
DSC 40A - Homework 3

due Friday, October 25th 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.

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.


Homework 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. We encourage you to type your solutions in L^AT_EX, using the Overleaf template on the course website.

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

Note: For Problems 6, you'll need to code your answers in Python. More detailed instructions are provided in Problem 6. Note that to submit the homework, you'll have to submit your answers in PDF to the Homework 3 assignment on Gradescope, and submit your completed notebook `hw03-code.ipynb` to the Homework 3, Problems 6(b) and 6(c) autograder on Gradescope.

Problem 1. Reflection and Feedback Form

 Make sure to fill out this Reflection and Feedback Form, linked [here](#), for two 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 2. Streaming for Stardom

In the music industry, the highest-earning artists are often those who generate the most streams on platforms like Spotify. Every time an artist releases a song, it collects streams as listeners engage with the track. For this problem, we will say an artist “collected a stream” each time their song is played.

Suppose we have access to a dataset containing information about a random sample of 50 music artists. For each artist, we have the number of streams their songs received in 2023, along with their revenue for that year. In the 2023 dataset, the number of streams for all artists has a mean of 50 million and a standard deviation of 15 million.

We minimize mean squared error to fit a linear hypothesis function,

$$H(x) = w_0 + w_1x,$$

to this dataset. We will use this hypothesis function to help other artists predict their 2023 revenue (in millions of dollars) based on their number of streams x .

One of the artists in our 2023 dataset was Drake. Suppose that in 2023, he had 65 million streams and his revenue was only \$2 million, the smallest in our sample.

In 2024, Drake signed a new record deal based on his performance. In 2023, he again collected 65 million streams, but his revenue shot up to \$10 million!

Suppose we create two linear hypothesis functions, one using the dataset from 2023 when Drake had a revenue of \$2 million, and another using the dataset from 2024 when Drake had a revenue of \$10 million. Assume that all other artists had the same number of streams and earned the same revenue in both datasets. That is, only Drake's data point is different between these two datasets.

Let $H^*(x)$ be the linear hypothesis function fit on the 2023 dataset, where the optimal slope and intercept fit on the first dataset (2023) are w_1^* and w_0^* , respectively (i.e. $H^*(x) = w_0^* + w_1^*x$). Let $H'(x)$ be the linear hypothesis function fit on the 2024 dataset, where the optimal slope and intercept fit on the second dataset (2024) are w_1' and w_0' , respectively (i.e. $H'(x) = w_0' + w_1'x$).

- a) 🥑🥑🥑🥑 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.

Solution: Write your solution here.

- b) 🥑🥑🥑 Consider two other artists, The Weeknd and Taylor Swift, neither of whom were part of our original sample of music artists in 2023. Suppose that in 2021, The Weeknd had 45 million streams and Taylor had 100 million streams.

Both The Weeknd and Taylor want to use one of our linear hypothesis functions to predict their revenue. Suppose they both first use $H^*(x)$ to determine their predicted revenue as per the first rule (when Drake had revenue of \$2 million). Then, they both use $H'(x)$ to determine predicted revenue as per the second rule (when Drake had revenue of \$10 million).

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

Solution: Write your solution here.

- c) 🥑🥑 In this problem, we'll consider how our answer to part (b) might have been different if Drake had fewer streams in both 2023 and 2024. Note you don't have to actually calculate the new slopes below, but given the information in the problem and the work you've already done, you should be able to answer the questions and give brief justification.

Suppose Drake instead had 50 million streams in both 2023 and 2024. If his revenue increased from 2023 to 2024, and everyone else's data stayed the same, which slope would be larger: $H^*(x)$ or $H'(x)$?

Solution: Write your solution here.

Suppose Drake instead had 30 million streams in both 2023 and 2024. If his revenue increased from 2023 to 2024, and everyone else's data stayed the same, which slope would be larger: $H^*(x)$ or $H'(x)$?

Solution: Write your solution here.

Problem 3. Correlation Bounds

In both this class and DSC 10, you were told that the correlation coefficient, r , ranges between -1 and 1 , where $r = -1$ implies a perfect negative linear association and $r = 1$ implies a perfect positive linear association. However, you were never given a proof of the fact that $-1 \leq r \leq 1$.

Here, you will prove this fact, using linear algebra. Before proceeding, you'll want to review [slide 19 onwards in Lecture 8](#). **Remember to show your work all throughout!**

- a) 🥑🥑 Determine the angle between the vectors $\vec{a} = \begin{bmatrix} 5 \\ -1 \\ 7 \end{bmatrix}$ and $\vec{b} = \begin{bmatrix} -3 \\ 2 \\ 12 \end{bmatrix}$. Your answer should involve the function \cos^{-1} (you do not have to find the angle in terms of degrees or radians).

