

---

**DSC 40A - Homework 2**  
due Friday, October 17th at 11:59 PM

---

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; you have four slip days to use in total throughout the quarter.


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. **Only handwritten solutions will be accepted (use of tablets is permitted). Do not typeset your homework (using L<sup>A</sup>T<sub>E</sub>X or any other software).**

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 **59 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](#), 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. Speed Measurement

A coastal research station records hourly surface current velocities

$$\vec{y}_i = (y_i^{(1)}, y_i^{(2)}) \in \mathbb{R}^2,$$

measured in meters per second, where  $y_i^{(1)}$  is the eastward component and  $y_i^{(2)}$  is the northward component. Oceanographers suspect that during the study period the motion is driven predominantly by a single tidal stream that always points in the direction

$$\vec{d} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

but whose *speed* (magnitude) is unknown. They therefore propose the vector-valued constant model

$$f(t) = h\vec{d}, \quad h \in \mathbb{R}.$$

Six velocity measurements (in  $\text{ms}^{-1}$ ) and their timepoints (in hours from the start of the experiment) are listed below:

$$\{(t_i, \vec{y}_i)\}_{i=1}^6 = \left\{ (0, (0.60, 1.21)), (1.1, (0.70, 1.47)), (1.9, (0.50, 0.93)), \right. \\ \left. (2.5, (0.65, 1.25)), (3.2, (0.55, 1.11)), (4.0, (0.72, 1.38)) \right\}.$$

- a) 🥑🥑🥑 Identify the predictor and response variables, state the hypothesis function, and write down a formula for the squared loss  $L_{sq}(h)$  in this setting.

**Solution:**

The hypothesis function is synonymous with the model or prediction function. In this case, it is

$$f(t) = h\vec{d}.$$

Predictor variables are the inputs our model uses to make predictions, and response variables are the variables we are predicting (if our model is any good, they should depend on or "respond to" changes in the predictor variables).

Examining our hypothesis function, we see that  $t$  is the predictor variable.

The output of  $f$ ,  $h\vec{d}$ , is the direction of the tidal stream scaled by a vector. In a supervised learning problem, we predict labels for unseen data based on the pattern we learn between the features and labels of our dataset. Since the labels of our dataset are the  $\vec{y}_i$  vectors, we can infer that we are scaling  $\vec{d}$  so that it represents a vector of *velocities* in the eastwards and northwards directions. Therefore we could either say that the response variable is a velocity vector or that there are two response variables, the eastwards velocity and the northwards velocity.

Notice that  $\vec{y}_i$  is a vector in  $\mathbb{R}^2$  and so is  $h\vec{d}$ , so the most natural loss function is

$$L_{sq}(h) = \|\vec{y} - h\vec{d}\|^2$$

for a vector  $y$ .

- b) 🥑🥑🥑🥑 Write the empirical risk function

$$R(h) = \frac{1}{6} \sum_{i=1}^6 L_{sq}(h)$$

for the data above and derive the critical point equation

$$h^* = \frac{\sum_{i=1}^n \vec{d}^\top \vec{y}_i}{\sum_{i=1}^n \|\vec{d}\|^2}.$$

**Solution:**

$$\begin{aligned}
R(h) &= \frac{1}{6} \sum_{i=1}^6 \|\vec{y}_i - h\vec{d}\|^2 \\
&= \frac{1}{6} \sum_{i=1}^n (\vec{y}_i - h\vec{d})^\top (\vec{y}_i - h\vec{d}) \\
&= \frac{1}{6} \sum_{i=1}^n (\vec{y}_i^\top - h\vec{d}^\top) (\vec{y}_i - h\vec{d}) \\
&= \frac{1}{6} \sum_{i=1}^n (\vec{y}_i^\top y_i - h\vec{y}_i^\top \vec{d} - h\vec{d}^\top \vec{y}_i + h^2 \vec{d}^\top \vec{d}) \\
&= \frac{1}{6} \sum_{i=1}^n \|y_i\|^2 - \frac{h}{3} \sum_{i=1}^n \vec{d}^\top \vec{y}_i + \frac{h^2}{6} \sum_{i=1}^n \|\vec{d}\|^2
\end{aligned}$$

Differentiating, we have

$$\frac{d}{dh} R(h) = -\frac{1}{3} \sum_{i=1}^n \vec{d}^\top \vec{y}_i + \frac{h}{3} \sum_{i=1}^n \|\vec{d}\|^2$$

