Lecture 5

More Simple Linear Regression

DSC 40A, Summer I 2024

Agenda

- Recap: Simple linear regression.
- Correlation.
- Interpreting the formulas.
- Connections to related models.
- Introduction to linear algebra.

Question 🤔

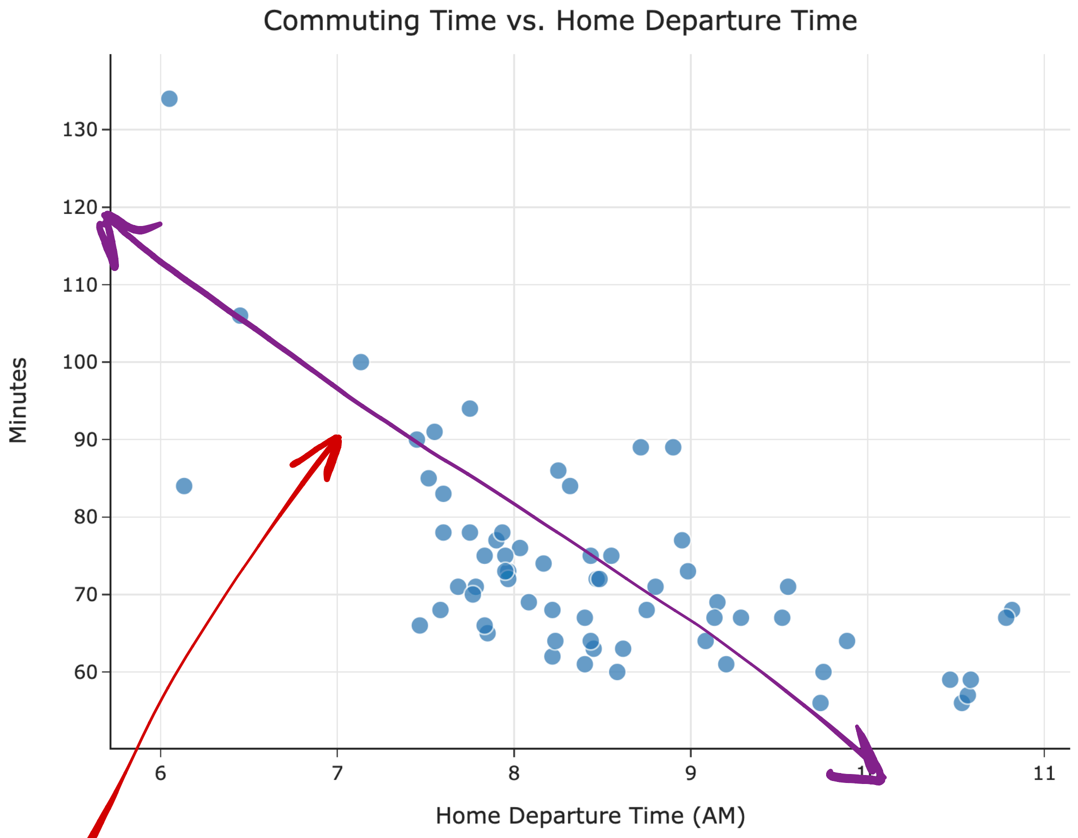
Take a moment to pause and reflect...

If you have any questions please post online to our forms/Q&A site.

Course staff will answer them ASAP!

Recap: Simple linear regression

Recap



Regression line!!!

- In Lecture 4, our goal was to fit a **simple linear regression** model, $H(x) = w_0 + w_1 x$, to our commute times dataset.
    - x_i: The ith home departure time (e.g. 8.5, for 8:30 AM).
    - y_i: The ith actual commute time (e.g. 76 minutes).
    - $H(x_i)$: The ith predicted commute time.
- To do so, we used squared loss.

The modeling recipe

1. Choose a model.

Intercept

slope

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

2. Choose a loss function.

(actual − predicted)^2

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

3. Minimize average loss to find optimal model parameters.

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

Least squares solutions

- Our goal was to find the parameters w_0^* and $w_1{}^*$ that minimized:

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

- To do so, we used calculus, and we found that the minimizing values are:

Best slope

Best Intercept

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

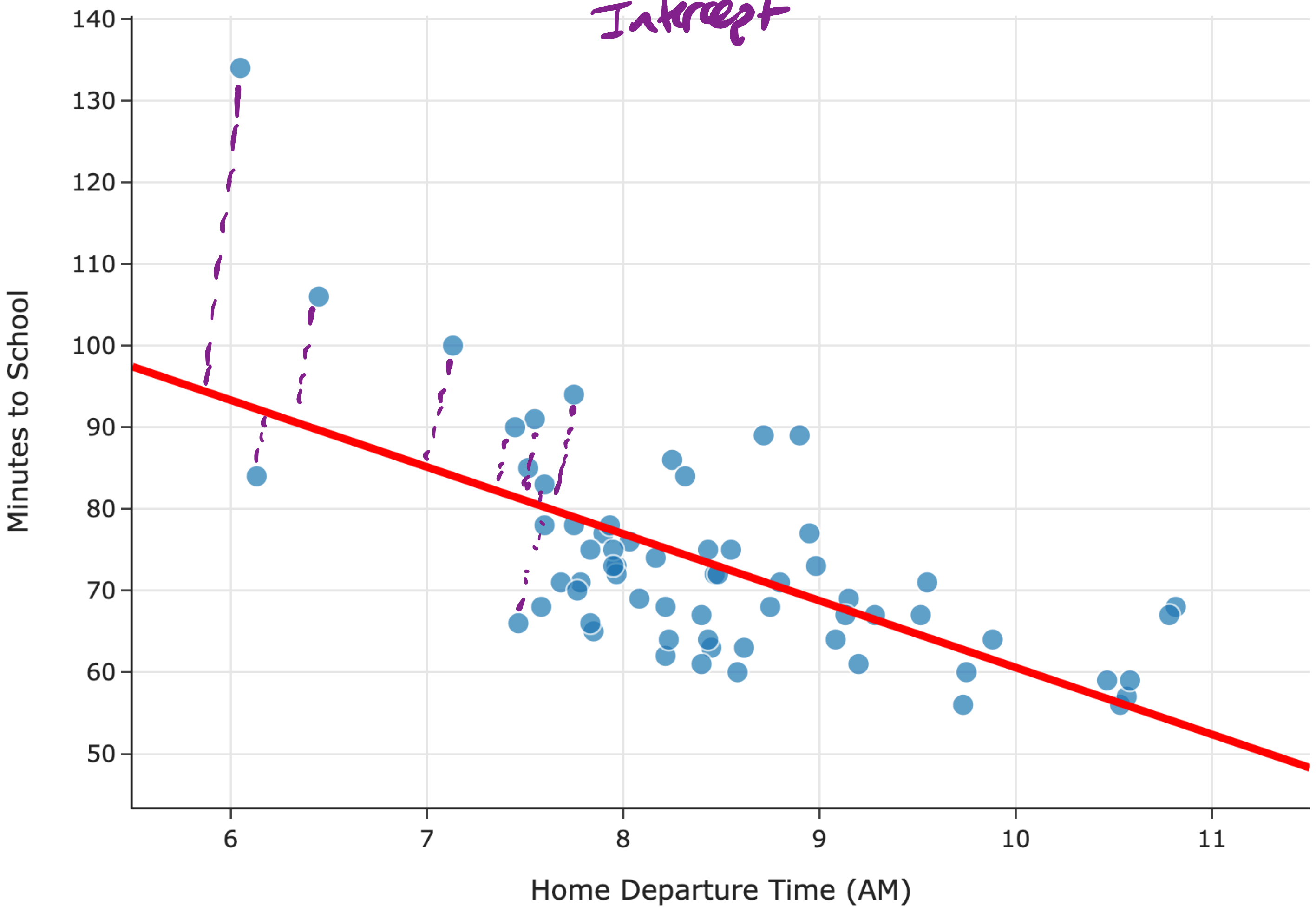
- We say w_0^* and w_1^* are **optimal parameters**, and the resulting line is called the **regression line**.

Slope

$\text{Predicted Commute Time} = \boxed{142.25} - \boxed{8.19} * \text{Departure Hour}$

Intercept

no other line exists!!! for this Dataset



Now what?

We've found the optimal slope and intercept for linear hypothesis functions using squared loss (i.e. for the regression line). Now, we'll:

- See how the formulas we just derived connect to the formulas for the slope and intercept of the regression line we saw in DSC 10.
  - They're the same, but we need to do a bit of work to prove that.
- Learn how to interpret the slope of the regression line.
- Understand connections to other related models.
- Learn how to build regression models with **multiple inputs**.
  - To do this, we'll need linear algebra!

Question 🤔

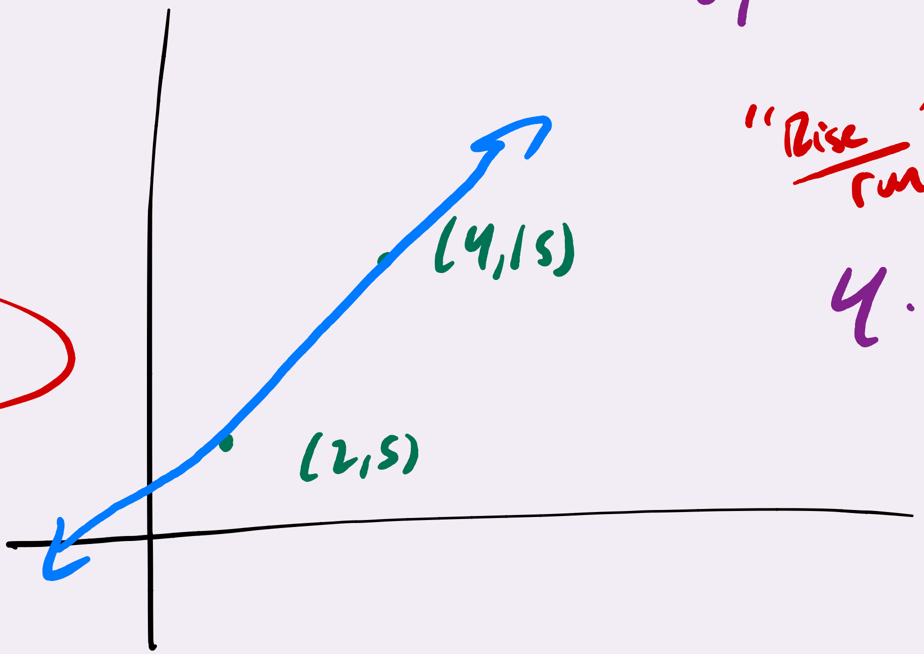
Take a moment to pause and reflect...

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. $w_0^* = -5, w_1^* = 5$

$w_1^* = \text{slope} = \frac{15-5}{4-2} = \frac{10}{2} = 5$

"$\frac{\text{Rise}}{\text{run}}$"



$4 \cdot 5 + w_0^* = 15$

$w_0^* = 15 - 20$

$= -5$

Correlation

Correlation = linear Pattern   Association = Any Pattern

Quantifying patterns in scatter plots

- In DSC 10, you were introduced to the idea of the **correlation coefficient**, r.
- It 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.

