Lecture 4

Simple Linear Regression

DSC 40A, Spring 2024

Agenda

- Recap: Center and spread.
- Simple linear regression.
- Minimizing mean squared error for the simple linear model.

Recap: Center and spread

The relationship between h^* and $R(h^*)$

- Recall, for a general loss function L and the constant model $H(x) = h$, empirical risk is of the form:

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

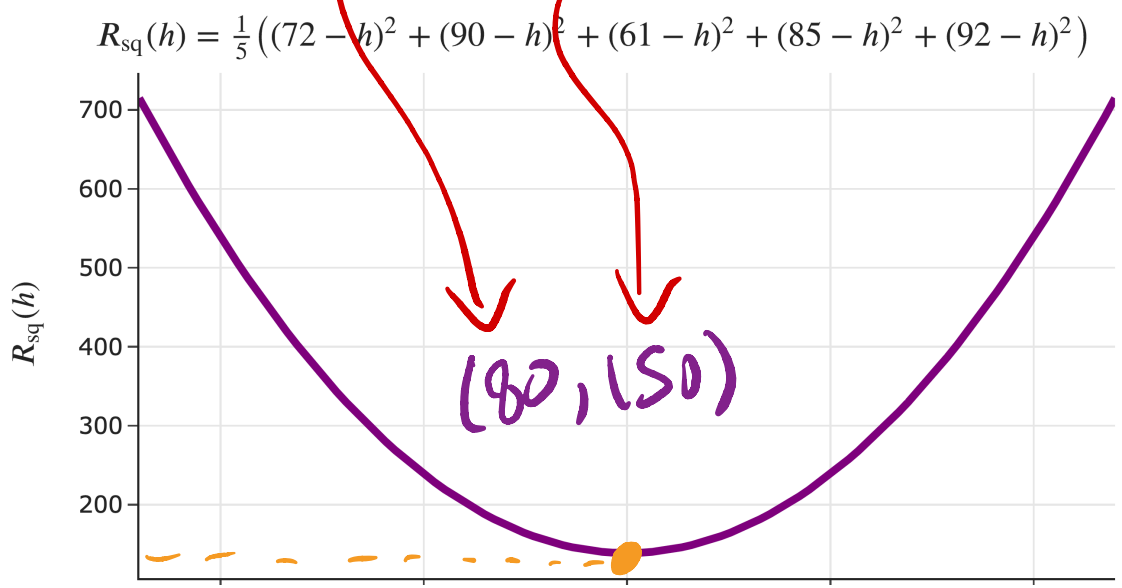
- h^*, the value of h that minimizes empirical risk, represents the **center** of the dataset in some way.
- $R(h^*)$, the smallest possible value of empirical risk, represents the **spread** of the dataset in some way.
- The specific center and spread depend on the choice of loss function.

Examples

mean Absolute Deviation: "how far from the median"

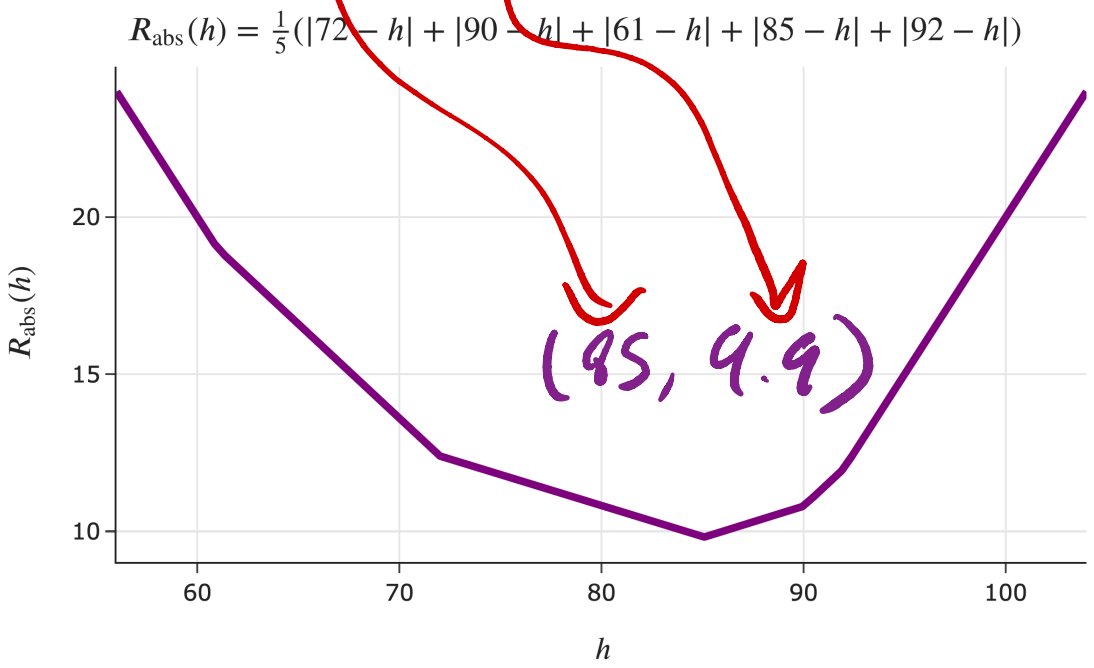
When using **squared loss**:

- $h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$.
- $R_{\text{sq}}(h^*) = \text{Variance}(y_1, y_2, \ldots, y_n)$.



When using **absolute loss**:

- $h^* = \text{Median}(y_1, y_2, \ldots, y_n)$.
- $R_{\text{abs}}(h^*) = \text{MAD from the median}$.



0-1 loss

- The empirical risk for the 0-1 loss is:

$$R_{0,1}(h) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0 & y_i = h\\ 1 & y_i \neq h\end{cases}$$

- This is the proportion (between 0 and 1) of data points not equal to h.
- $R_{0,1}(h)$ is minimized when $h^* = \text{Mode}(y_1, y_2, \ldots, y_n)$. (center)
- Therefore, $R_{0,1}(h^*)$ is the proportion of data points not equal to the mode. (spread)
- **Example**: What's the proportion of values not equal to the mode in the dataset $2, 3, 3, 4, 5$?

5 points (y_i s)
2 (3 s)
3!(3 s)
$= \frac{3}{5}$ measure of our spread

A poor way to measure spread

- The minimum value of $R_{0,1}(h)$ is the proportion of data points not equal to the mode.
- A higher value means less of the data is clustered at the mode.
- Just as the mode is a very basic way of measuring the center of the data, $R_{0,1}(h^*)$ is a very basic and uninformative way of measuring spread.

Summary of center and spread

- Different loss functions $L(y_i, h)$ lead to different empirical risk functions $R(h)$, which are minimized at various measures of **center**.
- The minimum values of empirical risk, $R(h^*)$, are various measures of **spread**.
- There are many different ways to measure both center and spread; these are sometimes called **descriptive statistics**.

Simple linear regression

What's next?

- 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**, $H(x) = w_0 + w_1 x$.
- This will allow us to make predictions that aren't all the same for every data point.

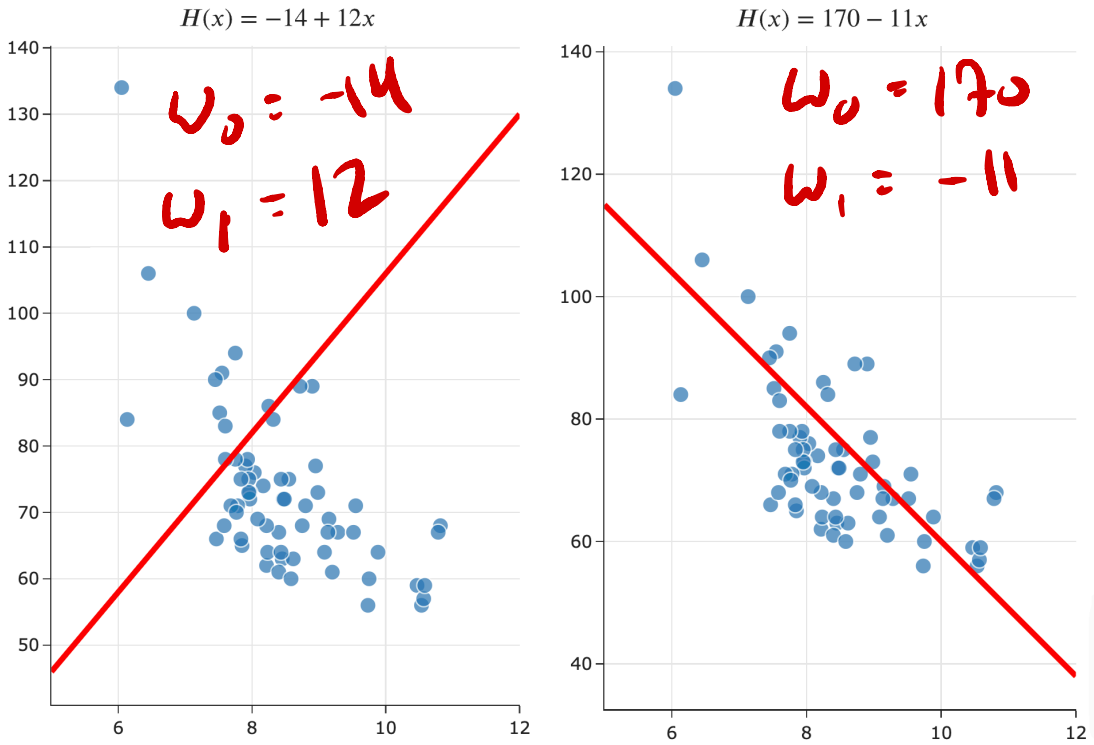
H(hour dep. time) → time(min.)