Setting the derivative to zero gives

$$\begin{aligned}
-\frac{1}{3} \sum_{i=1}^n \vec{d}^\top \vec{y}_i + \frac{h}{3} \sum_{i=1}^n \|\vec{d}\|^2 &= 0 \\
\Rightarrow \frac{h}{3} \sum_{i=1}^n \|\vec{d}\|^2 &= \frac{1}{3} \sum_{i=1}^n \vec{d}^\top \vec{y}_i \\
\Rightarrow h &= \frac{\sum_{i=1}^n \vec{d}^\top \vec{y}_i}{\sum_{i=1}^n \|\vec{d}\|^2}
\end{aligned}$$

- c) 🥑🥑🥑 Verify that the second derivative  $\frac{d^2 R}{dh^2}$  is positive and conclude that  $h^*$  is the unique global minimizer.

**Solution:**

$$\frac{d^2}{dh^2} R(h) = \frac{1}{3} \sum_{i=1}^n \|\vec{d}\|^2 = \frac{1}{3} n \|\vec{d}\|^2$$

Since  $\vec{d} = (1, 2)^\top$ ,  $\|\vec{d}\|^2 = 1 + 4 = 5$ . Hence  $\frac{d^2}{dh^2} R(h) > 0$ , proving by the second derivative test that  $h^*$  is a minimizer. Notice that the original risk function is quadratic with respect to  $h$ . Quadratic functions have one minimizer or maximizer. Hence this critical point is the unique global minimum.

- d) 🥑🥑 Evaluate the expressions in part (b) for the six data points and report the numerical value of  $h^*$ .

**Solution:**

The denominator is equal to  $n\|d\|^2 = 6 \cdot 5 = 30$ . For each  $\vec{y}_i$ ,  $d^\top \vec{y}_i = \vec{y}_i^{(1)} + 2\vec{y}_i^{(2)}$ . Hence

$$h^* = \frac{.6 + 2.42 + .7 + 2.94 + .5 + 1.86 + .65 + 2.5 + .55 + 2.22 + .72 + 2.76}{30} = .614$$

### Problem 3. Pop Quiz

Complete the following two concept checks.

- a) 🥑🥑🥑🥑 Determine whether each statement is true or false. Explain briefly.
- (a) If the training features  $x_1, \dots, x_n$  for a simple linear regression problem have mean zero then the intercept of the optimal linear model will equal zero.
  - (b) The line of best fit is found by minimizing the empirical risk.
  - (c) The slope of the line of best fit is always positive.
  - (d) A simple linear regression model has exactly two parameters.

#### Solution:

- a False. Consider a dataset  $x=(-1,0,1)$  and  $y=(10,10,10)$ . The training features ( $x$ ) have a mean of 0 but if you actually solve for the optimal intercept, it will be 10.
- b True. Remember the modeling recipe: pick a model, pick a loss function, then minimize the empirical risk. The "line that fits the best" is usually defined by the fact that it has the lowest empirical risk compared to other lines.
- c False. Intuitively, when using the squared loss, if two variables have a negative correlation, the optimal slope will be negative. Mathematically, when using squared loss, the optimal slope  $w_1^*$  is  $w_1^* = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$ . Consider the dataset  $x=(1,3)$  and  $y=(-1,-3)$ . When you calculate the slope, the result will be  $w_1^* = \frac{(1-2)(-1+2) + (3-2)(-3+2)}{(1-2)^2 + (3-2)^2} = -1$
- d True. Simple linear regressions are of the form  $H(x) = w_0 + w_1x$ . That's only 2 parameters:  $w_0$  and  $w_1$ . Note that when you have more than 2 parameters, that would become "Multiple Linear Regression".

- b) 🥑🥑 Match each term (a)–(e) with its correct definition (1)–(5) (no explanation required):

|     |                  |  |     |
|-----|------------------|--|-----|
| (a) | Empirical risk   | Equations obtained by setting partial derivatives of empirical risk to zero.                           | (1) |
| (b) | Intercept        | Average value of the loss function over the training data.   | (2) |
| (c) | Square loss      | Parameter indicating the amount that the predicted value changes if the feature increases by one unit. | (3) |
| (d) | Normal equations | Parameter indicating the predicted value when the feature is zero.                                     | (4) |
| (e) | Slope            | A loss function measuring the squared difference between prediction and actual value.                  | (5) |

#### Solution:

- a 2
- b 4
- c 5
- d 1

#### Problem 4. Potion Brewing

An alchemist measures gold created (in grams) against hours spent brewing philosopher's potions and obtains the following dataset.

|                     |     |      |     |     |     |     |
|---------------------|-----|------|-----|-----|-----|-----|
| Hours spent brewing | 2   | 3    | 5   | 7   | 8   | 11  |
| Gold created        | 0.5 | 0.75 | 1.2 | 1.7 | 2.1 | 2.9 |

- a) 🥑🥑 Determine the slope and intercept of the simple linear model  $y = w_1x + w_0$  which minimizes mean squared error for the given dataset.

