Lecture 12

Simple Linear Regression

EECS 398: Practical Data Science, Spring 2025

practicaldsc.org • github.com/practicaldsc/sp25 • 📣 See latest announcements **here on Ed**

Agenda 📆

- Empirical risk minimization.
- Towards simple linear regression.
- Minimizing mean squared error for the simple linear model.
- Correlation.
- Interpreting the formulas.

There are several important videos for Lectures 11 and 12; they are all in this YouTube playlist.

Empirical risk minimization

The modeling recipe

- Last lecture, we made two full passes through our modeling recipe.

1. Choose a model.

$H(x_i) = h$ constant model

2. Choose a loss function.

$L_{sq}(y_i, h) = (y_i - h)^2$ squared loss

$L_{abs}(y_i, h) = |y_i - h|$ absolute loss

3. Minimize average loss to find optimal model parameters.

$R_{sq}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2$ "mean squared error"

$R_{abs}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$ "mean absolute error"

Empirical risk minimization

from our data

- The formal name for the process of minimizing average loss is **empirical risk minimization**; another name for "average loss" is **empirical risk**.
- When we use the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$, the corresponding empirical risk is mean squared error:

$$R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2 \implies h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$$

- When we use the absolute loss function, $L_{\text{abs}}(y_i, h) = |y_i - h|$, the corresponding empirical risk is mean absolute error:

$$R_{\text{abs}}(h) = \frac{1}{n}\sum_{i=1}^{n}|y_i - h| \implies h^* = \text{Median}(y_1, y_2, \ldots, y_n)$$

Empirical risk minimization, in general

- **Key idea**: If L is **any** loss function, and H is any hypothesis function, the corresponding empirical risk is:

$$R(H) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, H(x_i))$$

loss

average

- In Homework 6 and tomorrow's discussion, there are several questions where:
  - You are given a new loss function L.
  - You have to find the optimal parameter h^* for the constant model $H(x_i) = h$.

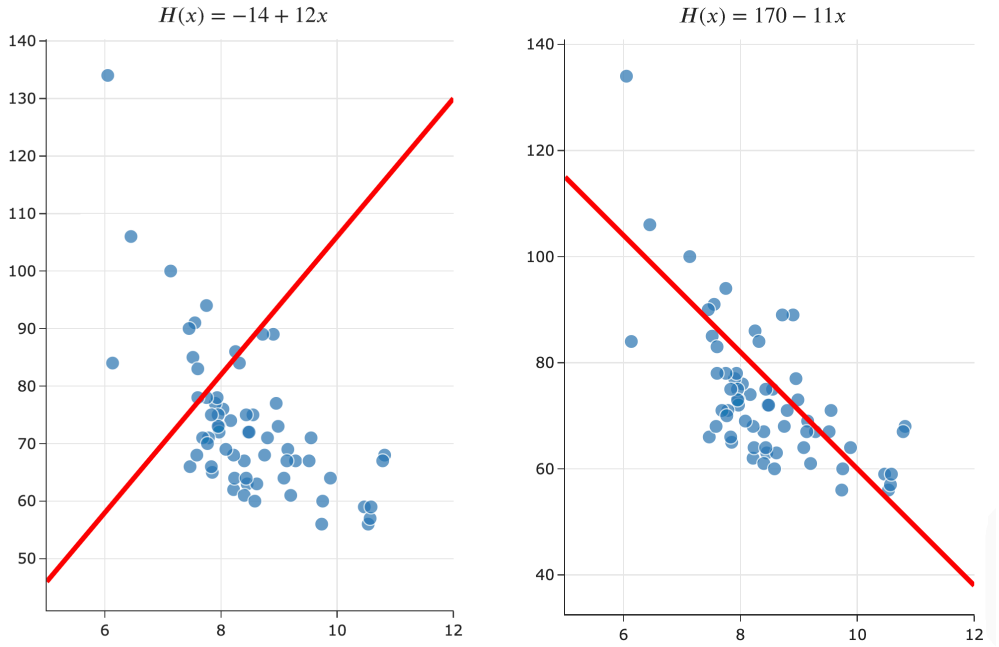
Towards simple linear regression

Recap: Hypothesis functions and parameters

- A hypothesis function, H, takes in an x_i as input and returns a predicted y_i.
- **Parameters** define the relationship between the input and output of a hypothesis function.
- **Example**: The simple linear regression model, $H(x_i) = w_0 + w_1 x_i$, has two parameters: w_0 and w_1.

slope

intercept



The modeling recipe

x_i : departure time (hours)

y_i : actual commute time (minutes)

$H(x_i)$: predicted commute time

1. Choose a model.

$$H(x_i) = w_0 + w_1 x_i$$

simple linear regression model

2. Choose a loss function.

$$L_{sq}(y_i, H(x_i)) = (y_i - H(x_i))^2$$

squared loss

3. Minimize average loss to find optimal model parameters.

$$R_{sq}(H) = \frac{1}{n}\sum_{i=1}^{n}(y_i - H(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$

$$R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$

mean squared error!

