Lecture 4

Simple Linear Regression

DSC 40A, Spring 2024

Announcements

- Homework 1 is due **tonight**.
  - Before working on it, watch the Walkthrough Videos on problem solving and using Overleaf.
  - Using the Overleaf template is required for Homework 2 (and only Homework 2).
- 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

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

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: Center and spread

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

- Recall, for a general loss function L and the constant model $H(x) = h$, $\underbrace{\text{empirical risk}}_{\text{"average loss"}}$ is of the form:

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

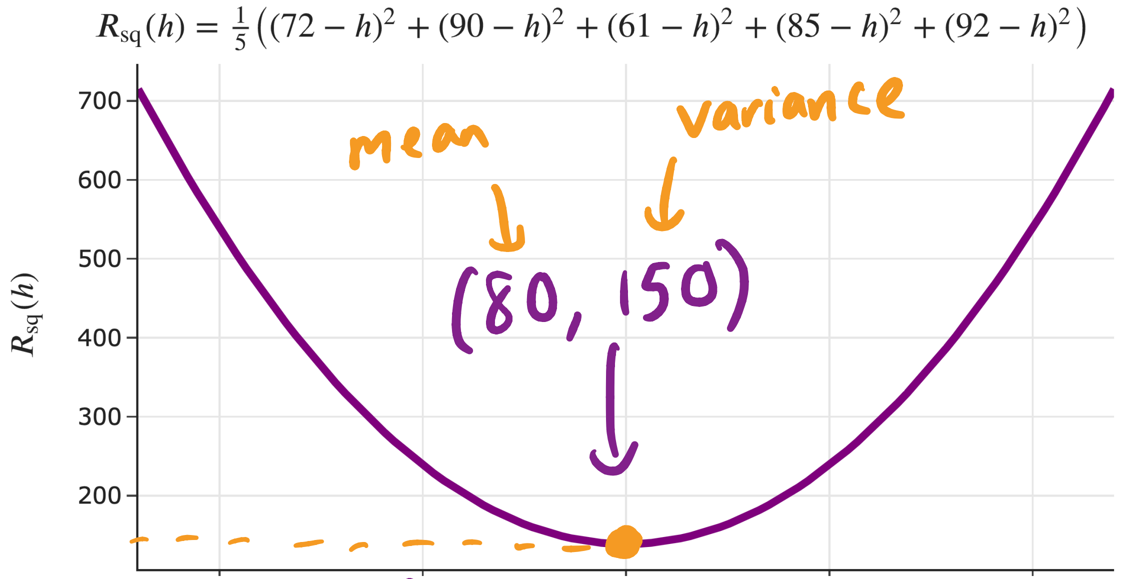
- h^*, the value of h that minimizes empirical risk, represents the **center** of the dataset in some way.
- $R(h^*)$, the smallest possible value of empirical risk, represents the **spread** of the dataset in some way.
- The specific center and spread depend on the choice of loss function.

"mean absolute deviation"
$\Rightarrow$ how far from the median

Examples

When using **squared loss**:

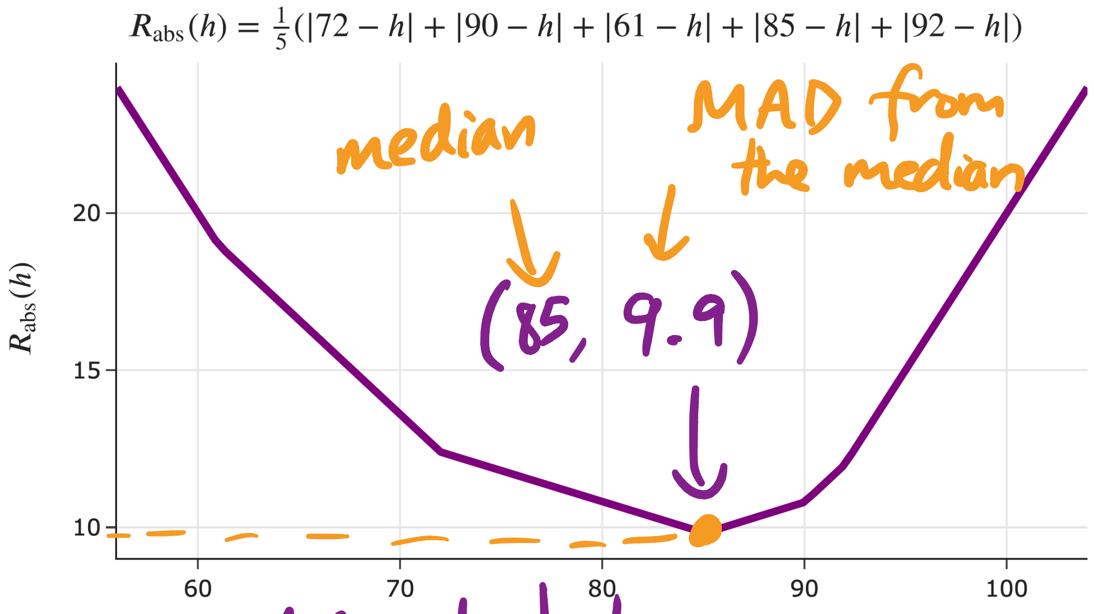
- $h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$.
- $R_{\text{sq}}(h^*) = \text{Variance}(y_1, y_2, \ldots, y_n)$.



