

Lectures 5-7

# Simple Linear Regression

DSC 40A, Fall 2025

## Announcements

- Homework 1 is due **Friday night**.
- Look at the office hours schedule [here](#) and plan to start regularly attending!
- Remember to take a look at the supplementary readings linked on the course website.

# Agenda

- Simple linear regression.
- Minimizing mean squared error for the simple linear model.
- Correlation.
- Interpreting the formulas.
- Connections to related models.
- What next? Linear algebra.

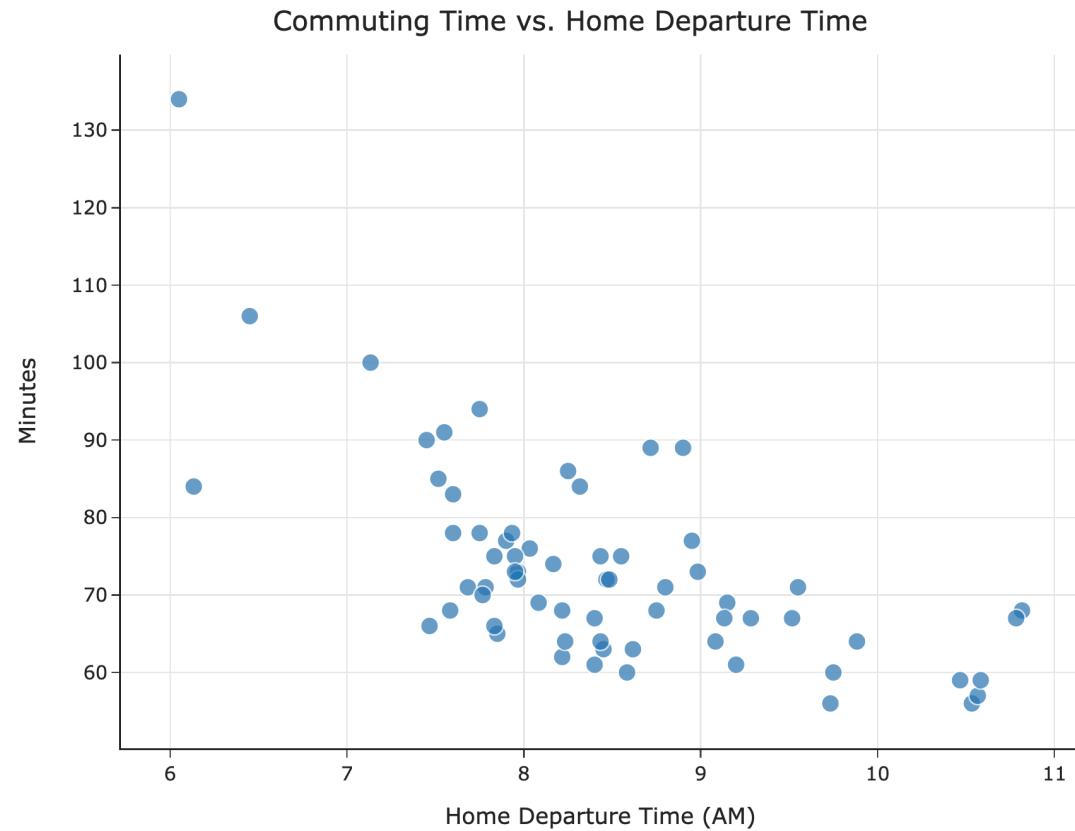
## Question 🤔

Answer at [q.dsc40a.com](https://q.dsc40a.com)

**Remember, you can always ask questions at [q.dsc40a.com!](https://q.dsc40a.com)**

If the direct link doesn't work, click the " Lecture Questions" link in the top right corner of [dsc40a.com](https://dsc40a.com).

# Towards simple linear regression



- In Lecture 1, we introduced the idea of a hypothesis function,  $H(x)$ .
- We've focused on finding the best constant model,  $H(x) = h$ .
- Now that we understand the modeling recipe, we can apply it to find the best simple linear regression model,  $H(x) = w_0 + w_1x$ .
- This will allow us to make predictions that aren't all the same for every data point.

## The modeling recipe

1. Choose a model.

Before : Constant  $f(x) = h$

Now: SLR  $H(\underline{x}) = \underline{w}_0 + \underline{w}_1 \underline{x}$

2. Choose a loss function.

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

actual  $\swarrow$   $\searrow$  prediction

$$L_{abs}(y_i, H(x_i)) = |y_i - H(x_i)|$$

3. Minimize average loss to find optimal model parameters.

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

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

## Finding the best linear model

- Goal: Out of all linear functions  $\mathbb{R} \rightarrow \mathbb{R}$ , find the function  $H^*$  with the smallest mean squared error.
  - Linear functions are of the form  $H(x) = w_0 + w_1 x$ .
  - They are defined by a slope ( $w_1$ ) and intercept ( $w_0$ ).
- That is,  $H^* = w_0^* + w_1^* x$  should be the linear function that minimizes

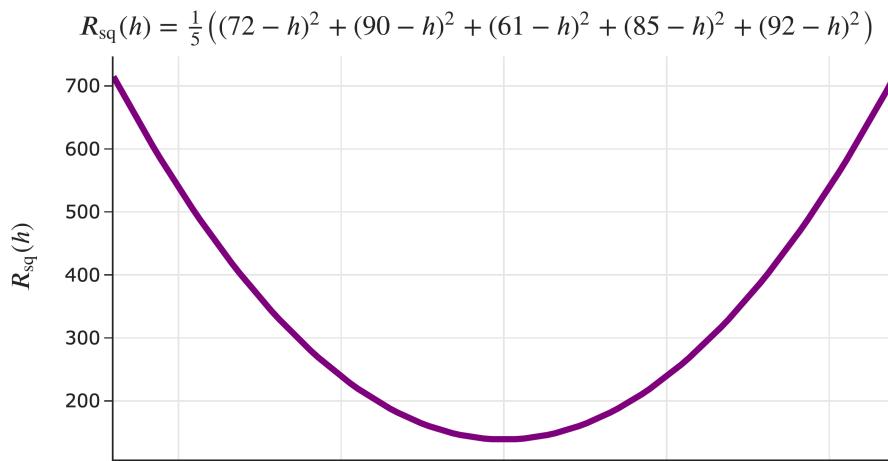
$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^n (y_i - H(x_i))^2$$