Recap: Hypothesis functions and parameters

A hypothesis function, H, takes in an x as input and returns a predicted y.

Parameters define the relationship between the input and output of a hypothesis function.

The simple linear regression model, $H(x) = w_0 + w_1 x$, has two parameters: w_0 and w_1.



↳ intercept → slope

w_0 "naught" + w_1 sub one

w_0^* w_1^*

$H(10) = 170 - 11(10)$
AM
$170 - 110 = 60$ min

The modeling recipe

1. Choose a model.

$$H(x) = h \implies H(x) = w_0 + w_1 x$$

2. Choose a loss function.

$$L_{sq}(y_i, H(x_i)) = (y_i - H(x_i))^2 \qquad L_{abs}(y_i, H(x_i)) = |y_i - H(x_i)|$$

(y_i: actual; $H(x_i)$: predicted)

3. Minimize average loss to find optimal model parameters.

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

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 $H^*(x)$ 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 $H(x) = w_0 + w_1 x$, we can re-write R_{sq} as a function of w_0 and w_1:

Intercept Slope

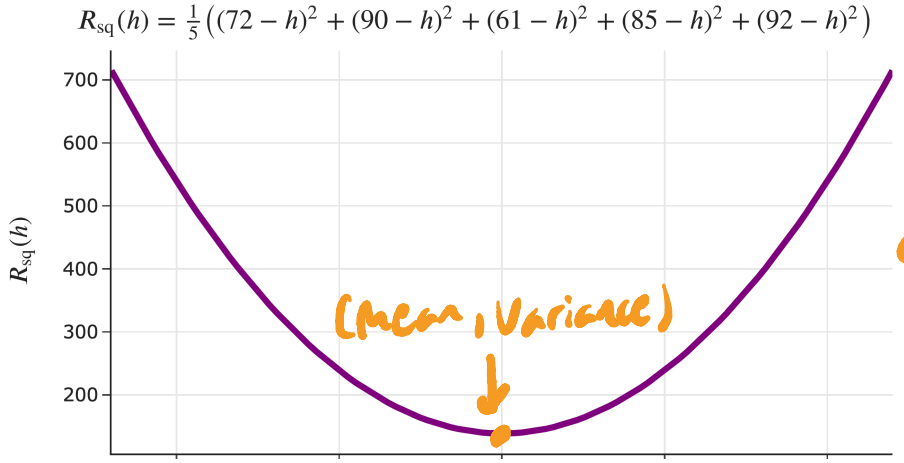
$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2$$

