DSC 40A - Homework 3

Due: Wednesday, Jan 31 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.

Problem 1. Reflection and Feedback Form

🥑🥑 Make sure to fill out this Reflection and Feedback Form, linked here with your ucsd account 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. Jensen Gap and Convexity

🥑🥑🥑🥑🥑🥑 Recall the definition of convex function, if f is convex, then $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2), \forall t \in [0,1]$. A general form of this is known as *Jensen's Inequality*: for a real convex function g, with positive weights $a_i, \sum_i a_i = 1$, we have

$$g\left(\sum_i a_i x_i\right) \leq \sum_i a_i g(x_i)$$

Recall that for a discrete random variable X, with possible outcome x_i and corresponding probability p_i, its expectation is defined as $E[X] = \sum_i p_i x_i$. The distance $g(E[X]) - E[g(X)]$ when g is convex is known as *Jensen gap.*

Use the above facts to prove that, if X is a discrete random variable with all possible outcomes are positive, then

$$\ln E[X] - E[\ln X] \geq 0$$

Solution: Note that $\sum_i p_i = 1$, by Jensen's inequality,

$$g\left(E[X]\right) = g\left(\sum_i p_i x_i\right) \leq \sum_i p_i g(x_i) = E[g(X)]$$

$g(x) = -\ln x$ is convex as $g''(x) = 1/x^2 > 0, x \neq 0$

Therefore, $-\ln(E[X]) \leq E[-\ln(X)] \Rightarrow \ln E[X] - E[\ln X] \geq 0$

Problem 3. 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.

Solution: This is a true statement. Suppose that $f_1(x)$ and $f_2(x)$ are convex functions. We wish to show that their sum, $f(x) = f_1(x) + f_2(x)$, is also a convex function. To do this, we must show that for any real numbers a and b and for all $t \in [0, 1]$,

$$(1-t)f(a) + tf(b) \geq f((1-t)a + tb).$$

We will start with $(1-t)f(a) + tf(b)$ and try to make it look like the right hand side using a chain of inequalities.

We have two pieces of information that we know must be used in the proof:

1. f is the sum of f_1 and f_2, and
2. f_1 and f_2 are convex.

We can't use the second piece of information yet, since we don't see f_1 or f_2, but we can use the first:

$$\begin{aligned}(1-t)f(a) + tf(b) &= (1-t)(f_1(a) + f_2(a)) + t(f_1(b) + f_2(b)) \\ &= (1-t)f_1(a) + (1-t)f_2(a) + tf_1(b) + tf_2(b) \\ &= [(1-t)f_1(a) + tf_1(b)] + [(1-t)f_2(a) + tf_2(b)]\end{aligned}$$

Now we can use the second piece of information. Since f_1 is convex, we know that $(1-t)f_1(a) + tf_1(b) \geq f_1((1-t)a + tb)$; a similar statement holds for f_2. Therefore:

$$\geq f_1((1-t)a + tb) + f_2((1-t)a + tb)$$

It helps to remember where we are trying to go. We want to make this look like $f((1-t)a + tb)$. We need to apply the first piece of information again, which says that f is the sum of f_1 and f_2. This gives the desired result:

$$= f((1-t)a + tb)$$

b) The difference of two convex functions must also be convex.

Solution: This statement is false. For example, $f_1(x) = x^2$ and $f_2(x) = 2x^2$ are both convex functions because they are upward-facing parabolas. However $f_1(x) - f_2(x) = x^2 - 2x^2 = -x^2$ is a downward-facing parabola, which is nonconvex.

Problem 4. Conditions of Linear Regression

Often times in data science, one can create a decently accurate model using linear regression! It also has the added benefit of being fast, easy to interpret, and simple (compared to more complicated models like neural networks). However, there are some assumptions/conditions that need to be met in order for linear regression to work well.

In general, linear regression can be expressed as $Y = \beta_0 + \beta_1 X + \epsilon_i$ where β_0, β_1 are our unknown parameters, and ϵ_i is a random variable that represents the error. Since this class only covers linear regression of the form $H(\vec{x}) = w_0 + w_1\vec{x}$, let us explore the conditions that are unrelated to error: linearity and outliers.