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

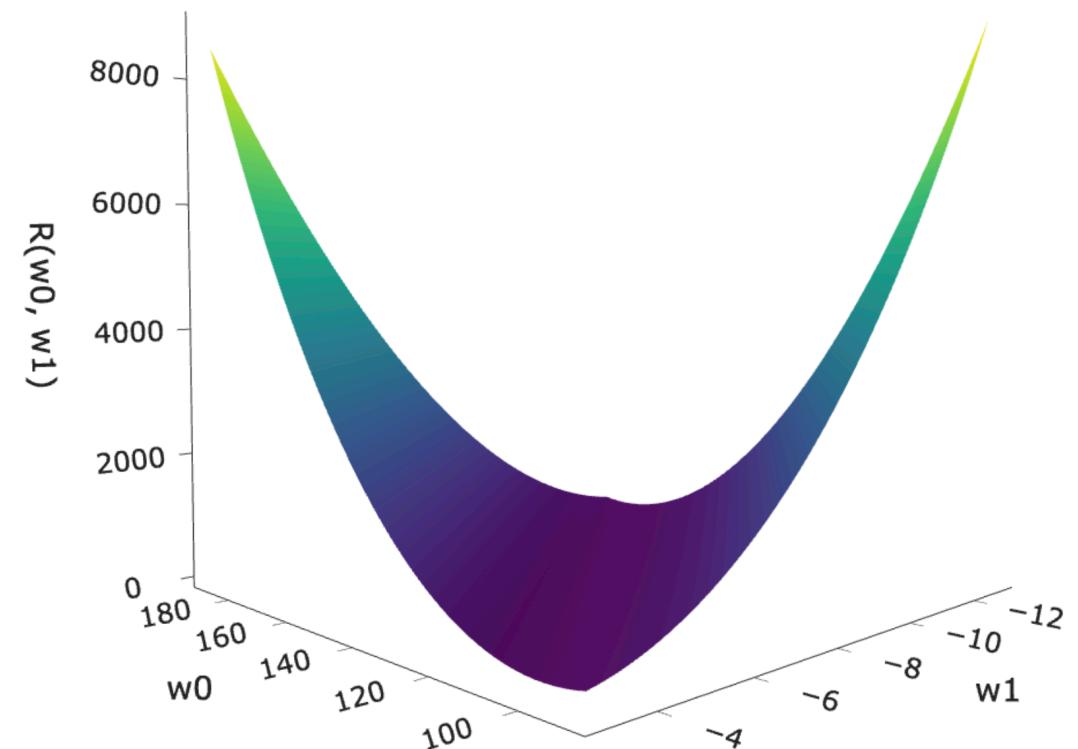
- We chose squared loss, since it's the easiest to minimize.
- How do we find the parameters  $w_0^*$  and  $w_1^*$  that minimize  $R_{\text{sq}}(w_0, w_1)$ ?

## Loss surface

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



What does the graph of  $R_{\text{sq}}(w_0, w_1)$  look like for the simple linear regression model?



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:

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

- $R_{\text{sq}}$  is a function of two variables:  $w_0$  and  $w_1$ .
- To minimize a function of multiple variables:
  - Take partial derivatives with respect to each variable.
  - Set all partial derivatives to 0.
  - 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).

## Example

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

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

# Minimizing multivariate functions

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

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

- $R_{\text{sq}}$  is a function of two variables:  $w_0$  and  $w_1$ .
- To minimize a function of multiple variables:
  - Take partial derivatives with respect to each variable.
  - Set all partial derivatives to 0.
  - 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).

## Minimizing mean squared error

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

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.

## Question 🤔

Answer at [q.dsc40a.com](http://q.dsc40a.com)

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

Which of the following is equal to  $\frac{\partial R_{\text{sq}}}{\partial w_0}$ ?

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

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

$$\frac{\partial R_{\text{sq}}}{\partial w_0} =$$

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

$$\frac{\partial R_{\text{sq}}}{\partial w_1} =$$

## Strategy

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

$$-\frac{2}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i)) = 0 \quad -\frac{2}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i)) x_i = 0$$

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  $w_0^*$**

$$-\frac{2}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i)) = 0$$

## Solving for $w_1^*$

$$-\frac{2}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i)) x_i = 0$$

## Least squares solutions

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

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

where:

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

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

## An equivalent formula for $w_1^*$

Claim:

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

Proof:

## Least squares solutions

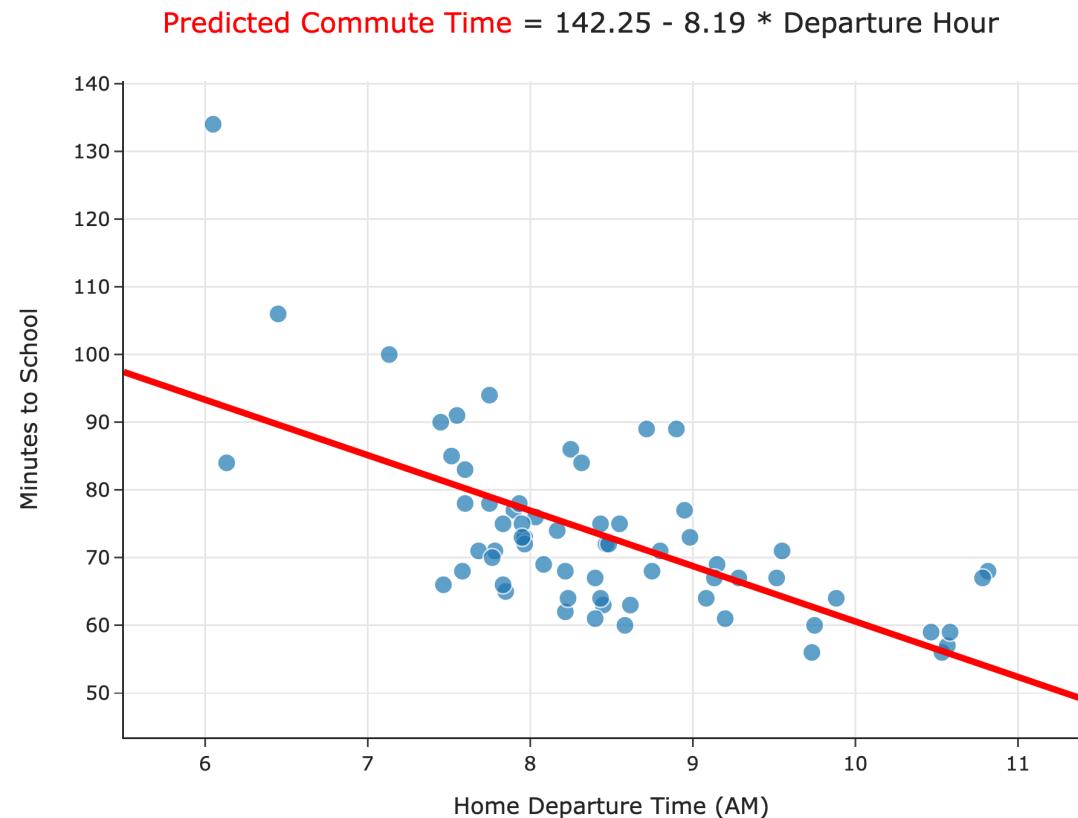
- The **least squares solutions** for the intercept  $w_0$  and slope  $w_1$  are:

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

- We say  $w_0^*$  and  $w_1^*$  are **optimal parameters**, and the resulting line is called the **regression line**.
- The process of minimizing empirical risk to find optimal parameters is also called "fitting to the data."
- To make predictions about the future, we use  $H^*(x) = w_0^* + w_1^* x$ .

# Causality

Solving for best linear model for commute



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

## What's next?

We now know how to find the optimal slope and intercept for linear hypothesis functions. Next, 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.
- Discuss *causality*.
- Learn how to build regression models with **multiple inputs**.
  - To do this, we'll need linear algebra!

# Agenda

- Simple linear regression.
- Correlation.
- Interpreting the formulas.
- Connections to related models.

## Least squares solutions

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

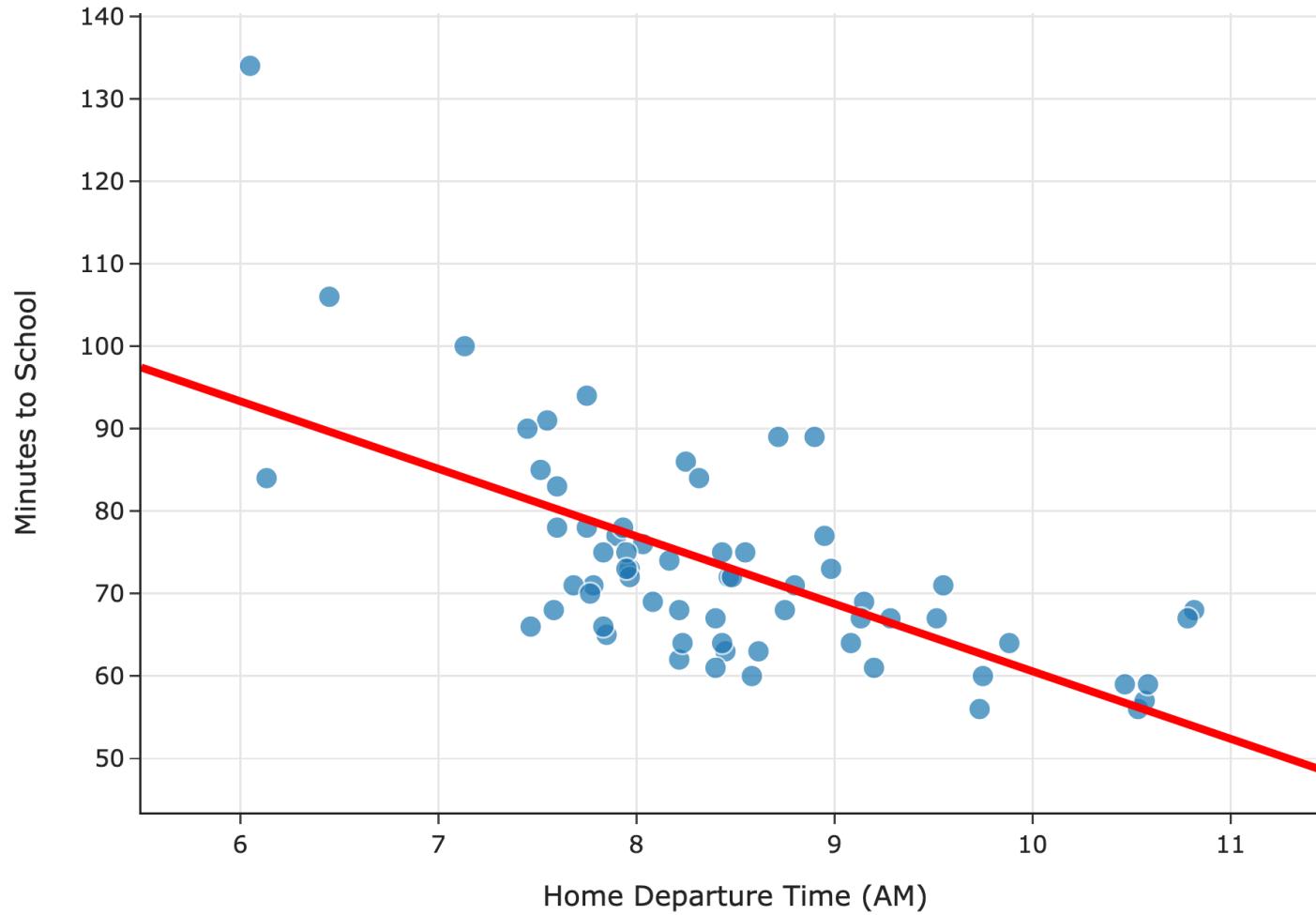
$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$$