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

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

Loss surface

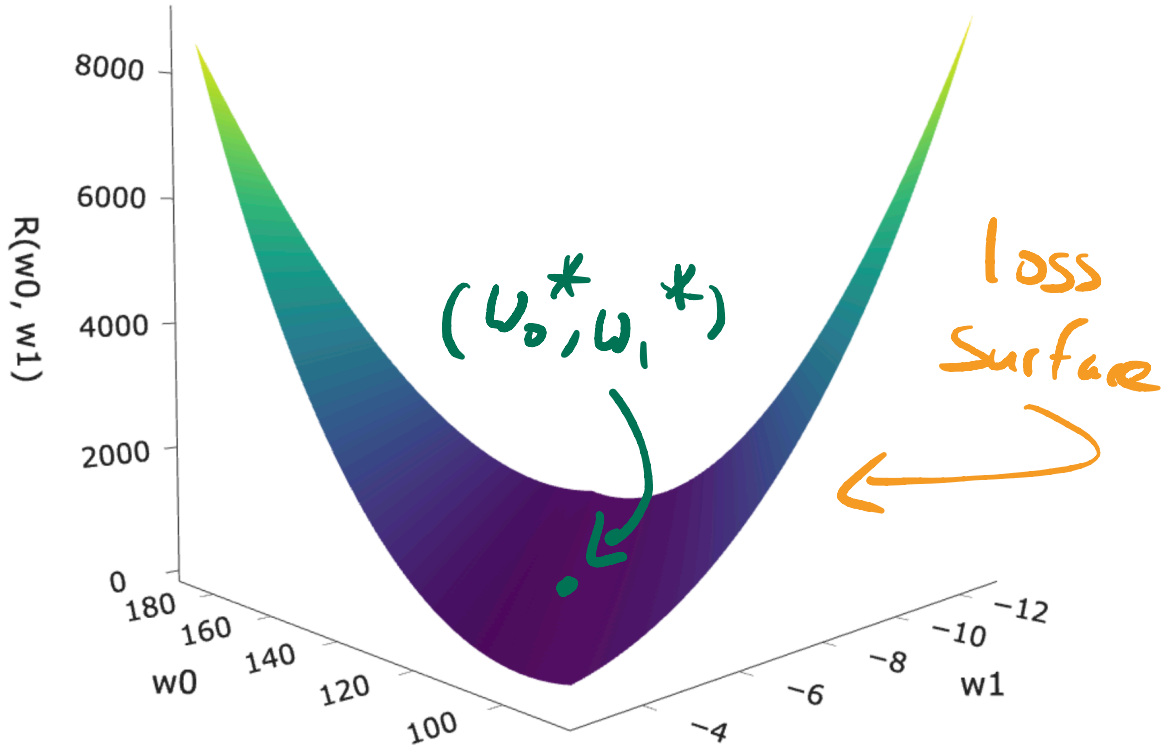
For the constant model, the graph of $R_{\text{sq}}(h)$ looked like a parabola.



Loss curve

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

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



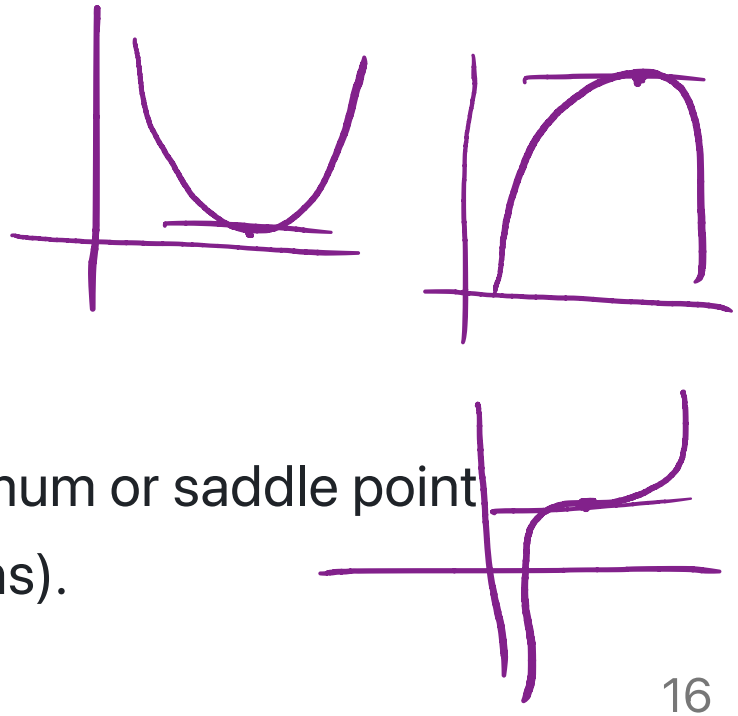
Minimizing mean squared error for the simple linear model

Minimizing multivariate functions

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

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

- R_{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.

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

$f_x = \frac{\partial f}{\partial x} = 2x - 8 \rightarrow 2x - 8 = 0 \rightarrow \boxed{x = 4}$

$f_y = \frac{\partial f}{\partial y} = 2y + 6 \rightarrow 2y + 6 = 0 \rightarrow \boxed{y = -3}$

} minimized at $x = 4$, $y = -3$

$f(x, y) = (x - 4)^2 - 16 + (y + 3)^2 - 9 - 7$

$= (x - 4)^2 + (y + 3)^2 - 32$

$(4 - 4)^2 + (-3 + 3)^2 - 32$

} minimized @ $x = 4$, $y = -3$ $\quad (4, -3, -32)$

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.

