5	10
y	0.1	0.2	0.2	0.25	0.32

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

$$y = \frac{x}{a + bx}.$$

Round the parameters a and b to three decimal places.

In addition, turn in a plot of the data points in the table above along with the curve that you found to make sure it looks to be a good fit. You can use Desmos (linked) or a similar plotting tool.

Problem 3. Lego Builds

a) 🥑🥑 Yang loves building challenging Lego sets. For several Lego sets that he built, he recorded the number of pieces in the set, x, and the number of hours it took him to build the set, y.

Lego Set	pieces (x)	hours (y)
Space Shuttle Discovery	2400	10
Hogwarts Castle	6000	22
Millennium Falcon	7500	26
Taj Mahal	6000	24
Minecraft Mountain Cave	2900	13

What linear relationship $y = c_0 + c_1 x$ best describes the number of hours for a build as a function of the number of pieces in the set? Give exact values for c_0 and c_1 (do not round).

b) 🥑🥑🥑 Now, let's interpret the meaning of the linear function $y = c_0 + c_1 x$ that you found in part (a).

- What does the intercept c_0 represent in terms of Yang's Lego builds?
- What does $1000 * c_1$ represent in terms of Yang's Lego builds?
- What does the reciprocal of the slope, $\frac{1}{c_1}$ represent in terms of Yang's Lego builds?

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) 🥑🥑 Yang knows he's invested a lot of money into his Lego habit, so he decides to try to quantify the value of his investment in terms of entertainment value provided. For each of the Lego sets, he records the amount he paid, in dollars, z and the number of hours it took him to build, y.

Lego Set	price (z)	hours (y)
Space Shuttle Discovery	\$ 290	10
Hogwarts Castle	\$ 650	22
Millennium Falcon	\$ 800	26
Taj Mahal	\$ 650	24
Minecraft Mountain Cave	\$ 340	13

What linear relationship $y = d_0 + d_1 z$ best describes the number of hours for a build as a function of the price? 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.

f) 🥑🥑🥑🥑🥑 You should have found that $MSE_x = MSE_z$, which says that for this data, the mean squared error is the same if we use the predictor x or the predictor z to make our regression line. This happens because the price of a Lego set z is linearly related to the number of pieces in the set x by the formula

$$z = 0.1x + 50.$$

Next, we'll show some general properties concerning the scenario where we predict some variable y based on x, as compared to predicting y based on z, when z is a linear transformation of x.

For the remaining parts of this problem, we'll no longer use the Lego data given above, but we'll prove properties in general.

First, suppose we have a data set $\{x_1, x_2, \ldots, x_n\}$ and we define a data set $\{z_1, z_2, \ldots, z_n\}$ by the linear transformation

$$z_i = ax_i + b_i.$$

Suppose also we have a data set $\{y_1, y_2, \ldots, y_n\}$.

Let c_0 and c_1 be the intercept and slope of the regression line for y with x as the predictor variable,

$$y = c_0 + c_1 x.$$

Similarly, let d_0 and d_1 be the intercept and slope of the regression line for y with z as the predictor variable,

$$y = d_0 + d_1 z.$$

Express d_0 and d_1 in terms of c_0, c_1, a, and b.

g) 🥑🥑🥑🥑 Let MSE_x be the mean squared error for the data set $\{y_1, y_2, \dots, y_n\}$ using the regression line

$$y = c_0 + c_1 x.$$

Similarly, let MSE_z be the mean squared error for the data set $\{y_1, y_2, \dots, y_n\}$ using the regression line

$$y = d_0 + d_1 z.$$

Show that $MSE_x = MSE_z$.

Problem 4. Meet Billy

Suppose that in 2020 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 90 and a standard deviation of 20. We then use least squares to fit a linear prediction rule $H(x) = w_0 + w_1 x$, which we will use to help other farmers predict their avocado yield based on the number of trees they have.

a) 🥑 Why is this an appropriate scenario for fitting a linear function $H(x)$?

b) 🥑🥑🥑🥑 Billy was one of the 200 farmers surveyed in 2020. Billy is a really poor farmer. In 2020, his 10 avocado trees yielded only 120 avocados, the smallest total number reported by any of the survey participants.

Billy later found out that avocado yield can be increased by additional watering. He began watering his avocado trees more frequently, and in the year 2021, his 10 trees yielded 320 avocados – what an improvement!

Suppose we create two linear prediction rules, one using the dataset from 2020 when Billy's trees yielded 120 avocados and another using the dataset from 2021 when Billy's trees yielded 320 avocados. Assume that all other farmers had the same number of trees and same avocado yield in both 2020 and 2021. That is, only this one data point is different between these two datasets.

Suppose the optimal slope and intercept fit on the first dataset (2020) are w_1^* and w_0^*, respectively, and the optimal slope and intercept fit on the second dataset (2021) 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.

Hint: There are many equivalent formulas for the slope of the regression line. We recommend using the formula from Problem 1(a).

c) 🥑🥑🥑 Let $H^*(x)$ be the linear prediction rule fit on the 2020 dataset (i.e. $H^*(x) = w_0^* + w_1^*$) and $H'(x)$ be the linear prediction rule fit on the 2021 dataset (i.e. $H'(x) = w_0' + w_1' x$).

Consider two other farmers, Dev and Maria, neither of whom were part of the survey data in 2020 or 2021. Dev has 20 avocado trees and Maria has 40 avocado trees.

Both Dev and Maria 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 120 avocados). Just to see what happens, they both then use $H'(x)$ to determine predicted yields as per the second rule (when Billy yielded 320 avocados).

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

d) 🥑🥑 In this problem, we'll consider how our answer to part (b) might have been different if Billy had had more avocado trees on his farm.

- If Billy instead had 150 avocado trees, and his yield increased from 2020 to 2021, which slope would be larger: $H^*(x)$ or $H'(x)$?
- If Billy instead had 90 avocado trees, and his yield increased from 2020 to 2021, 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 5. Least Absolute Deviation Regression

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

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

Notice 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 instead?

In this problem, 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

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

Since absolute value functions are not differentiable, 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, as we did to minimize R_{sq}. 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. Since our data will be in $\mathbb{R}^2$, we will generate all possible unique pairs of points and calculate the intercept w_0 and slope w_1 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) 🥑🥑🥑🥑 First, we'll find the regular least squares regression line. In this supplementary notebook (linked) fill in the `least_squares_regression` function. You'll need to implement the formulas for the slope and intercept of the least squares regression line (see Problem 1) into a Python function which takes in the x and y values as an input and returns a tuple (w_0, w_1) with the intercept and slope of the least squares regression line.

c) 🥑🥑🥑🥑🥑 Now, let's find the LAD line. Recall from the problem description the procedure outlined to generate an optimal LAD regression line. In the same supplementary notebook, 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.

d) 🥑🥑 Now that we have calculated the least squares regression line and the least absolute deviation regression line for our data, let's try plotting them together to see the difference! In the same supplementary notebook, generate a scatterplot with the data in black, the least squares line in blue, and the LAD line in red. Turn in a picture of your plot.

e) 🥑🥑 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.