Solution: Write your solution here.

- b) 🥑🥑 Let $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$. We define the “mean-centered” version of \vec{x} to be $\vec{x}_c = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$, where \bar{x} is the mean of the components of \vec{x} .

The mean-centered version of \vec{y} , named \vec{y}_c , is defined similarly. Express $\vec{x}_c \cdot \vec{y}_c$ using summation notation.

Solution: Write your solution here.

- c) 🥑🥑🥑 Prove that:

$$r = \frac{\vec{x}_c \cdot \vec{y}_c}{\|\vec{x}_c\| \|\vec{y}_c\|}$$

Solution: Write your solution here.

- d) 🥑🥑 Argue why the result in (c) implies that $-1 \leq r \leq 1$.

Hint: If you're completely stuck on how to proceed, try to think about what the purpose of part (a) was — it's in some way related to this part.

Solution: Write your solution here.

Problem 4. Making Connections... and Projections

Suppose we have a dataset of n points, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. In Groupwork 3, we proved that the optimal parameter m^* that minimizes mean squared error for the hypothesis function $H(x) = mx$ is:

$$m^* = \frac{\sum_{i=1}^n t_i y_i}{\sum_{i=1}^n t_i^2}$$

(There, we used the variable w instead of m ; we've used m above to avoid conflicting with a different definition of w below.)

In this problem, we'll derive the same result using our knowledge of vector projections from Lectures 9 and 10, to start making the connections between linear algebra and empirical risk minimization more clear.

Moving forward, consider the dataset of two points, $(2, 1)$ and $(3, 2)$. We can store the x and y coordinates of our two points in vectors, \vec{x} and \vec{y} , as follows:

$$\vec{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \vec{y} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- a) 🥑🥑🥑 Our goal is to find the vector in $\text{span}(\vec{x})$ that is closest to y . The answer is a vector of the form $w\vec{x}$, where $w \in \mathbb{R}$ is some scalar. The w that we choose is one that minimizes the length, $\|\vec{e}\|$ of the error vector, \vec{e} :

$$\vec{e} = \vec{y} - w\vec{x}$$

What is w^* , the value w that minimizes $\|\vec{e}\|$? In other words, what value of w minimizes projection error? (Note that the *vector* projection of \vec{y} onto $\text{span}(\vec{x})$ is not w^* , but $w^*\vec{x}$ — however, here we're just asking you for the value of w^* , and of course, to show your work).

Solution: Write your solution here.

- b) 🥑🥑 What is the error vector, \vec{e} , you found in part (a), and what is its length, $\|\vec{e}\|$?

Solution: Write your solution here.

- c) 🥑🥑 The value of w^* you found in part (a) should be equal to the value you find using the formula for m^* . In general, the w^* that minimizes $\|\vec{y} - w\vec{x}\|$ is equal to m^* , the m that minimizes $\frac{1}{n} \sum_{i=1}^n (y_i - mx_i)^2$. Explain why this is the case.

Hint: $\|\vec{y} - w\vec{x}\|$ and $\frac{1}{n} \sum_{i=1}^n (y_i - mx_i)^2$ are related, but not exactly the same.

Solution: Write your solution here.

In parts (a) through (c), we projected \vec{y} onto the span of a single vector, \vec{x} . But in Lecture 10, we looked at how to project a vector \vec{y} onto the span of two or more vectors. Let's explore that concept here.

- d) 🥑🥑 Consider the vectors $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$, defined as follows:

$$\vec{x}^{(1)} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \quad \vec{x}^{(2)} = \begin{bmatrix} 5 \\ 23 \end{bmatrix}$$

Again, let $\vec{y} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$.

What is the vector projection of \vec{y} onto $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ — that is, what vector in $\text{span}(\vec{x}^{(1)}, \vec{x}^{(2)})$ is closest to \vec{y} ? Give your answer in the form of a vector.

Solution: Write your solution here.

- e) 🥑🥑🥑 Let \vec{h} be your answer to the previous part. Find scalars w_1 and w_2 such that:

$$w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)} = \vec{h}$$

Solution: Write your solution here.

