Lab 2: Empirical Risk and Simple Linear Regression

EECS 245, Fall 2025 at the University of Michigan

due by the end of your lab section on Wednesday, September 3rd, 2025

Name: ______________________________

uniqname: ______________________________

Each lab worksheet will contain several activities, some of which will involve writing code and others that will involve writing math on paper. To receive credit for a lab, you must complete all activities and show your lab TA by the end of the lab section.

While you must get checked off by your lab TA **individually**, we encourage you to form groups with 1-2 other students to complete the activities together.

Activity 1: Relative Squared Loss

Suppose we'd like to find the optimal parameter, w^*, for the constant model $h(x_i) = w$. To do so, we use the following loss function, called the **relative squared loss**:

$$L_{\text{rsq}}(y_i, h(x_i)) = \frac{(y_i - h(x_i))^2}{y_i}$$

a) What value of w minimizes the average loss (i.e. empirical risk) when using the relative squared loss function – that is, what is w^*? Your answer should only be in terms of the variables $n, y_1, y_2, \dots, y_n$, and any constants.

The next page is left blank for scratch work, in case you need more space.

Solution:
Since $h(x_i) = w$ for the constant model, relative squared loss for the constant model is:

$$L_{\text{rsq}}(y_i, w) = \frac{(y_i - w)^2}{y_i}$$

and so average relative squared loss for the constant model is:

$$R_{\text{rsq}}(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - w)^2}{y_i}$$

To find the value of w that minimizes $R_{\text{rsq}}(w)$, we'll first find its first derivative and set it to zero. The first derivative of $R_{\text{rsq}}(w)$ is:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}w}R_{\text{rsq}}(w) &= \frac{\mathrm{d}}{\mathrm{d}w}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - w)^2}{y_i}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}w}\left(\frac{(y_i - w)^2}{y_i}\right)\end{aligned}$$

At this point, it'll be useful to step aside and find the derivative of $L_{\text{rsq}}(y_i, w)$ with respect to w, as this is the expression being summed. The derivative of $L_{\text{rsq}}(y_i, w)$ with respect to w is:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}w}L_{\text{rsq}}(y_i, w) &= \frac{\mathrm{d}}{\mathrm{d}w}\frac{(y_i - w)^2}{y_i}\\ &= \frac{1}{y_i}\cdot\frac{\mathrm{d}}{\mathrm{d}w}(y_i - w)^2\\ &= \frac{1}{y_i}\cdot 2(y_i - w)\cdot(-1)\\ &= -2\cdot\frac{y_i - w}{y_i}\\ &= \boxed{2\cdot\frac{w}{y_i} - 2}\end{aligned}$$

Back to $\frac{\mathrm{d}}{\mathrm{d}w}R_{\text{rsq}}(w)$, we have:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}w}R_{\text{rsq}}(w) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}w}\left(\frac{(y_i - w)^2}{y_i}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(2\cdot\frac{w}{y_i} - 2\right)\\ &= \frac{2w}{n}\sum_{i=1}^{n}(\frac{1}{y_i}) - \frac{1}{n}\sum_{i=1}^{n}2\\ &= \frac{2w}{n}\sum_{i=1}^{n}(\frac{1}{y_i}) - 2\end{aligned}$$

Solution: (continued) Setting this equal to 0 yields:

$$\begin{aligned}
\frac{2w}{n}\sum_{i=1}^{n}(\frac{1}{y_i}) - 2 &= 0 \\
\frac{w}{n}\sum_{i=1}^{n}(\frac{1}{y_i}) &= 1 \\
w^* &= \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(\frac{1}{y_i})} \\
w^* &= \boxed{\frac{n}{\sum_{i=1}^{n}\frac{1}{y_i}}}
\end{aligned}$$

This is known as the **harmonic mean** of $y_1, y_2, ..., y_n$.

$$L_{\mathrm{rsq}}(y_i, w) = \frac{(y_i - w)^2}{y_i}$$

b) Let $C(y_1, y_2, ..., y_n)$ be your minimizer w^* from the previous part. That is, for a particular dataset $y_1, y_2, ..., y_n$, $C(y_1, y_2, ..., y_n)$ is the value of w that minimizes empirical risk for relative squared loss on that dataset.