Question 🤔

Take a moment to pause and reflect...

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

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

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

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

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

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

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

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

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

Part of the Summation

Strategy

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

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

Partial Der. with respect to w_0

Par. Der. with respect to w_1

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^*

Isolate this!!

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

$-\frac{n}{2}$ $-\frac{n}{2}$

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

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

$$\rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} w_0 - \sum_{i=1}^{n} w_1 x_i = 0$$

$$\sum_{i=1}^{n} w_0 = w_0 + w_0 + \cdots + w_0 = n w_0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n w_0 - \sum_{i=1}^{n} w_1 x_i = 0$$

$$\rightarrow \sum_{i=1}^{n} y_i - n w_0 - w_1 \sum_{i=1}^{n} x_i = 0$$

$$\rightarrow \sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i = n w_0$$

$$\rightarrow w_0 = \frac{\sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i}{n}$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\bar{y}} - w_1 \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\bar{x}}$$

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

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

Solving for w_1^*

Isolate

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

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

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

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

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

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

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

Least squares solutions

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

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

where:

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

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

An equivalent formula for w_1^*

Key Idea: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Claim:

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

$\rightarrow (x_1 - \bar{x})(x_1 - \bar{x})$

Proof:

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

Distribute

$$= \sum_{i=1}^{n}(y_i - \bar{y})x_i - \bar{x}\sum_{i=1}^{n}(y_i - \bar{y}) \quad \left(\sum_{i=1}^{n}(y_i - \bar{y}) = 0\right)$$

