DSC 40A Fall 2024 - Group Work Session 3

due Monday, October 14th 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 No intercept!

In the lecture this week you encountered the **linear model**, which seeks to model a collection of input data $\{x_i\}$ and corresponding observed output data $\{y_i\}$ with a linear function $H(x_i) = w_0 + w_1 x_i$. In this problem we will take a look at a simplified version, the **intercept-free linear model** which takes the form $H(x_i) = wx_i$ for some scalar $w \in \mathbb{R}$.

It is important to node that the intercept-free linear model is the same as the linear model when $w_0 = 0$. Therefore, we should consider it as a weaker or less generalizable type of model and it should be used in a smaller set of circumstances compared to the traditional linear model.

Nevertheless, it can still find its way into some interesting use cases.

Problem 1.

Look over the lecture slides and review the **modeling recipe**. Write the three steps ("ingredients") here. The next four problems form a walkthrough of these steps.

Problem 2.

For each of the following hypothetical situations, identify the input (or independent) variable x and the output (or response) variable y. Then, explain whether the **linear model** or the **intercept-free linear model** would be more appropriate.

a) Every day for three months, a student collected the price of a gallon of gasoline in US Dollars at the gas station across from campus and now they want to try and understand the data with a model.

b) Julie owns a flower shop and, starting from opening time, she writes down the cumulative amount of sales in Euros for each hour of the day. She now wants to estimate the rate of dollars she earns per hour on a typical day.

c) Tom is a botanist at UCSD and is researching new subspecies of cherry tomato that are more drought- and heat-resistant. On the first day of an experiment, he plants a tomato seedling. Then, for each of the 100 subsequent days, he records its height in centimeters. He now wants to develop a model to predict the height of the plant on a given number of days after planting.

d) Arnie is a farmer at UC Davis and is researching the impact of certain dietary supplements on the growth of newborn cattle. He has ten different dosages of vitamin D that he is adding to various calfs' food troughs. For each dosage level, he feeds a particular calf the same amount of vitamin D each day for a month. At the end of the month, he records their weight. He now wants to understand the relationship between dosage amounts and calf weight one month after birth.

Problem 3.

Now assume that we are working with Tom at UCSD (see **(c)** above). He gives us access to the first four days of data so that we can get started on the model while the plant grows.

Days after planting	Plant height (cm)
0.0	0.0
1.0	0.4
2.0	1.3
3.0	1.7

a) Suppose we want to model this data using a **linear model** combined with the **square loss**. Write down the hypothesis function $H(t_i)$ where t_i is a time in days after planting. Then, write down the empirical risk function $R_{\text{sq}}(H(t_i), y_i)$ for the data in the table above.

b) Suppose we want to model this data using a **intercept-free linear model** combined with the **absolute loss**. Write down the hypothesis function $H(t_i)$ where t_i is a time in days after planting. Then, write down the empirical risk function $R_{\text{abs}}(H(t_i), y_i)$ for the data in the table above.

Problem 4.

a) After taking a closer look at the data in the last problem, suppose we wish to continue by using the **intercept-free linear model** combined with the **absolute loss** (as in part **(b)**). Use Desmos to plot the empirical risk function $R_{\text{abs}}(H(t_i), y_i)$ in terms of the parameter w for the hypothesis function. Sketch the plot here and find the value w^* which is a minimizer for the empirical risk.

b) In the previous part, how should the researcher interpret the value of w^* within the context of the experiment and data? What are some advantages and disadvantages of using $R_{\text{abs}}(H(t_i), y_i)$ compared to, for example, $R_{\text{sq}}(H(t_i), y_i)$?

Problem 5.