Mean squared error, average squared loss, empirical risk (for squared loss) $\Rightarrow$ the same!

When using **absolute loss**:

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



mean absolute error, average absolute loss, empirical risk (for absolute loss)

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 $h^* = \text{Mode}(y_1, y_2, \ldots, y_n)$.

  the most common value

- Therefore, $R_{0,1}(h^*)$ is the proportion of data points not equal to the mode.

  a measure of spread!

- **Example**: What's the proportion of values not equal to the mode in the dataset $2, 3, 3, 4, 5$?

  5 points,
  2 are equal to the mode (3),
  3 are not: $\boxed{\frac{3}{5}}$

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**.

  h^*

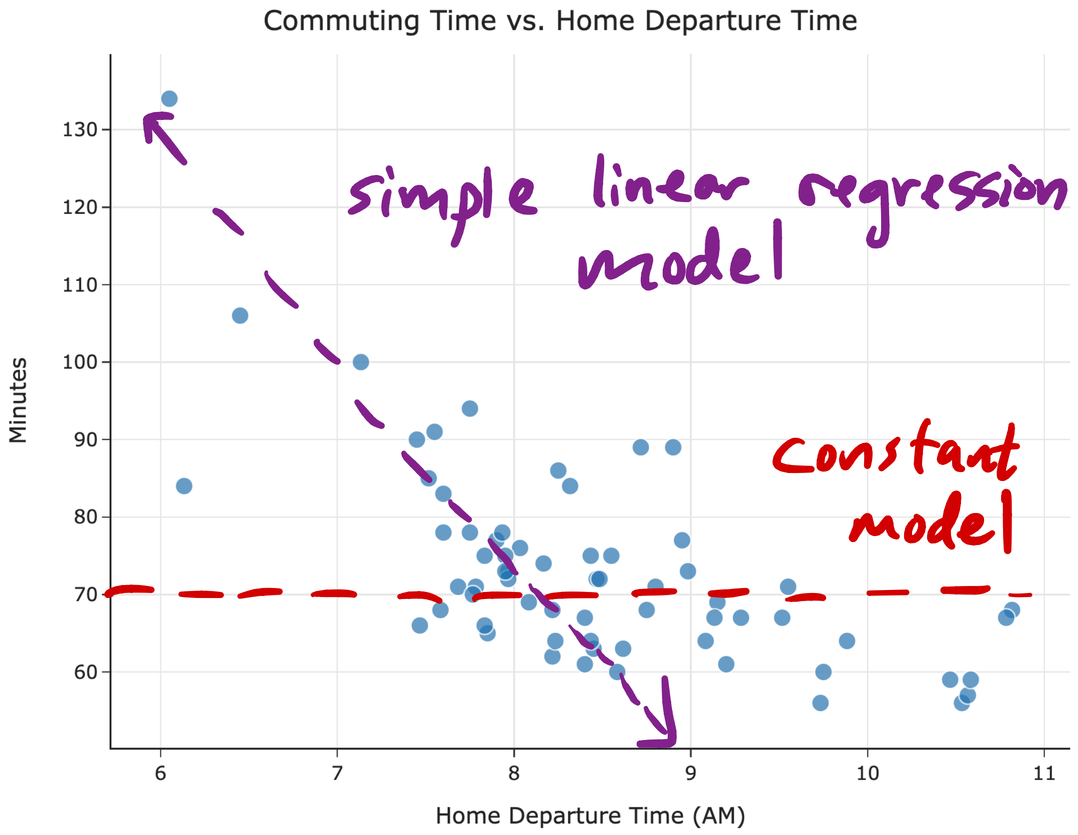
- The minimum values of empirical risk, $R(h^*)$, are various measures of **spread**.
- There are many different ways to measure both center and spread; these are sometimes called **descriptive statistics**.