**Solution:** From lecture we know that

$$w_1^* = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \quad \text{and} \quad w_0^* = \bar{y} - w_1^* \bar{x}.$$

First we calculate the means:

$$\bar{x} = \frac{2 + 3 + 5 + 7 + 8 + 11}{6} = 6, \quad \bar{y} = \frac{0.5 + 0.75 + 1.2 + 1.7 + 2.1 + 2.9}{6} = 1.525.$$

Next, calculate the numerator and denominator:

$$\begin{aligned} \sum_i (x_i - \bar{x})(y_i - \bar{y}) &= (-4)(-1.025) + (-3)(-0.775) + (-1)(-0.325) + (1)(0.175) + (2)(0.575) + (5)(1.375) \\ &= 14.95 \end{aligned}$$

$$\sum_i (x_i - \bar{x})^2 = (-4)^2 + (-3)^2 + (-1)^2 + (1)^2 + (2)^2 + (5)^2 = 56$$

Therefore,

$$w_1^* = \frac{14.95}{56} = 0.26696, \quad w_0^* = 1.525 - 0.26696(6) = -0.0768.$$

And so the fitted regression line is

$$\hat{y} = 0.267x - 0.0768.$$

- b) 🥑🥑 Predict the amount of gold the alchemist will create after 10 hours of brewing philosopher's potions.

**Solution:** Using  $\hat{y} = 0.267x - 0.0768$ , for  $x = 10$ :

$$\hat{y} = 0.267(10) - 0.0768 = 2.5932.$$

So the alchemist is predicted to create approximately 2.59 grams of gold after 10 hours.

#### Problem 5. Tour de France

Two bicyclists are training to compete in the Tour de France. One lives in San Diego, US and the other lives in Stuttgart, Germany. Each week, they record their training distances in miles and kilometers, respectively.

For each week  $i$ , let

$$x_i = \text{Athlete A's distance (in miles)}, \quad y_i = \text{Athlete B's distance (in kilometers)}.$$

The paired measurements are:

| Athlete A (miles) | Athlete B (km) |
|-------------------|----------------|
| 186.4             | 300            |
| 199.5             | 320            |
| 211.3             | 340            |
| 223.7             | 360            |
| 236.1             | 380            |
| 248.5             | 400            |
| 260.9             | 420            |
| 273.4             | 440            |

Your goal is to use simple linear regression to model Athlete B's distance as a linear function of Athlete A's distance:

$$y = w_1 x + w_0.$$

You may also find it convenient to convert Athlete A's distances to kilometers using

$$1 \text{ mile} = 1.60 \text{ km} \quad \Rightarrow \quad x_{\text{km}} = 1.60 x. \quad (\star)$$

a) 🥑🥑🥑🥑 Your friend Zoe suggests two equivalent approaches:

**Workflow A:** Convert each  $x_i$  to  $x_{i,\text{km}} = 1.60 x_i$  and model  $y$  on  $x_{\text{km}}$  to get coefficients:  $\tilde{w}_1$  (in units of *Athlete B km per Athlete A km*), and  $\tilde{w}_0$  (in units of *km*).

**Workflow B:** Model  $y$  on  $x$  (in *miles*) to get coefficients:  $w_1$  (in units of *Athlete B km per Athlete A mi*), and  $w_0$  (in units of *km*); then convert  $w_1$  (possibly shifting  $w_0$  as needed) using  $(\star)$ .

Is Zoe correct that both workflows lead to the same fitted line in kilometers? Show your calculations for computing  $(w_1, w_0)$  and  $(\tilde{w}_1, \tilde{w}_0)$ .

