Lecture 5

More Simple Linear Regression

DSC 40A, Summer 2024

Announcements

- Homework 2 is due **tomorrow**. Remember that using the Overleaf template is *required* for Homework 2 (and only Homework 2).
- Groupwork 1 solutions are available on Ed. Homework 1 solutions coming this afternoon.
- Reminder to check out the FAQs page and the tutor-created supplemental resources on the course website, if you'd like extra practice or review.
- Please turn your camera on when working with tutors in virtual office hours.
- Grace period for Groupwork 1: submit by 11:59p tonight, if you haven't yet.

Agenda

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

Question 🤔

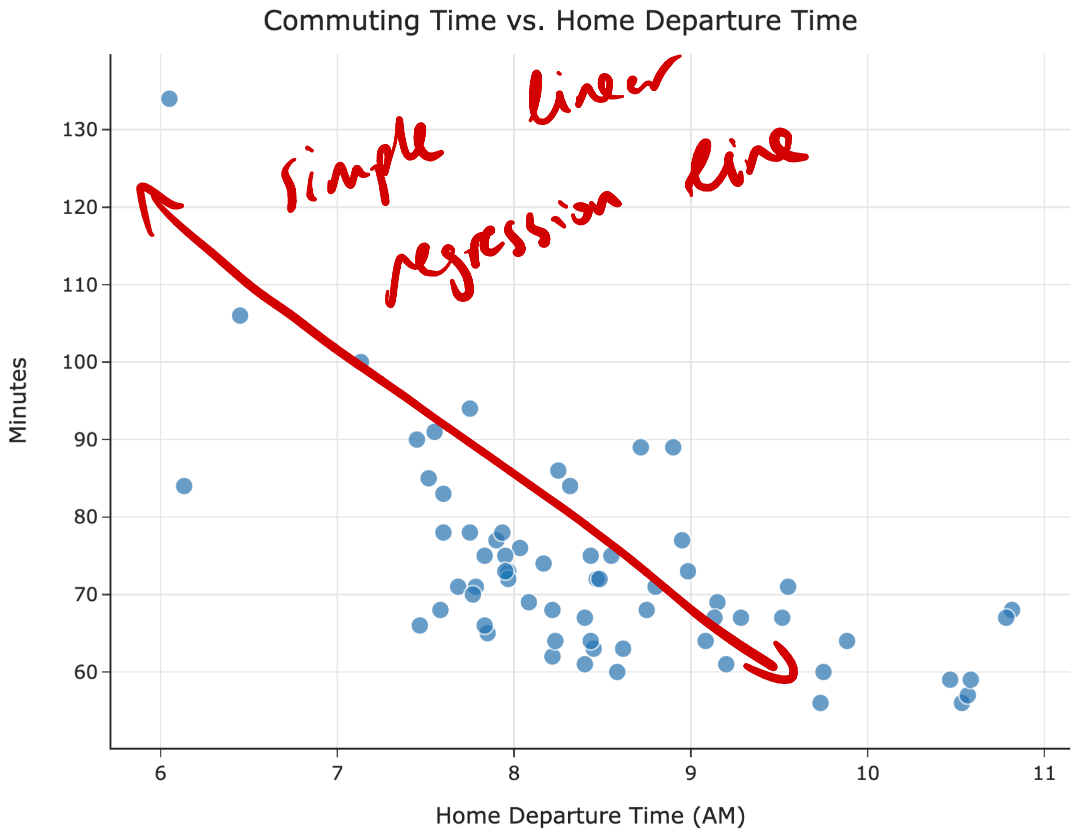
Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

If the direct link doesn't work, click the "🤔 Lecture Questions" link in the top right corner of dsc40a.com.

Recap: Simple linear regression

Recap



- 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.

$$H(x) = \underbrace{w_0}_{\text{intercept}} + \underbrace{w_1}_{\text{slope}} x$$

2. Choose a loss function.

Squared loss

$$L_{sq}(y_i, H(x_i)) = (\underbrace{y_i - H(x_i)}_{\text{actual} - \text{predicted}})^2$$

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:

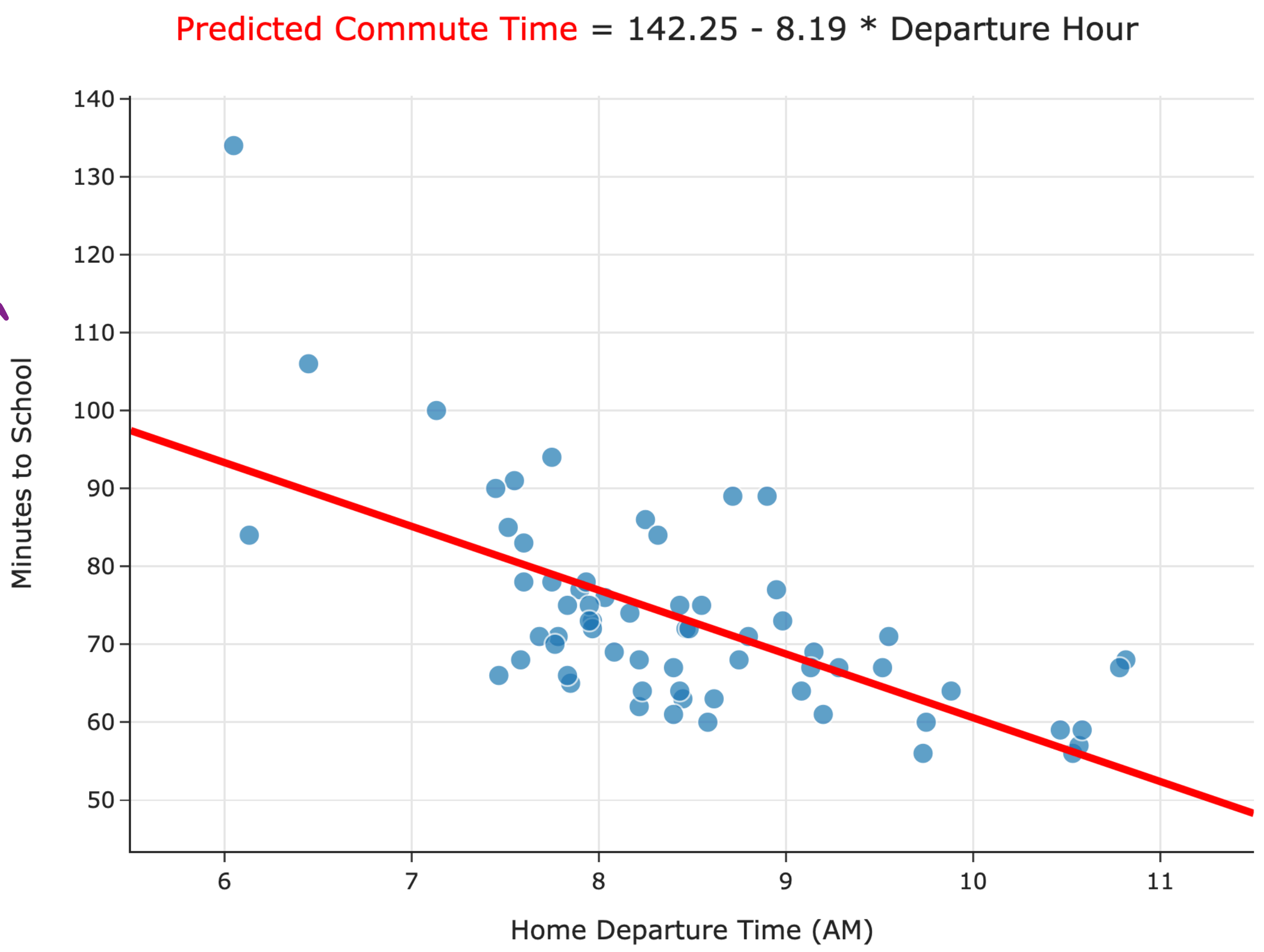
best slope

best intercept

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

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

There is no other line for this data with a smaller mean squared error!



Now what?

$H(x) = w_0 + w_1 x$

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!

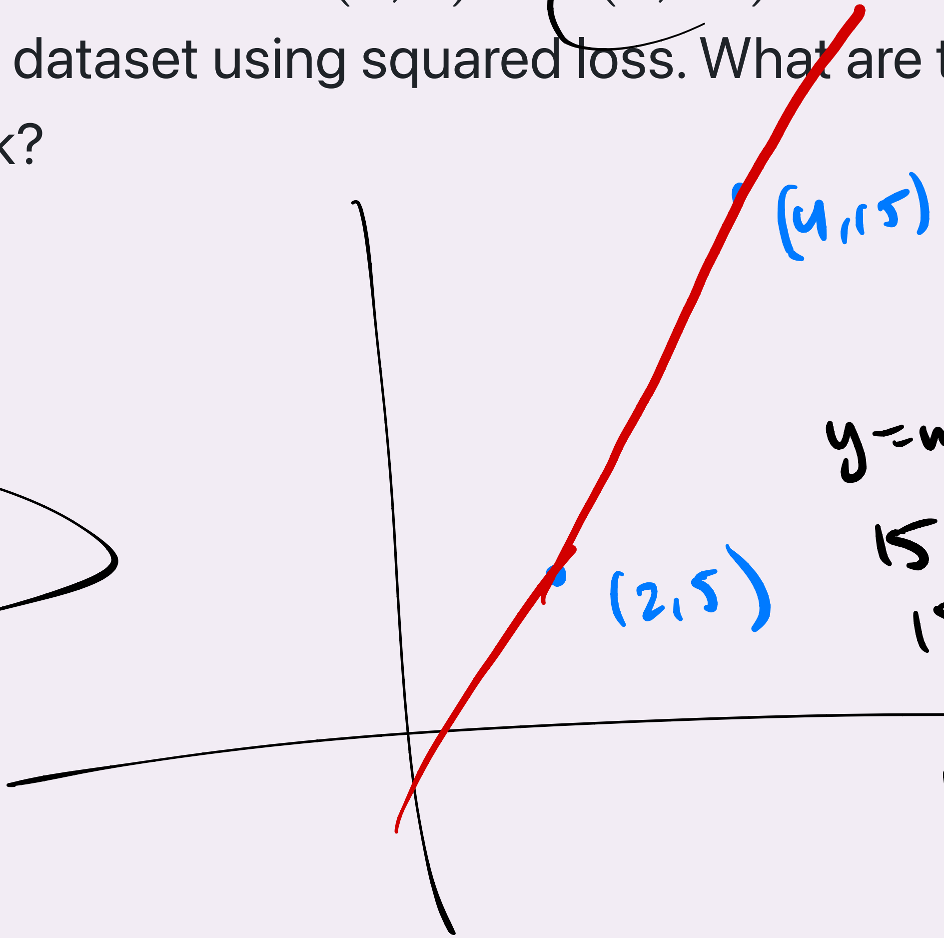
$+ w_2 x^2$

Question 🤔

Answer at q.dsc40a.com

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.~~ $\cancel{w_0^* = 3, w_1^* = 10}$
- C. $w_0^* = -2, w_1^* = 5$
- D. $w_0^* = -5, w_1^* = 5$



