DSC 40A Fall 2025 - Group Work Session 3
due Monday, October 13th 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.

Solution:
1. Choose a model.
2. Choose a loss function.
3. Minimize average loss to find optimal model parameters.

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.

Solution: The input variable x is the date measured in days, and the response variable y is the price of a gallon of gasoline measured in USD. One expects that on day "0" (i.e. $x = 0$), the price of gasoline is positive (for example, $y = 4.00$). Therefore, $H(0) \neq 0$ for this situation and a **linear model** is more appropriate.

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.

> **Solution:** The input variable x is each hour of the day, and the response variable y is the cumulative sale amount in Euro. The sale amount will be expected to be 0 Euro at the start of the day, therefore the intercept-free linear model will be more appropriate.

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.

> **Solution:** The input variable x is the date of every 100 days after the seeds have been planted, and the response variable y is the height of the plant in centimeters. On day 0, we will expect the height of the plant to be 0, the intercept free linear model will be the better option.

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.

> **Solution:** The input variable x will be the dosage, and the response variable here, y, will be the weight of the cows after one month. In this case, because on day 0 we expect cows to have weight to start with, the linear model (with intercept) will be more appropriate.

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.

> **Solution:** In this case $H(t_i) = w_0 + w_1 t_i$ for some $w_0, w_1 \in \mathbb{R}$. Using the square loss, we have
>
> $$R_{\text{sq}}(H(t_i), y_i) = \frac{1}{4}\left(|0.0 - w_0|^2 + |0.4 - w_0 - w_1|^2 + |1.3 - w_0 - 2w_1|^2 + |1.7 - w_0 - 3w_1|^2\right)$$

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.

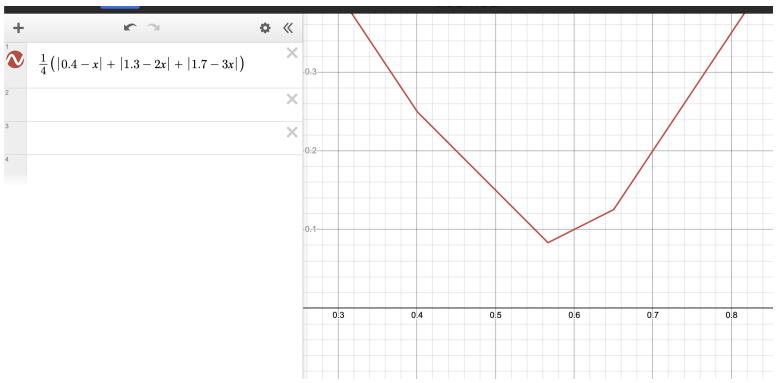
Solution: With no intercept, the equation we will use becomes $H(t_i) = wt_i$ for some $w \in \mathbb{R}$. Using absolute loss, we have

$$R_{\text{abs}}(H(t_i), y_i) = \frac{1}{4}\left(|0.4 - w| + |1.3 - 2w| + |1.7 - 3w|\right)$$

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.

Solution:



The minimizer occurs when $w^* \approx 0.567$ according to the plot.

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, $R_{\text{sq}}(H(t_i), y_i)$?

Solution: The value of w* should represent the minimizer of the intercept-free linear model, meaning it is the best-fit growth rate, a prediction of how much the plant should grow per day. Advantages of using mean square error ($R_{\text{sq}}(H(t_i), y_i)$) is that you will be able to find its derivative when finding a minimizer, therefore easy to optimize; mean absolute error $R_{\text{abs}}(H(t_i), y_i)$ is the best when your data has lots of outliers, since outliers can skew the mean.

2 Least Squares Regression

Recall that for a simple linear hypothesis function $H(x) = w_0 + w_1 x$ trained on data (x_i, y_i), the parameters minimizing mean squared error are given by:

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

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 5.

a) Let $c \in \mathbb{R}$ with $c \neq 0$ 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.

Solution: We'll use the slope equation above:

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

When we multiply all the y-values by the same constant c, notice that the average y value also gets multiplied by c. Then the slope of the new regression line becomes

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(cy_i - c\overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{c * \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = cw_1^*,$$

which shows that the slope of the regression line gets multiplied by c.

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 $(x_i, y_i + c)$?

Solution: Now, when the same constant value c is added to each of the y-values, we would notice that the average y value also increases by c. Then, the slope of the new regression line becomes:

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

which shows that the slope of the regression line does not change.

Problem 6.

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 along the lines of* $z_i = \sin(x_i)$*?*

Solution: Since H is linear in terms of b_0 and b_1, we can substitute $\sin(x_i)$ for x_i.

$$b_1^* = \frac{\sum_{i=1}^{n} (\sin(x_i) - (1/n)\sum_{i=1}^{n} \sin(x_i))(y_i - \bar{y})}{\sum_{i=1}^{n} (\sin(x_i) - (1/n)\sum_{i=1}^{n} \sin(x_i))^2}$$

$$b_0^* = \bar{y} - b_1^* \frac{1}{n} \sum_{i=1}^{n} \sin(x_i)$$