Neg. Corr. (-1 ⟷ 0)

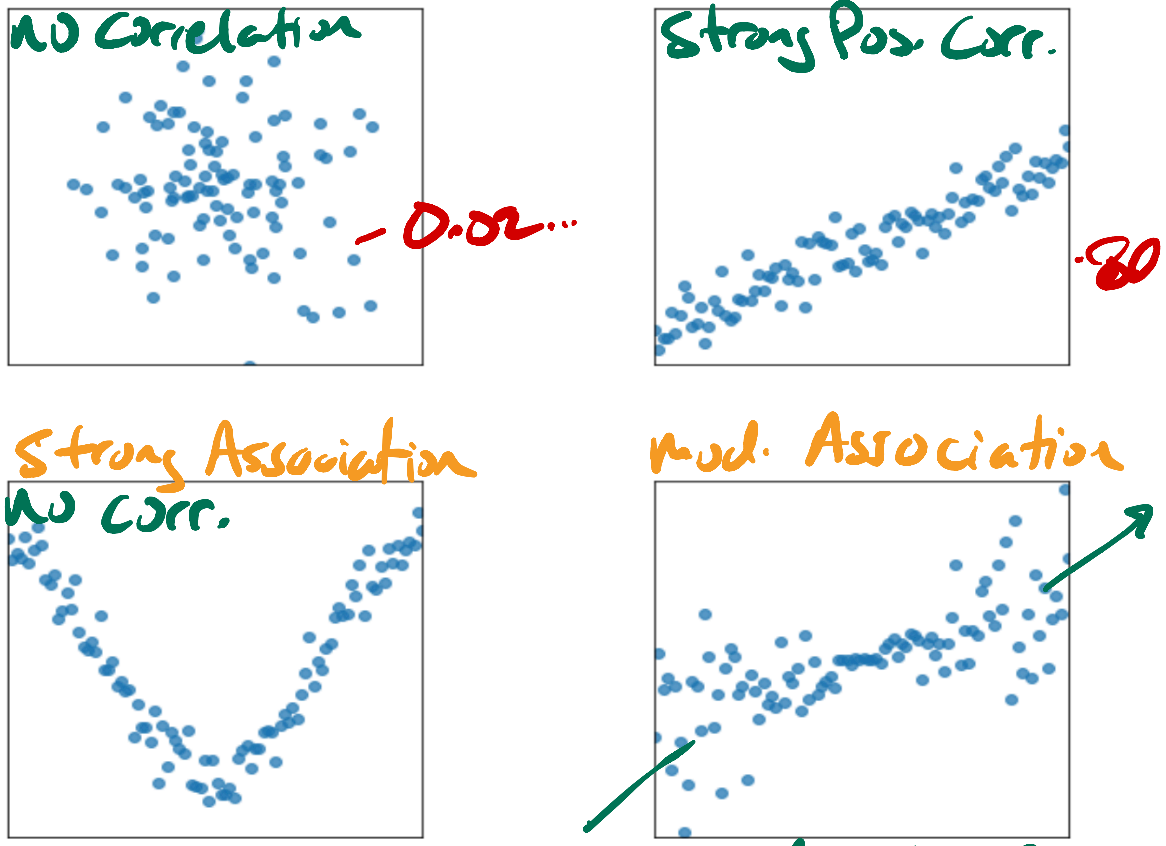
Pos. Corr. (0, 1)

No Association
No Correlation
-0.02...

Strong Association
Strong Pos. Corr.
.80

Strong Association
No Corr.

Mod. Association



Moderate Pos. Corr.

in DSC40A it is "Pearson"

Standardized or normalized.

-20 -10 10 20

mean 0

$\sigma = 1$

$(x, y) \rightarrow (0, 1)$

The correlation coefficient

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

sigma σ

$$r = \underbrace{\frac{1}{n} \sum_{i=1}^{n}}_{\text{avg.}} \underbrace{\left(\frac{x_i - \bar{x}}{\sigma_x} \right)}_{x_i \text{ in standard units}} \underbrace{\left(\frac{y_i - \bar{y}}{\sigma_y} \right)}_{y_i \text{ " " "}}$$

Why the product of x and y

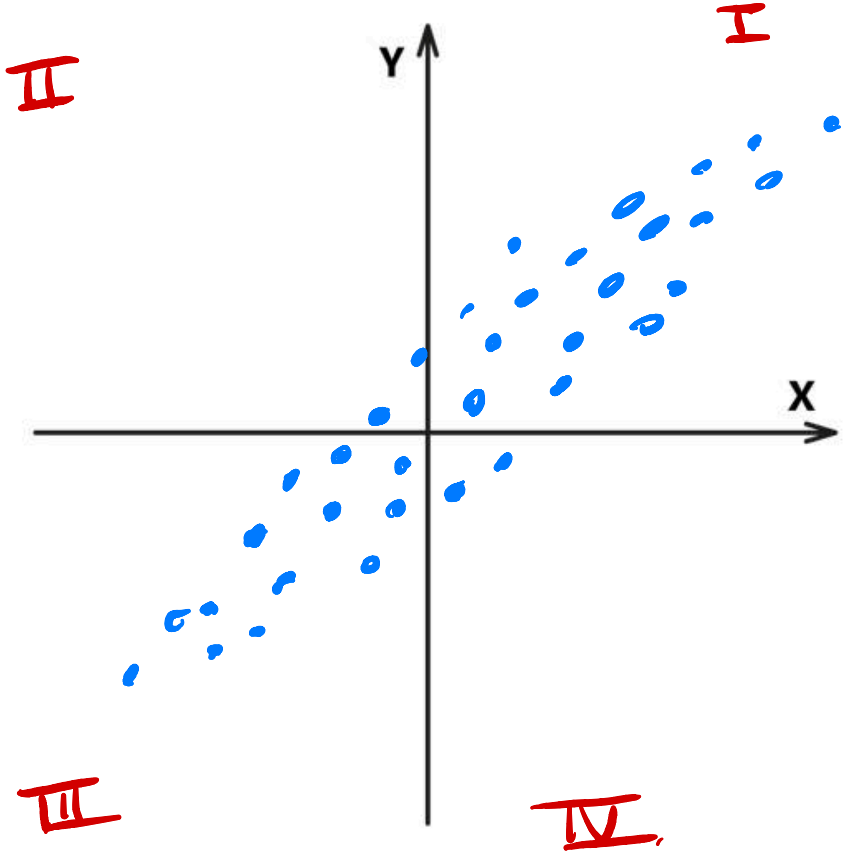
- We really care about the sign (i.e. $+$ or $-$)

Strong Pos. Corr.