slope $= \frac{\text{rise}}{\text{run}} = \frac{10}{2}$

$= 5$

$y = mx + b$

$15 = 5 \cdot 4 + b$

$15 = 20 + b$

$b = -5$

Correlation

association: any pattern

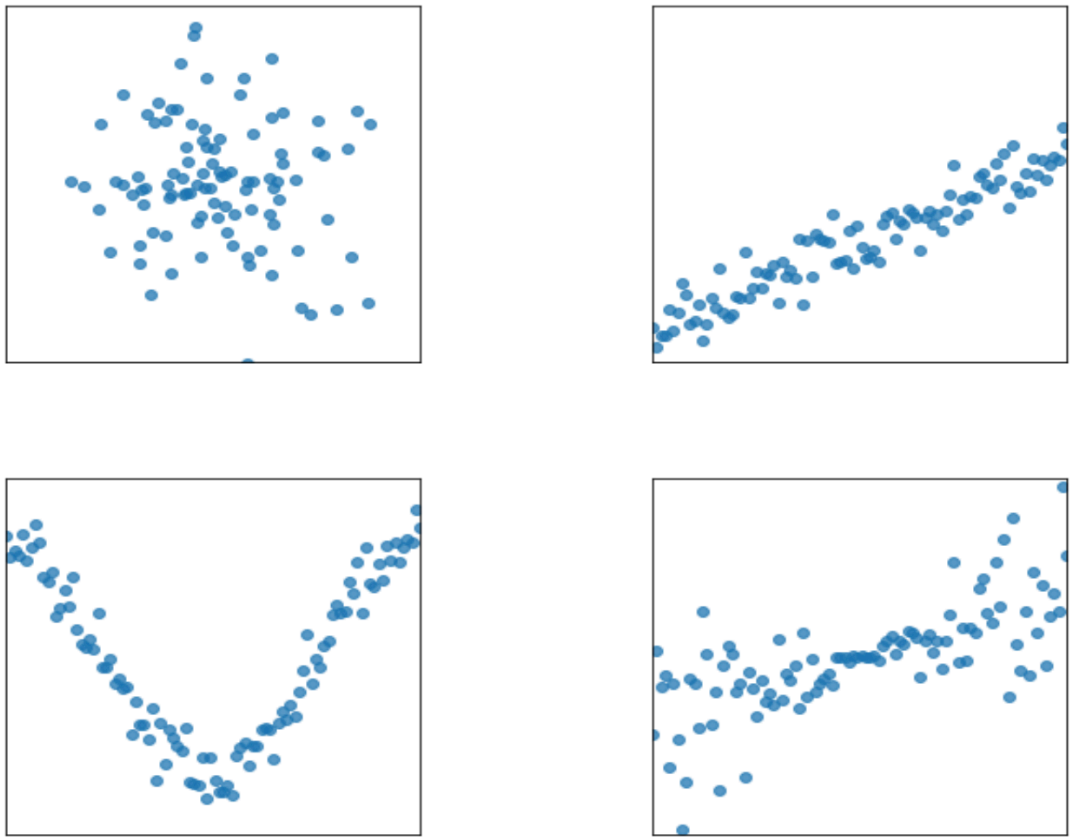
correlation: linear pattern $\subseteq$ association

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.

$r = -1$

$r = 1$



Pearson's (there are others)