What is the value of $\lim\limits_{y_4 \to \infty} C(1, 3, 5, y_4)$ in terms of $C(1, 3, 5)$? Your answer should involve the function C and/or one or more constants.

Hint: To notice the pattern, evaluate $C(1, 3, 5, 100)$*,* $C(1, 3, 5, 10000)$*, and* $C(1, 3, 5, 1000000)$*.*

Solution:

$$\begin{aligned}
\lim_{y_4 \to \infty} C(1, 3, 5, y_4) &= \lim_{y_4 \to \infty} \frac{4}{\frac{1}{1} + \frac{1}{3} + \frac{1}{5} + \frac{1}{y_4}} \\
&= \frac{4}{\frac{1}{1} + \frac{1}{3} + \frac{1}{5} + 0} \\
&= \frac{4}{3} \cdot \frac{3}{\frac{1}{1} + \frac{1}{3} + \frac{1}{5}} \\
&= \frac{4}{3} \cdot C(1, 3, 5)
\end{aligned}$$

c) What is the value of $\lim\limits_{y_4 \to 0} C(1, 3, 5, y_4)$? Again, your answer should involve the function C and/or one or more constants.

Solution:

$$\begin{aligned}
\lim_{y_4 \to 0} C(1, 3, 5, y_4) &= \lim_{y_4 \to 0} \frac{4}{\frac{1}{1} + \frac{1}{3} + \frac{1}{5} + \frac{1}{y_4}} \\
&= \frac{4}{\frac{1}{1} + \frac{1}{3} + \frac{1}{5} + \infty} \\
&= \frac{4}{\infty} = 0
\end{aligned}$$

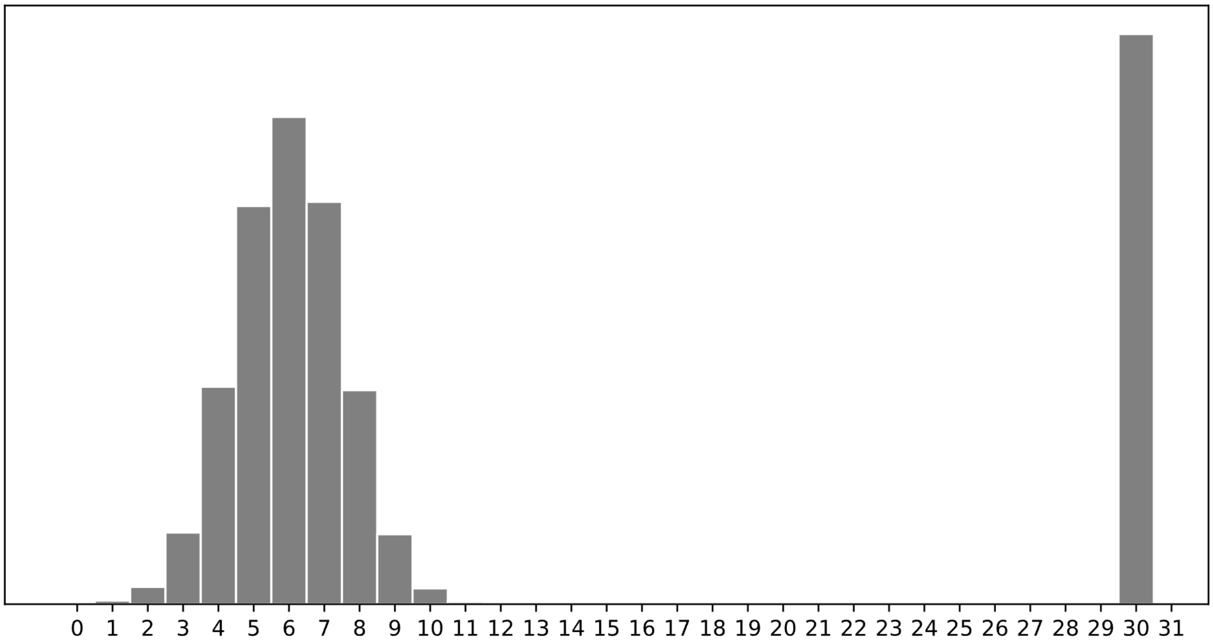
d) Based on the results of the previous two parts, when is the prediction $C(y_1, y_2, ..., y_n)$ robust to outliers? When is it not robust to outliers?