Suppose you work on an avocado farm and construct the following dataset of month, average high temperature (in °F), and avocado yield (in hundreds), shown in Table 1.

Month	Average High Temperature (°F)	Avocado Yield (hundreds)
1	66	18
2	64	22
3	64	19
4	66	17
5	67	14
6	70	10
7	74	6
8	75	2
9	75	3
10	73	6
11	69	10
12	65	18

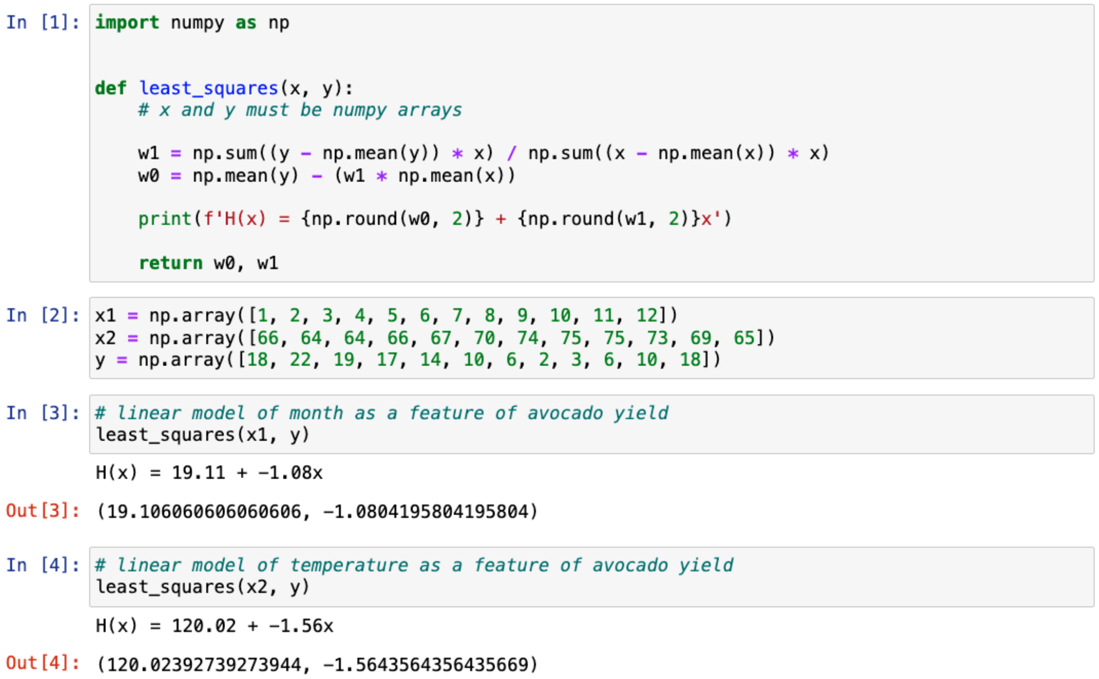
Table 1: Avocado Yield

Answer the following sub questions based on Table 1. Calculations by hand or by code are both acceptable as long as sufficient work is shown.

a) Arguably, the most important condition for linear regression is linearity in the data, specifically between the y_i's and the x_i's. In general, it does not really make sense to fit a linear model to data that does not exhibit a linear relationship.

Perform linear regression for the linear models $H(x^{(1)}) = a + bx^{(1)}$ and for $H(x^{(2)}) = c + dx^{(2)}$, where $x^{(1)}$ is Month, $x^{(2)}$ is Average High Temperature, and y is Avocado Yield. In the context of the data, explain the meaning of the model parameters you found for the two models.