Minimizing mean squared error for the simple linear model

- We'll choose squared loss, since it's the easiest to minimize.
- Our goal, then, is to find the linear hypothesis function $H^*(x_i)$ that minimizes empirical risk:

$$R_{\text{sq}}(H) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - H(x_i)\right)^2$$

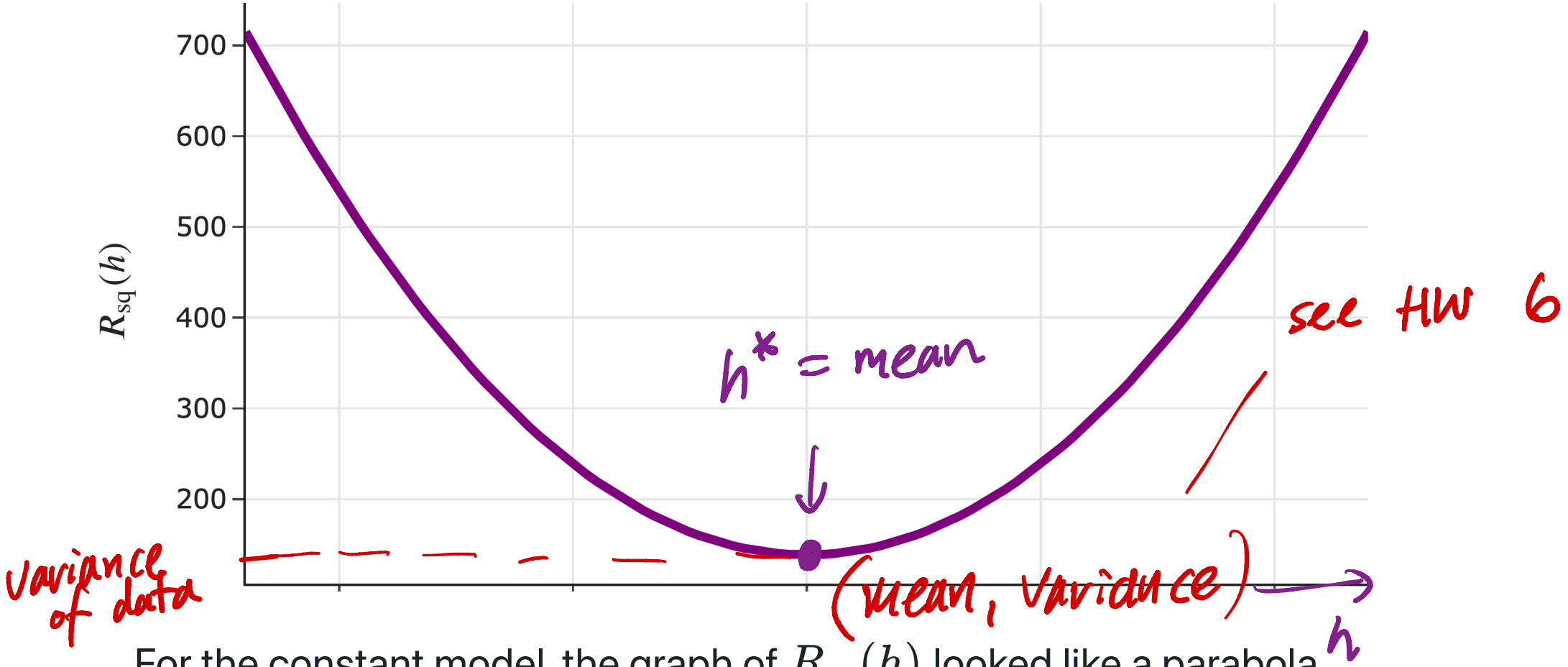
- Since linear hypothesis functions are of the form $H(x_i) = w_0 + w_1 x_i$, we can re-write R_{sq} as a function of w_0 and w_1:

mean squared error

$$\boxed{R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2}$$

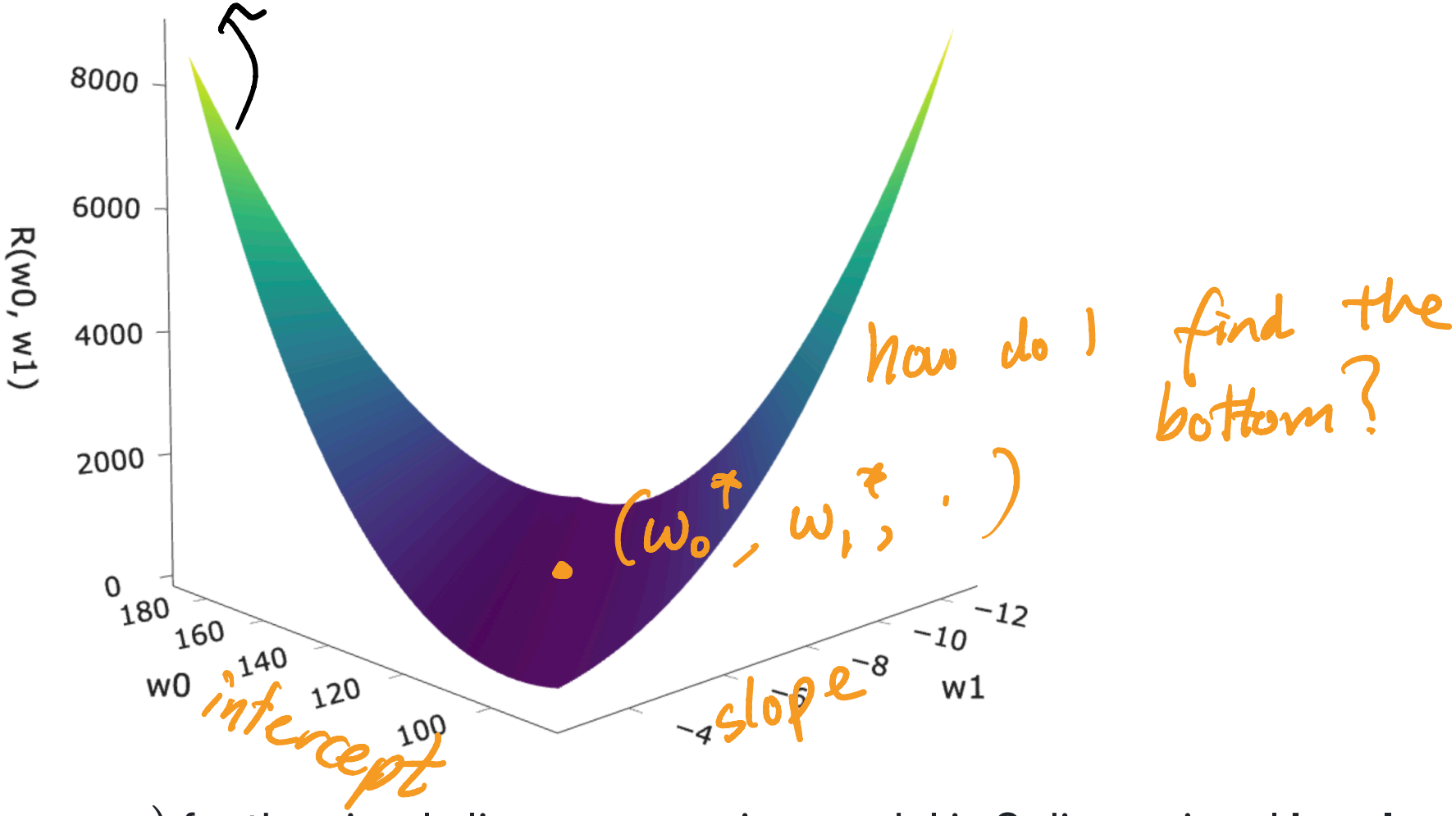
- **How do we find the parameters w_0^* and w_1^* that minimize $R_{\text{sq}}(w_0, w_1)$?**