*Note: In both workflows, the units of  $w_0$  are always in km since they should match the output of the model. In workflow B we do not convert them directly after training the model, but it is possible that they shift by a constant (see part (b)).*

**Solution:**

Let  $x'_i = 1.6x_i$  and  $\bar{x}' = \frac{1}{n} \sum_{i=1}^n x'_i$ . Notice that

$$\bar{x}' = \frac{1}{n} \sum_{i=1}^n 1.6x_i = 1.6\bar{x}$$

Using the equations for the optimal slope and the optimal intercept, it follows that

$$\begin{aligned}
 \tilde{w}_1 &= \frac{\sum_{i=1}^n (x'_i - \bar{x}')(y_i - \bar{y})}{\sum_{i=1}^n (x'_i - \bar{x}')^2} \\
 &= \frac{\sum_{i=1}^n (1.6x_i - 1.6\bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (1.6x_i - 1.6\bar{x})^2} \\
 &= \frac{\sum_{i=1}^n 1.6(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n 1.6^2(x_i - \bar{x})^2} \\
 &= \frac{1}{1.6} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 &= \frac{1}{1.6} w_1
 \end{aligned}$$


and that

$$\begin{aligned}
 \tilde{w}_0 &= \bar{y} - \tilde{w}_1 \bar{x}' \\
 &= \bar{y} - \frac{1}{1.6} w_1 \cdot 1.6\bar{x} \\
 &= \bar{y} - w_1 \bar{x} \\
 &= w_0.
 \end{aligned}$$

Examining these equalities, we see that Zoe's hypothesis is correct. It does not matter whether we predict on the scaled or unscaled data. Letting  $\tilde{h}(x')$  and  $h(x)$  be the hypothesis functions corresponding to workflow A and B respectively, we have

$$\tilde{h}(x') = \tilde{w}_1 x' + \tilde{w}_0 = \frac{w_1}{1.6} \cdot 1.6x + w_0 = w_1 x + w_0 = h(x)$$

Both approaches predict the same values.

- b)  More generally, consider the simple linear regression fit of a response  $y$  on features  $x$  is given by  $y = w_1 x + w_0$ . Now replace  $x$  by a linear transformation  $z = ax + b$  (with constants  $a, b \in \mathbb{R}$ ,  $a \neq 0$ ) and refit  $y$  on  $z$ .

Express the new slope and intercept  $(\hat{w}_1, \hat{w}_0)$  in terms of  $w_1, w_0, a, b$ . Prove your formulas (e.g., rewrite the original fitted line using  $z$  and match coefficients).

**Solution:** Let  $\hat{w}_1$  be the optimal slope found from fitting  $y$  on  $z$  and  $w_1$  be the optimal slope found from fitting  $y$  on  $x$ . First, we'll prove a property about  $\bar{z}$ :

$$\begin{aligned}
 \bar{z} &= \frac{1}{n} \sum_{i=1}^n z_i \\
 &= \frac{1}{n} \sum_{i=1}^n (ax_i + b) \\
 &= \frac{a}{n} \sum_{i=1}^n x_i + \frac{nb}{n} \\
 &= a\bar{x} + b
 \end{aligned}$$

Next, we'll solve for the optimal slope

$$\begin{aligned}
 \hat{w}_1 &= \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} \\
 &= \frac{\sum_{i=1}^n (ax_i + b - a\bar{x} - b)(y_i - \bar{y})}{\sum_{i=1}^n (ax_i + b - a\bar{x} - b)^2} \\
 &= \frac{\sum_{i=1}^n a(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (a(x_i - \bar{x}))^2} \\
 &= \frac{a}{a^2} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 &= \frac{w_1}{a}
 \end{aligned}$$

Similarly, we can solve for  $\hat{w}_0$ :

$$\begin{aligned}
 \hat{w}_0 &= \bar{y} - \hat{w}_1 \bar{z} \\
 &= \bar{y} - \frac{w_1}{a} (a\bar{x} + b) \\
 &= \bar{y} - w_1 \bar{x} - \frac{b}{a} w_1 \\
 &= w_0 - \frac{b}{a} w_1
 \end{aligned}$$

- c) 🥑🥑🥑 For any data set  $(x_i, y_i)_{i=1}^n$ , define  $z_i = ax_i + b$ . Show that the *minimum* mean squared error (MSE) from regressing  $y$  on  $x$  equals the minimum MSE from regressing  $y$  on  $z$ .