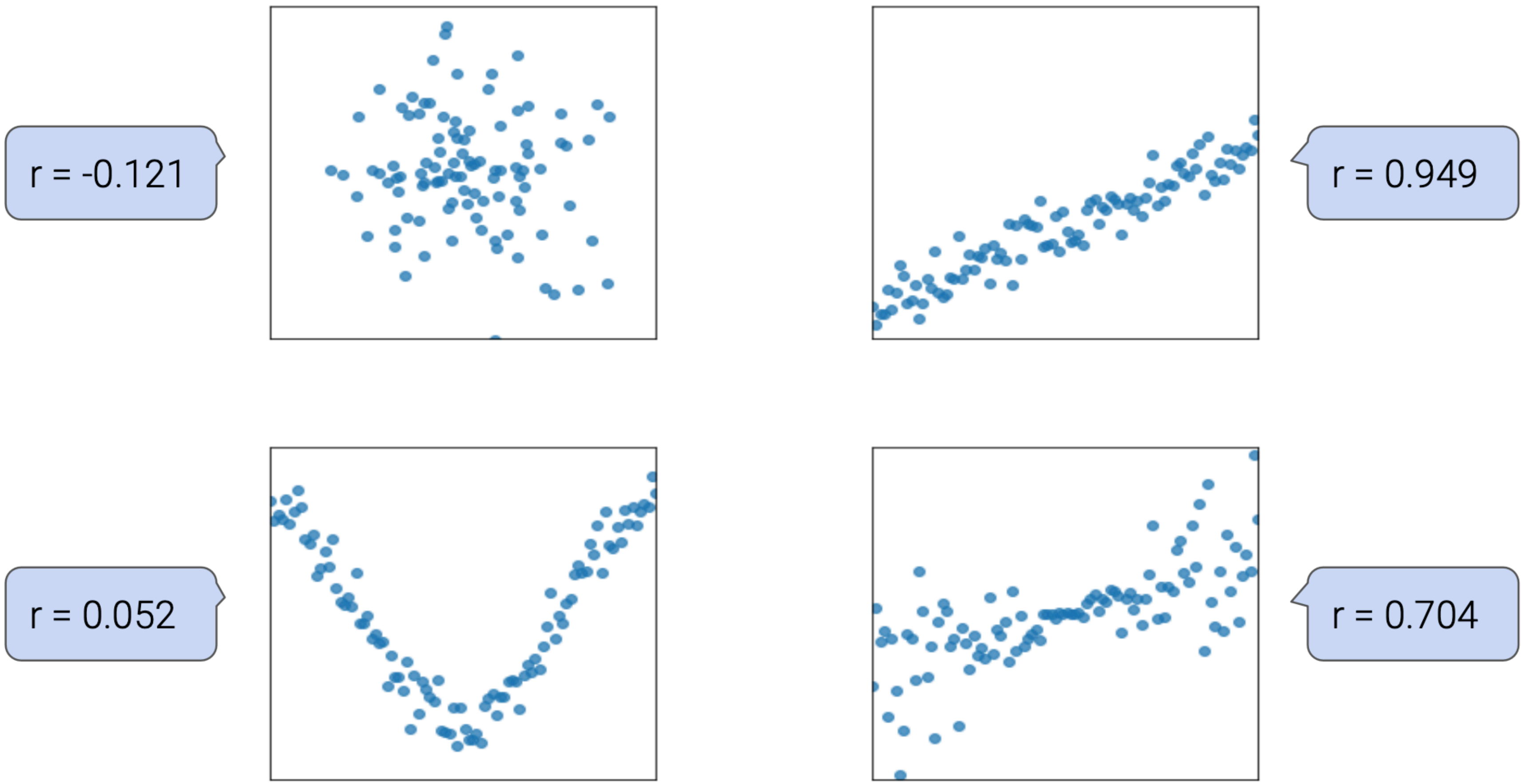
The correlation coefficient

- 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 x"

$$r = \underbrace{\frac{1}{n} \sum_{i=1}^{n}}_{\text{average}} \underbrace{\left(\frac{x_i - \bar{x}}{\sigma_x} \right)}_{x_{su}} \cdot \underbrace{\left(\frac{y_i - \bar{y}}{\sigma_y} \right)}_{y_{su}}$$

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!

$$w_1^* = \overbrace{\frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2}}^{\text{last week}} = \overbrace{r\frac{\sigma_y}{\sigma_x}}^{\text{DSC 10}}$$

- 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^*:

$$w_1^* = r\frac{\overset{\text{SDy}}{\sigma_y}}{\underset{\text{SDx}}{\sigma_x}} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

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

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

$$= \frac{r \cdot \cancel{n} \cdot \cancel{\sigma_x} \sigma_y}{\cancel{n}\, \sigma_x^{\cancel{2}}}$$

$$= (r)\left(\frac{\sigma_y}{\sigma_x}\right)$$

Aside

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

constants?

$$= \frac{1}{n \cdot \sigma_x \cdot \sigma_y}\sum (x_i-\bar{x})(y_i-\bar{y})$$

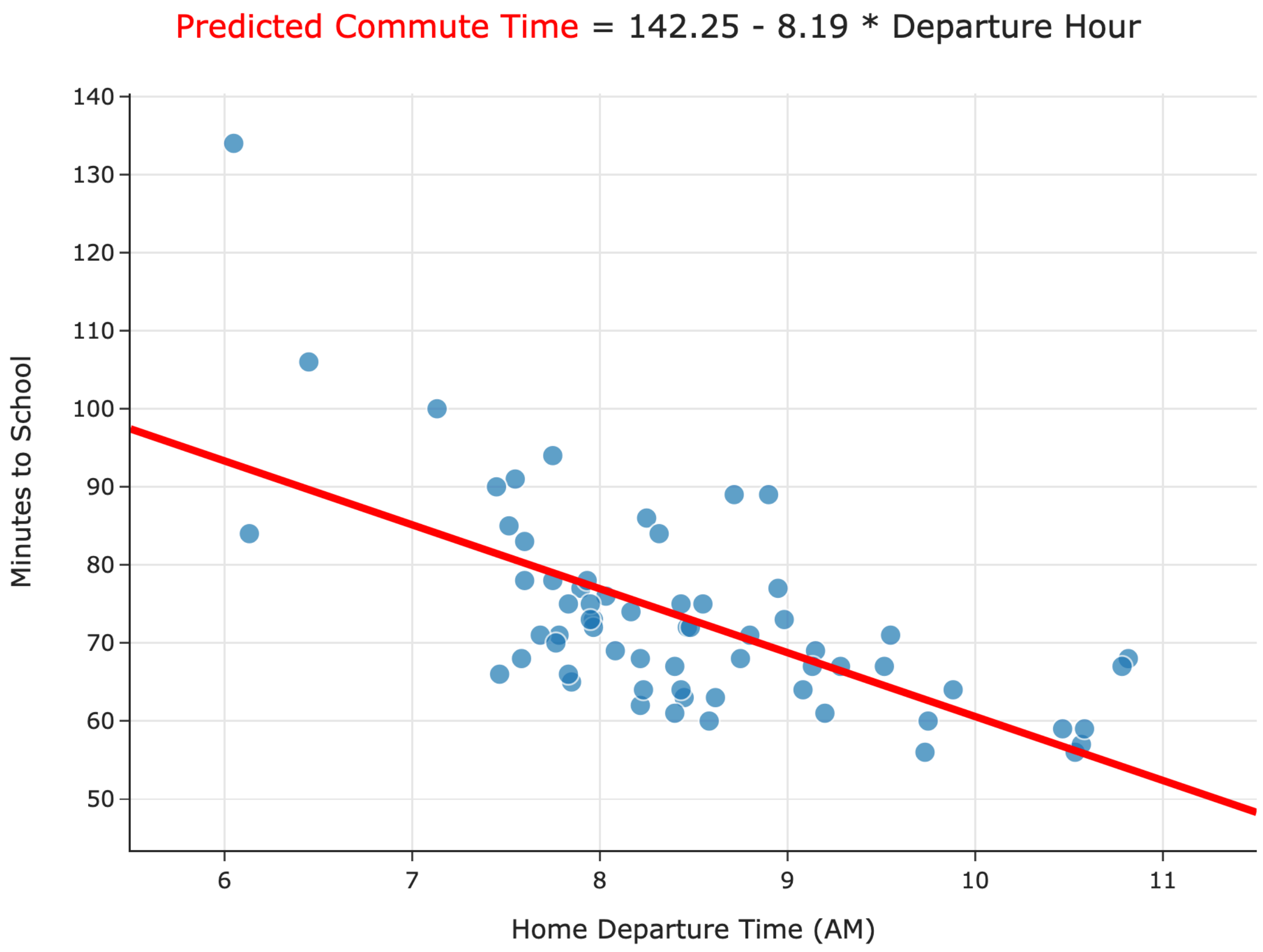
(1) $$\boxed{r \cdot n \cdot \sigma_x \cdot \sigma_y = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}$$