$$R_{\text{sq}}(h) = \frac{1}{5}\left((72-h)^2 + (90-h)^2 + (61-h)^2 + (85-h)^2 + (92-h)^2\right)$$



For the constant model, the graph of $R_{\text{sq}}(h)$ looked like a parabola.

graph of $R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$



The graph of $R_{\text{sq}}(w_0, w_1)$ for the simple linear regression model is 3 dimensional **bowl**, and is called a **loss surface**.

Minimizing mean squared error for the simple linear model

Minimizing multivariate functions

- Our goal is to find the parameters w_0^* and w_1^* that minimize mean squared error:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i)\right)^2$$

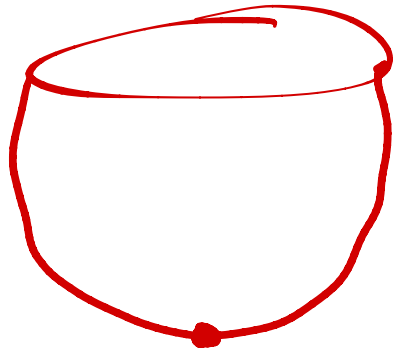
- R_{sq} is a function of two variables: w_0 and w_1, and is a bowl-like shape in 3D.
- To minimize a function of multiple variables:
  - Take partial derivatives with respect to each variable.
  - Set all partial derivatives to 0 and solve the resulting system of equations.
  - Ensure that you've found a minimum, rather than a maximum or saddle point (using the second derivative test for multivariate functions).
- To save time, we won't do the derivation live in class, but you are responsible for it!

  Here's a video of me walking through it, and the slides will be annotated with it.

Example

Find the point (x, y, z) at which the following function is minimized.

$$f(x, y) = x^2 - 8x + y^2 + 6y - 7$$



$(4, -3, -32)$

$\frac{\partial f}{\partial x} = 2x - 8$

This is the partial derivative w.r.t. x; to take it, treat y as constant

$\frac{\partial f}{\partial y} = 2y + 6$

To solve for x^*, y^*, we set both $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y} = 0$.

$2x - 8 = 0$ ①
$2y + 6 = 0$ ②

$\Rightarrow$ $x = 4$
$y = -3$
$z = f(4, -3) = -32$

Minimizing mean squared error

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

To find the w_0^* and w_1^* that minimize $R_{\text{sq}}(w_0, w_1)$, we'll:

1. Find $\frac{\partial R_{\text{sq}}}{\partial w_0}$ and set it equal to 0.
2. Find $\frac{\partial R_{\text{sq}}}{\partial w_1}$ and set it equal to 0.
3. Solve the resulting system of equations.

unlike on previous slide, BOTH partial derivatives will involve both variables, w_0 and w_1.

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

$$\frac{\partial R_{\text{sq}}}{\partial w_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w_0}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

chain rule!

$$= \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - (w_0 + w_1 x_i)\right)\cdot\frac{\partial}{\partial w_0}\left(y_i - (w_0 + w_1 x_i)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - (w_0 + w_1 x_i)\right)(-1)$$

$$= \boxed{\frac{-2}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)}$$

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

$$\frac{\partial R_{\text{sq}}}{\partial w_1} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial w_1} (y_i - (w_0 + w_1 x_i))^2$$

chain rule!

$$= \frac{1}{n} \sum_{i=1}^{n} 2(y_i - (w_0 + w_1 x_i)) \cdot \frac{\partial}{\partial w_1} (y_i - (w_0 + w_1 x_i))$$

$$= \frac{1}{n} \sum_{i=1}^{n} 2(y_i - (w_0 + w_1 x_i))(-x_i)$$

$$= \boxed{-\frac{2}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i)) x_i}$$

Strategy

- We have a system of two equations and two unknowns (w_0 and w_1):

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i-(w_0+w_1x_i))=0 \qquad -\frac{2}{n}\sum_{i=1}^{n}(y_i-(w_0+w_1x_i))x_i=0$$

this tells us that the sum of (actual - predicted) must be 0!