**Solution:**

Remember that  $\hat{w}_1, \hat{w}_0$  were chosen because they are the values that minimize MSE when regressing  $y$  on  $z$ . Similarly,  $w_1$  and  $w_0$  were chosen to be the slope and intercept values that minimize MSE when regressing  $y$  on  $x$ . Therefore the minimum MSE when regressing  $y$  on  $z$  is

$$\begin{aligned}
 \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{w}_1 z_i + \hat{w}_0))^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \frac{w_1}{a} (ax_i + b) - w_0 + \frac{b}{a} w_1)^2 \\
 &= \frac{1}{n} \sum_{i=1}^n (y_i - w_1 x_i - \frac{b}{a} w_1 - w_0 + \frac{b}{a} w_1)^2 \\
 &= \frac{1}{n} \sum_{i=1}^n (y_i - (w_1 x_i + w_0))^2
 \end{aligned}$$

which is the minimum MSE when regressing  $y$  on  $x$ .

- d) 🥑🥑 Determine if the following is true or false by either proving the statement or finding a counterexample.

*The simple linear regression line always passes through  $(\bar{x}, \bar{y})$ .*

**Solution:** The statement is true. Let  $w_1^*$  and  $w_0^*$  be the slope and intercept chosen to minimize MSE for our dataset. We know that  $w_0^* = \bar{y} - w_1^* \bar{x}$ . Hence if  $h$  is our hypothesis function,

$$h(x) = w_1^* x + \bar{y} - w_1^* \bar{x}$$



and so

$$h(\bar{x}) = w_1^* \bar{x} + \bar{y} - w_1^* \bar{x} = \bar{y}$$

As you attempt this problem, it might be helpful to review [Example 2.1.4 in the course notes](#).

### Problem 6. Explain The Plot

The four plots below contain a training dataset  $\{(x_i, y_i)\}_{i=1}^{30}$  of thirty points in the  $xy$ -plane (blue), along with the line of best fit for the associated simple linear regression problem (red).

- a) 🥑🥑 For each plot below, determine whether the simple linear model is appropriate. Write a short explanation of your reasoning.

#### Solution:

1. This is a noisy dataset, and we do not want our model to fit to the noise. Any model more complex than a linear model is therefore not a good match for (i). Since there is no clear linear trend, however, the constant model is likely a better model than the linear one.
2. The data clearly follows a linear trend (and with very little noise). A linear model is a great choice.
3. The data does not follow a linear trend. It would be much better to fit a sine wave to this dataset or perhaps a cubic function.
4. Like in (ii), the data clearly follows a linear trend. In most situations, the best choice here is to ignore the outlier and fit a line to the rest of the data. There is a clear natural trend that our model should capture precisely.

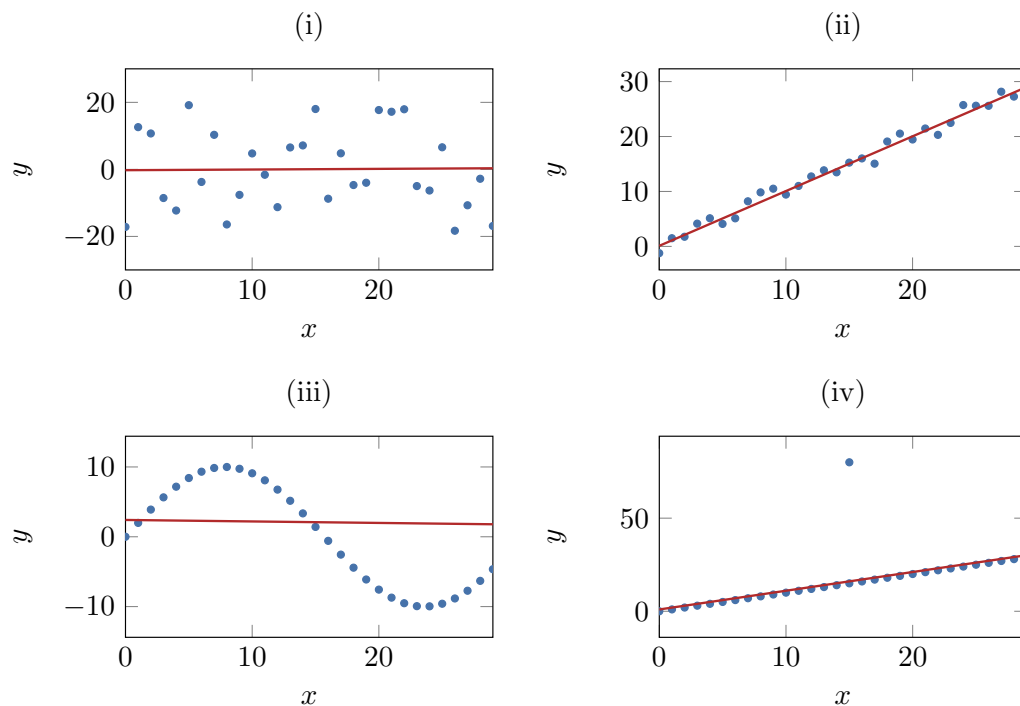
- b) 🥑🥑🥑 For each plot, consider the value  $R(w_0^*, w_1^*)$  of the empirical risk associated to the optimal parameters for the simple linear model and given training dataset. Rank the plots from least to greatest value of  $R(w_0^*, w_1^*)$ .