After talking to Tom, he suggested we use the **square loss** alongside the **intercept-free linear model** and asked us to find a generic formula that he use every morning as he expands and updates his dataset.

a) Suppose after n days $\{t_1, \ldots, t_n\}$ he records plant heights $\{y_1, \ldots, y_n\}$. Write the empirical risk function $R_{\text{sq}}(H(t_i), y_i)$ for the data collected up to this point.

b) Show that the parameter w^* which minimizes the empirical risk for the square loss of the intercept-free linear model $H(t_i) = wt_i$ is given by the formula:

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

c) After a few days, Tom updates our dataset with new entries below. Use a calculator/computer and the formula in the previous part to find w^*, and give Tom a prediction for the height of the plant after $t = 10$ days.

Days after planting	Plant height (cm)
0.0	0.0
1.0	0.4
2.0	1.3
3.0	1.7
4.0	2.0
5.0	2.5
6.0	3.2

2 Least Squares Regression

Recall that the least squares solutions to the problem of fitting a straight line, $H(x) = w_0 + w_1 x$, to the data (x_i, y_i) are:

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

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Problem 6.

a) Let $c \in \mathbb{R}$ be a fixed scalar and suppose we transform the data (y_i) by multiplying by the scalar c. Show that the slope of the least squares regression line for the transformed dataset (x_i, cy_i) is given by cw_1^*. That is, the slope of the regression line is transformed by c as well.

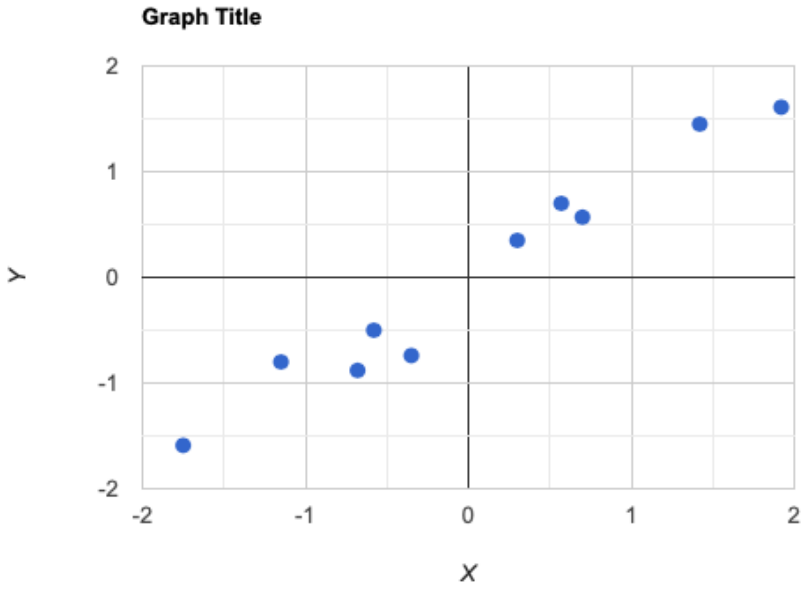
b) Suppose that instead of multiplying by c, we transform the dataset by translating the data (y_i) by c. What will be the slope of the least squares regression line for the transformed dataset $(x_i, y_i + c)$?

Problem 7.

Consider the data set given below. Fit a line $y = a + bx$ by the method of least squares, where x is the predictor variable and y is the response variable, and sketch it on the plot. Fit a line $x = c + dy$ by the method of least squares, where y is the predictor variable and x is the response variable, and sketch it on the plot. (You can use a calculator or Python to find a, b, c, and d.)

Are the lines the same or different? Explain why.

x	0.30	1.42	-0.58	0.70	1.92	-0.68	-0.35	-1.15	0.57	-1.75
y	0.35	1.45	-0.50	0.57	1.61	-0.88	-0.74	-0.80	0.70	-1.59



Problem 8.

Consider the problem of fitting a function of the form $H(x) = b_0 + b_1 \sin(x)$ to the data $(x_1, y_1), \ldots, (x_n, y_n)$. What are the least squares solutions for b_0 and b_1?

Hint: While this looks different than what we've studied in lecture, it turns out that it's quite similar. What if we define a change of variables, such that $z_i = \sin(x_i)$?