- To proceed, we'll:

  1. Solve for w_0 in the first equation.

  The result becomes w_0^*, because it's the "best intercept."

  2. Plug w_0^* into the second equation and solve for w_1.

  The result becomes w_1^*, because it's the "best slope."

Solving for w_0^*

optimal intercept!

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i-(w_0+w_1x_i))=0$$

$$\sum_{i=1}^{n}(y_i-(w_0+w_1x_i))=0$$

$$\sum_{i=1}^{n}(y_i-w_0-w_1x_i)=0$$

$$\sum_{i=1}^{n}y_i-\underbrace{\sum_{i=1}^{n}w_0}_{w_0+w_0+\cdots+w_0}-\sum_{i=1}^{n}w_1x_i=0$$

$$\sum_{i=1}^{n}y_i-nw_0-w_1\sum_{i=1}^{n}x_i=0$$

$$\frac{\sum_{i=1}^{n}y_i-w_1\sum_{i=1}^{n}x_i}{n}=\frac{nw_0}{n}$$

$$\Rightarrow \underbrace{\frac{\sum_{i=1}^{n}y_i}{n}}_{\text{mean of } y!}-w_1\frac{\sum_{i=1}^{n}x_i}{n}=w_0$$

$$\Rightarrow \boxed{\bar{y}-w_1\bar{x}=w_0}$$

optimal intercept!

optimal intercept is defined in terms of optimal slope!

Solving for w_1^*

optimal slope

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i-(w_0+w_1x_i))x_i=0$$

$$\sum_{i=1}^{n}(y_i-(w_0+w_1x_i))x_i=0$$

We know that $w_0^* = \bar{y} - w_1^*\bar{x}$

$$\sum_{i=1}^{n}(y_i-(\bar{y}-w_1^*\bar{x}+w_1^*x_i))x_i=0$$

grouping like terms

$$\sum_{i=1}^{n}(y_i-\bar{y}-w_1^*(x_i-\bar{x}))x_i=0$$

$$\sum_{i=1}^{n}(y_i-\bar{y})x_i-\sum_{i=1}^{n}w_1^*(x_i-\bar{x})x_i=0$$

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

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

formula for optimal slope!

Least squares solutions

- We've found that the values w_0^* and w_1^* that minimize R_{sq} are:

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

optimal slope

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

optimal intercept

where:

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

- **These formulas work, but let's re-write w_1^* to be a little more symmetric.**

An equivalent formula for w_1^*

another equivalent formula!

- Claim:

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

"the sum of deviations is 0"

- Proof: Start with the fact that

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

$$= n\bar{x} - n\bar{x} = 0$$

Then, on the numerator, starting from the left side:

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

$$= \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \bar{x}\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})}_{0} = \sum_{i=1}^{n}(y_i - \bar{y})x_i$$

numerator ✓ same logic for denom

line with the LEAST mean SQUARED error

Least squares solutions

- The **least squares solutions** for the intercept w_0 and slope w_1 are:

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

"line of best line"

"regression line"

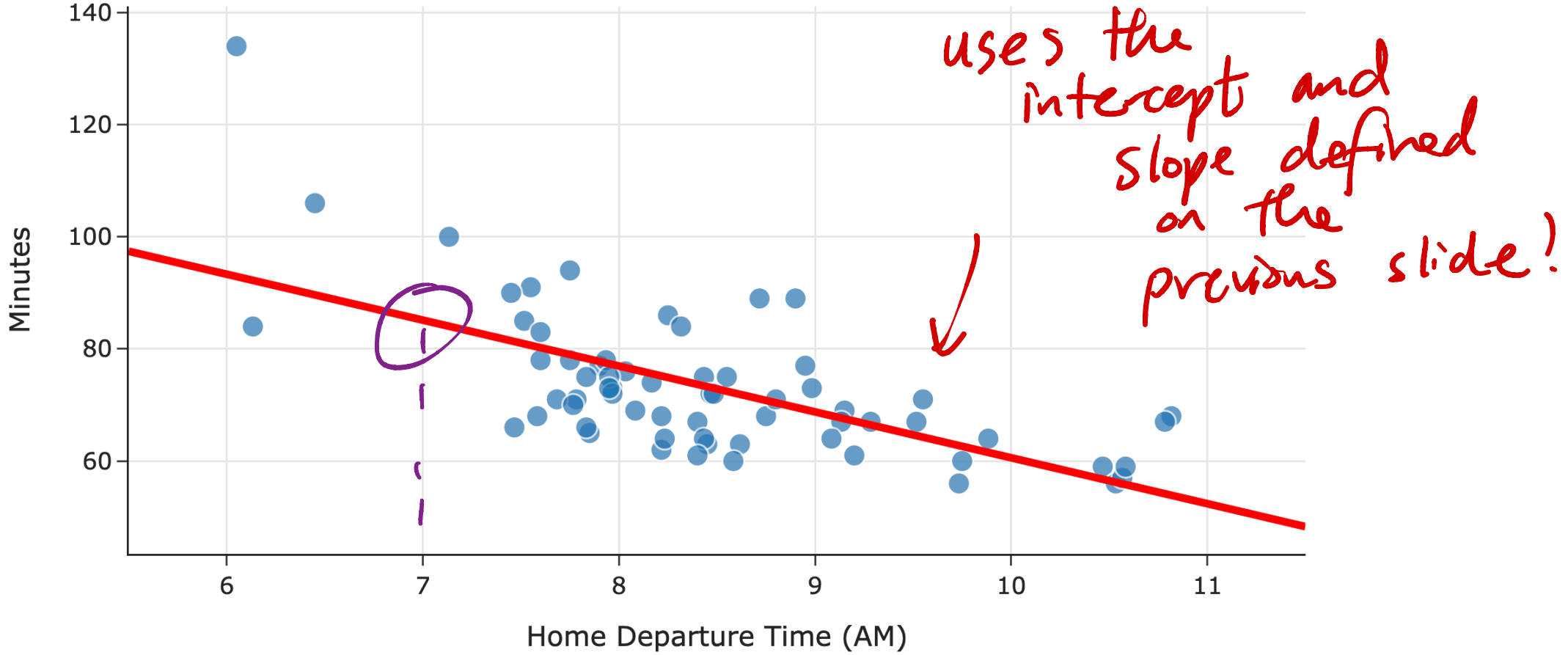
- We say w_0^* and w_1^* are **optimal parameters**, and the resulting line is called the **regression line**.
- The process of minimizing empirical risk to find optimal parameters is also called "**fitting to the data.**"