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

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

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

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



## 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 🤔

Answer at [q.dsc40a.com](https://q.dsc40a.com)

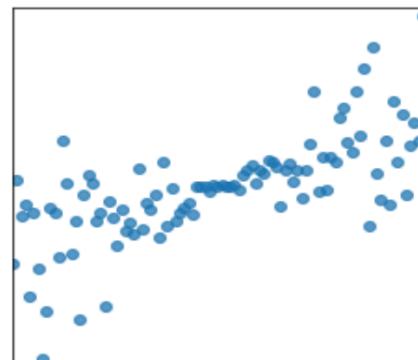
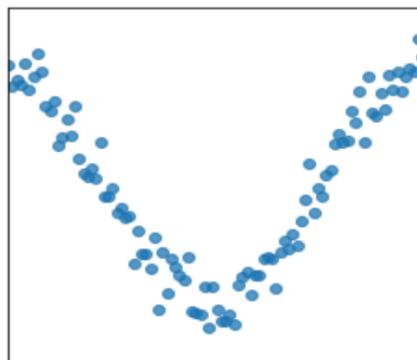
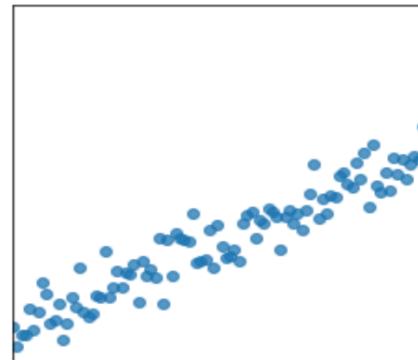
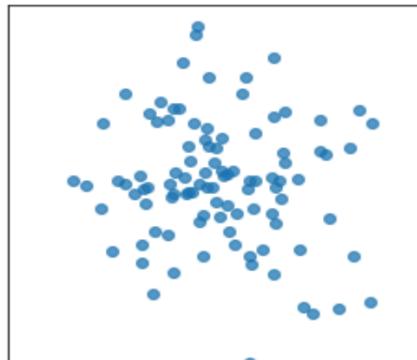
Consider a dataset with just two points,  $(2, 5)$  and  $(4, 15)$ . Suppose we want to fit a linear hypothesis function to this dataset using squared loss. What are the values of  $w_0^*$  and  $w_1^*$  that minimize empirical risk?

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

# Correlation

# 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.



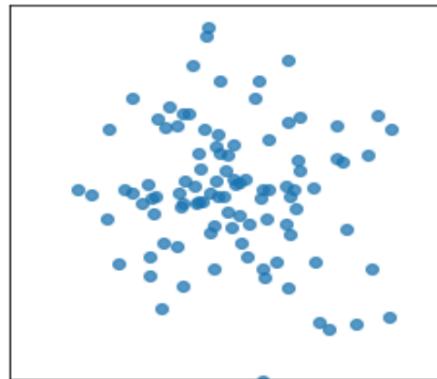
## The correlation coefficient

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

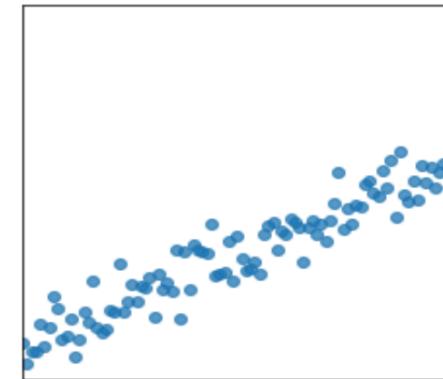
$$r = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma_x} \right) \left( \frac{y_i - \bar{y}}{\sigma_y} \right)$$