100 points. {40% in I & III, 10% in II, IV}

I: x_i (su) positive

y_i (su) positive

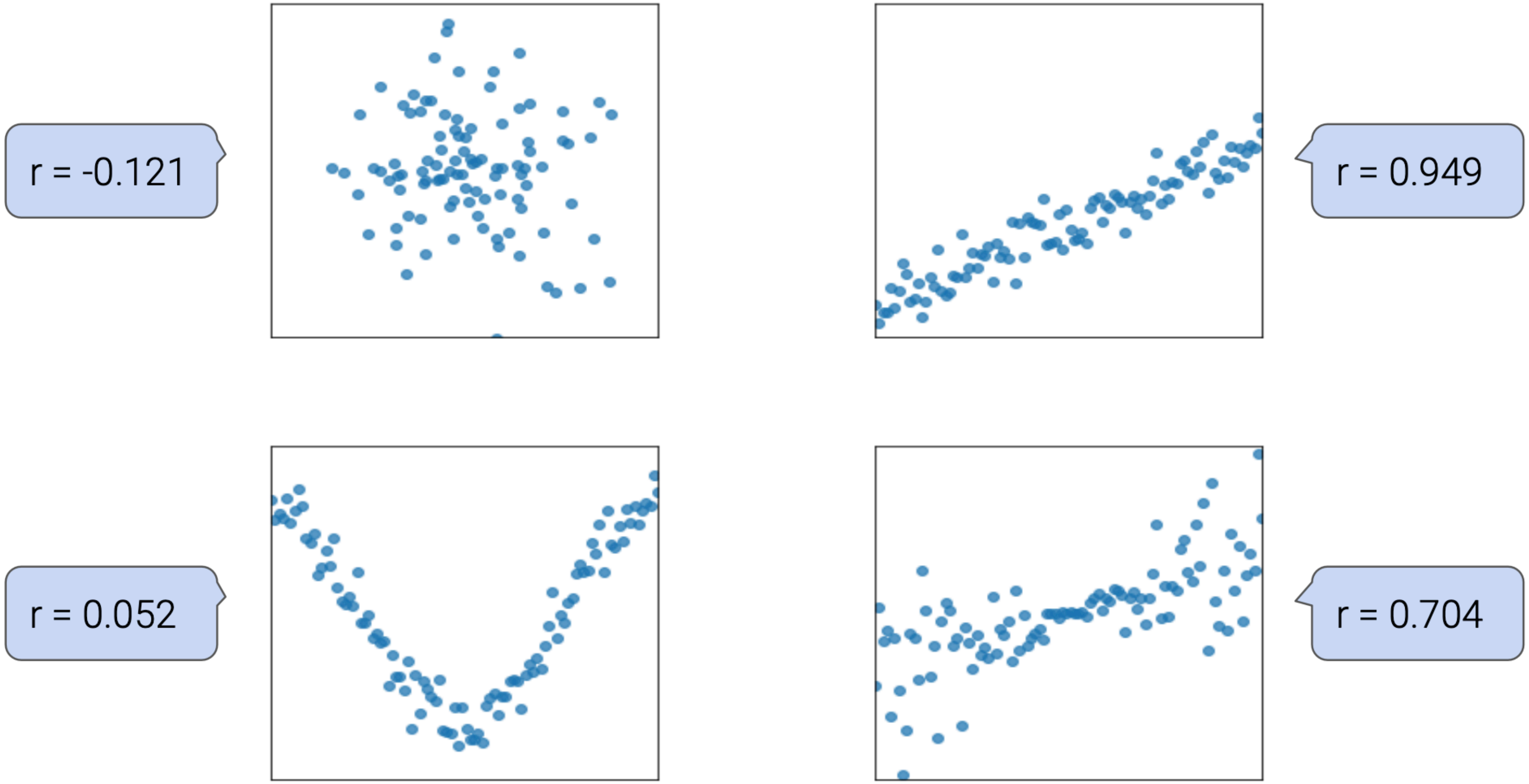


III.

x_i (su) negative

y_i (su) negative

The correlation coefficient, visualized



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!