"training the model"

- To make predictions about the future, we use $H^*(x_i) = w_0^* + w_1^* x_i$.

$$-8.19 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Predicted Commute Time = 142.25 - 8.19 * Departure Hour



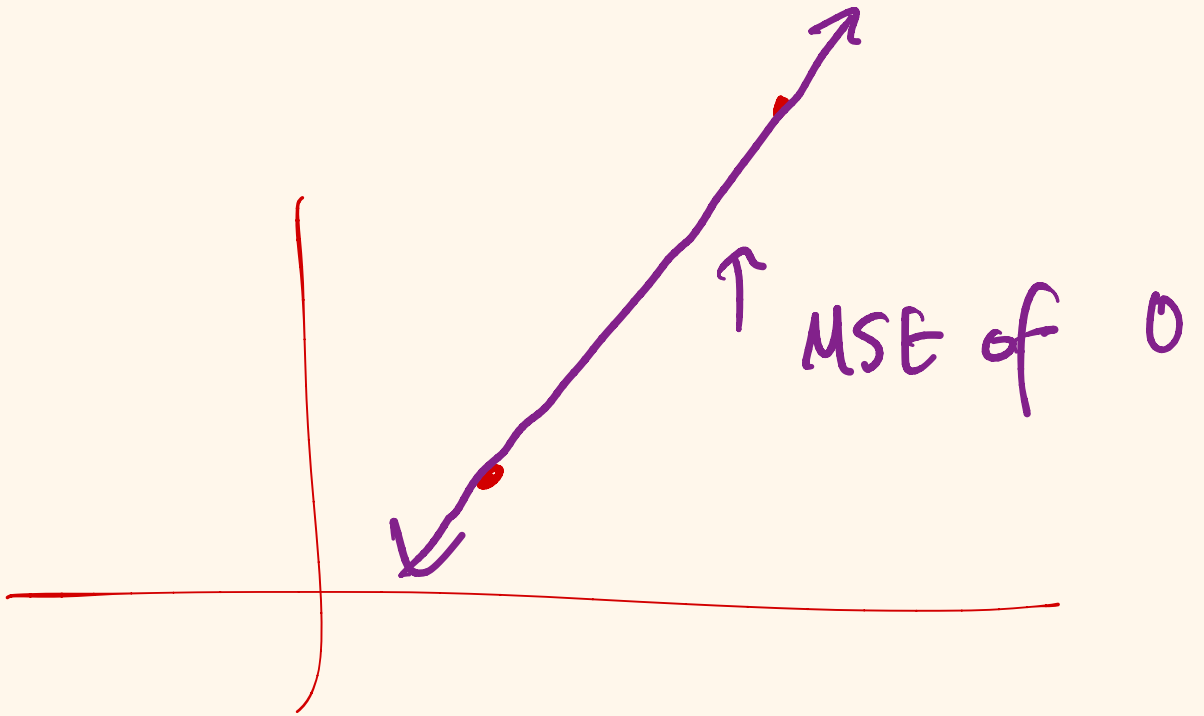
uses the intercept and slope defined on the previous slide!

Question 🤔

Answer at practicaldsc.org/q

Consider a dataset with just two points, $(2, 5)$ and $(4, 15)$. Suppose we want to fit a linear hypothesis function to this dataset using squared loss. What are the values of w_0^* and w_1^* that minimize empirical risk?

- A. $w_0^* = 2, w_1^* = 5$
- B. $w_0^* = 3, w_1^* = 10$
- C. $w_0^* = -2, w_1^* = 5$
- D. $w_0^* = -5, w_1^* = 5$