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

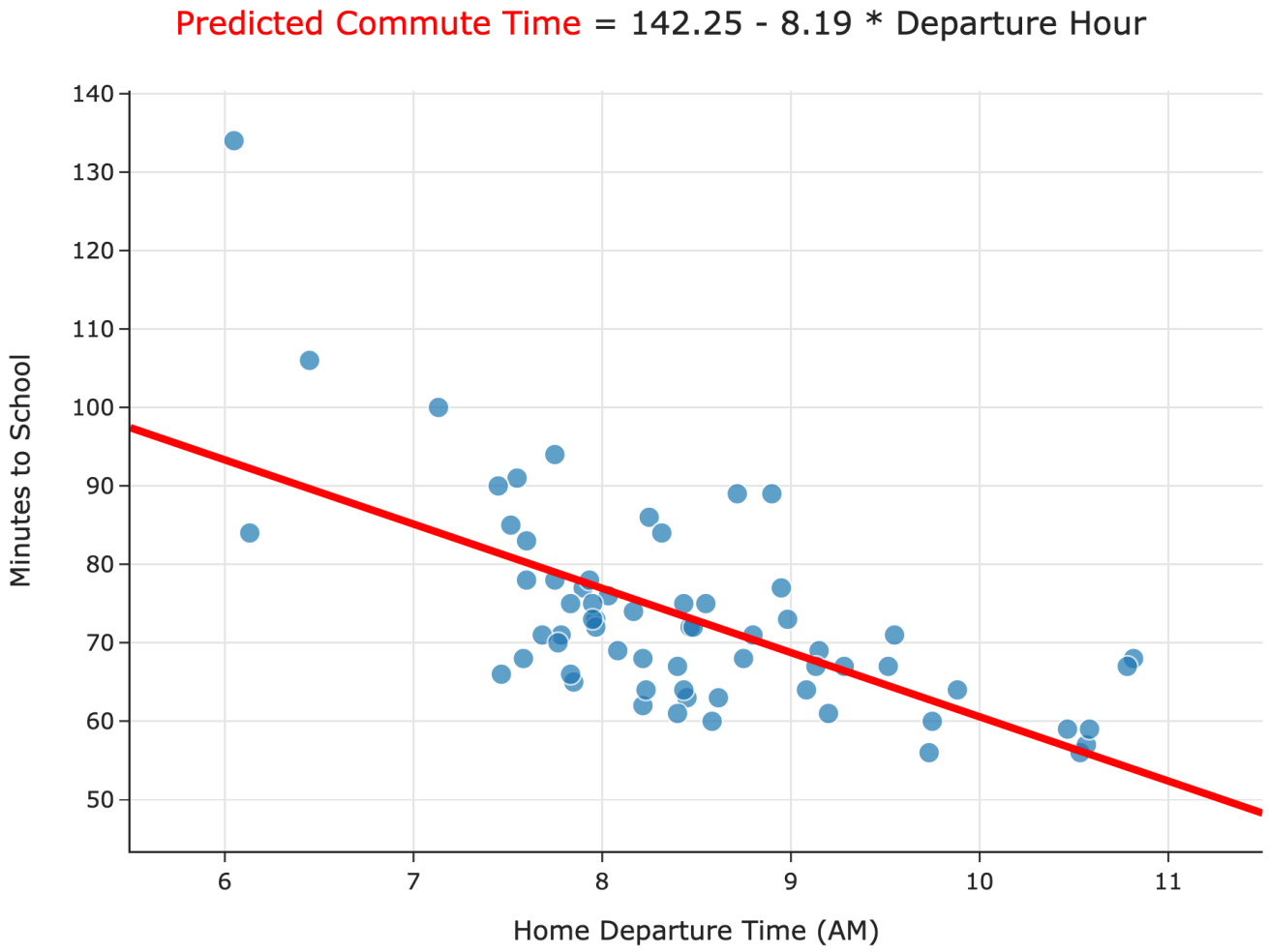
Least squares solutions

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

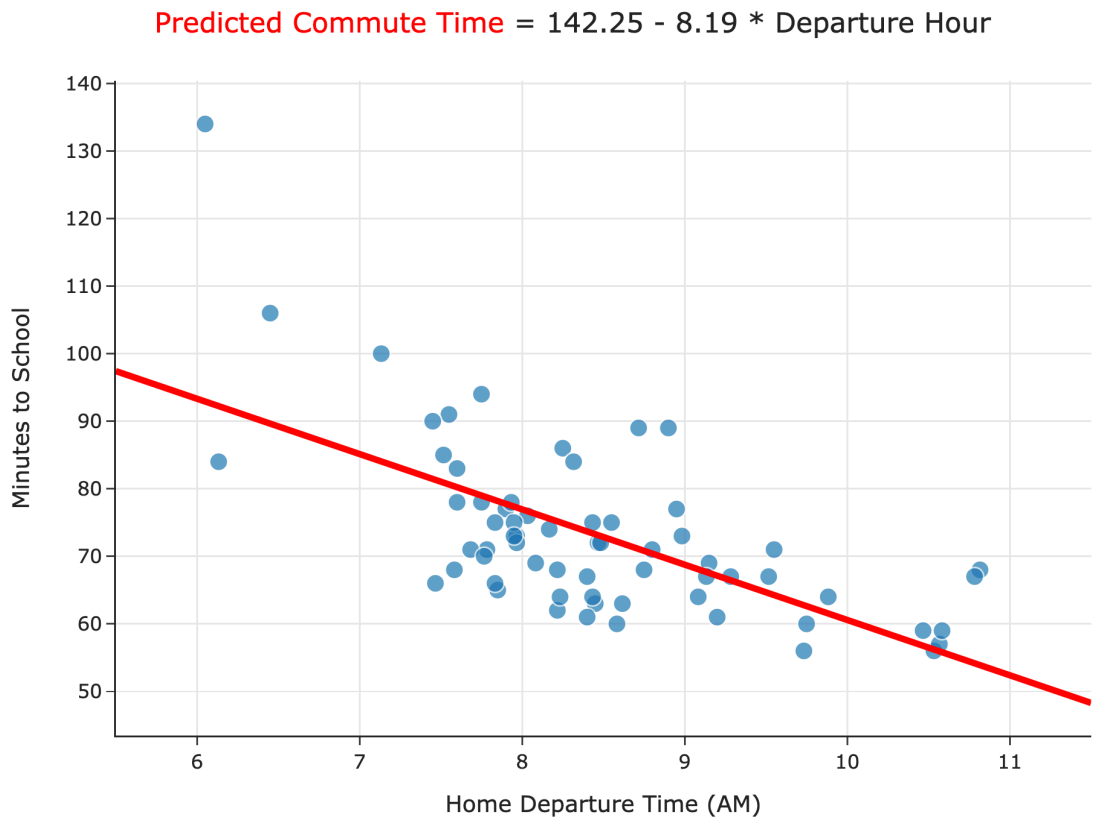
$$\boxed{w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad w_0^* = \bar{y} - w_1^* \bar{x}}$$

- We say w_0^* and w_1^* are **optimal parameters**, and the resulting line is called the **regression line**.
- 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 $\boxed{H^*(x) = w_0^* + w_1^* x}$.

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



Causality



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

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!