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

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

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

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

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

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

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

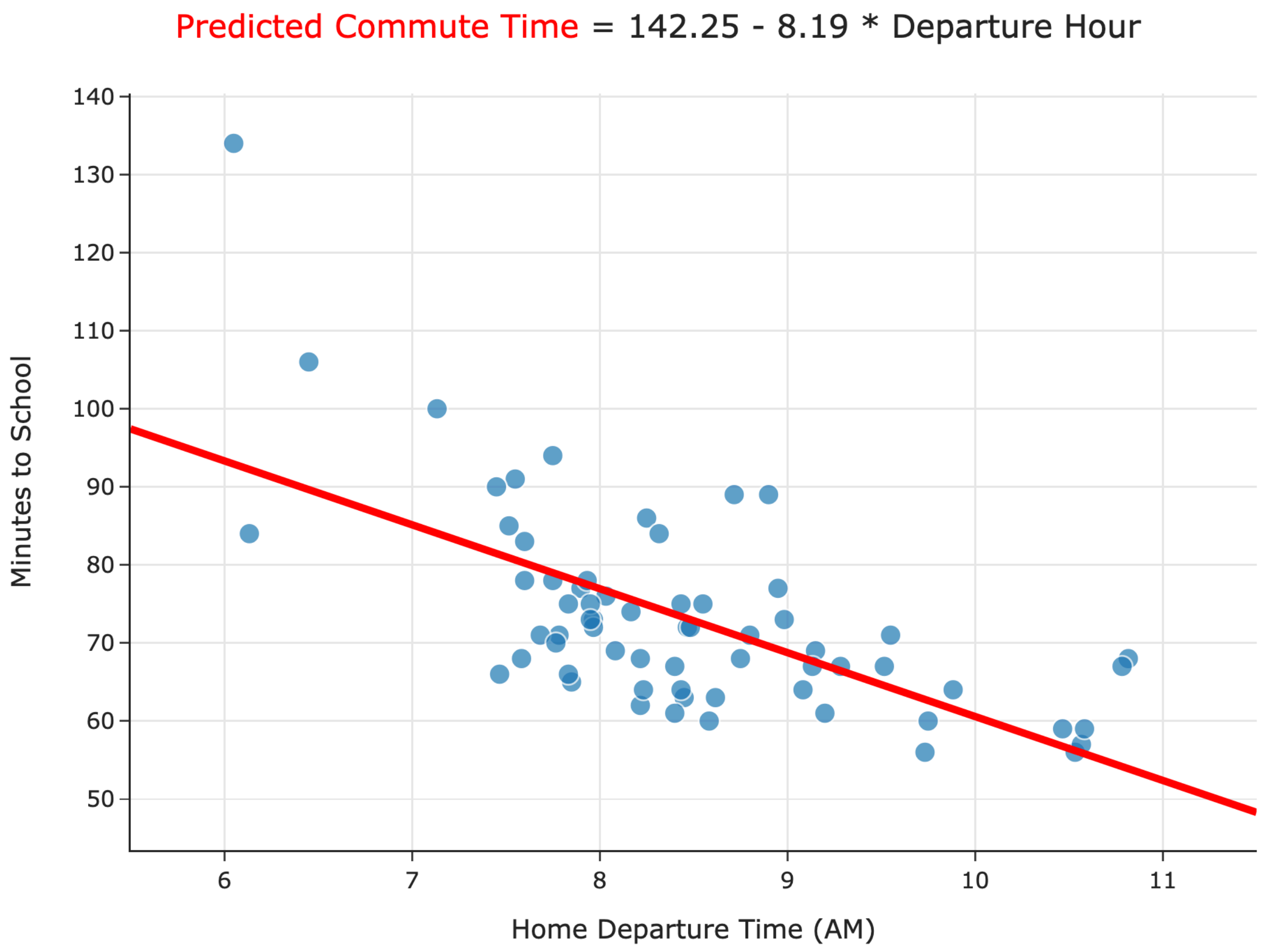
$$w_1^* = \frac{r \cancel{n} \cancel{\sigma_x} \sigma_y}{\cancel{n} \sigma_x^{\cancel{2}}}$$ ① ②

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

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

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

Let's test these new formulas out in code! Follow along by using the link for this week on the course home page.



Interpreting the formulas

r: has no units

σ_y : units of y

σ_x : units of x

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 $H(x) = 142.25 - 8.19x$, our predicted commute time **decreases by 8.19 minutes per hour**.

x_i : Departure time in hours

y_i : Commute time in minutes

8.19 m/h

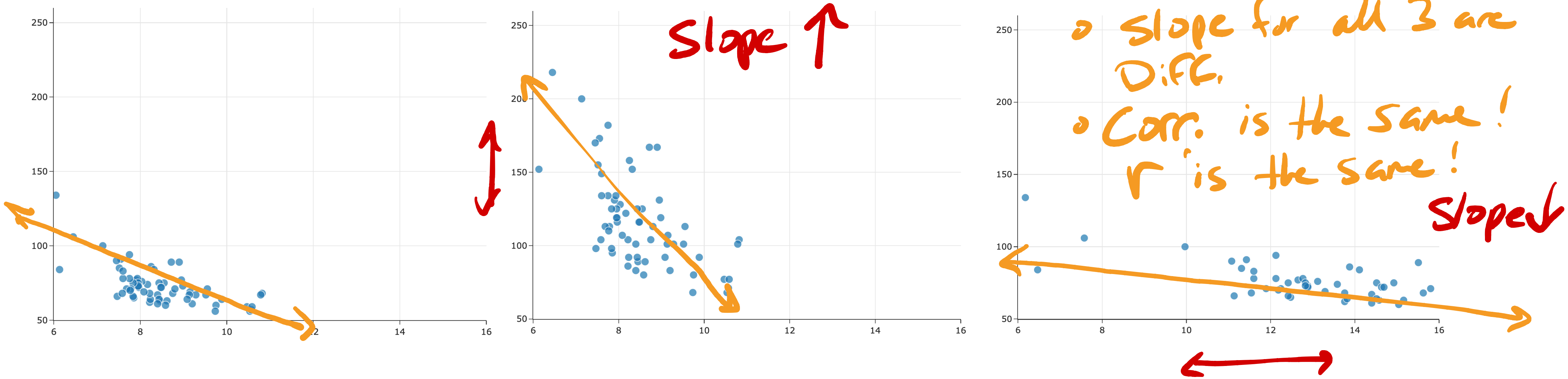
or

8 min & ~12 sec. decrease every hour later you leave

if: Low slope
then: this is no correlation ✗ not true
slope ≠ corr.