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

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

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

Let's test these new formulas out in code! Follow along here.



Interpreting the formulas

Interpreting the slope

unitless

units of y

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

units of x

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

xi: departure times in hours

yi: commute times in minutes

r is the same for all graphs

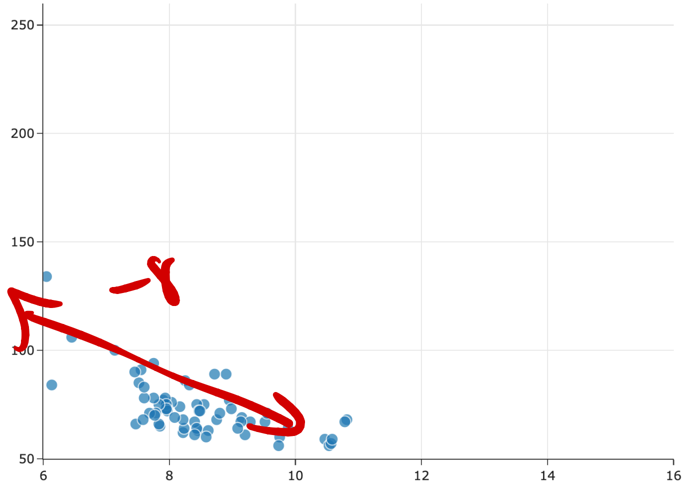
Interpreting the slope

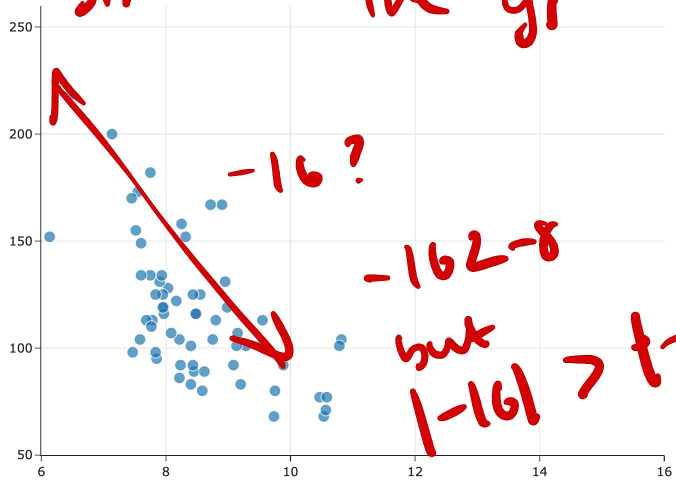
$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$

original

stretched the ys

stretched the xs

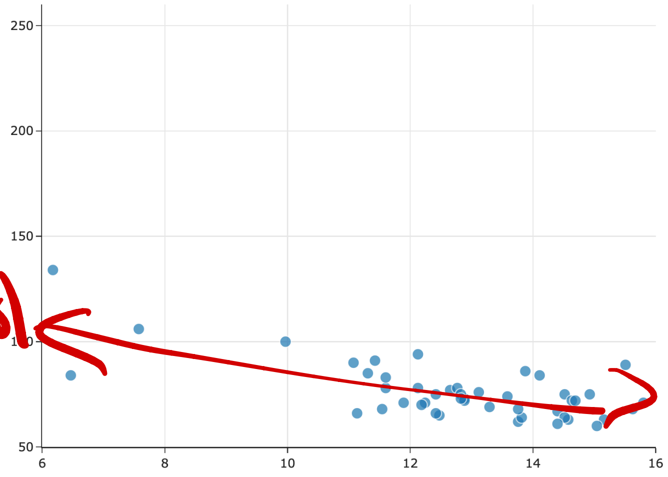




$-16 < -8$

but

$|-16| > |-8|$



- 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. → increases in magnitude
- As the x values get more spread out, σ_x increases, so the slope gets shallower. decreases in magnitude