# The correlation coefficient, visualized

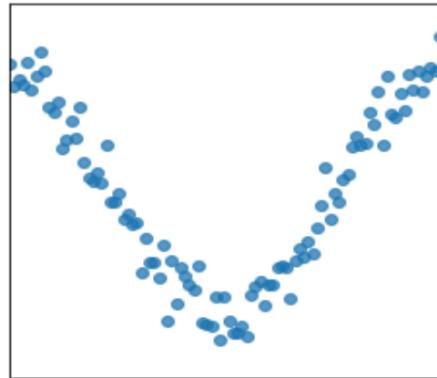
$r = -0.121$



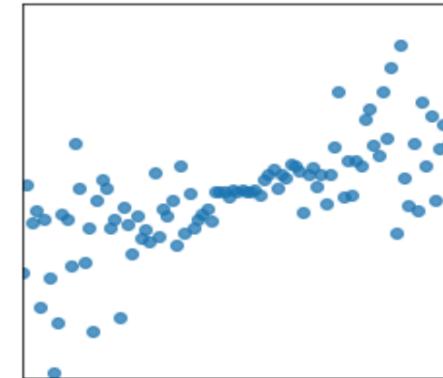
$r = 0.949$



$r = 0.052$



$r = 0.704$



## 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  $r$ !

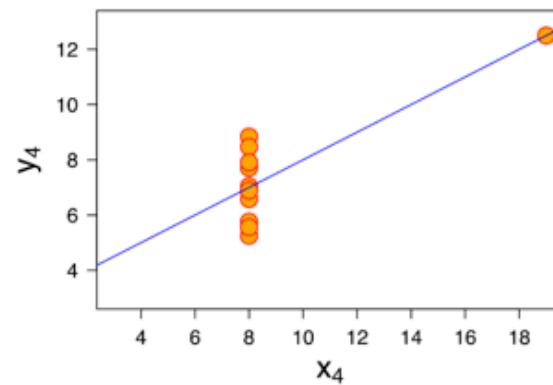
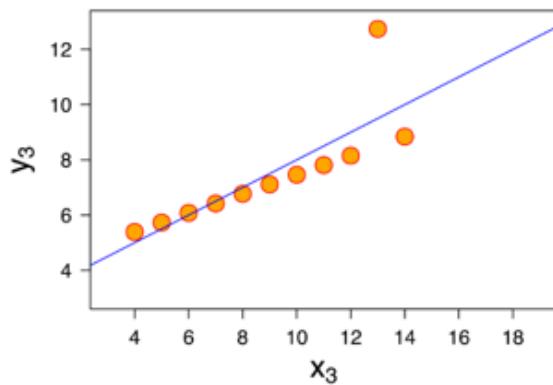
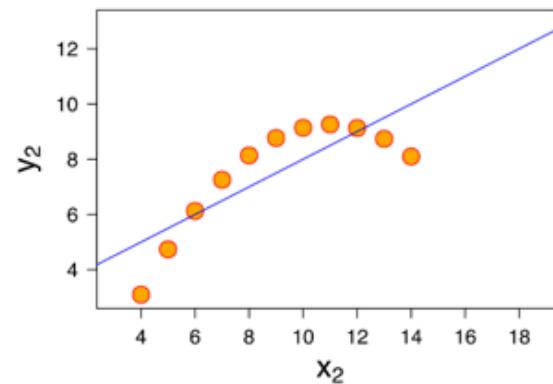
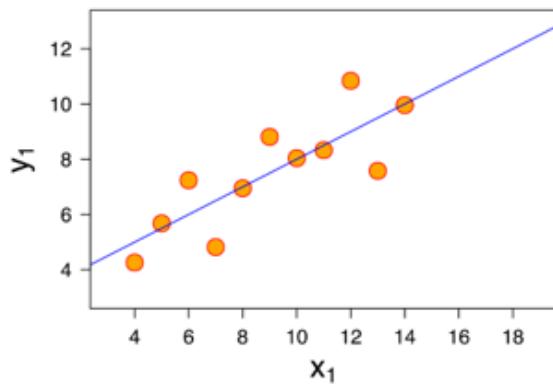
$$w_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \frac{\sigma_y}{\sigma_x}$$

- 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  $w_1^*$ :

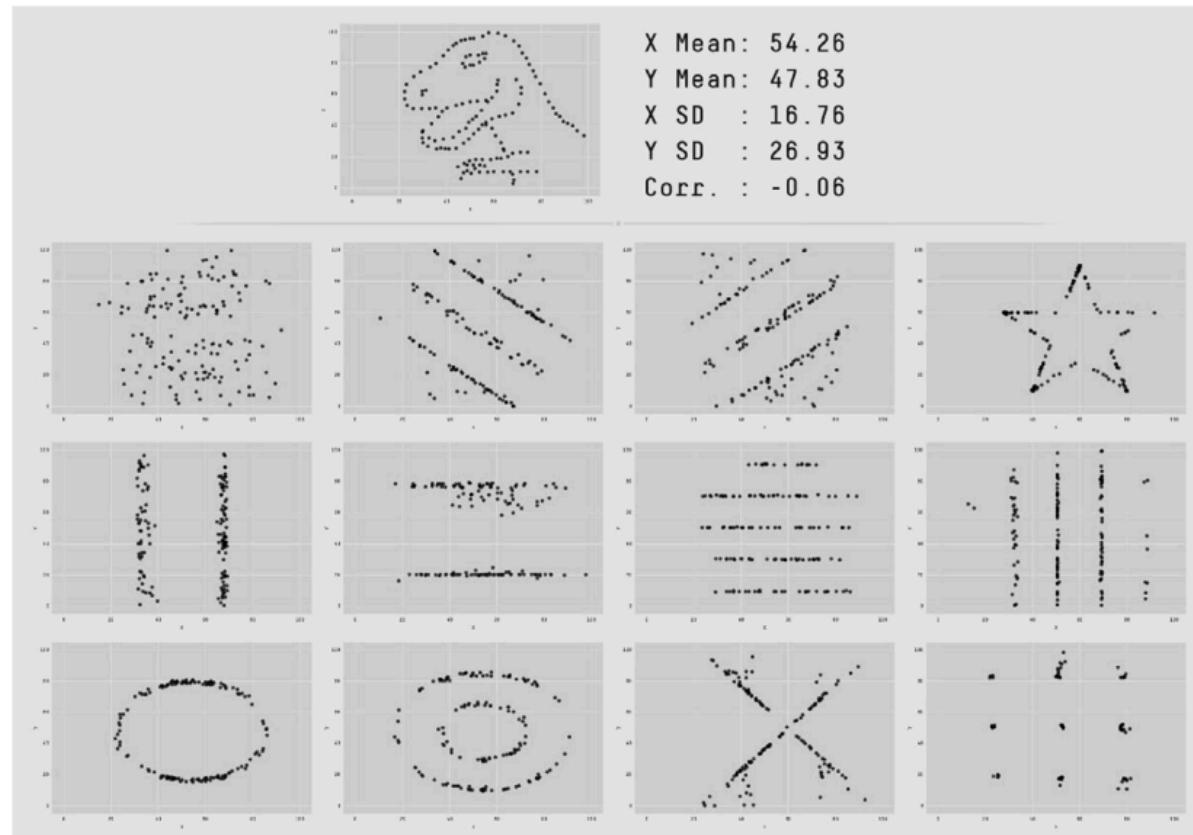
$$w_1^* = r \frac{\sigma_y}{\sigma_x} \quad w_0^* = \bar{y} - w_1^* \bar{x}$$