larger values of spread: data is <u>more</u> spread out.

only uses one "input variable" or "feature" for making predictions!

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.

$H\left(\text{time in the morning}\right) \longrightarrow$ predicted commute time

Recap: Hypothesis functions and parameters

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

till now: $H(x) = h$

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

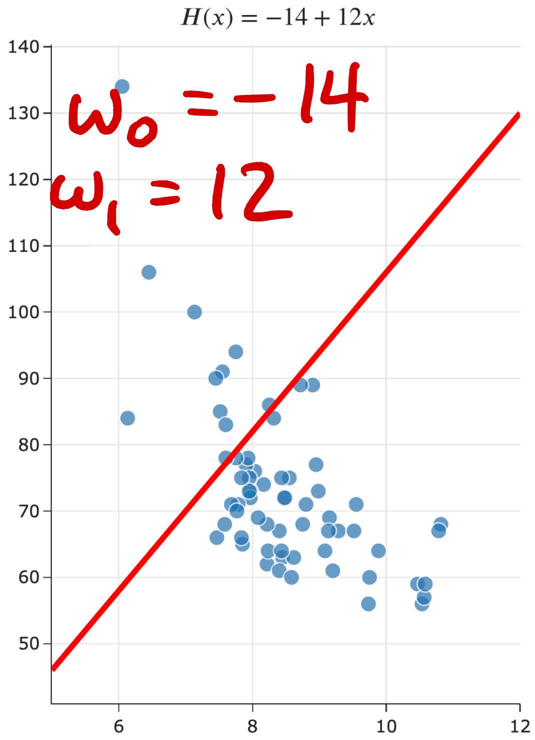
The simple linear regression model, $H(x) = w_0 + w_1 x$, has two parameters: w_0 and w_1.

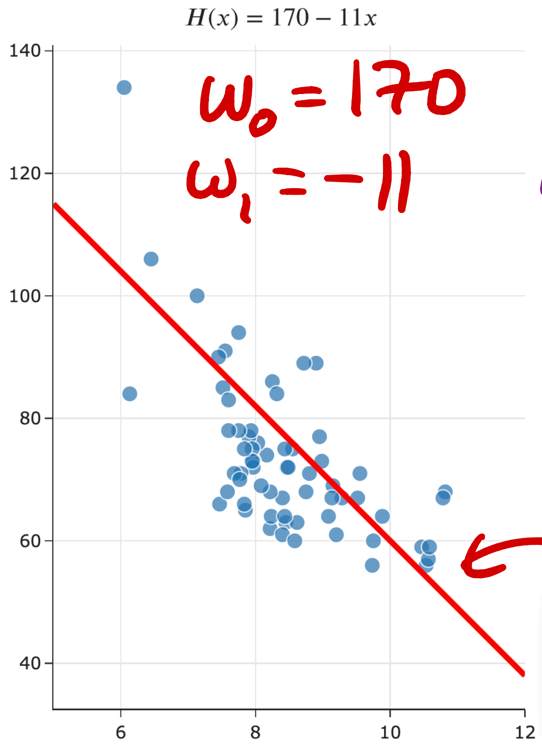
intercept

slope

w_0: "w naught"

We need to find the best slope, w_1^*, and the best intercept, w_0^*!





$$\begin{aligned} H(9) &= 170 - 11 \cdot 9 \\ &= 170 - 99 \\ &= 170 - 100 + 1 \\ &= 70 + 1 = \boxed{71} \end{aligned}$$

predicted commute time

The modeling recipe

1. Choose a model.

Before : $H(x) = h$

Now : $\underbrace{H(x) = w_0 + w_1 x}$

2. Choose a loss function.

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

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

3. Minimize average loss to find optimal model parameters.

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

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

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:

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

- 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:

slope

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

intercept

only unknowns are w_0, w_1!

