
DSC 40A - Homework 3

Due: Tuesday, April 25 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 available, but it will be graded out of 50 points. This means that it's possible to 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 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.
- 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 0. Reflection and Feedback Form

 Make sure to fill out this [Reflection and Feedback Form, linked here](#) for three 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 1. 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 2. 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 3. 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 2018.

- If Haaland instead had 31 goals, and his market value increased from 2018 to 2019, which slope would be larger: $H^*(x)$ or $H'(x)$?
- If Haaland instead had 45 goals, and his market value increased from 2018 to 2019, 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 4. Smurfs’ Village

a) 🥑🥑 Smurfs’ Village is a simulation game where players help the Smurfs to build and farm across five magical areas. In the game, players use Smurfs’ coins to purchase seeds, plant crops, and harvest crops. Harvesting crops earns players experience points, which are valuable because they allow players to level up and unlock more mini-games, as well as obtain better rewards. Yutian loves playing Smurfs’ Village. For several crops she planted, she recorded the number of Smurfs’ coins required to plant one unit of crop, x , and the experience points she gained from harvesting one unit of crop, y .

Crops	coins (x)	experience points (y)
Corn	23	200
Chocolate tulips	20	190
Onion	32	260
Pineapples	50	320
Daffodils	5	110

What linear relationship $y = c_0 + c_1x$ best describes the experience points gained from harvesting the crops as a function of the number of coins required for planting the crops? 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 $50 * c_1$ represent in terms of Yutian’s crops?
- What does the reciprocal of the slope, $\frac{1}{c_1}$ represent in terms of Yutian’s crops?

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) 🥑🥑 Yutian knows that growing crops in Smurfs’s Village takes a significant amount of time, so she decides to quantify the value of planting time in terms of experiences points gained. For each of the crops she planted, Yutian recorded the number of hours to grow one unit of the crop, z and the experience points she gained from harvesting one unit of the crop, y .

Crops	growing hours (z)	experience points (y)
Corn	9	200
Chocolate tulips	8	190
Onion	12	260
Pineapples	18	320
Daffodils	3	110

What linear relationship $y = d_0 + d_1 z$ best describes the experience points for harvesting one unit of crop as a function of the growing time? 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 number of hours to plant one unit of crop (z) is linearly related to the number of coins required for planting one unit of crop (x) by the formula

$$z = \frac{1}{3}x + \frac{4}{3}.$$

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 crops data given above, but we'll prove properties in general.

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

$$z_i = ax_i + b.$$

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

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 .

Hint: You'll need to use a result about linear transformations from a previous homework assignment. Make sure to cite the specific problem number when using its result.

- 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 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 scatter plot 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.