#### Solution:

From least to greatest risk, the plots should be ordered as

$$(2) \rightarrow (3) \rightarrow (1) \rightarrow (4)$$

(ii) has the least risk since the largest squared error looks to be approximately 4, and most squared errors are less than 1. Looking at the units of graphs (i) and (iii), we see that the points of (iii) are generally much closer to the line of best fit than the points of (i). Finally, the outlier in (iv) is so far from the rest of the data, that once its error is squared, it will dramatically increase the MSE.



### Problem 7. Does more data help (for linear regression)?

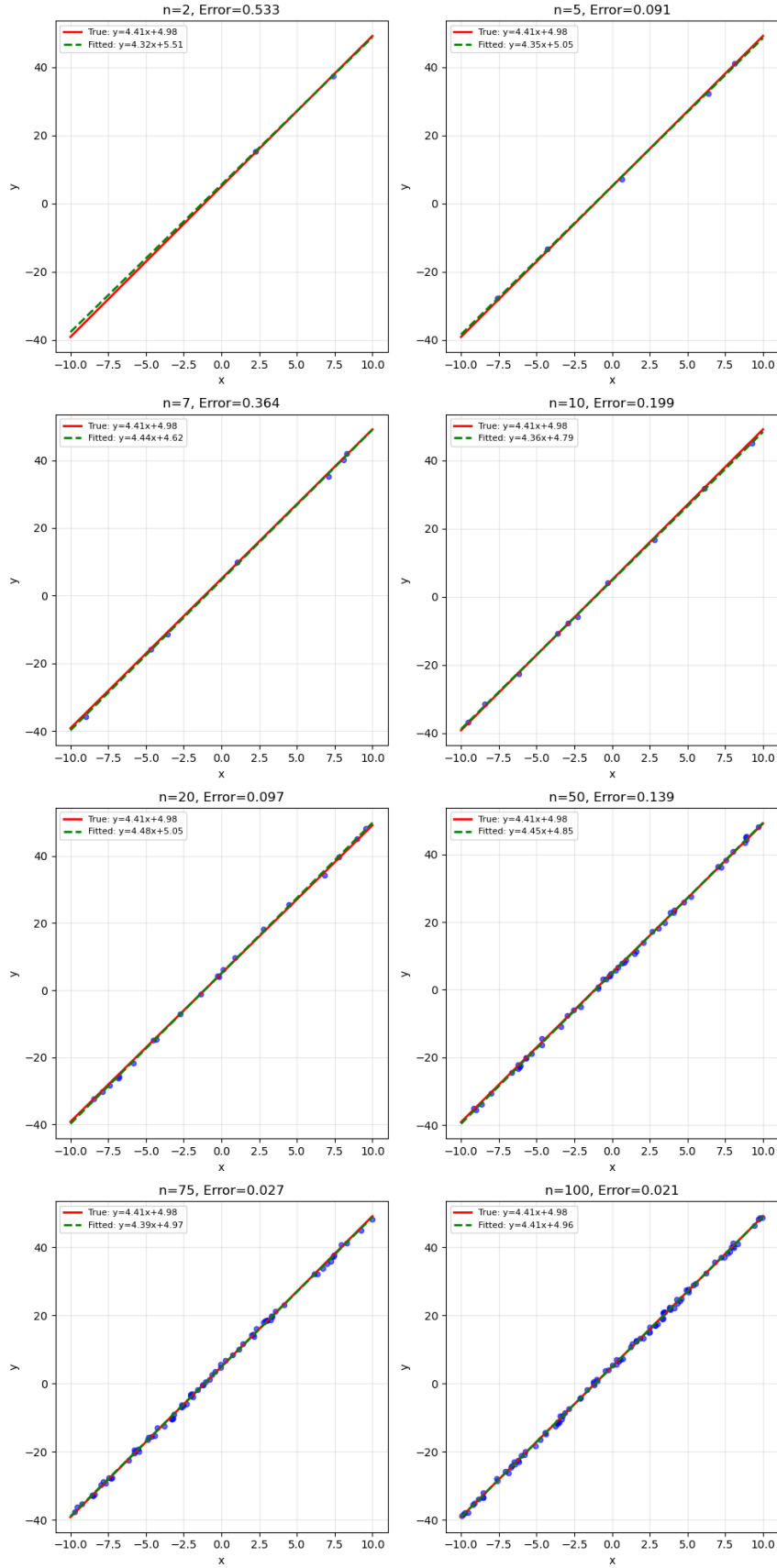
In this problem we will generate a dataset and implement simple linear regression.

