**DSC 40A - Group Work Session 3**
due Wednesday, April 18 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 **tag 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.

**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.

# 2 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 $\sigma_x$ and $\sigma_y$ 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:

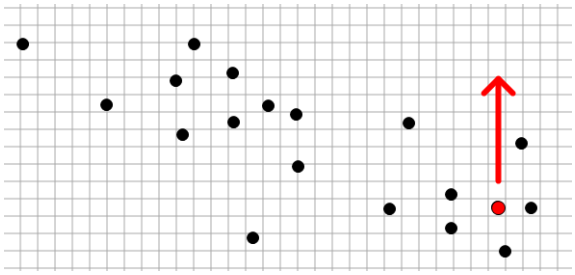
$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

# 3 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).

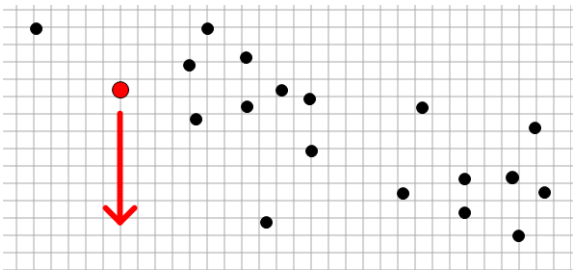
**Problem 4.**

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?



**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?



**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?

**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?

**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?