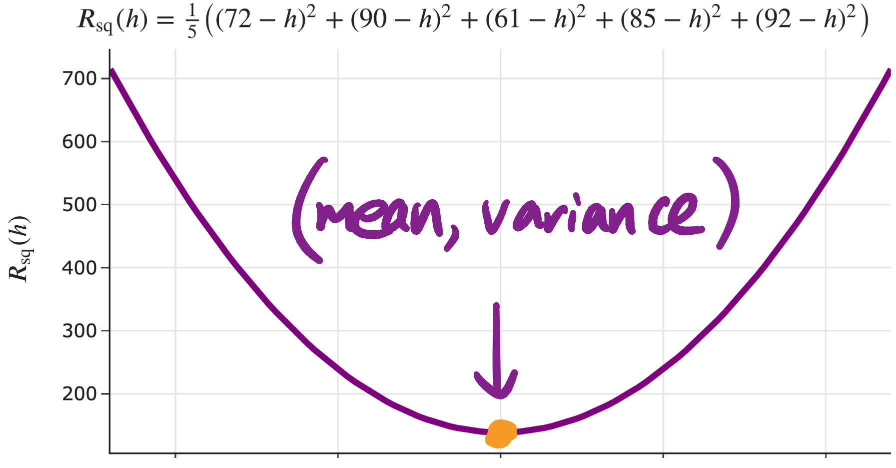
- **How do we find the parameters w_0^* and w_1^* that minimize $R_{\text{sq}}(w_0, w_1)$?**

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

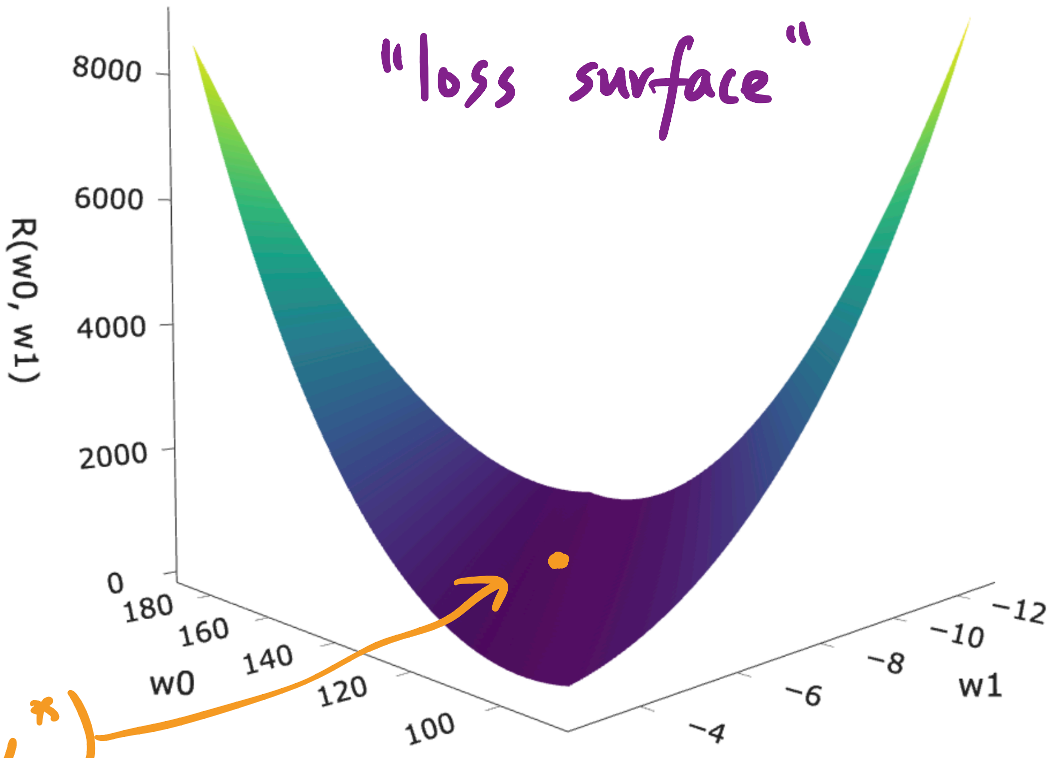
Loss surface

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?