Solution:
$C(y_1, y_2, ..., y_n)$ is great at ignoring large outliers. No matter how large you make any particular value, $C(y_1, y_2, ..., y_n)$ is upper-bounded by $\frac{n}{n-1}$ multiplied by the value of C applied to all data points excluding the large outlier. This is as opposed to the regular "arithmetic mean", where if you make a single data point arbitrarily large, the mean also becomes arbitrarily large (i.e. if $y_n \to \infty$, then $\text{Mean}(y_1, y_2, ..., y_n) \to \infty$ too).

However, $C(y_1, y_2, ..., y_n)$ is not robust to small outliers. As a particular data point approaches 0, the value of $C(y_1, y_2, ..., y_n)$ also approaches 0 no matter how large the other data points are.

Activity 2: Rapid Fire

Consider a dataset of n **integers**, $y_1, y_2, \dots, y_n$, whose histogram is given below:



a) Which of the following is closest to the constant prediction w^* that minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}\begin{cases}0 & y_i = w\\ 1 & y_i \neq w\end{cases}$$

○ 1 ○ 5 ○ 6 ○ 7 ○ 11 ○ 15 ○ 30

Solution: 30**.**
The minimizer of average 0-1 loss is the **mode**.

See: **Chapter 1.3: Beyond Absolute and Squared Loss**

b) Which of the following is closest to the constant prediction w^* that minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - w|$$

○ 1 ○ 5 ○ 6 ○ 7 ○ 11 ○ 15 ○ 30

Solution: 7**.**
The minimizer of average absolute loss is the **median**. The outliers near 30 shift it from 6 to 7.

c) Which of the following is closest to the constant prediction w^* that minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w)^2$$

○ 1 ○ 5 ○ 6 ○ 7 ○ 11 ○ 15 ○ 30

Solution: 11.
The minimizer of average squared loss is the **mean**, pulled upward by the heavy right tail, so it's above the median (7) and closest to 11.

d) Which of the following is closest to the constant prediction w^* that minimizes:

$$\lim_{p\to\infty}\frac{1}{n}\sum_{i=1}^{n}|y_i - w|^p$$

○ 1 ○ 5 ○ 6 ○ 7 ○ 11 ○ 15 ○ 30

Solution: 15.
As $p \to \infty$, the minimizer is the **midrange**, halfway between min and max.

Activity 3: Slope of Mean Absolute Error

Consider a dataset of 8 points, $y_1, y_2, \ldots, y_8$ that are in sorted order, i.e. $y_1 < y_2 < \ldots < y_8$.

Recall that mean absolute error, $R_{\text{abs}}(w)$, is defined as:

$$R_{\text{abs}}(w) = \frac{1}{n}\sum_{i=1}^{n}|y_i - w|$$

This is a piecewise linear function that changes slope at each data point. The slope of $R_{\text{abs}}(w)$ at any w that is not a data point is:

$$\frac{\mathrm{d}}{\mathrm{d}w}R_{\text{abs}}(w) = \frac{\text{\# left of } w - \text{\# right of } w}{n}$$

Suppose that $y_4 = 10$, $y_5 = 14$, $y_6 = 22$, and $R_{\text{abs}}(11) = 9$. What is $R_{\text{abs}}(22)$?

Solution:
$R_{\textbf{abs}}(22) = 11$.