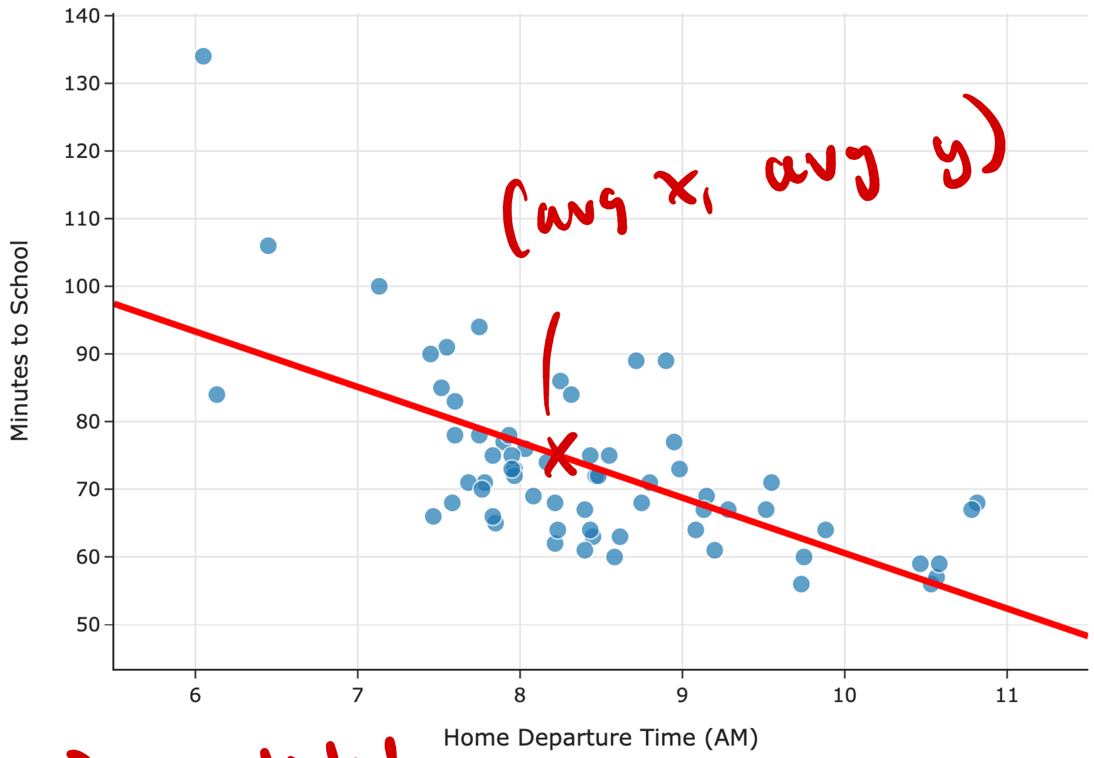
from 10: $y_{su}^{pred} = r \cdot x_{su}$

$0 = r \quad 0$

Interpreting the intercept

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

Predicted Commute Time = 142.25 - 8.19 * Departure Hour



H(0) = predicted commute time at midnight

- What are the units of the intercept?

Units of y: minutes

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

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

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

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

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

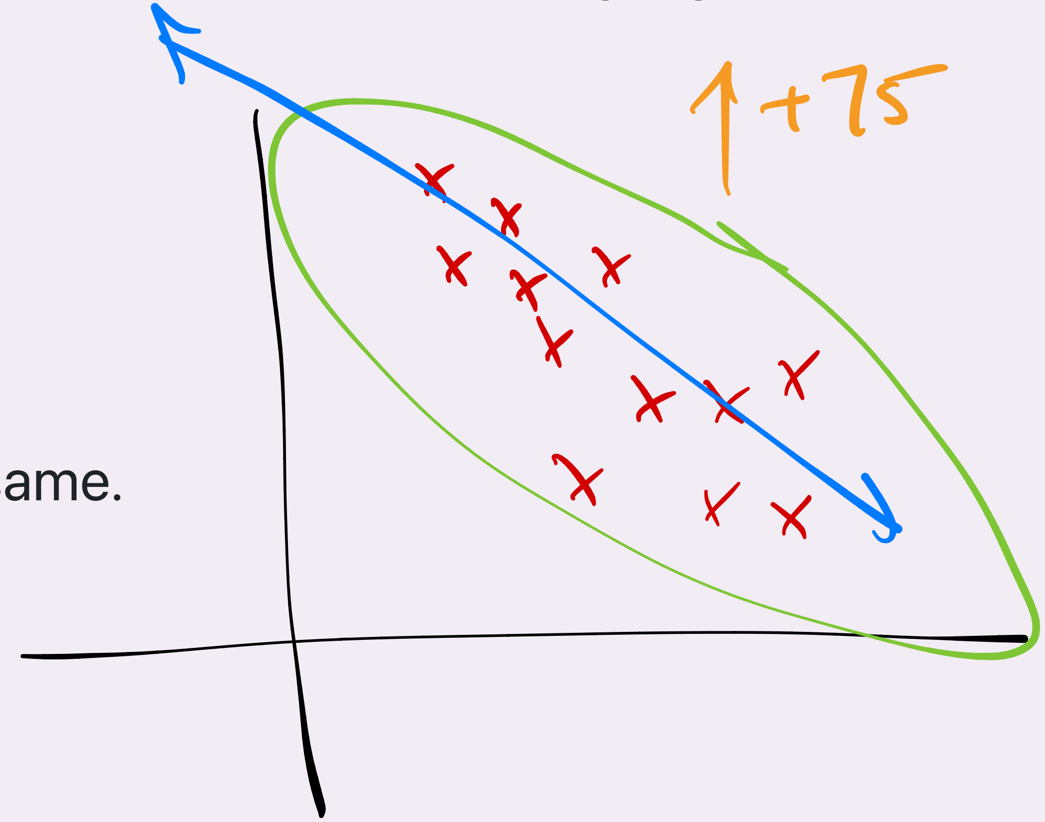
$= \bar{y}$

Question 🤔

Answer at q.dsc40a.com

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

|r| closer to 1

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

"better" the correlation, the lower the MSE

MSE equiv

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

- That is, the mean squared error of the regression line's predictions and the correlation coefficient, r, always satisfy the relationship above.
- For more, find the proof in our FAQs (link). But 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, 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 🤔

Answer at q.dsc40a.com

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

What is the optimal model parameter, w_1^*?