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



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.

$$f(x, y) = \underbrace{x^2 - 8x}_{} + \underbrace{y^2 + 6y}_{} - 7$$

we'll use calculus:

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

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

$2x - 8 = 0 \Rightarrow x = 4$

$2y + 6 = 0 \Rightarrow y = -3$

minimized @ $x = 4$, $y = -3$.

could complete the square

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

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

$\Rightarrow$ minimized at $(4, -3, -32)$

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

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

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

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

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

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

the coefficient on w_0 if you expand is -1.

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

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

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

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

the coefficient on w_1 when we expand is $-x_i$.

Strategy

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

$$\underbrace{-\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i)) = 0}_{\text{partial wrt } w_0} \qquad \underbrace{-\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))x_i = 0}_{\text{partial wrt } w_1}$$

"with respect to"

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^*

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

Goal: Isolate w_0.

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

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

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

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

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

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

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

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

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

Solving for w_1^*

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

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

Goal: Isolate w_1^*.

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

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

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

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

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

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

Least squares solutions

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

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

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.

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

showed last class, and in HW1!

An equivalent formula for w_1^*

Claim:

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

n·variance$(x_1, x_2, \ldots, x_n)$

Proof:

right numerator

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

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

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

Least squares solutions

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

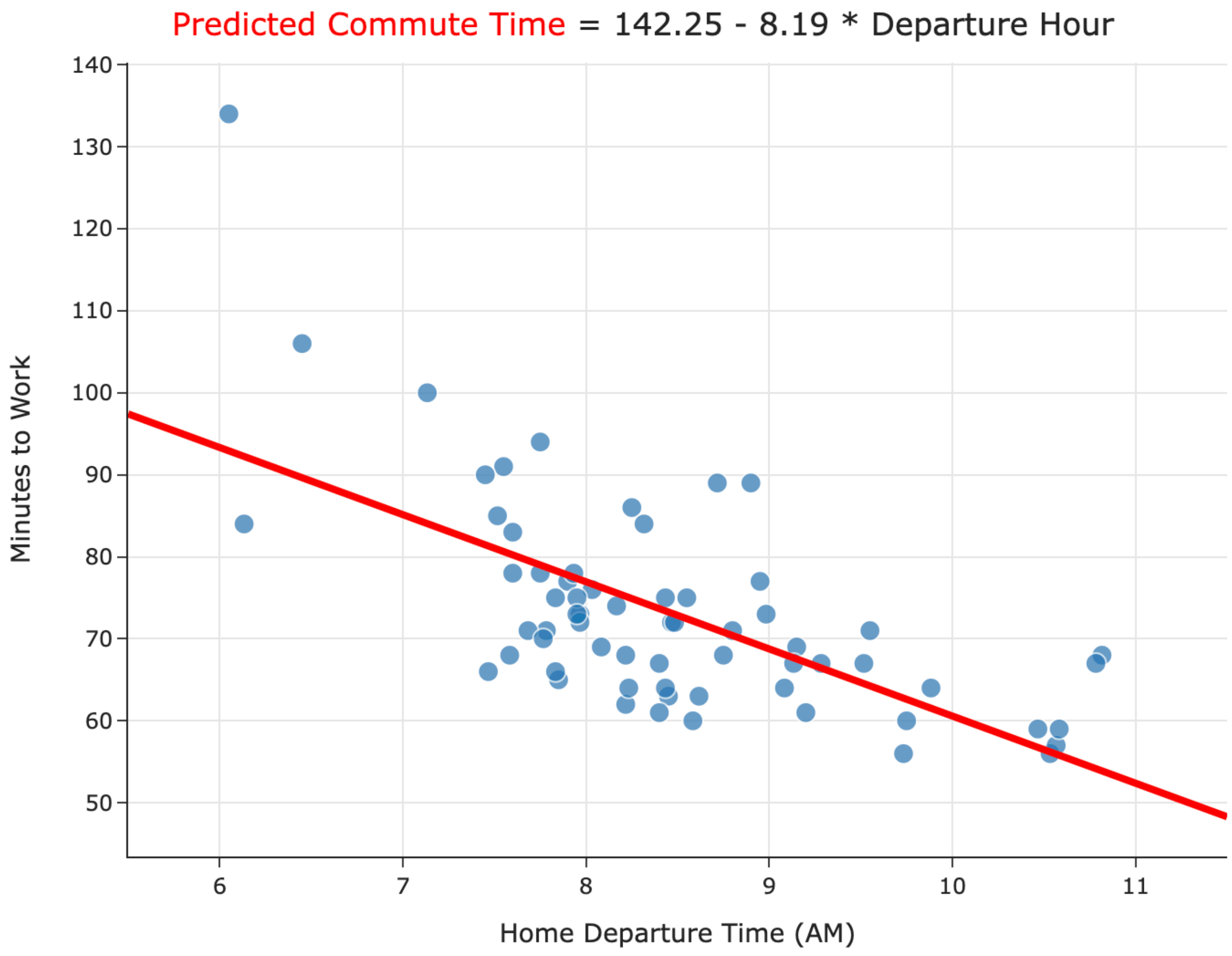
$$\boxed{w_1^* = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad\qquad w_0^* = \bar{y} - w_1\bar{x}}$$