The question is provided at [this supplementary Jupyter Notebook](#). The code that you write in that notebook will **not** be graded and you do not need to submit the code on Gradescope. You **do** need to add the plots you generate in the notebook below and answer the two questions.

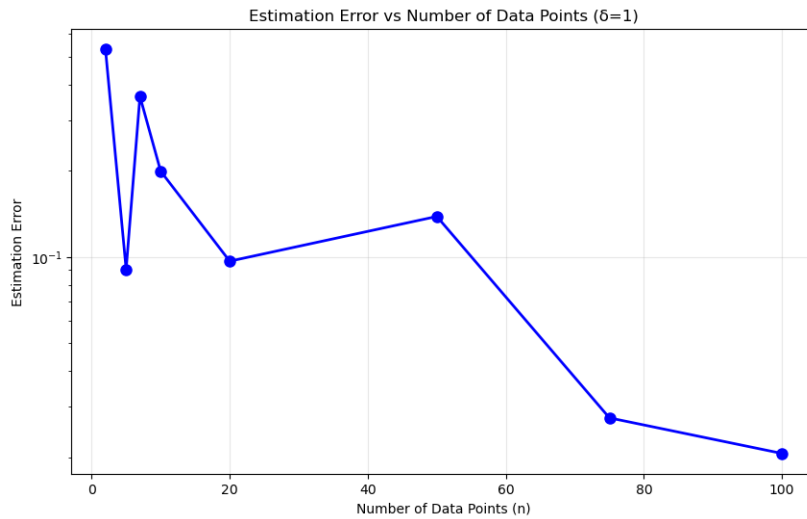
- a) 🥑🥑🥑🥑 Add the plots you generated for fixed  $\delta$  and varying  $n$ .

**Solution:** Note that the actual numbers in student solutions don't have to be exactly equal to this. But the graphs should look something like this. For the error plot, it's actually possible to get unlucky and have the error increase when  $n$  reaches a big number.

# Experiment 1: Effect of Number of Data Points ( $\delta=1$ )



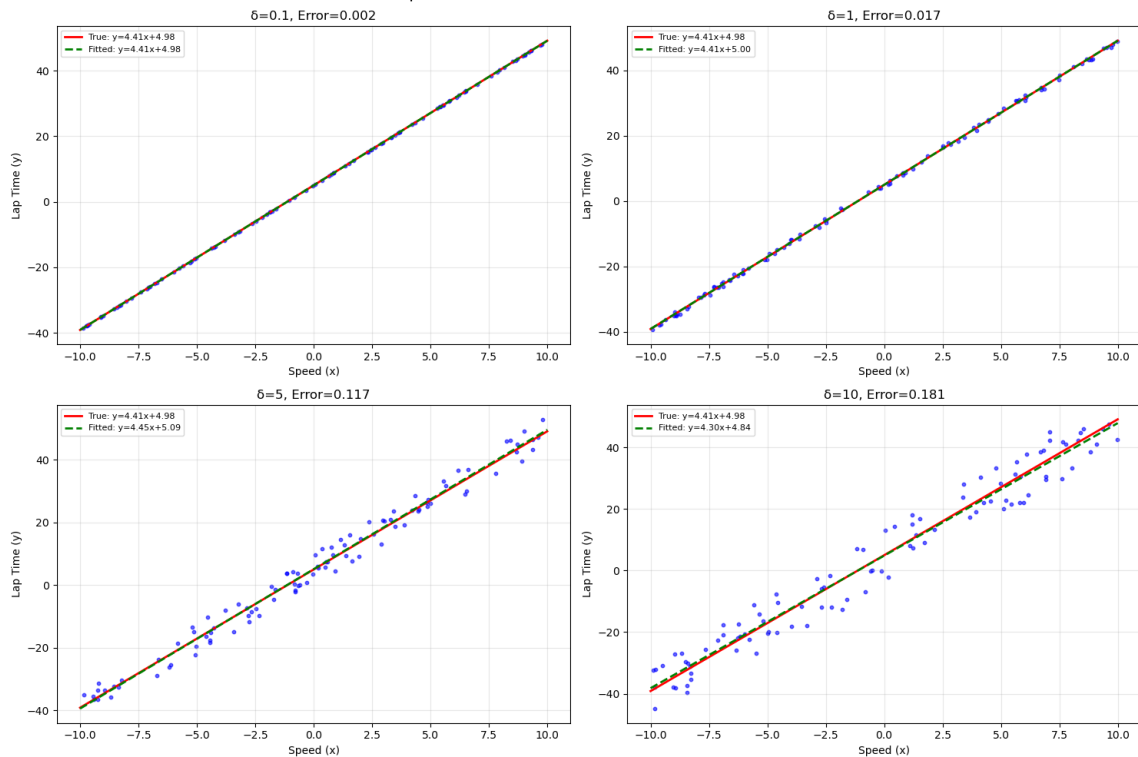
Plot of error (y-axis) relation to  $n$  (x-axis)