Solution:



```
In [1]: import numpy as np


        def least_squares(x, y):
            # x and y must be numpy arrays

            w1 = np.sum((y - np.mean(y)) * x) / np.sum((x - np.mean(x)) * x)
            w0 = np.mean(y) - (w1 * np.mean(x))

            print(f'H(x) = {np.round(w0, 2)} + {np.round(w1, 2)}x')

            return w0, w1

In [2]: x1 = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12])
        x2 = np.array([66, 64, 64, 66, 67, 70, 74, 75, 75, 73, 69, 65])
        y = np.array([18, 22, 19, 17, 14, 10, 6, 2, 3, 6, 10, 18])

In [3]: # linear model of month as a feature of avocado yield
        least_squares(x1, y)

        H(x) = 19.11 + -1.08x

Out[3]: (19.106060606060606, -1.0804195804195804)

In [4]: # linear model of temperature as a feature of avocado yield
        least_squares(x2, y)

        H(x) = 120.02 + -1.56x

Out[4]: (120.02392739273944, -1.5643564356435669)
```

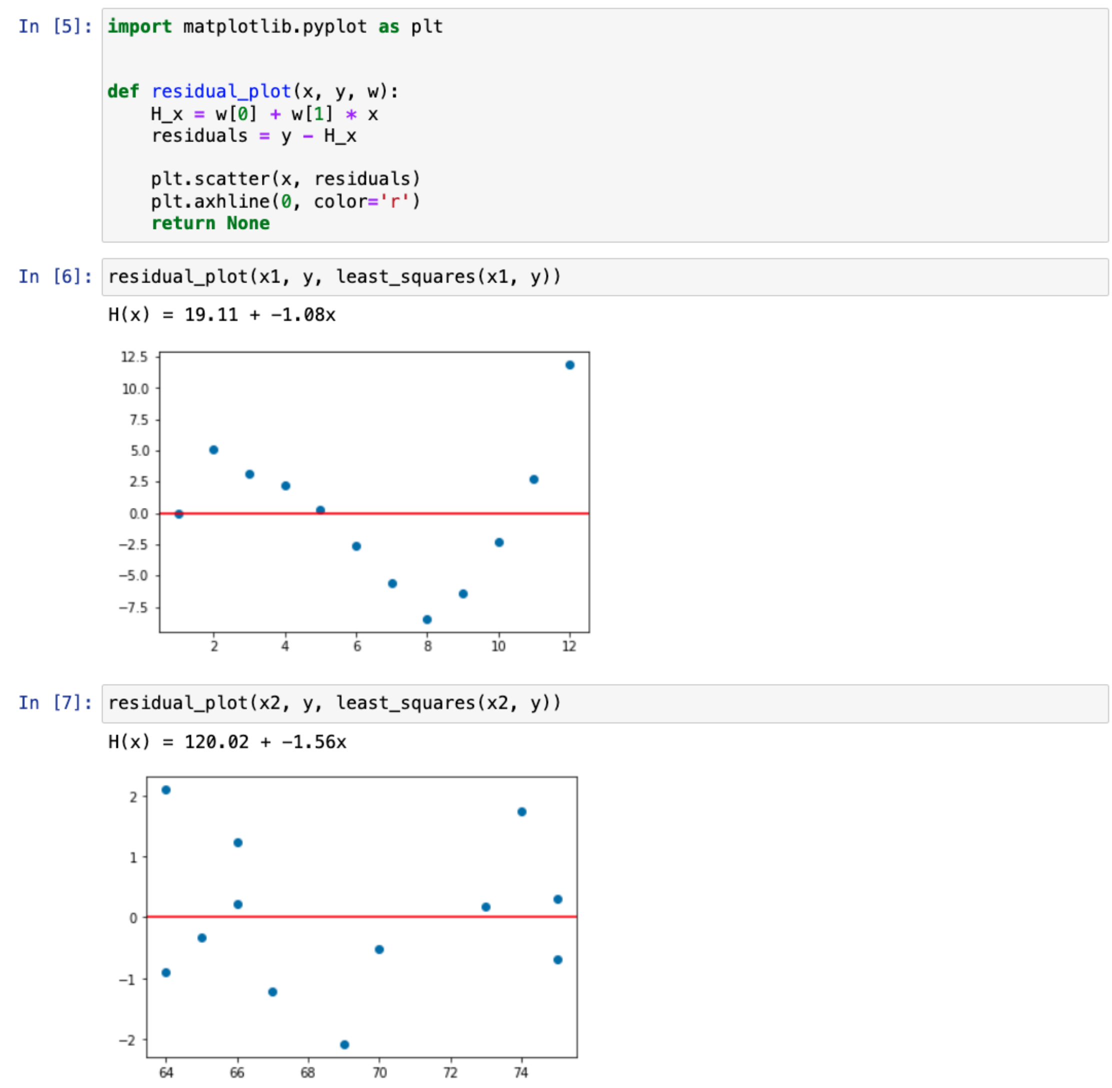
We can see that $a = 19.11$, $b = -1.08$, $c = 120.02$, and $d = -1.56$.

