DSC 40A - Group Work Session 3
due Monday, Jan 22 at 11:59pm

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. **One person** from each group should submit your solutions to Gradescope and **include all group members** so everyone gets credit.

This worksheet won't be graded on correctness, but rather on good-faith effort. Even if you don't solve any of the problems, you should include some explanation of what you thought about and discussed, so that you can get credit for spending time on the assignment.

In order to receive full credit, you must work in a group of two to four students for at least 50 minutes in your assigned discussion section. You can also self-organize a group and meet outside of discussion section for 80 percent credit. You may not do the groupwork alone.

1 Error in a Prediction Rule

The problems in this section test your understanding of definitions only. You should be able to write down the answers to these questions without referring to any notes or resources.

Problem 1.

Consider the data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and the linear prediction rule $y = 3x + 7$. Write down the expression for the mean squared error of this prediction rule on the data set.

Solution: The mean squared error is

$$\frac{1}{n}\sum_{i=1}^{n}(3x_i + 7 - y_i)^2.$$

Notice that we could equivalently write each term as $(y_i - (3x_i + 7))^2$.

Problem 2.

Consider the data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and the quadratic prediction rule $y = 2x^2 - 4x + 1$. Write down the expression for the mean absolute error of this prediction rule on the data set.

Solution: The mean absolute error is

$$\frac{1}{n}\sum_{i=1}^{n}|2x_i^2 - 4x_i + 1 - y_i|.$$

Notice that we could equivalently write each term as $|y_i - (2x_i^2 - 4x_i + 1)|$.

Equivalent Formulas for Linear Regression

In class, we showed that the slope and intercept of the regression line $H^*(x) = w_0^* + w_1^* x$ are given by

$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$w_0^* = \overline{y} - w_1^* \overline{x},$$

where $\bar{x}$ and $\bar{y}$ represent the mean of the x's and y's, respectively.

We also showed an equivalent form of the slope:

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

where σ_x and σ_y represent the standard deviations of the x's and y's, respectively.

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

Problem 3.

Show that

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i.$$

Substituting this into the numerator of w_1^* gives an equivalent formulation of the slope of the regression line:

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

Solution:

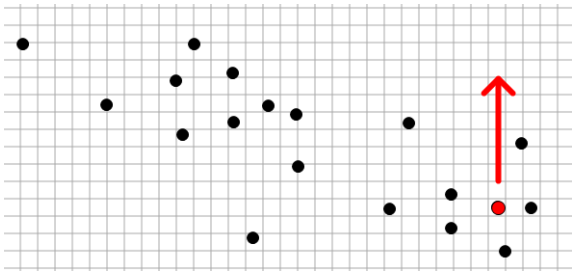
$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i - \bar{x})y_i - (x_i - \bar{x})\bar{y} \\
&= \sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}\right) \\
&= \sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}(n\bar{x} - n\bar{x}) \\
&= \sum_{i=1}^{n}(x_i - \bar{x})y_i
\end{aligned}$$

Visualizing Changes in the Data

The problems in this section will help you visualize how changes in the data affect the regression line. Assume all data is in the first quadrant (positive x and y coordinates).

Problem 4.

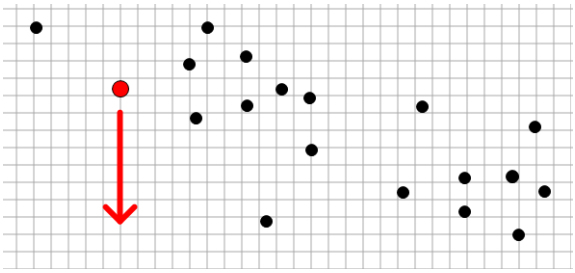
For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?



Solution: The regression line will tilt, becoming more shallow. This means the slope will increase (becoming less negative) and the intercept will decrease (assuming the data is located in the first quadrant).

Problem 5.

For the data set shown below, how will the slope and intercept of the regression line change if we move the red point in the direction of the arrow?

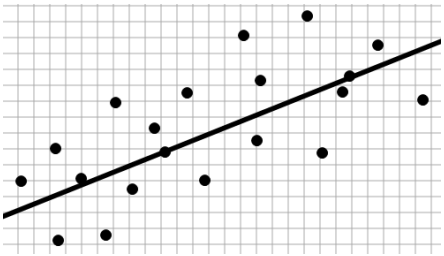


Solution: Again, the regression line will tilt, becoming more shallow. This means the slope will increase (becoming less negative) and the intercept will decrease (assuming the data is located in the first quadrant).

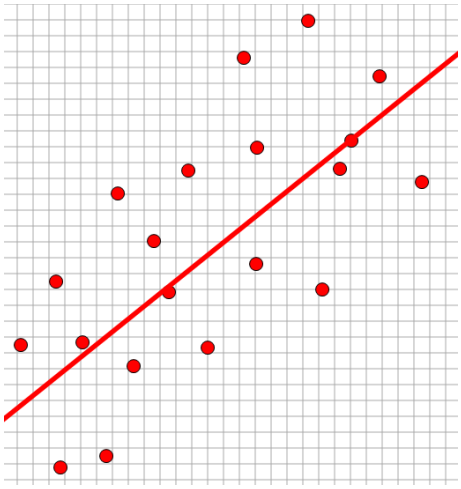
Problem 6.

Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each y-value, creating a transformed data set $\{(x_i, 2y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

Solution: The slope of the regression line fit to the transformed data should be twice the slope of the regression line fit to the original data. We can see this visually because the transformation stretches the points (and the regression line) vertically by a factor of 2, which doubles the slope. For example, if the original data $\{(x_i, y_i)\}$ and its regression line look like this:



then the transformed data $\{(x_i, 2y_i)\}$ and its regression line will look like this:



We can use the result of Homework 2 Problem 1 to show that when we apply the linear transformation $f(y_i) = 2y_i$, then $\overline{f(y)} = 2\bar{y}$. Then the slope of the new regression line becomes

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

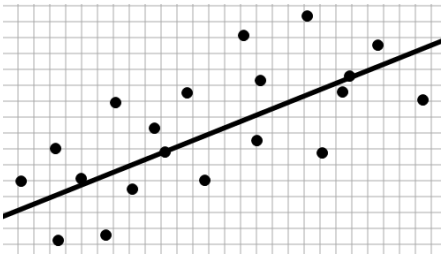
$$= 2 * \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

which shows that the slope of the regression line gets doubled.

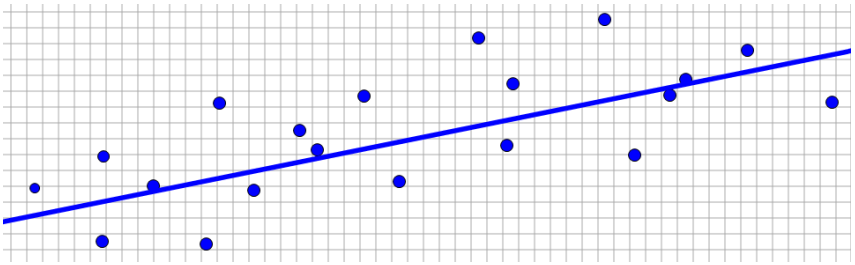
Problem 7.

Suppose we transform a data set of $\{(x_i, y_i)\}$ pairs by doubling each x-value, creating a transformed data set $\{(2x_i, y_i)\}$. How does the slope of the regression line fit to the transformed data compare to the slope of the regression line fit to the original data? Can you prove your answer from the formula for the slope of the regression line?

Solution: The slope of the regression line fit to the transformed data should be half the slope of the regression line fit to the original data. We can see this visually because the transformation stretches the points (and the regression line) horizontally by a factor of 2, which halves the slope. For example, if the original data $\{(x_i, y_i)\}$ and its regression line look like this:



then the transformed data $\{(2x_i, y_i)\}$ and its regression line will look like this:



We can use the result of Homework 2 Problem 1 to show that when we apply the linear transformation $f(x_i) = 2x_i$, then $\overline{f(x)} = 2\bar{x}$. Then the slope of the new regression line becomes

$$\begin{aligned}
\frac{\sum_{i=1}^{n}(2x_i - 2\bar{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(2x_i - 2\bar{x})^2} &= \frac{\sum_{i=1}^{n} 2(x_i - \bar{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(2(x_i - \bar{x}))^2} \\
&= \frac{\sum_{i=1}^{n} 2(x_i - \bar{x})(y_i - \overline{y})}{\sum_{i=1}^{n} 4(x_i - \bar{x})^2} \\
&= \frac{1}{2} * \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

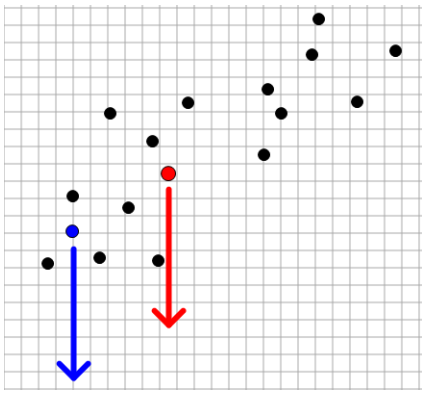
which shows that the slope of the regression line gets halved.

Problem 8.

Compare two different possible changes to the data set shown below.

- Move the red point down c units.
- Move the blue point down c units.

Which move will change the slope of the regression line more? Why?



Solution: Moving the blue point will change the slope of the regression line more. We can see this using the following formula for the slope.

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

Notice that when we move either point down by c units, $\bar{x}$ and each x_i stay the same. So the slope of the regression line before and after such a move has the same denominator. We can tell which move has a larger effect on w_1 by comparing the change in the numerator only.

Say the coordinates of the red point are (x_{red}, y_{red}). If we move it down by c units, the coordinated become $(x_{red}, y_{red} - c)$. In the formula for the numerator of w_1, each term in the sum is the same before and after this move, except for the one term corresponding to the red point, which is

$$(x_{red} - \bar{x}) * y_{red}$$

before the move and

$$(x_{red} - \bar{x}) * (y_{red} - c)$$

after the move, which means the move changes the numerator of w_1 by

$$\begin{aligned}(x_{red} - \bar{x}) * (y_{red} - c) - (x_{red} - \bar{x}) * y_{red} &= -(x_{red} - \bar{x}) * c \\ &= (\bar{x} - x_{red}) * c\end{aligned}$$

Similarly, if the coordinates of the blue point are (x_{blue}, y_{blue}), then moving the blue point down c units changes the numerator of w_1 by $(\bar{x} - x_{blue}) * c$.

Since the red point is closer to the average x, we have $(\bar{x} - x_{red}) * c < (\bar{x} - x_{blue}) * c$, which means moving the blue point changes the slope of w_1 more than moving the red point.

That is, moving a point further away from the mean has a more substantial effect on the regression line than moving a point closer to the mean.