We can write the points given to us as:

$$y_1, y_2, y_3, 10, 14, 22, y_7, y_8$$

Since there are an even number of data points ($n = 8$), the minimizer of absolute error is not a single point but the entire interval between the two middle points. Here, the middle two are 10 and 14, so every $w \in [10, 14]$ minimizes $R_{\text{abs}}(w)$. This explains why the error is *flat* inside that interval: the number of points on the left equals the number on the right, so shifting w around does not change the error. As a result, $R_{\text{abs}}(11) = 9$ and $R_{\text{abs}}(14) = 9$. Once we move beyond 14, the balance breaks. There are now five points to the left and only three to the right, so the slope of $R_{\text{abs}}(w)$ becomes positive. The slope formula tells us:

$$\frac{d}{dw} R_{\text{abs}}(w) = \frac{\text{\# left of } w - \text{\# right of } w}{n}$$

so for any $w \in (14, 22)$ we have

$$\frac{d}{dw} R_{\text{abs}}(w) = \frac{5-3}{8} = \tfrac{1}{4}.$$

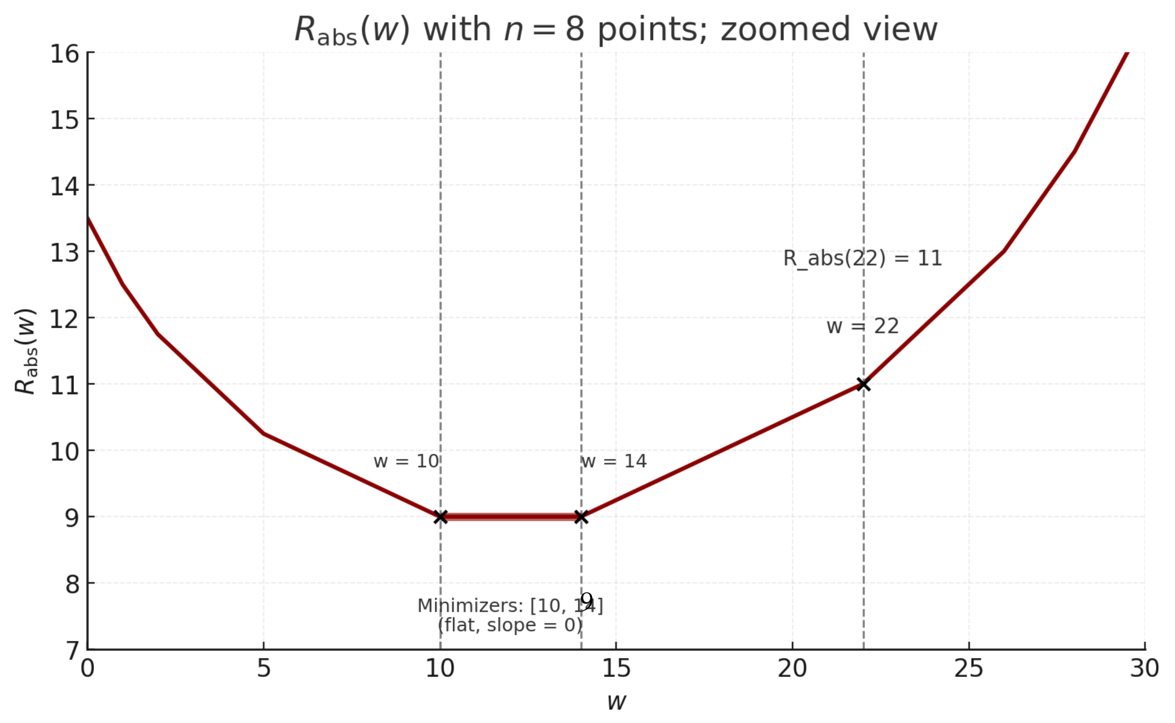
This means that for every one unit we move to the right of $w = 14$, the error increases by $\frac{1}{4}$. Moving from $w = 14$ to $w = 22$ is a distance of $22 - 14 = 8$ units, so the error increases by

$$8 \cdot \tfrac{1}{4} = 2.$$

Adding this to the baseline error of $R_{\text{abs}}(14) = 9$, we get:

$$\begin{aligned} R_{\text{abs}}(22) &= R_{\text{abs}}(14) + (22 - 14) \cdot \tfrac{1}{4} \\ &= 9 + 2 = 11. \end{aligned}$$

Here is a visualization of the solution to this problem:



Activity 4: Programming

Complete the tasks in the `lab02.ipynb` notebook, which you can either access through the DataHub link on the course homepage or by pulling our GitHub repository. To receive credit for Activity 4, you'll need to submit your completed `lab02.ipynb` notebook to Gradescope and show your lab TA that all test cases have passed. Instructions on how to do this are in the lab notebook.