For the model involving Month, this means that as we increase Month by one unit, the number of avocados decreases by 108 (1.08 hundred) avocados, and at Month $= 0$, we predict that the farm yields 1,911 (19.11 hundred) avocados. However, this doesn't really make sense since there is no Month $= 0$.

For the model involving Average High Temperature, this means that as we increase the temperature by $1°F$, the number of avocados decreases by 156 (1.56 hundred) avocados, and at Average High Temperature $= 0°F$, we predict that the farm yields 12,002 (120.02 hundred) avocados. This also doesn't make much sense practically because $0°F$ is pretty cold and likely not the best climate for avocados.

b) Additionally, you may recall the idea of a residual plot from DSC 10, where a residual is defined as $y_i - H(x_i)$. (Notice that least squares regression minimizes the mean squared residuals!) Draw the residual plots for the two models from part a), and use them to determine if there is linearity in the data.

Solution:



```
In [5]: import matplotlib.pyplot as plt


        def residual_plot(x, y, w):
            H_x = w[0] + w[1] * x
            residuals = y - H_x

            plt.scatter(x, residuals)
            plt.axhline(0, color='r')
            return None
```

```
In [6]: residual_plot(x1, y, least_squares(x1, y))
        H(x) = 19.11 + -1.08x
```

```
In [7]: residual_plot(x2, y, least_squares(x2, y))
        H(x) = 120.02 + -1.56x
```

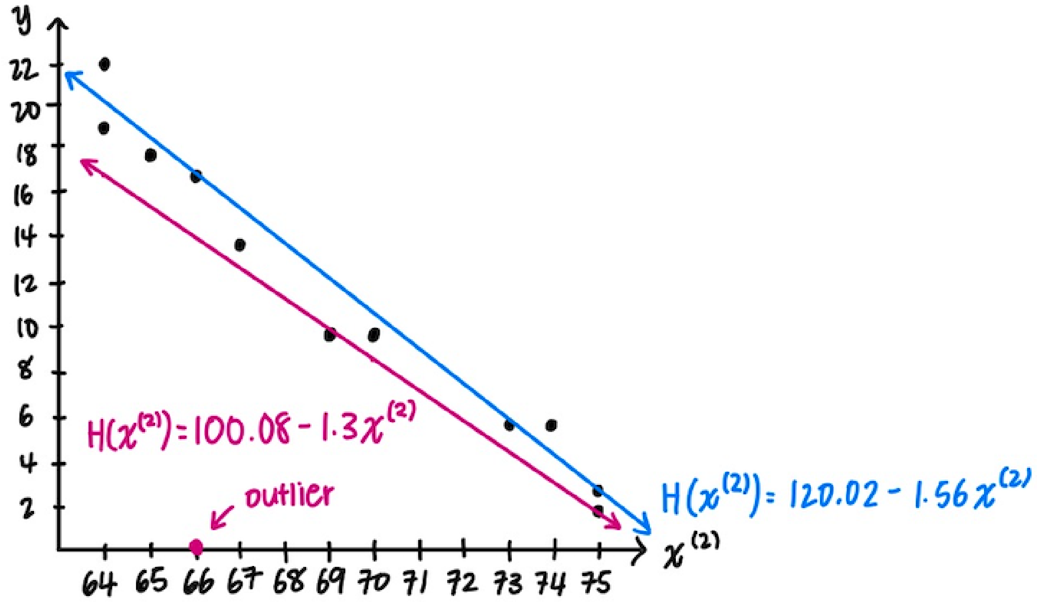
We can see that the residual plot of $x^{(1)}$ has a trend that is not being captured by our model. Thus, there is no linearity between Month and Avocado Yield. Intuitively, time variables like month are cyclical, so non-linear.