Interpreting the slope

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



Slope ↑

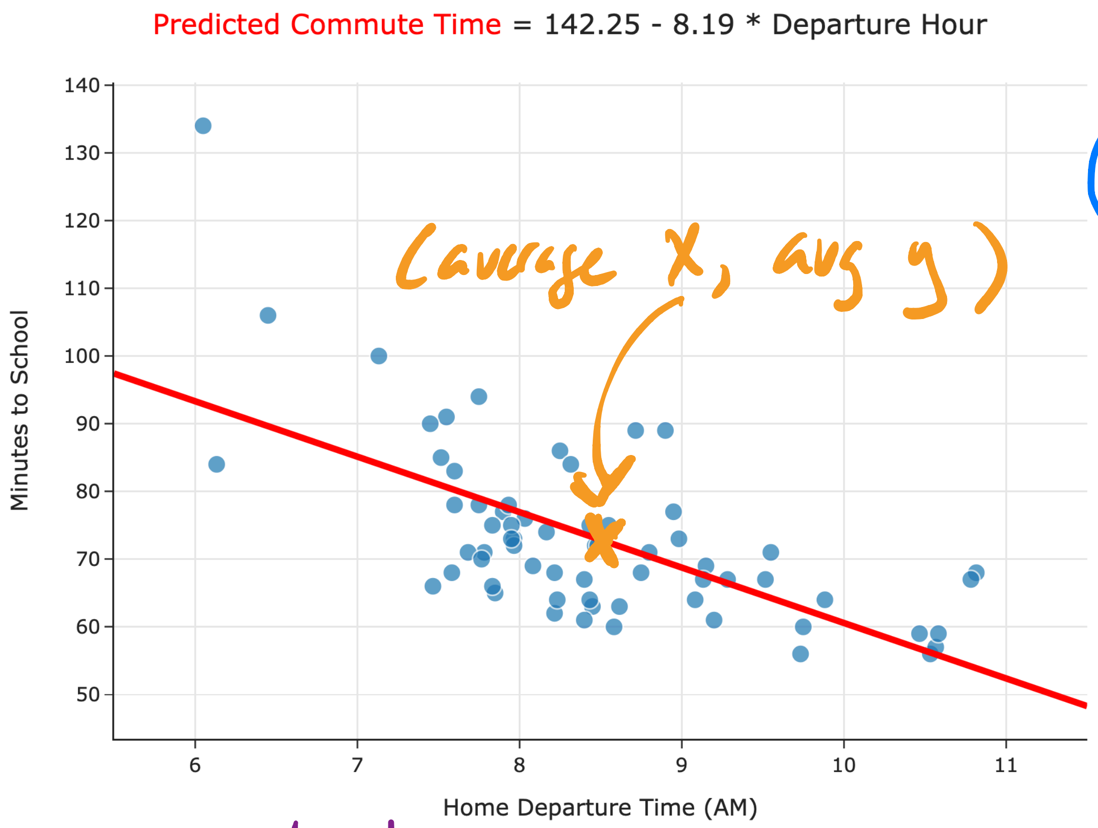
- slope for all 3 are DiFF.
- Corr. is the same!
  r is the same!

Slope ↓

- 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

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



$H(0)$ = Commute time at 2400 "midnight"

- What are the units of the intercept?

units of y : minutes

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

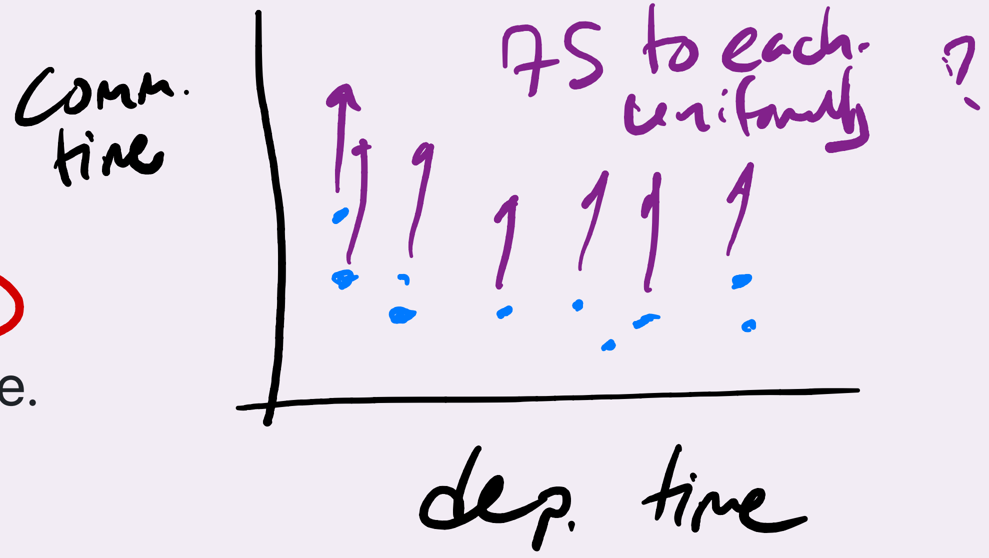
$$H^*(x_i) = w_0^* + w_1^* x_i$$
$$= \bar{y} - w_1^* \bar{x} + w_1^* x_i$$
$$= \bar{y} - w_1^* (\bar{x} - x_i)$$
$$H^*(\bar{x}) = \bar{y} + w_1^* (x_i - \bar{x})$$
$$= \bar{y} + w_1^* \cancel{(\bar{x} - \bar{x})} \rightarrow 0 \quad = \bar{y}$$

Question 🤔

Take a moment to pause and reflect...

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.



Correlation and mean squared error

- **Claim**: Suppose that w_0^* and w_1^* are the optimal intercept and slope for the regression line. Then,

$$R_{\text{sq}}(w_0^*, w_1^*) = \sigma_y^2(1 - r^2)$$

- That is, the mean squared error of the regression line's predictions and the correlation coefficient, r, always satisfy the relationship above.
- Even if it's true, why do we care?
  - In machine learning, we often use both the mean squared error and r^2 to compare the performances of different models.
  - If we can prove the above statement, we can show that **finding models that minimize mean squared error** is equivalent to **finding models that maximize r^2**

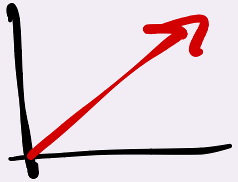
.

Connections to related models

Question 🤔

Take a moment to pause and reflect...

line going through the origin!



Suppose we chose the model $H(x) = w_1 x$ and squared loss.

What is the optimal model parameter, w_1^*?

- A. $\cancel{\dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
- B. $\dfrac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$
- [redacted]
- D. $\cancel{\dfrac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}}$

Exercise

Suppose we chose the model $H(x) = w_1 x$ and squared loss.

What is the optimal model parameter, w_1^*?