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

# Dangers of correlation



# Dangers of correlation



# Interpreting the formulas

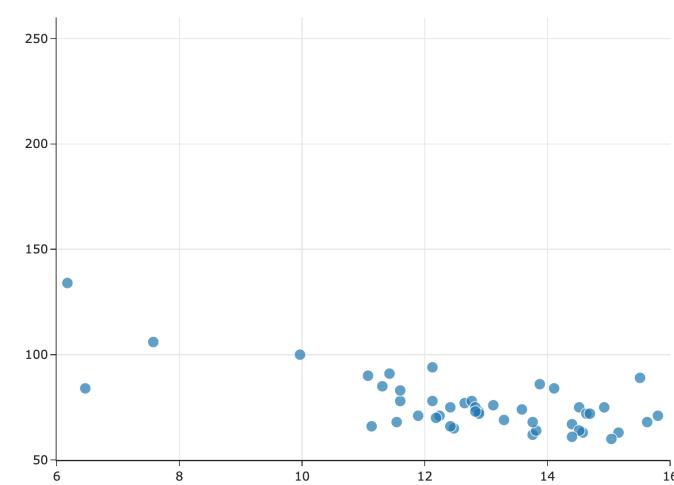
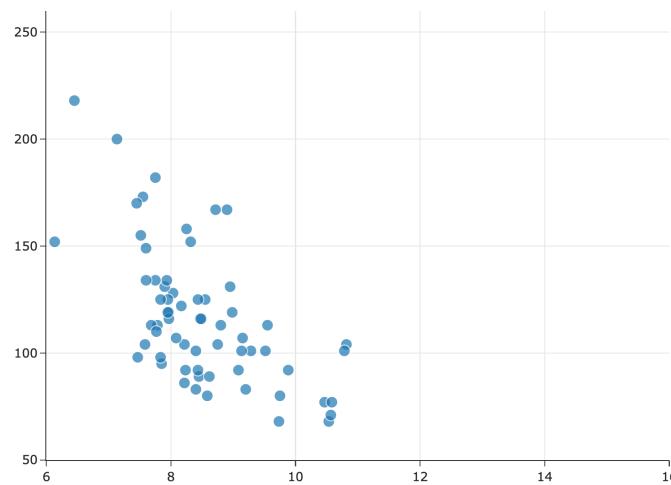
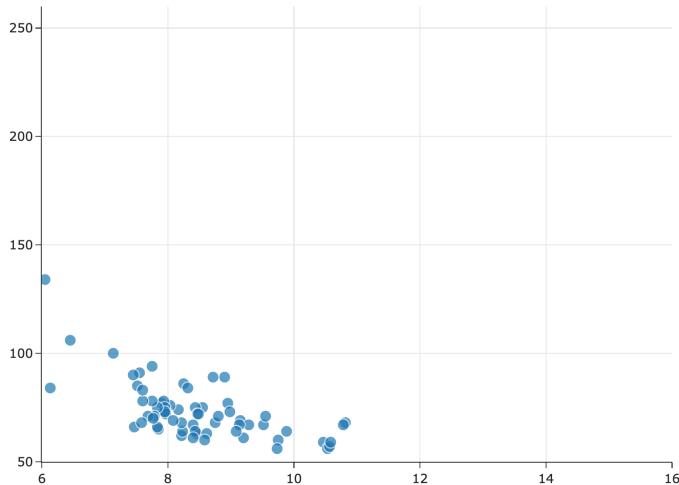
## Interpreting the slope

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

- 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**.

# Interpreting the slope

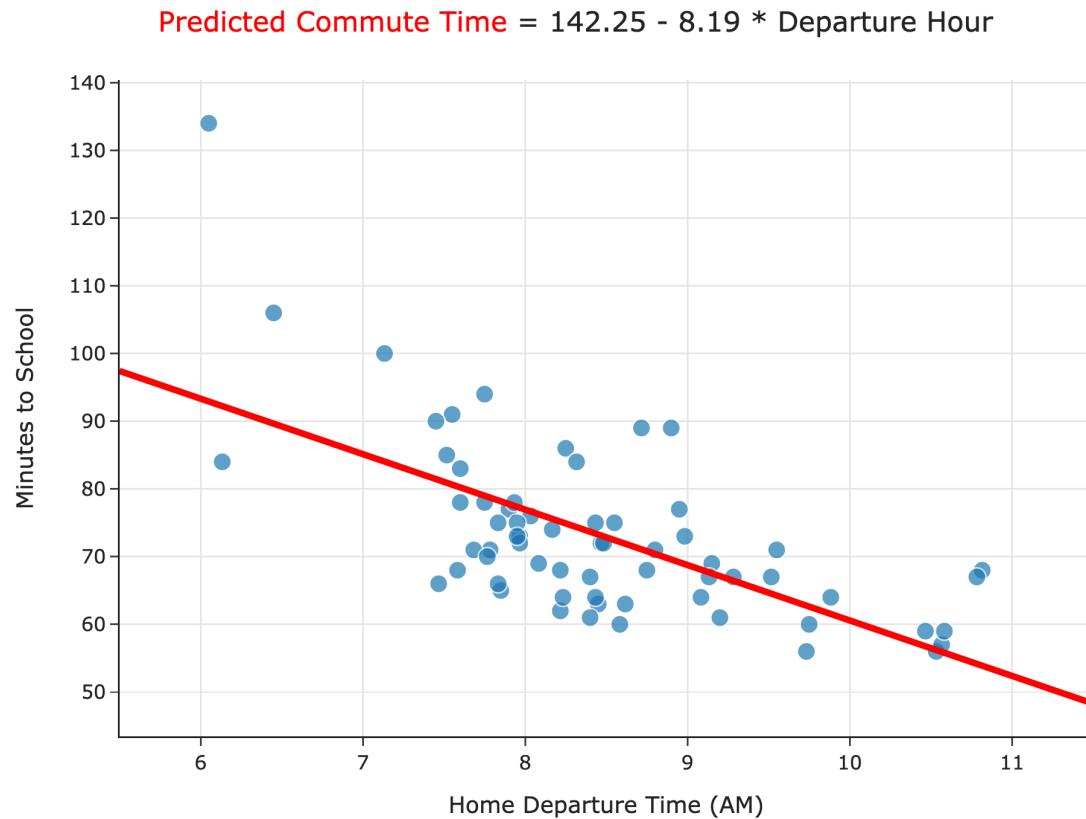
$$w_1^* = r \frac{\sigma_y}{\sigma_x}$$



- 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,  $\sigma_y$  increases, so the slope gets steeper.
- As the  $x$  values get more spread out,  $\sigma_x$  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 [q.dsc40a.com](https://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