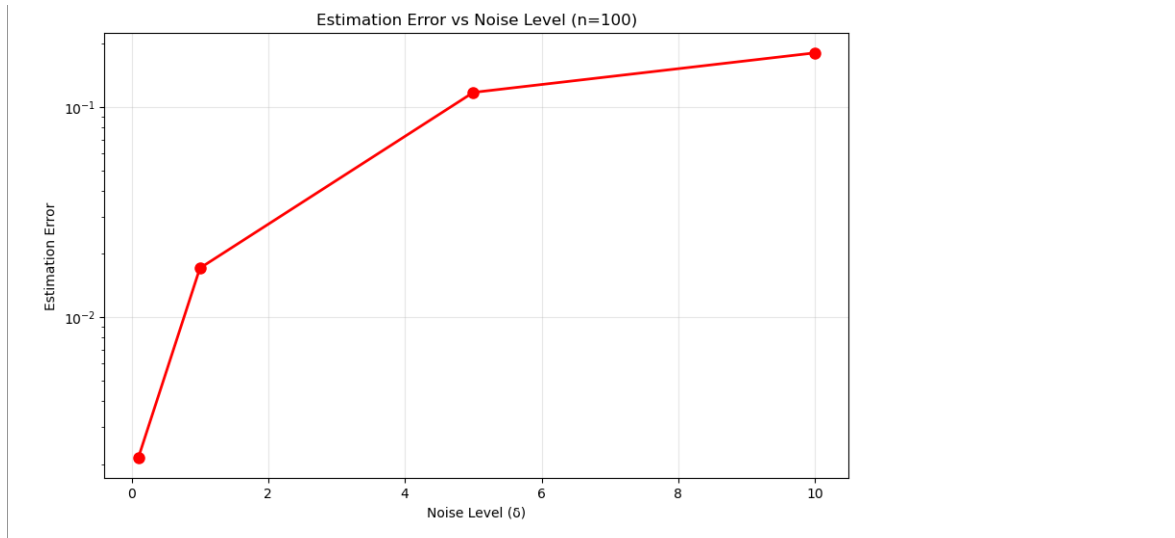
b) 🥑🥑🥑 Add the plots you generated for fixed  $n$  and varying  $\delta$ .

**Solution:** Again note that the actual numbers in student solutions don't have to be exactly equal to this. But the graphs should look something like this.

Experiment 2: Effect of Noise Level ( $n=100$ )



Plot of error (y-axis) relation to noise (x-axis)



c) 🥑🥑 How does adding more points to the dataset improve the fit of the model?

**Solution:** As we can see from the plots, the fitted line gets closer and closer to the true line the more data points we add. The estimation error also gets smaller as we add more  $n$ , indicating our model's accuracy is getting better as we add more data points. **So adding more points improves the fit of the model.** Mathematically, this makes sense because the more data points with noise are added, the more the model tries to "balance" out the error from each direction in order to minimize MSE, and therefore gets closer and closer to the true line. Intuitively, our model will naturally become more accurate as we collect more simple random samples.

d) 🥑🥑 How does increasing noise impact the model's accuracy?

**Solution:** As we can see, the fitted line gets farther away from the true line the more noise we add. The estimation error also gets larger as we add more noise. **This means the model becomes less accurate with more noise.** This makes sense because **adding noise makes the data points farther away from the true line**, and since the fitted line tries to find the minimal MSE among the data points (effectively "following" the data points), it also makes the fitted line farther away from the true line.