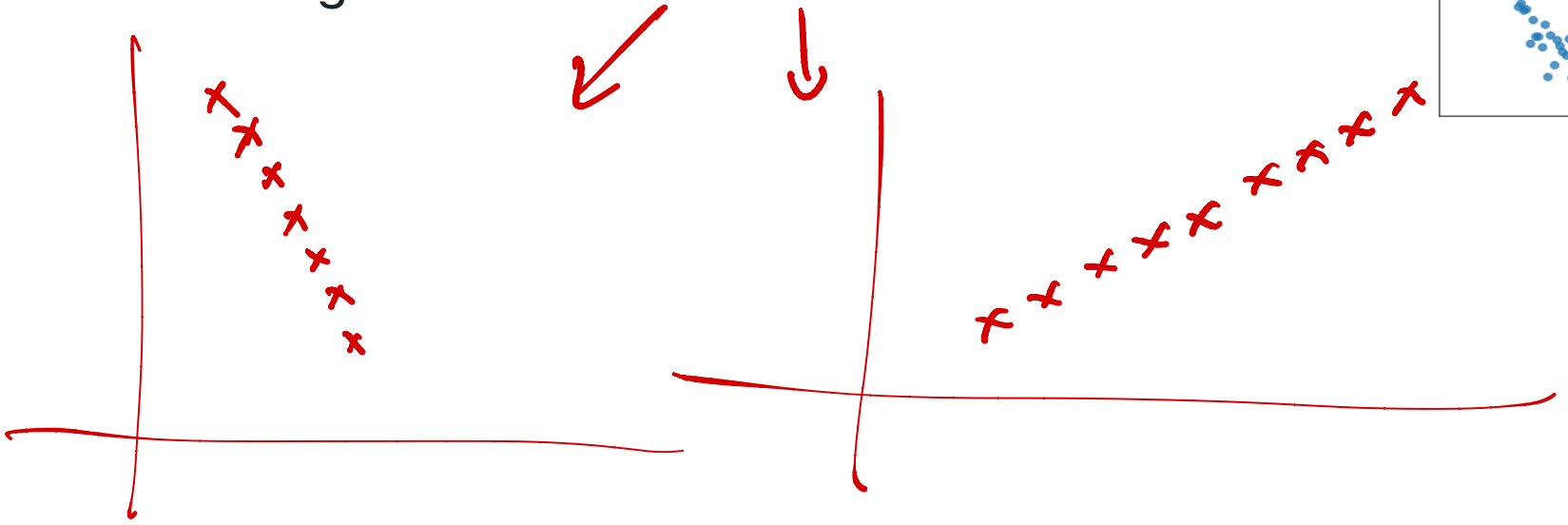


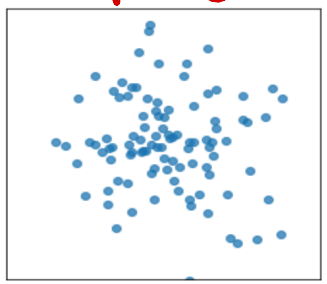
Correlation

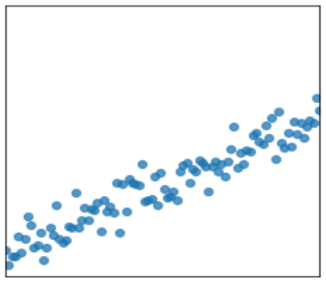
correlation ≠ causation

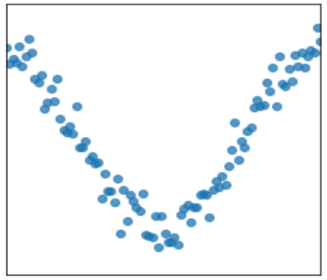
Quantifying patterns in scatter plots

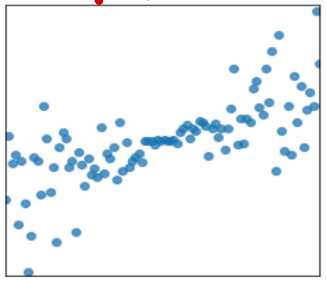
- The **correlation coefficient**, r, is a measure of the strength of the **linear association** of two variables, x and y.
- Intuitively, it measures how tightly clustered a scatter plot is around a straight line.
- It ranges between -1 and 1.











"Pearson's correlation"

The correlation coefficient

- The correlation coefficient, r, is defined as the **average of the product of x and y, when both are *standardized***.
- Let σ_x be the standard deviation of the x_is, and $\bar{x}$ be the mean of the x_is.
- x_i standardized is $\frac{x_i - \bar{x}}{\sigma_x}$.
- The correlation coefficient, then, is:

subtract the mean, divide by SD
"z-scores"

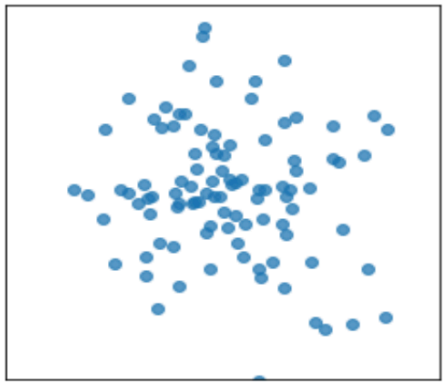
$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