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

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

$$= \frac{-2}{n} \sum_{i=1}^{n} (x_i y_i - w_1 x_i^2) = 0$$

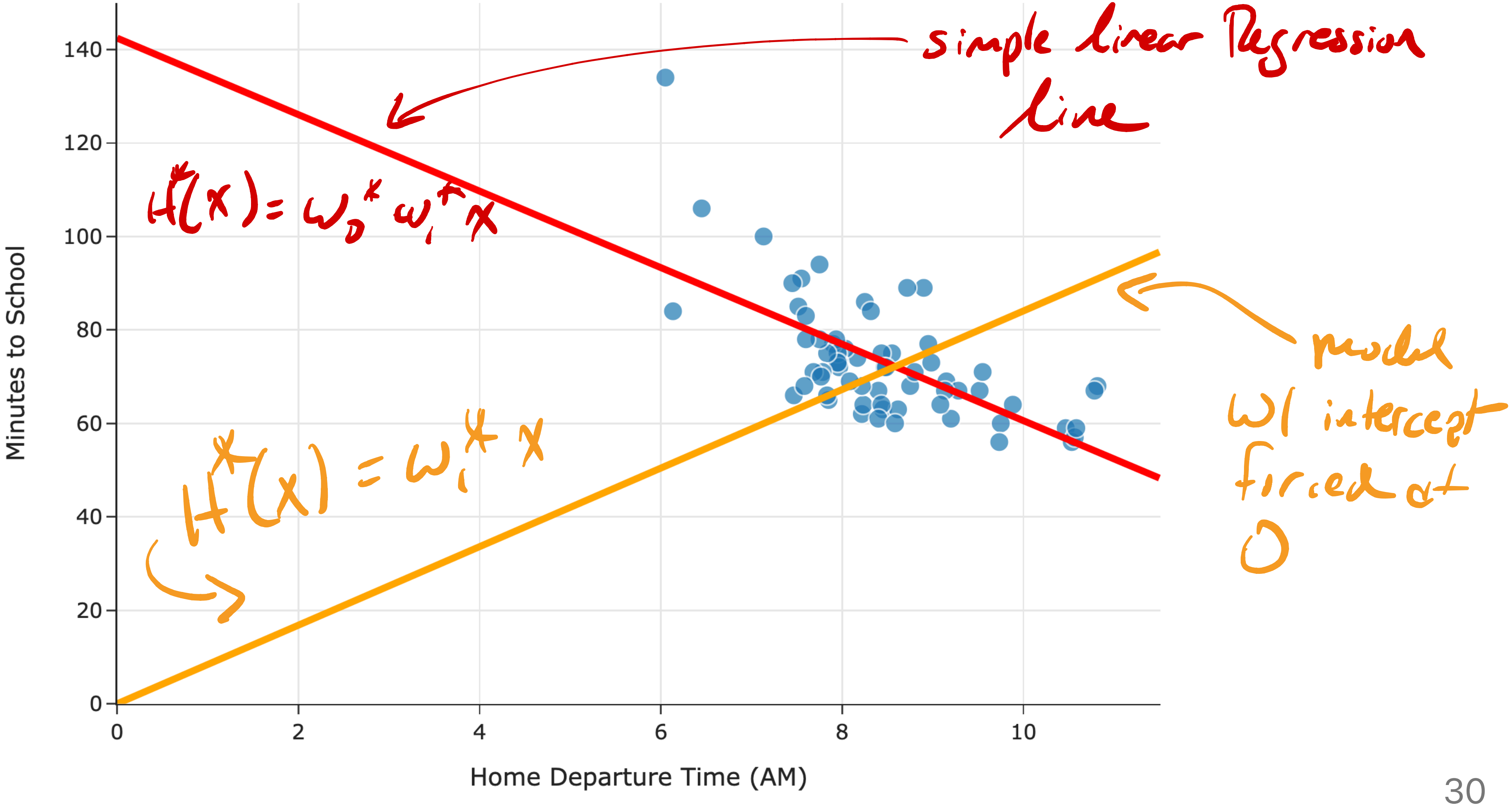
$$= \sum_{i=1}^{n} (x_i y_i - w_1 x_i^2) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} (x_i^2 w_1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = w_1 \sum_{i=1}^{n} x_i^2$$

$$w_1^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Predicted Commute Time = 142.25 - 8.19 * Departure Hour
Predicted Commute Time = 8.41 * Departure Hour



Exercise

Suppose we choose the model $H(x) = w_0$ and squared loss.

What is the optimal model parameter, w_0^*?