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

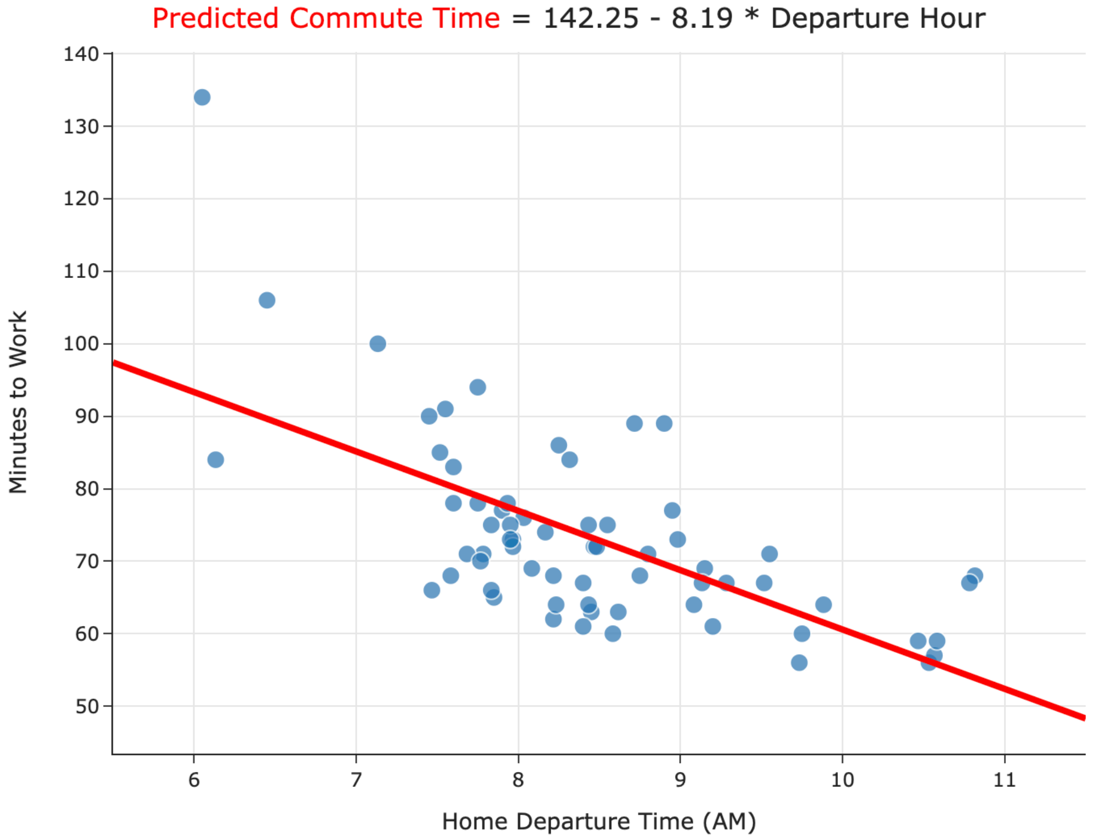
when using squared loss

- 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~~ quicker?

No! This is just an observed pattern.

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!