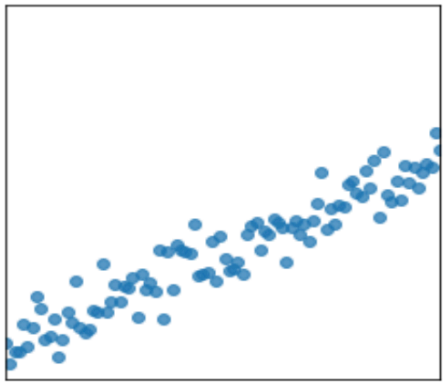
average

product

The correlation coefficient, visualized

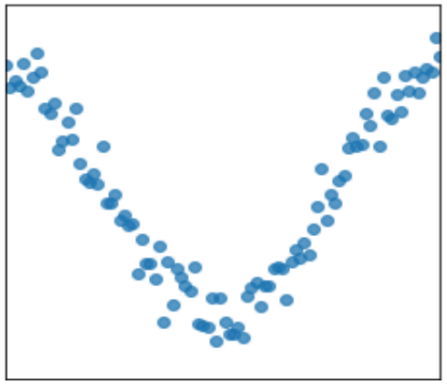
r = -0.121

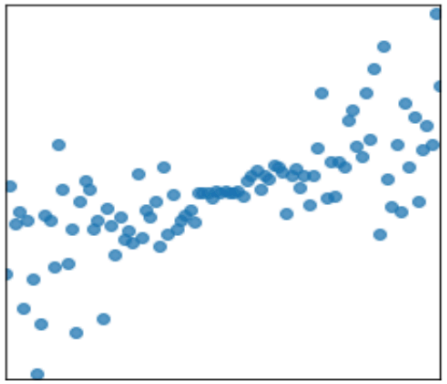




r = 0.949

r = 0.052





r = 0.704

end of midterm scope.

Another way to express w_1^*

- It turns out that w_1^*, the optimal slope for the linear hypothesis function when using squared loss (i.e. the regression line), can be written in terms of r!

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

sign (w_1^*) = sign (r)

- It's not surprising that r is related to w_1^*, since r is a measure of linear association.
- Concise way of writing w_0^* and w_1^*:

$$w_1^* = r\frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

Proof that $w_1^* = r\frac{\sigma_y}{\sigma_x}$

$$r\frac{\sigma_y}{\sigma_x} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right)\frac{\sigma_y}{\sigma_x}$$

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

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

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

Remember,

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

so

$$n\sigma_x^2 = n\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \sum_{i=1}^{n}(x_i-\bar{x})^2$$

Recap: Simple linear regression

- **Goal**: Use the modeling recipe to find the "best" simple linear hypothesis function.

  1. **Model**: $H(x_i) = w_0 + w_1 x_i$.
  2. **Loss function**: $L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$.
  3. **Minimize empirical risk**: $R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$.

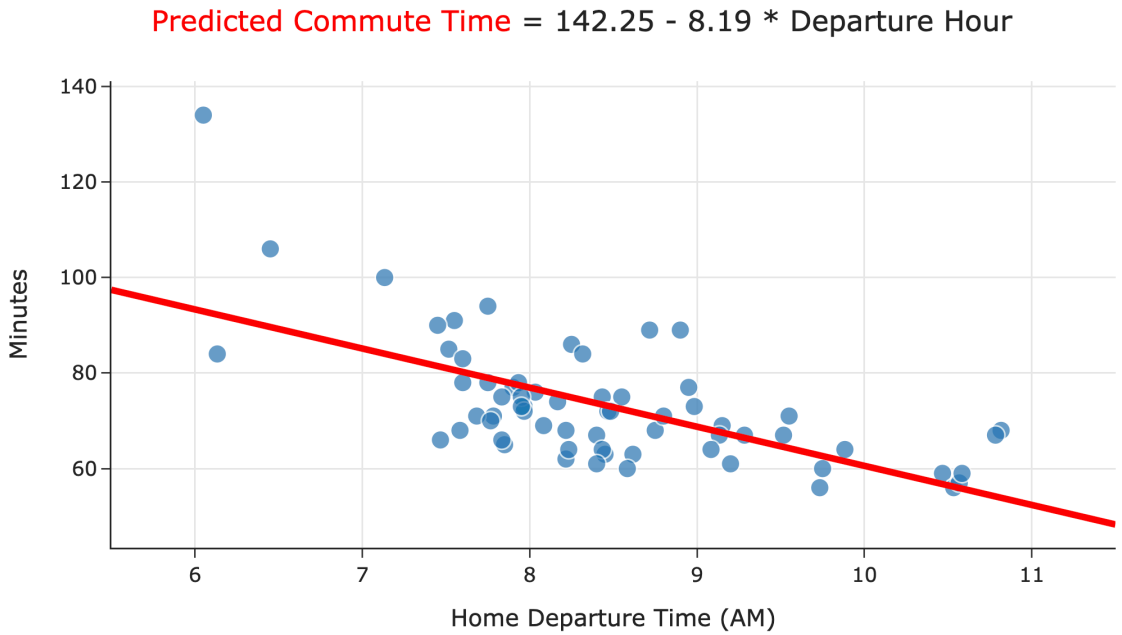
$$\implies w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{\sigma_y}{\sigma_x} \qquad\qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

- The resulting line, $H^*(x_i) = w_0^* + w_1^* x_i$, is the line that minimizes mean squared error. It's often called the **least squares regression line**.

Interpreting the formulas

Causality

- Can we conclude that leaving later **causes** you to get to school earlier?