- A. $\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- B. $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$
- C. $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i}$
- D. $\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$

Exercise

minimize empirical risk

best for THIS model

Suppose we chose the model $H(x) = w_1 x$ and squared loss.
What is the optimal model parameter, w_1^*?

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

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

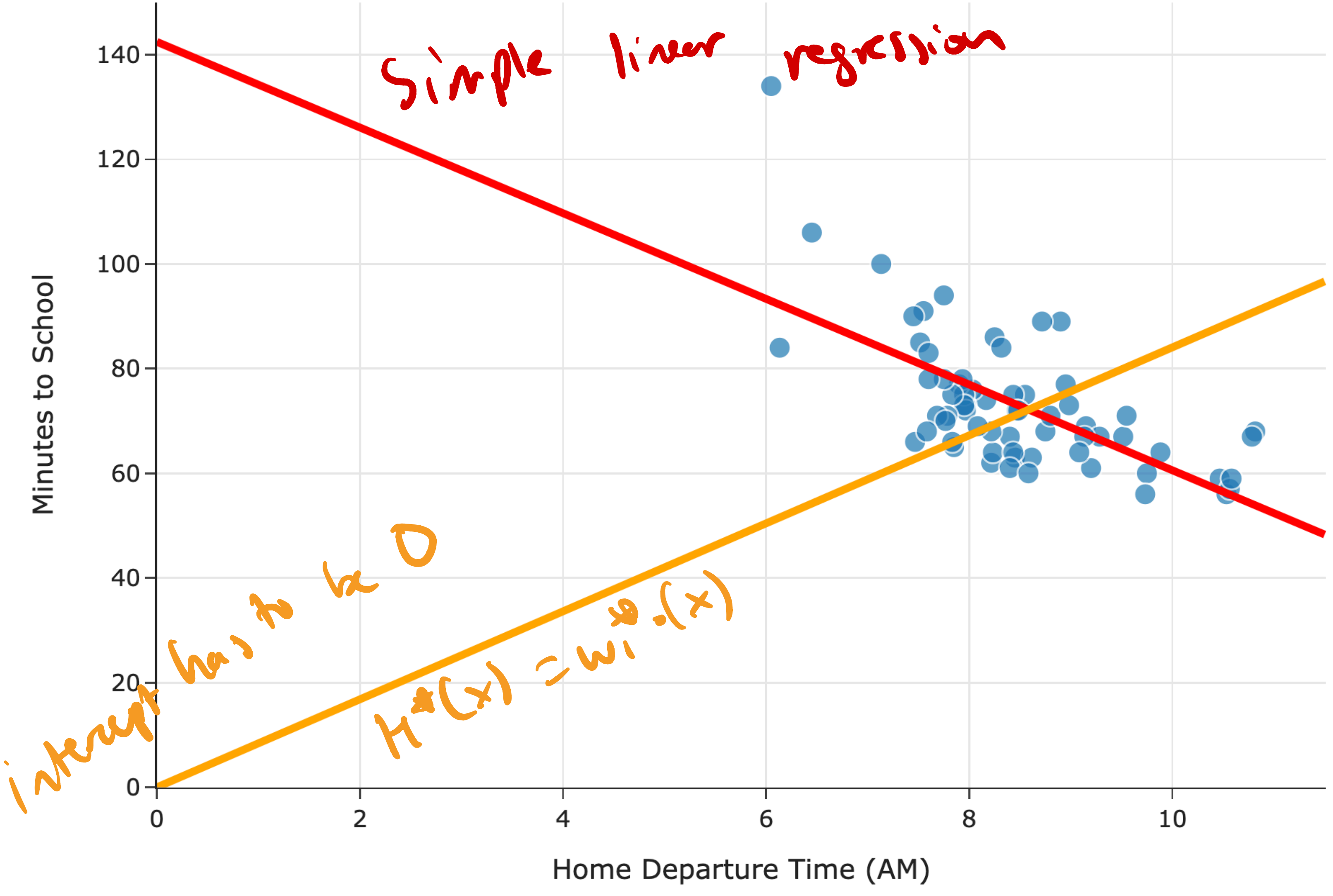
$$\frac{dR_{sq}}{dw_1} = \frac{1}{n}\cdot 2 \sum_{i=1}^{n} (y_i - w_1 x_i)(-x_i) = 0$$

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

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

Predicted Commute Time = 142.25 - 8.19 * Departure Hour

Predicted Commute Time = 8.41 * Departure Hour



Exercise

"h" just a new name!