The residual plot of $x^{(2)}$ is more randomly scattered, so there may be some linearity that our model captured.

c) Another condition to consider checking when performing linear regression is outliers that may strongly affect the linear model. Outliers that strongly affect the model are called influential points.

Suppose that the recent storm caused your farm to flood, destroying all of your avocado trees. Alas, you have 0 avocados for January (Month = 1). Explain whether this outlier is an influential point in terms of the model parameters for the linear model $y = c + dx_2$ where x_2 is Average High Temperature and y is Avocado Yield.

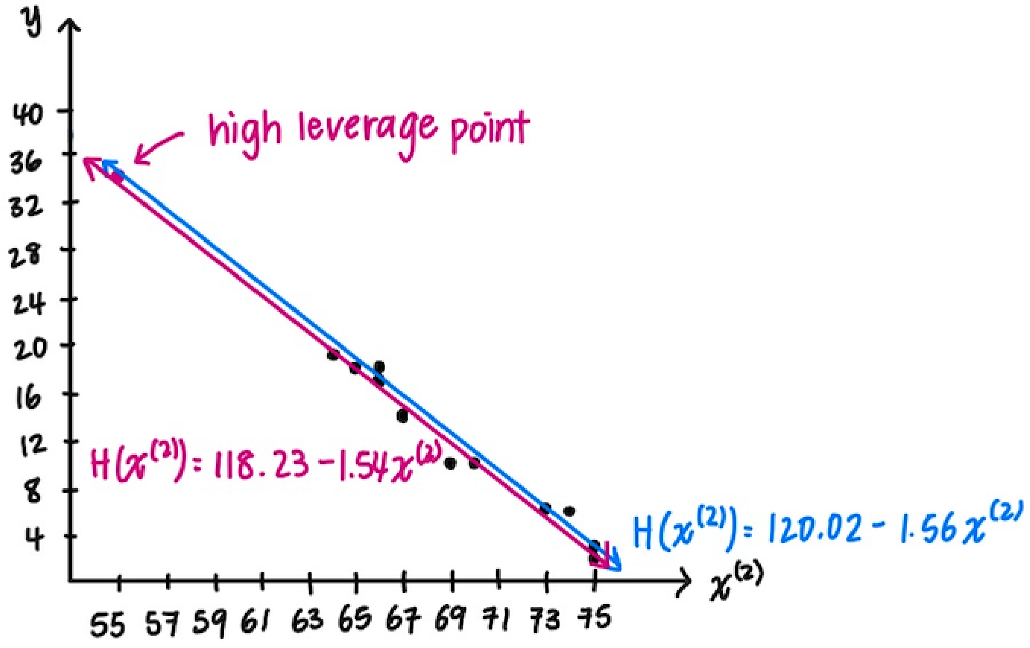
Solution:



We can see that linear regression on this new data with an outlier changes our model parameters on the order of 2,000 avocados for the intercept term and 26 avocados for the slope term. This outlier is an influential point because it affected the model parameters.

d) Outliers in the x-direction are called high leverage points. Suppose that February (Month = 2) turned out to be especially cold this winter, with an average high temperature of 55. Miraculously, your avocado trees managed to produce 3,400 avocados (Avocado Yield = 34)! Explain whether this high leverage point is an influential point.

Solution:



We can see that linear regression on this new dataset with the high leverage point did not change our model parameters by much. Quantitatively, the intercept term decreased by 200 avocados and the slope term increased by 2 avocados. Thus, this high leverage point is not an influential point.

Problem 5. 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

$$\overline{y}_1 = \frac{y_1 + y_2}{2}, \quad \overline{y}_2 = \frac{y_3 + y_4}{2}, \quad \overline{y}_3 = \frac{y_5 + y_6}{2}.$$

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 $(5, \overline{y}_1)$, $(10, \overline{y}_2)$, $(15, \overline{y}_3)$.