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

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

- That is, the **mean squared error** of the regression line's predictions and the correlation coefficient,  $\mathbf{r}$ , 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  $\mathbf{r}^2$  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  $\mathbf{r}^2$** .

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



# Connections to related models

## Question 🤔

Answer at [q.dsc40a.com](http://q.dsc40a.com)

Suppose we chose the model  $H(x) = w_1x$  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.  $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$
- D.  $\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$

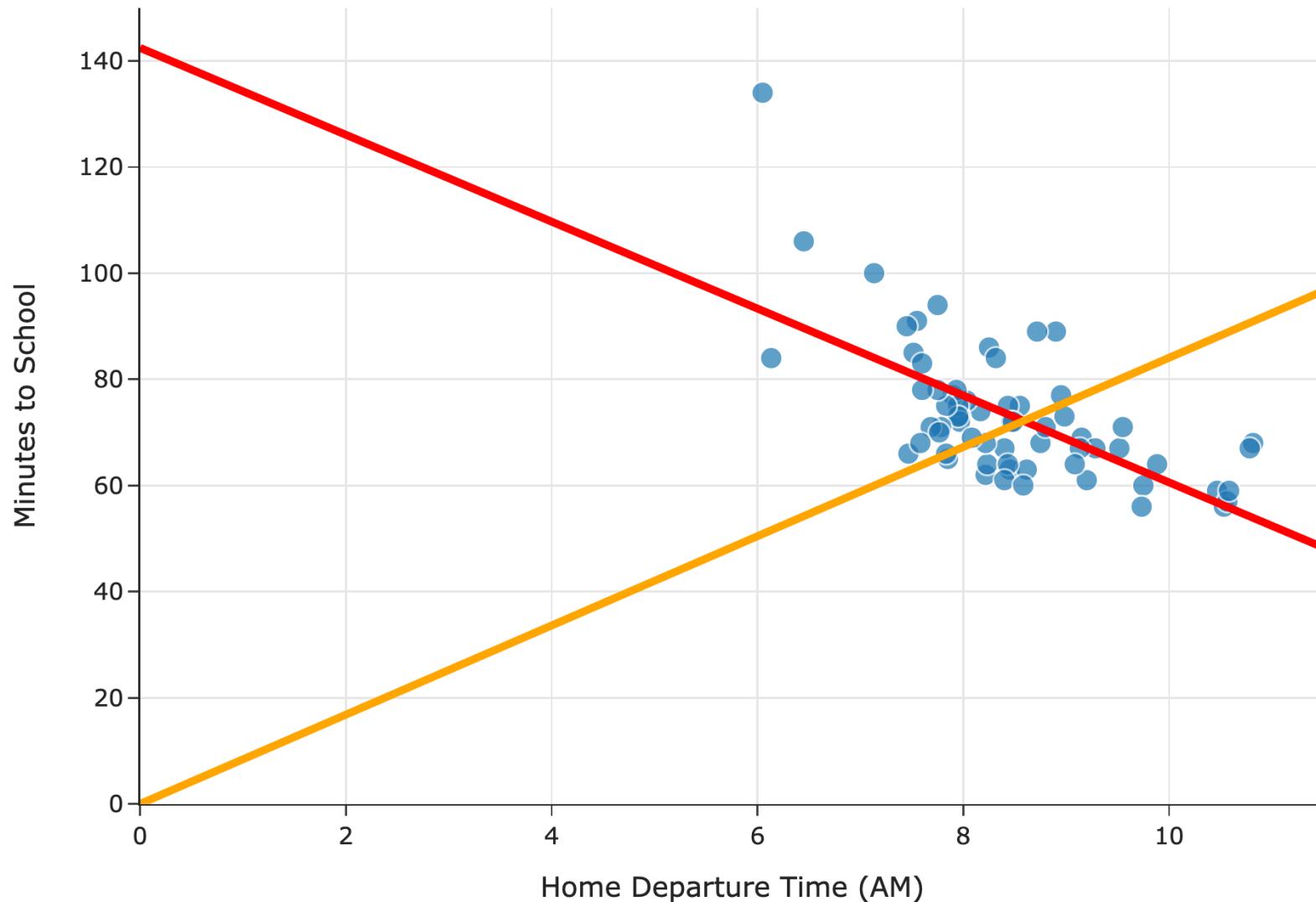
## Exercise

Suppose we chose the model  $H(x) = w_1x$  and squared loss.

What is the optimal model parameter,  $w_1^*$ ?

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

Predicted Commute Time =  $8.41 * \text{Departure Hour}$



## Exercise

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

What is the optimal model parameter,  $w_0^*$ ?