Activity 5: Visualizing Changes in the Data

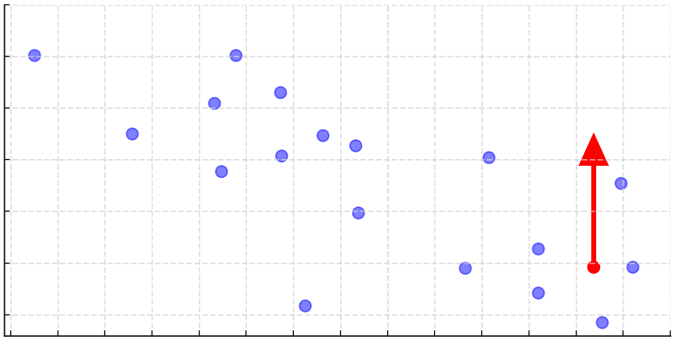
The problems in this final activity will help you visualize how changes in the data affect the optimal simple linear regression line. To recap, this is the line $h(x_i) = w_0 + w_1 x_i$ defined by:

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

r is the correlation coefficient between x and y, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

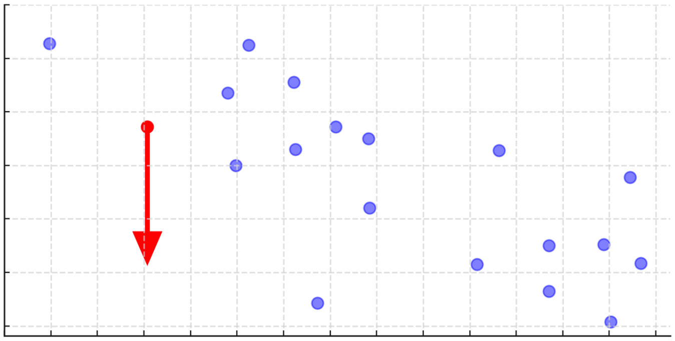
Assume all data is in the first quadrant, i.e. all x_i and y_i are positive.

a) For the dataset 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:**
> **Moving this point upward increases both the slope and the intercept of the regression line.**
>
> When this point is moved upward, its y-value increases while its x-value stays the same. Because this point lies to the right of $\bar{x}$, raising its y-value increases the term
>
> $$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$
>
> which makes the slope w_1^* larger.
>
> At the same time, the mean $\bar{y}$ also increases, and since
>
> $$w_0^* = \bar{y} - w_1^* \bar{x},$$
>
> both a larger $\bar{y}$ and a larger w_1^* result in a higher intercept.

b) For the dataset 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:
Moving the red point downward decreases both the slope and the intercept of the regression line.

When this point is moved downward, its y-value decreases while its x-value stays the same. Because this point lies to the *left* of $\bar{x}$, lowering its y-value increases the term

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

in the negative direction, which makes the slope w_1^* **decrease**.

At the same time, the mean $\bar{y}$ also decreases. Since

$$w_0^* = \bar{y} - w_1^* \bar{x},$$

a smaller $\bar{y}$ lowers the intercept, but the effect of the decreased slope partially offsets this. Overall, the intercept w_0^* will also **decrease**.

c) Suppose we transform a dataset of (x_i, y_i) pairs by doubling each y-value, creating a transformed dataset $(x_i, 2y_i)$. 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:
Doubling all y-values doubles the slope of the regression line.

We are asked how the slope of the regression line changes if every y_i is doubled. Recall the formula for the slope of the regression line:

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

Now consider the transformed dataset $\{(x_i, 2y_i)\}$. The new slope is

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

Factor out the 2:

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

Thus,

$$w_1' = 2w_1^*.$$

d) Suppose we transform a dataset of (x_i, y_i) pairs by doubling each x-value, creating a transformed dataset $(2x_i, y_i)$. 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:
Doubling all x-values halves the slope of the regression line.