Solution:

Since the average of the six x values is 10, the regression line fitted to the original six data points has slope

$$w_1 = \frac{\sum_{i=1}^{6}(x_i - \overline{x})y_i}{\sum_{i=1}^{6}(x_i - \overline{x})^2} = \frac{-5y_1 + -5y_2 + 0y_3 + 0y_4 + 5y_5 + 5y_6}{(-5)^2 + (-5)^2 + 0^2 + 0^2 + 5^2 + 5^2}$$

The three data points $(5, \overline{y}_1)$, $(10, \overline{y}_2)$, $(15, \overline{y}_3)$ also have an average x value of 10, so the regression line fitted to these three data points has slope

$$
\begin{aligned}
w_1 &= \frac{\displaystyle\sum_{i=1}^{3}(x_i - \overline{x})\overline{y}_i}{\displaystyle\sum_{i=1}^{3}(x_i - \overline{x})^2} \\
&= \frac{-5\overline{y}_1 + 0\overline{y}_2 + 5\overline{y}_3}{(-5)^2 + 0^2 + 5^2} \\
&= \frac{-5\left(\frac{y_1+y_2}{2}\right) + 0\left(\frac{y_3+y_4}{2}\right) + 5\left(\frac{y_5+y_6}{2}\right)}{(-5)^2 + 0^2 + 5^2}
\end{aligned}
$$

We can see that multiplying the top and bottom of this expression both by 2 does not change the value of the expression, and makes it match the slope of the regression line fitted to the original six data points. So the two regression lines have the same slope.

We have already observed that the average of the x values is the same whether we consider six or three points. Similarly the average of the y values is also the same whether we have six or three points because

$$
\frac{y_1 + y_2 + y_3 + y_4 + y_5 + y_6}{6} = \frac{\frac{y_1+y_2}{2} + \frac{y_3+y_4}{2} + \frac{y_5+y_6}{2}}{3}.
$$

Since $w_0 = \overline{y} - w_1\overline{x}$ and we have shown that each of $\overline{y}$, w_1, and $\overline{x}$ are the same regardless of whether we use six or three points, then w_0 is also the same when we use six or three points. This shows that the two regression lines have the same slope and intercept, so they are the same line.

Problem 6. 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:

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

Solution: Suppose Haaland was the jth player in our sample, so $(x_j, y_j) = (16, 20)$ is the data point corresponding to him in the original 2018 dataset, and $(x_j, y_j) = (16, 80)$ is the point corresponding to him in the new 2019 dataset.

Then, using the form of w_1^* given in the hint, we have

$$w_1^* = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \text{(least squares solution for linear model)}$$

$$w_1^* = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})y_i}{n\sigma_x^2} \qquad \text{(replacing denominator using definition of variance)}$$

$$= \frac{(16 - \bar{x}) \cdot 20 + \displaystyle\sum_{i \neq j}(x_i - \bar{x})y_i}{n\sigma_x^2} \qquad \text{(separate Haaland)}$$

Similarly,

$$w_1' = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \text{(least squares solution for linear model)}$$

$$w_1' = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})y_i}{n\sigma_x^2} \qquad \text{(replacing denominator using definition of variance)}$$

$$= \frac{(16 - \bar{x}) \cdot 80 + \displaystyle\sum_{i \neq j}(x_i - \bar{x})y_i}{n\sigma_x^2} \qquad \text{(separate Haaland)}$$

We need to find the difference between the two slopes.

$$\begin{aligned}
w_1' - w_1^* &= \frac{(16-\bar{x})\cdot 80 - (16-\bar{x})\cdot 20 + \sum_{i\neq j}(x_i-\bar{x})y_i - \sum_{i\neq j}(x_i-\bar{x})y_i}{n\sigma_x^2} \\
&= \frac{(16-\bar{x})\cdot 80 - (16-\bar{x})\cdot 20}{n\sigma_x^2} \\
&= \frac{(16-31)\cdot 80 - (16-31)\cdot 20}{200\cdot 6^2} && \text{(Substitute in } \bar{x}=31, \sigma_x^2 = 6^2, n=200\text{)} \\
&= \frac{(16-31)(80-20)}{200\cdot 6^2} \\
&= \frac{-1}{8} \\
w_1' - w_1^* &= -\frac{1}{8}
\end{aligned}$$