## Comparing mean squared errors

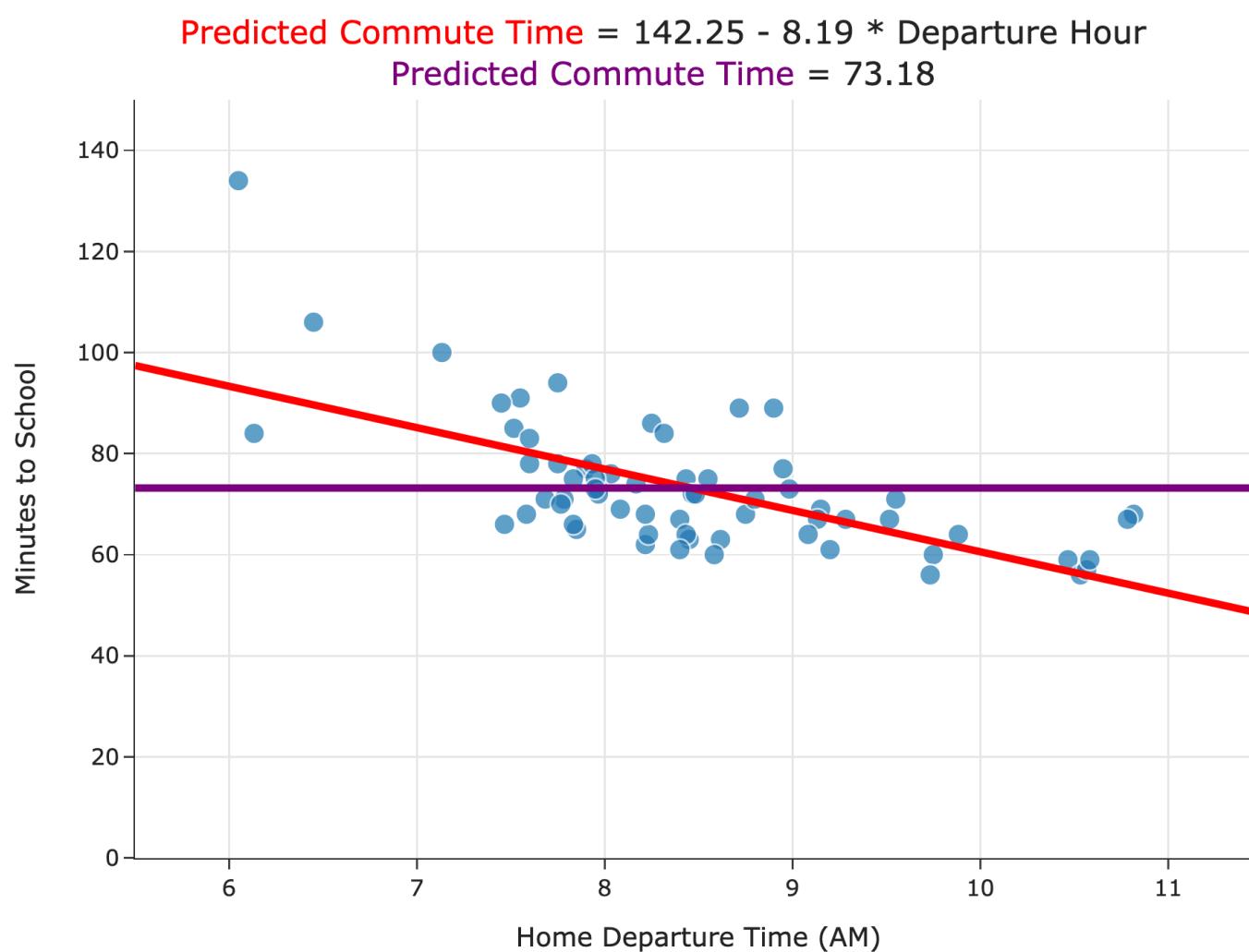
- With both:
  - the constant model,  $H(x) = h$ , and
  - the simple linear regression model,  $H(x) = w_0 + w_1x$ ,

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

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

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

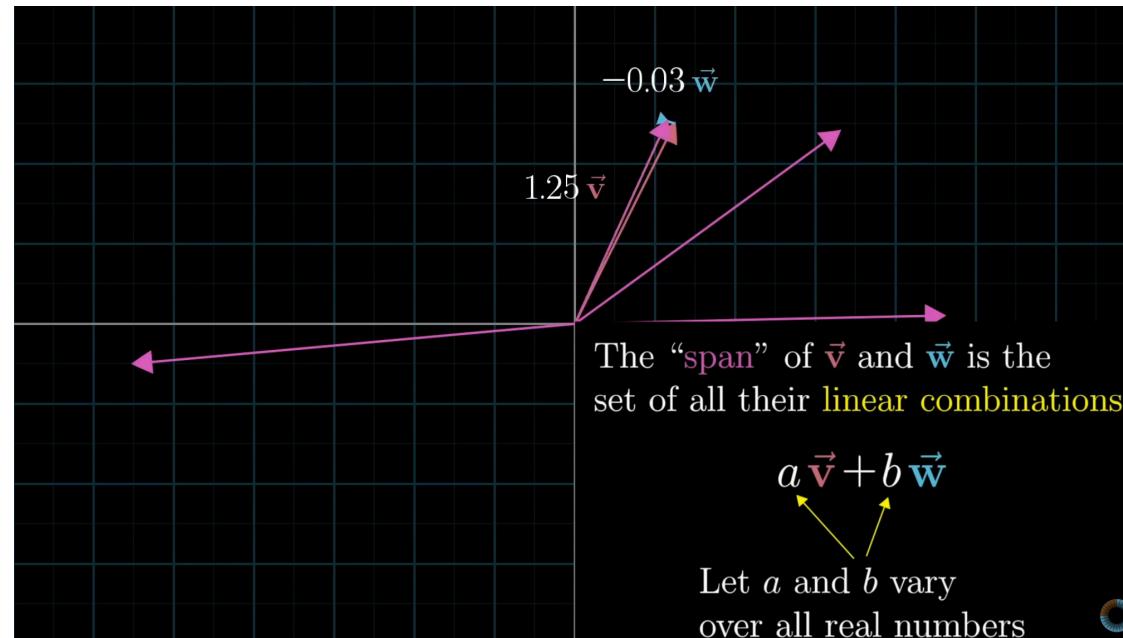
# Linear algebra review

## 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.  $H(x) = w_0 + w_1x + w_2x^2$ .
- 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.