We are asked how the slope of the regression line changes if every x_i is doubled. Recall the formula for the slope of the regression line:

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

Now consider the transformed dataset $\{(2x_i, y_i)\}$. Let $\bar{x}' = 2\bar{x}$. Then the new slope is

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

Factor out the constants:

$$w_1' = \frac{2\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{4\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Simplify:

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

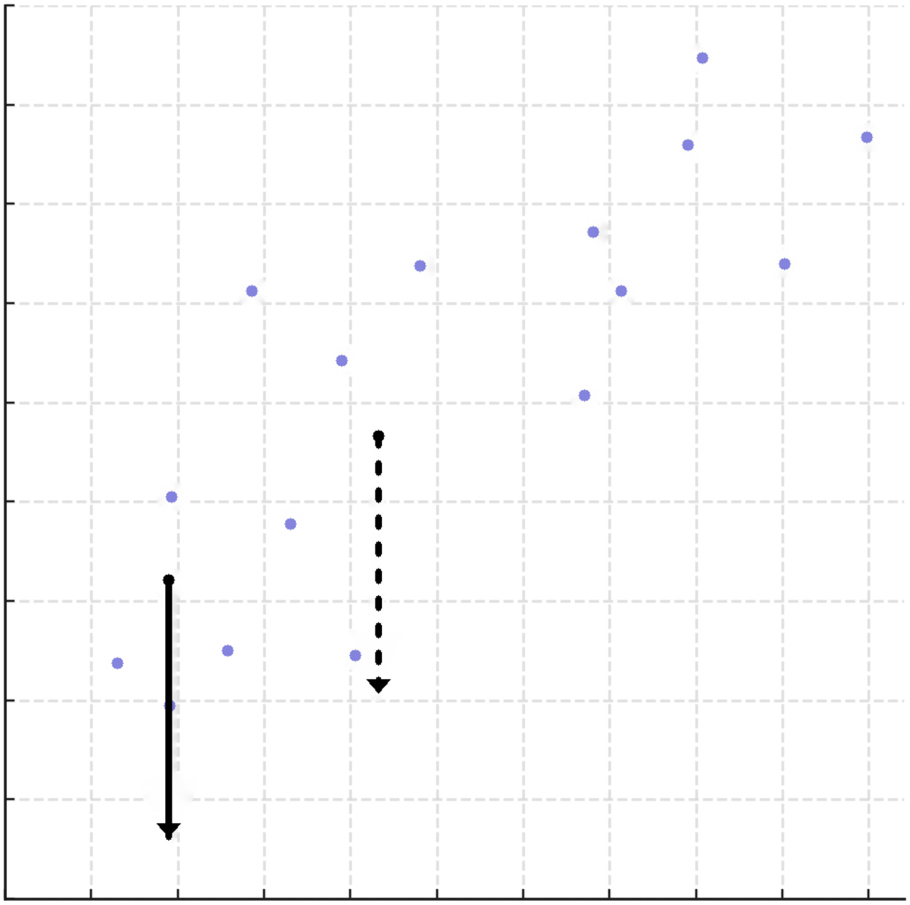
Thus,

$$w_1' = \tfrac{1}{2} w_1^*.$$

e) Compare two different possible changes to the dataset shown below.

- Move the dashed point down c units.
- Move the solid point down c units.

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



Solution:
The slope of the regression line is given by

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$$

When we move a single point down by c units, the only part of the slope formula that changes is the numerator

$$\sum(x_i - \bar{x})(y_i - \bar{y}),$$

which measures the covariance between x and y. The change in covariance due to moving one point is proportional to $(x_i - \bar{x})(-c)$.

- The dashed point is closer to the mean $\bar{x}$, so $|x_i - \bar{x}|$ is relatively small. Moving this point has only a modest effect on the slope.
- The solid point is farther from the mean $\bar{x}$, so $|x_i - \bar{x}|$ is larger. Moving this point has a much larger effect on the slope.

$$\Delta\text{slope} \propto (x_i - \bar{x})(-c).$$

Therefore, **moving the solid point down c units will change the slope of the regression line more than moving the dashed point.** This is because points farther from the mean of x have greater *leverage* on the slope.