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

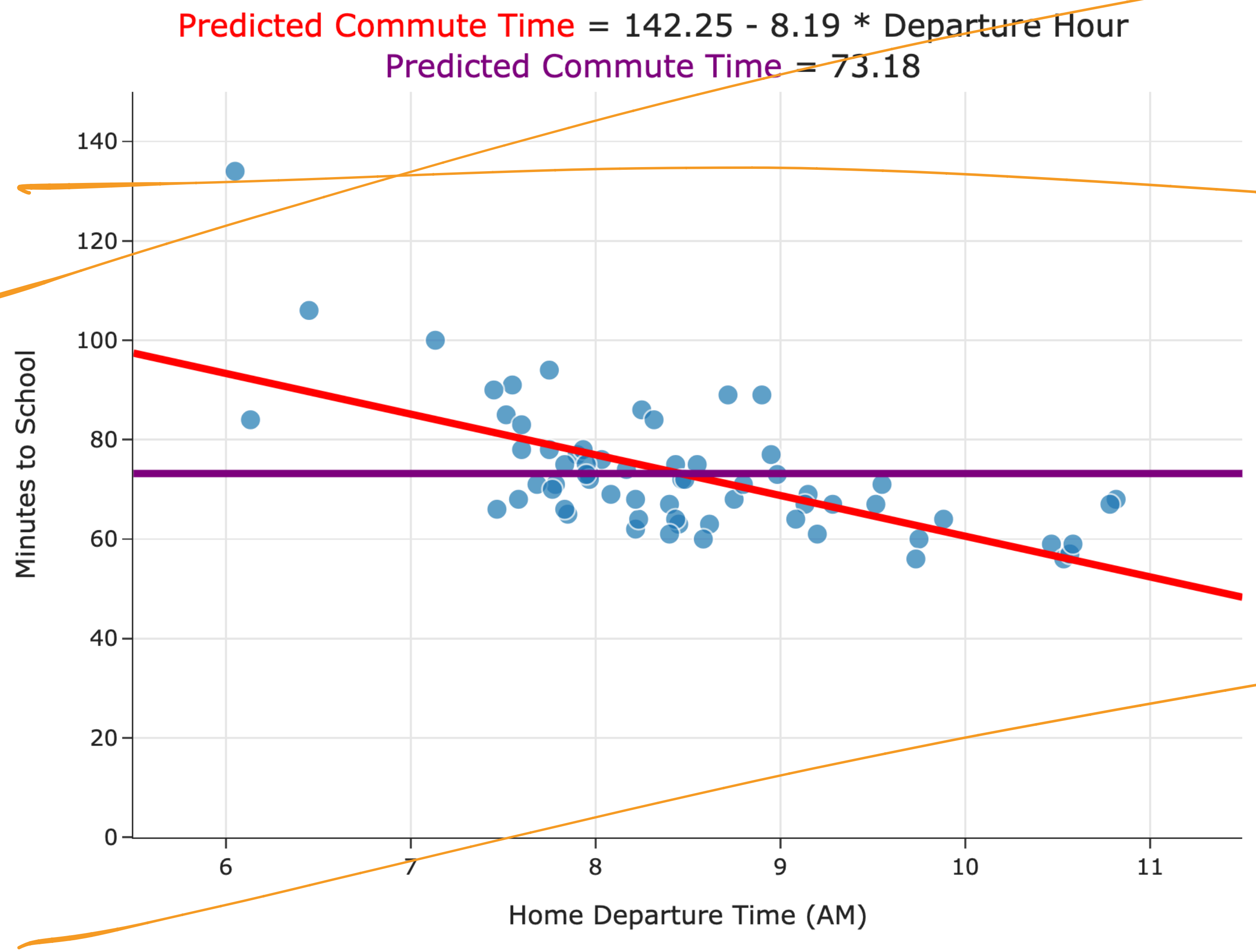
- With both:
    - the constant model, $H(x) = h$, and
    - the simple linear regression model, $H(x) = w_0 + w_1 x$,

  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.

Variance!

- The simple linear regression model is a more flexible version of the constant model.

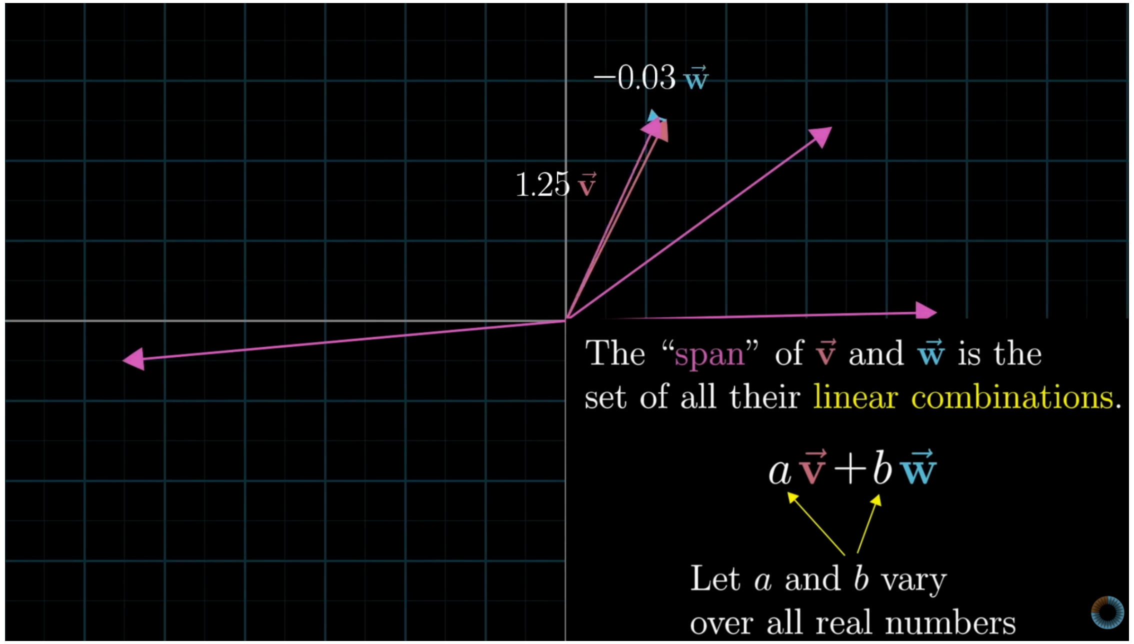
Linear algebra review

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**.
- We'll send some relevant linear algebra review videos on Ed.