Common misconceptions: Many students implicitly assumed that Haaland was player 1 or player 200. While this gives the same result, it can't be assumed by default that Haaland is player 1 or player 200; you'd need to formally state this as an assumption for this to be valid. (It works out to the same answer and is a valid assumption to state because the ordering of the data points is irrelevant.)

Many other students also said something like $x_i = 10$ or $x_i = 100$. However, when we write sums from $i = 1$ to n, we're using i as a placeholder index. Therefore, x_i represents the number of goals scored by player i, not by any one specific player.

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).

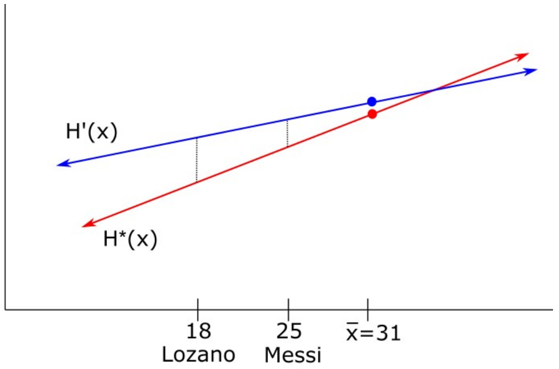
Whose prediction changed more by switching from $H^*(x)$ to $H'(x)$ – Lozano's or Messi's?

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?

Solution: Lozano's prediction changed more by switching from $H^*(x)$ to $H'(x)$.

As we saw in lecture, we know the point $(\bar{x}, \bar{y})$ always falls on the regression line.

Let $\overline{y}_{\text{old}}$ be the average market value of all players surveyed in 2018, when Haaland was worth 20 million. Let $\overline{y}_{\text{new}}$ be the average market value in 2019, when Haaland was worth 80 million. Since $80 > 20$, we know $\overline{y}_{\text{new}} > \overline{y}_{\text{old}}$. This means that $H'(\bar{x}) > H^*(\bar{x})$. Since $H^*(x)$ has a smaller value at $\bar{x}$ and a steeper slope than $H'(x)$, we know that the lines $H^*(x)$ and $H'(x)$ intersect to the right of $\bar{x}$, as in the picture.



This means as we move further left away from the intersection point, the difference between the two lines becomes more pronounced. Therefore, $H'(18) - H^*(18) > H'(25) - H^*(25)$, which means that the player with 18 goals (Lozano) is more affected by the change in prediction rule.

c) In this problem, we'll consider how our answer to part (b) might have been different if Haaland had more goals in both 2018 and 2019.

- Suppose Haaland instead had 31 goals in both 2018 and 2019. If his market value increased from 2018 to 2019, and everyone else's data stayed the same, which slope would be larger: $H^*(x)$ or $H'(x)$?
- Suppose Haaland instead had 45 goals in both 2018 and 2019. If his market value increased from 2018 to 2019, and everyone else's data stayed the same, 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.

Solution:

Note that when calculating $w_1' - w_1^*$, one of the immediate steps in our calculation said this:

$$w_1' - w_1^* = \frac{(16-31)(80-20)}{200 \cdot 6^2}$$

Here, 16 represents the number of goals Haaland had. Note that if instead Haaland had the average number of goals, 31, the difference in slopes would be 0. If he instead had more than the average number of goals, the difference in slopes would be positive, since the (number of goals$-$31) term would be positive and all other terms would remain the same.

In general, if Haaland has fewer than the average number of goals, then increasing his market value reduces the slope, whereas if Haaland has an above-average number of goals, then increasing his market value increases the slope. If Haaland happens to have the average number of goals, then increasing his market value doesn't change the slope.