In Lecture 10 we expressed $w_1\vec{x}^{(1)} + w_2\vec{x}^{(2)}$ as the matrix-vector product $X\vec{w}$, where $X = \begin{bmatrix} | & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} \\ | & | \end{bmatrix} = \begin{bmatrix} 0 & 5 \\ 3 & 23 \end{bmatrix}$ and $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$. This allowed us to more efficiently solve for the values w_1, w_2, \dots, w_d that minimize projection error when we have several spanning vectors, $\vec{x}^{(1)}, \vec{x}^{(2)}, \dots, \vec{x}^{(d)}$. Moving forward we will use this approach for multiple linear regression.

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, starting in part (b).

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

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

Note that the product $A^T \vec{b}$ involves taking the dot product of each row in A^T with \vec{b} . If $A^T \vec{b}$ is a vector of all 0s, i.e. the 0 vector, 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 .)

- a) 🥑🥑 In the example above, verify that \vec{b} is orthogonal to the columns of A .

Solution: Write your solution here.

- b) 🥑🥑 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} ?

Your answer may involve a summation symbol. Explain what it means in words.

Hint: This subpart should not take much time, so let us know if you're stuck on it. Try making up an example \vec{b} and see what $\vec{1}^T \vec{b}$ evaluates to, before generalizing your result to arbitrary \vec{b} .

Solution: Write your solution here.

- c) 🥑🥑 Now, consider a multiple regression scenario where X is a $n \times (d + 1)$ matrix where the first column is the all-ones column vector $\vec{1}$ and each column of the remaining d columns is a feature vector (this is also called a *design* matrix), $\vec{y} \in \mathbb{R}^n$ is our vector of targets (i.e. response variable) we are trying to fit, and $w^* \in \mathbb{R}^{d+1}$ is the optimal parameter vector. Our hypothesis is then

$$H(\vec{x}) = \vec{x} \cdot \vec{w} = \sum_{j=0}^d x^{(j)} w_j = w_0 + x^{(1)} w_1 + x^{(2)} w_2 + \dots + x^{(d)} w_d,$$

where the parameter w_0 corresponds to the intercept, and the mean squared error is

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2.$$

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

Hint: Again, this should not take very long. Start with the normal equations, $X^T X \vec{w}^* = X^T \vec{y}$, use the distributive property of matrix multiplication, and use what you learned in part (a).

Solution: Write your solution here.

- d) 🥑🥑🥑 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

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

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

Using what you learned in parts (a), (b), and (c), show that the **residuals of a multiple linear regression prediction rule with an intercept term sums to 0**, i.e. that $\sum_{i=1}^n e_i = 0$.

Solution: Write your solution here.

- e) 🥑🥑🥑 Now suppose our multiple linear regression prediction rule

does not have an intercept term, i.e. that our prediction rule is of the form $H(\vec{x}) = w_1 x^{(1)} + w_2 x^{(2)} + \dots + w_d x^{(d)}$.

1. Is it still guaranteed that $\sum_{i=1}^n e_i = 0$? Why or why not?
2. Is it still possible that $\sum_{i=1}^n e_i = 0$? If you believe the answer is yes, come up with a simple example where a prediction rule without an intercept has residuals that sum to 0. If you believe the answer is no, state why.

Solution: Write your solution here.

Problem 6. Least Absolute Deviation Regression

In lecture, 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 mean squared error:

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

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

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

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 dataset 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 $R_{\text{abs}}(w_0, w_1)$. This is guaranteed by the theorem to be an optimal LAD regression line.

Parts (b) and (c) of this problem will require you to write code in [this supplementary Jupyter Notebook](#). The code that you write in that notebook is autograded, both using public test cases that you can see in the notebook and hidden test cases that will only be run after you submit on Gradescope.

To submit your homework, in addition to submitting your answers PDF to the Homework 3 assignment on Gradescope, also submit `hw03-code.ipynb` to the Homework 3, Problems 6(b) and 6(c) autograder on Gradescope and wait until you see all public test cases pass!

- 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) .

Solution: Write your solution here.

- b) 🥑🥑🥑 First, we'll find the intercept and slope of the regular least squares regression line. In [the linked supplementary notebook](#), read the problem statement and complete the implementation of the function `least_squares_regression`.

Solution: Write your solution here.

- c) 🥑🥑🥑🥑 Next, we'll find the intercept and slope of the least absolute deviations line. In [the linked supplementary notebook](#), read the problem statement and complete the implementations of the functions `mean_absolute_error` and `find_best_mad_line`.

Solution: Write your solution here.

- 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 linked supplementary notebook](#), generate a scatter plot with the data in black, the least squares line in blue, and the least absolute deviation line in red. **Include a picture of your plot in your PDF; this problem is not autograded.**

Solution: Write your solution here.

- 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.

Solution: Write your solution here.