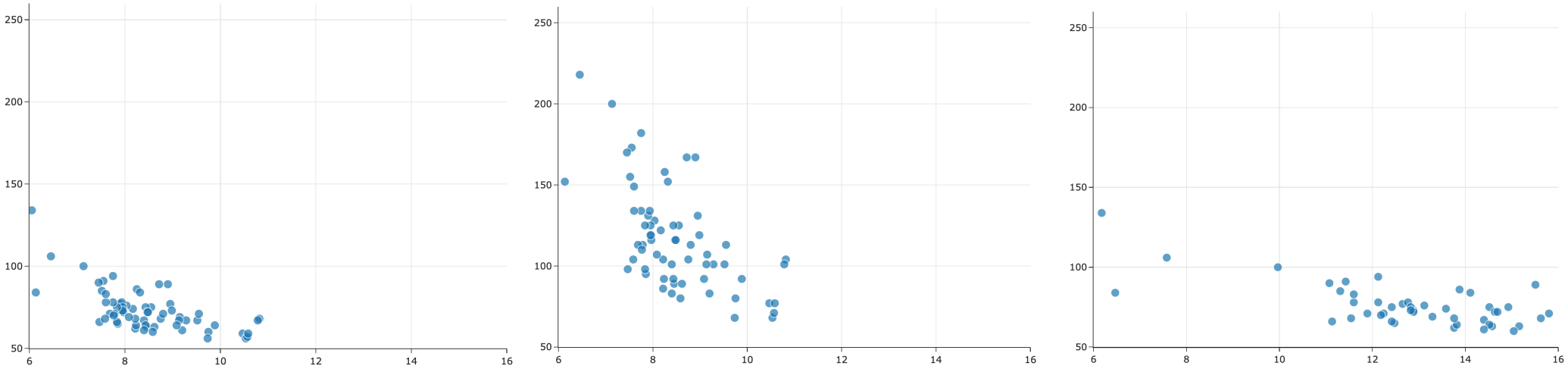
Interpreting the slope

$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$

- The units of the slope are **units of y per units of x**.
- In our commute times example, in $H^*(x_i) = 142.25 - 8.19x_i$, our predicted commute time **decreases by 8.19 minutes per hour**.

Interpreting the slope

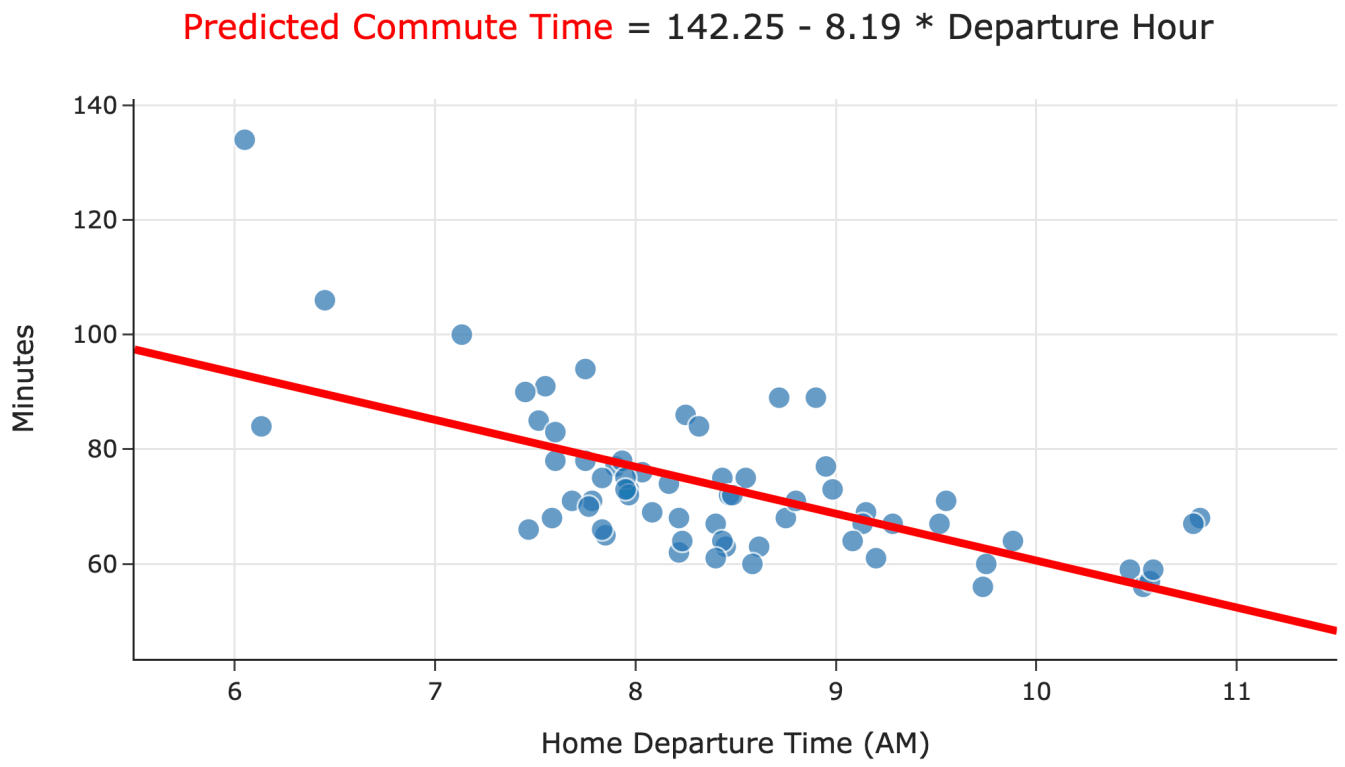
$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$



- Since $\sigma_x \geq 0$ and $\sigma_y \geq 0$, the slope's sign is r's sign.
- As the y values get more spread out, σ_y increases, so the slope gets steeper.
- As the x values get more spread out, σ_x increases, so the slope gets shallower.

Interpreting the intercept

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



- What are the units of the intercept?
- What is the value of $H^*(\bar{x})$?

Question 🤔

Answer at practicaldsc.org/q

We fit a regression line to predict commute times given departure hour. Then, we add 75 minutes to all commute times in our dataset. What happens to the resulting regression line?

- A. Slope increases, intercept increases.
- B. Slope decreases, intercept increases.
- C. Slope stays the same, intercept increases.
- D. Slope stays the same, intercept stays the same.