$$w_0^* = \text{Mean}(y_1, y_2, \ldots y_n)$$

Comparing mean squared errors

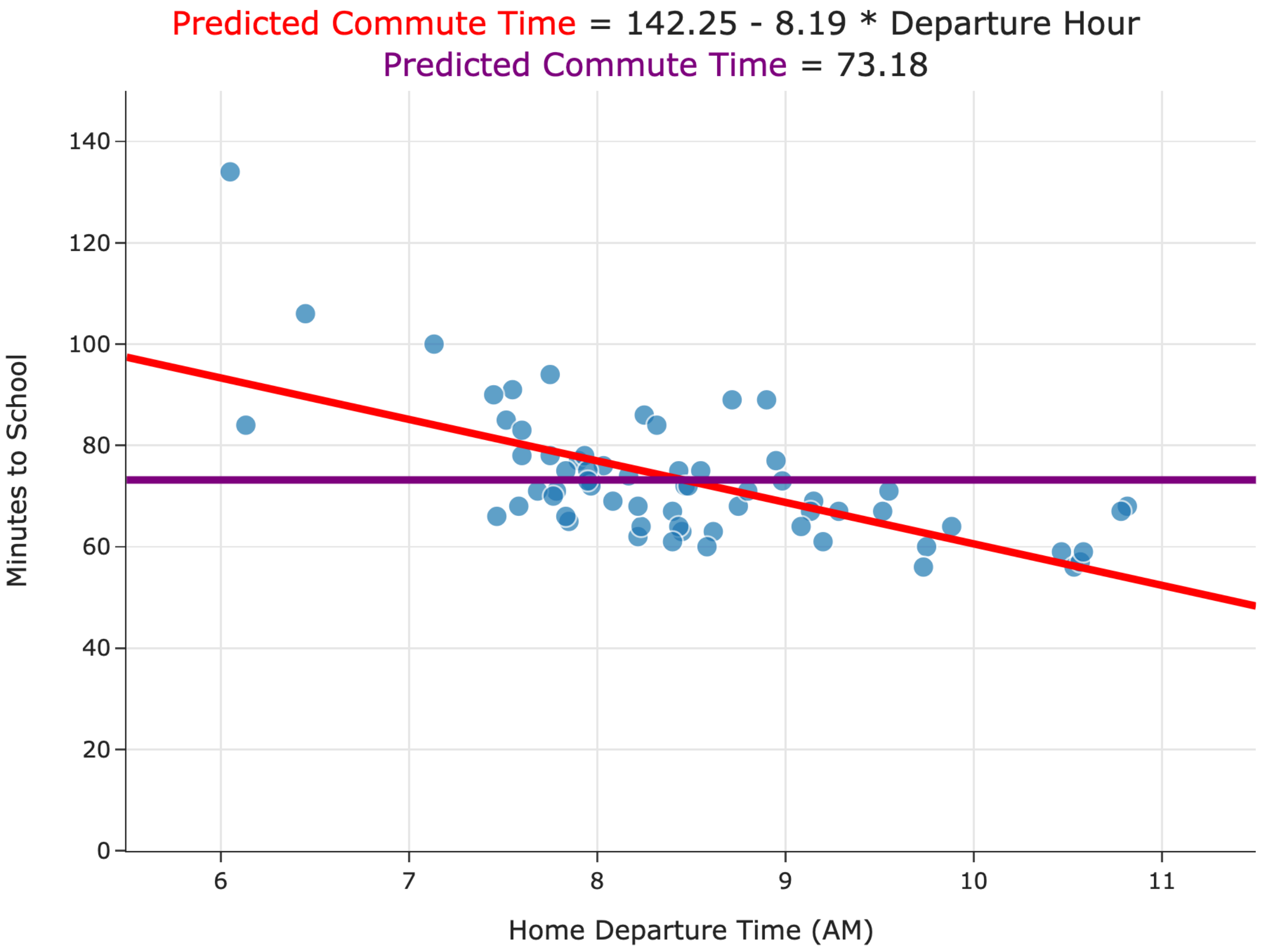
- With both:
    - the constant model, $H(x) = h$, and
    - the simple linear regression model, $H(x) = w_0 + w_1 x$,

  when we chose squared loss, we minimized mean squared error to find optimal parameters:

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

- **Which model minimizes mean squared error more?**

Comparing mean squared errors



$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - H(x_i))^2$$

- The MSE of the best simple linear regression model is ≈ 97.
- The MSE of the best constant model is ≈ 167.
- The simple linear regression model is a more flexible version of the constant model.

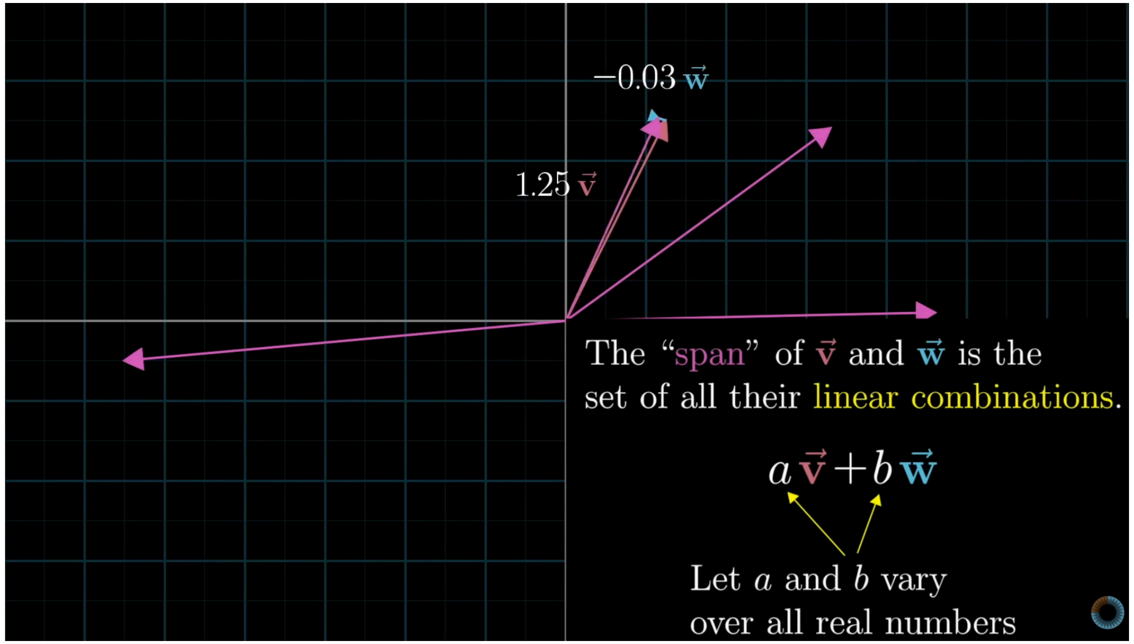
Linear algebra review

Wait... why do we need linear algebra?

- Soon, we'll want to make predictions using more than one feature.
  - Example: Predicting commute times using departure hour and temperature.
- Thinking about linear regression in terms of **matrices and vectors** will allow us to find hypothesis functions that:
  - Use multiple features (input variables).
  - Are non-linear, e.g. $H(x) = w_0 + w_1 x + w_2 x^2$.
- Before we dive in, let's review.

Spans of vectors

- One of the most important ideas you'll need to remember from linear algebra is the concept of the **span** of two or more vectors.
- To jump start our review of linear algebra, let's start by watching 🎥 this video by 3blue1brown.



Next time

- We'll review the necessary linear algebra prerequisites.
- We'll then start to formulate the problem of minimizing mean squared error for the simple linear regression model **using matrices and vectors**.