Lecture 4

Simple Linear Regression

DSC 40A, Summer 2024

Announcements

Friday 11:59p

- Homework 1 is due **tomorrow night**.
  - Before working on it, watch the Walkthrough Videos on problem solving and using Overleaf.

    due Tuesday night
  - 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 $H(x) = h$, empirical risk is of the form:

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

mode
mean
median
midrange

- 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.

variance

- The specific center and spread depend on the choice of loss function.

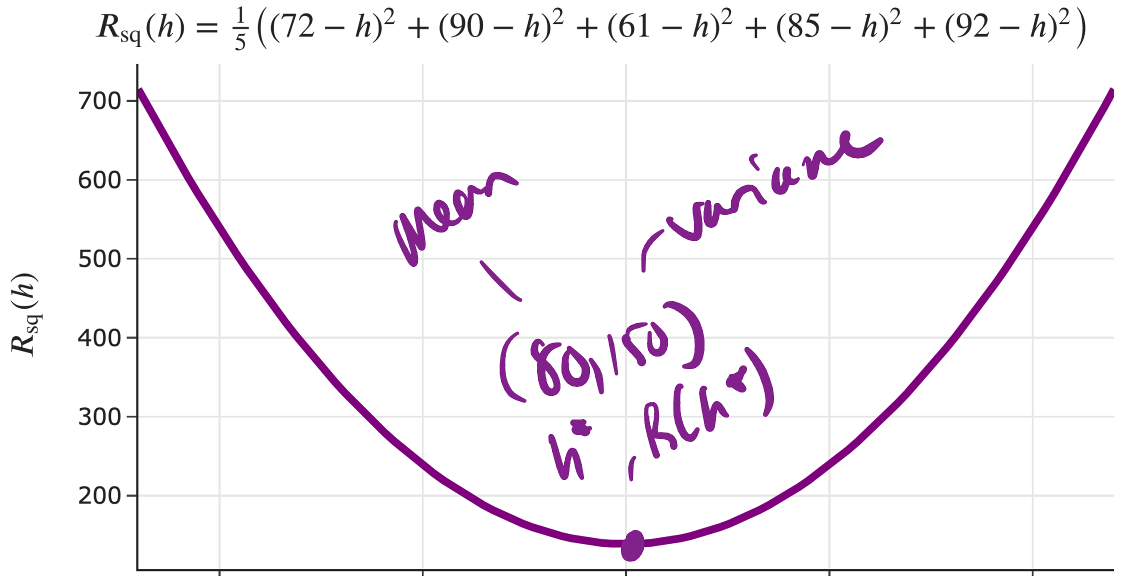
Examples

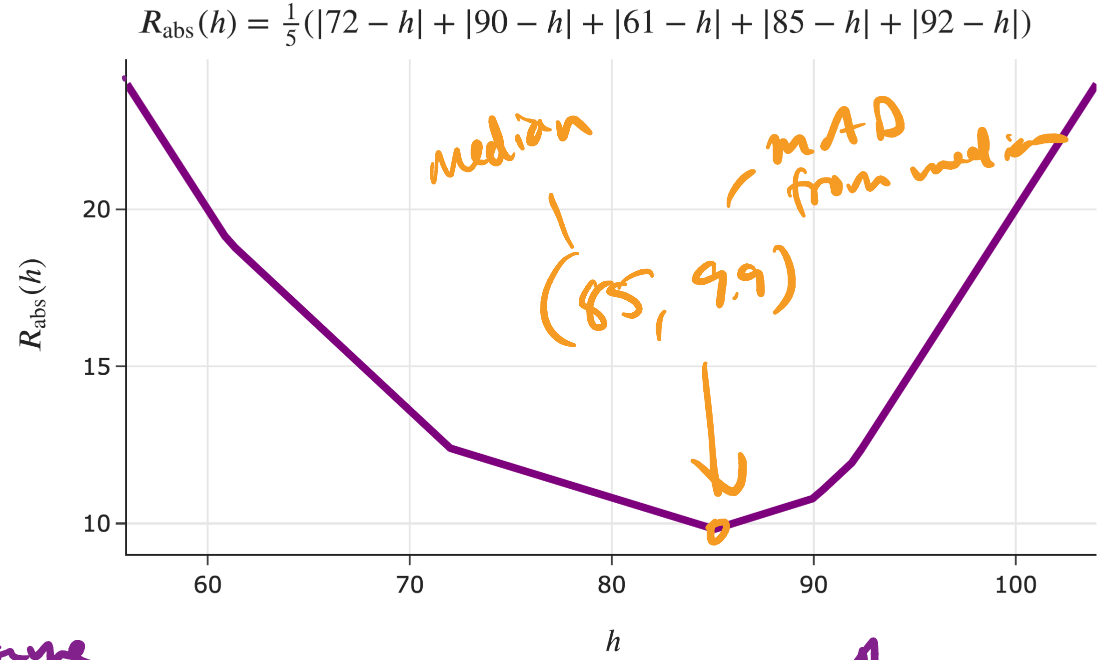
When using **squared loss**:

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

When using **absolute loss**:

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





mean squared error
average squared loss
empirical risk (for squared loss)
same

3 names for ↑

only use 1 input variable to make predictions

Simple linear regression

H(time in morning) → predict commute time

Recap: Hypothesis functions and parameters

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

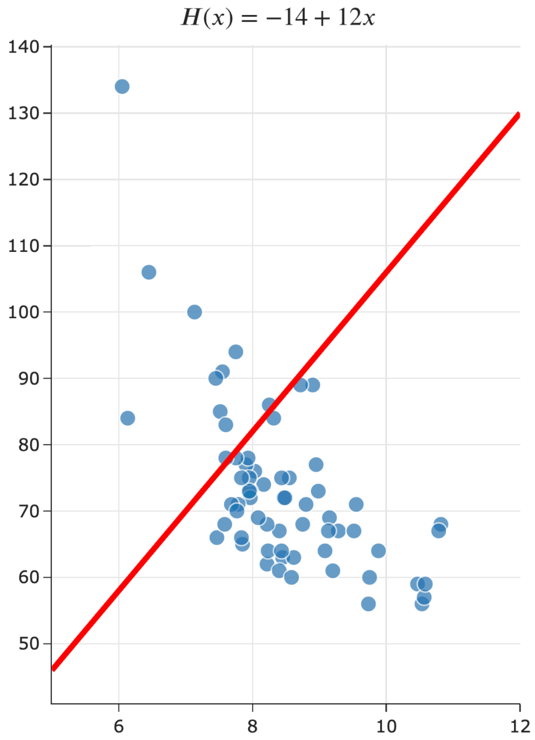
Parameters define the relationship between the input and output of a hypothesis function.

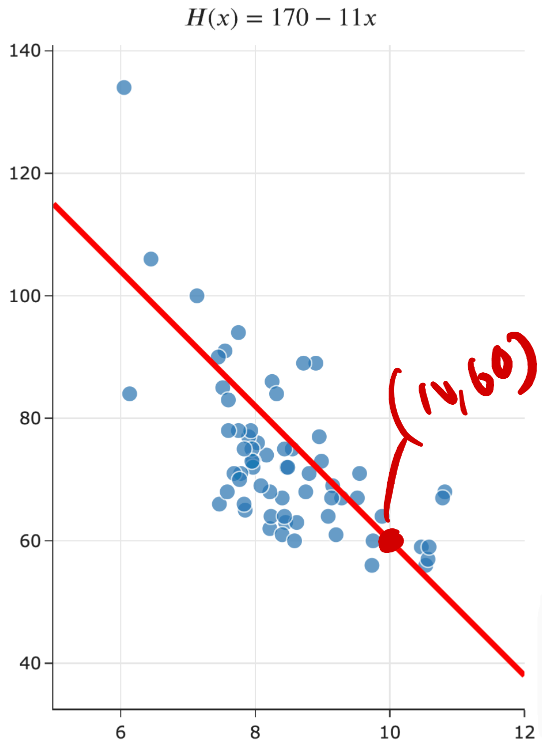
before $H(x) = h$

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"





find the "best" slope w_1^*,
and "best" intercept w_0^*

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

The modeling recipe

1. Choose a model.

Before: $H(x) = h$

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

2. Choose a loss function.

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

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

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:

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

- 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:

slope

only unknowns are w_0, w_1

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

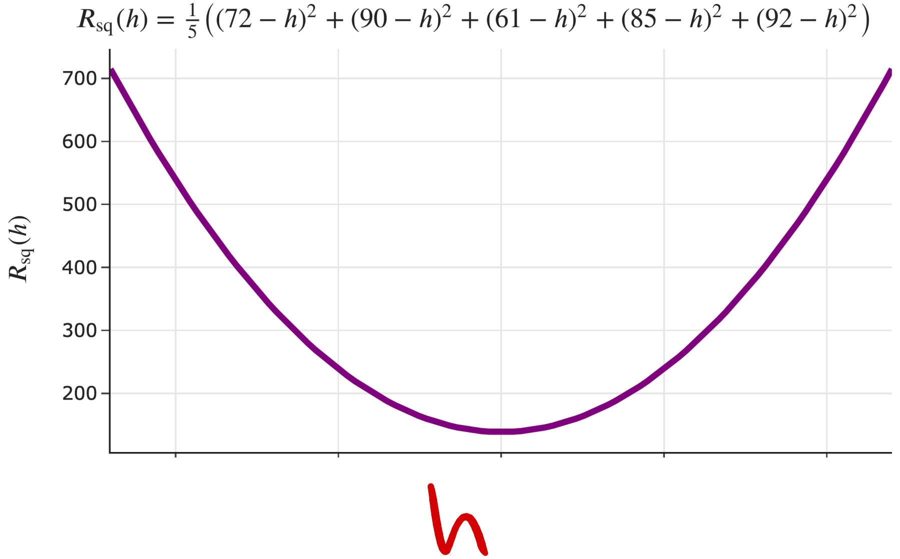
intercept

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

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

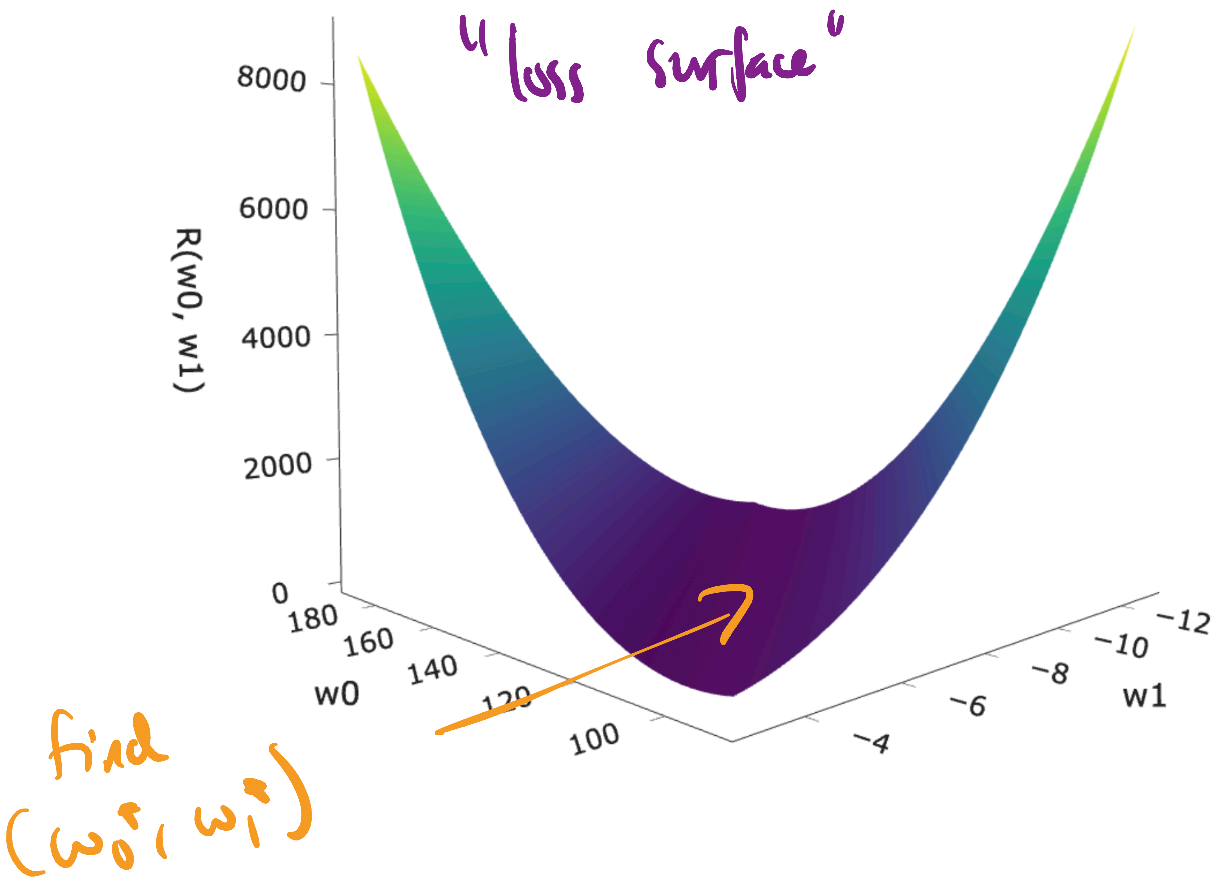
Loss surface

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



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

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:

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

- 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

plug back in

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

$-16 \quad -9 \quad -7 = -32$

$16 - 32 + 9 - 18 - 7$

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

$\frac{\partial f}{\partial x} = 2x - 8 \Rightarrow 2x - 8 = 0 \Rightarrow 2x = 8 \Rightarrow x = 4$

$\frac{\partial f}{\partial y} = 2y + 6 \Rightarrow 2y = -6 \Rightarrow y = -3$

minimized at $x^*, y^* = (4, -3, -32)$

Completing the square

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

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

min. at $(4, -3, -32)$

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 🤔

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

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

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

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

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

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

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

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."

Goal: isolate w_0

Solving for w_0^*

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

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

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

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

$$+ n w_0 \qquad + n w_0$$

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

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

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

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

Solving for w_1^*

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

Goal: isolate w_1^*

plug in here

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

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

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

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

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

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

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.

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

shown before, and in HW 1 algebraically!

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:

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

slope (in DSC 10) $= r \cdot \frac{SD(y)}{SD(x)}$

Least squares solutions

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

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

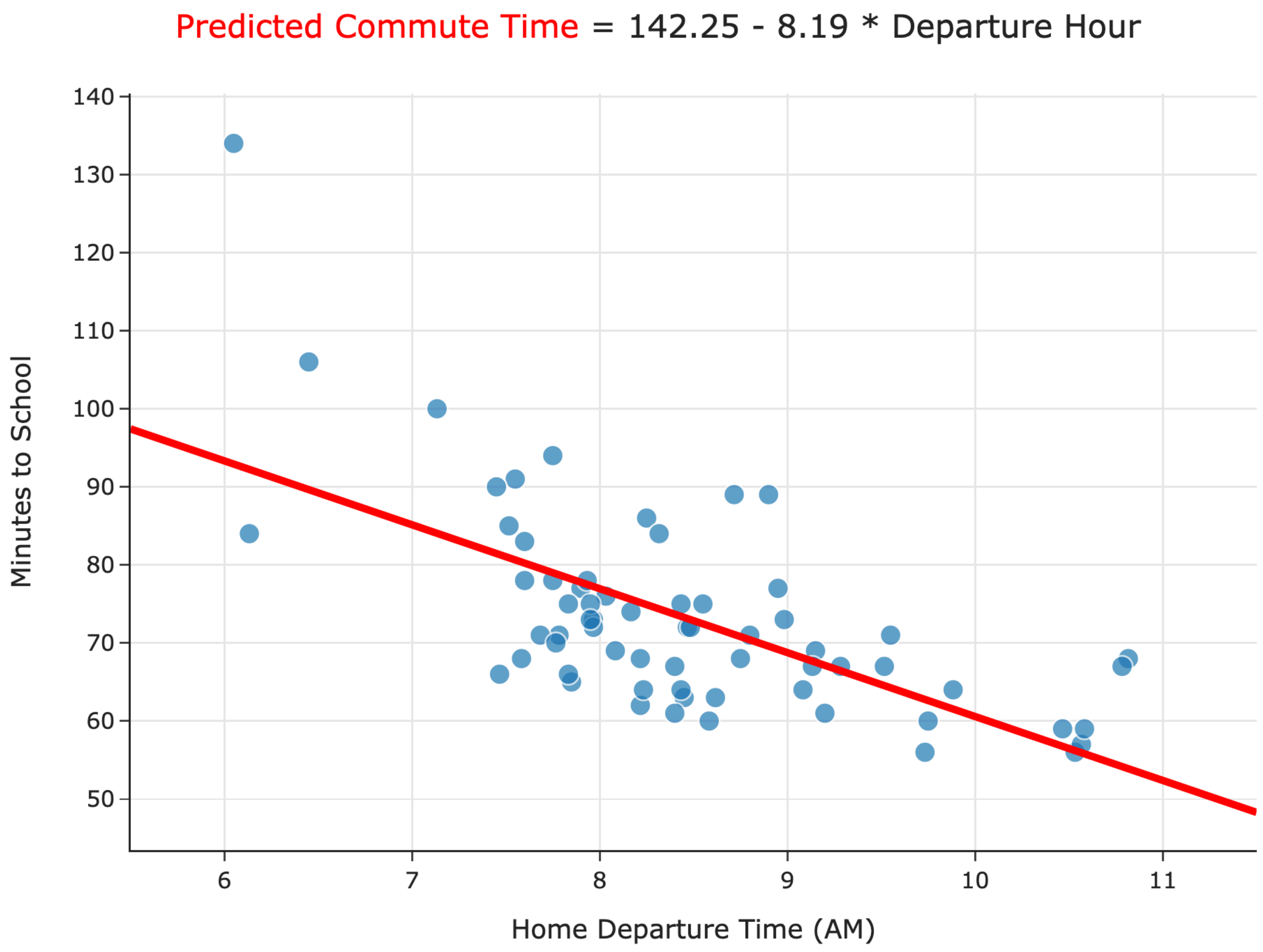
n· variance $(x_1, x_2, \ldots x_n)$

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

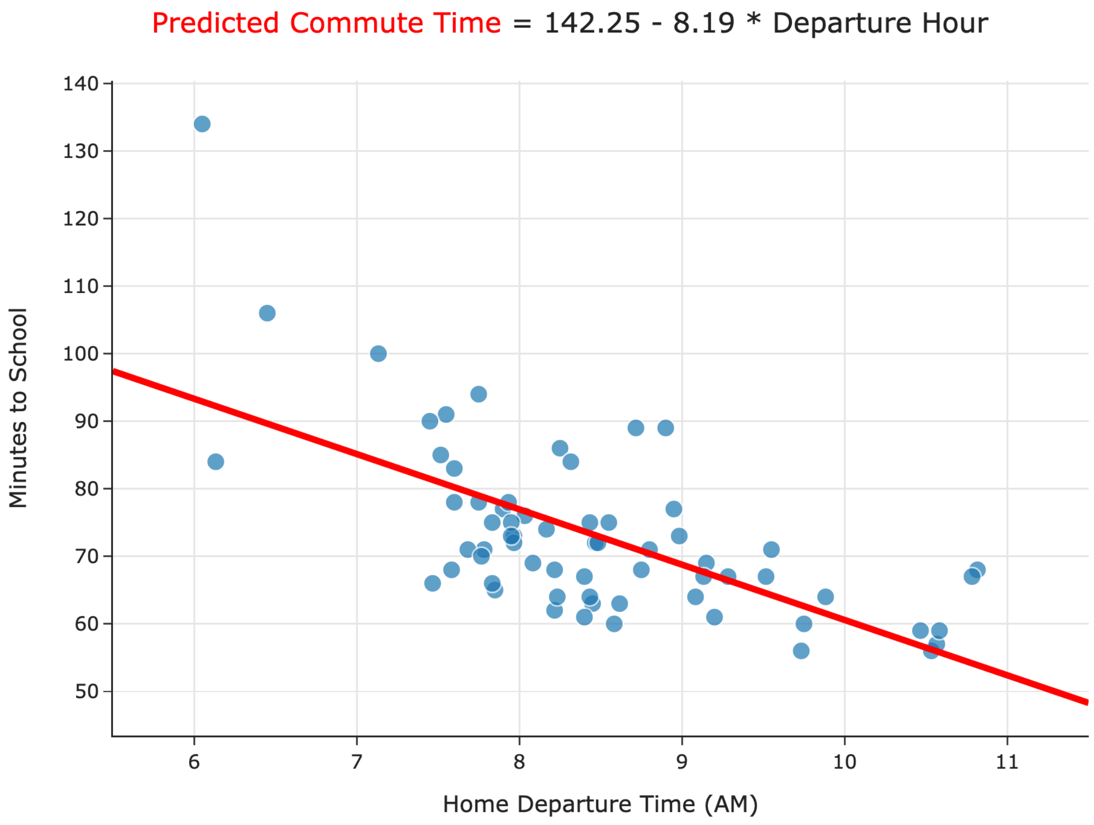
↳ when we use squared loss

- 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 quicker?

No: Just an observed pattern.

What's next?

minimize MSE $\rightarrow R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \overbrace{(w_0 + w_1 x_i)}^{H(x_i)}\Big)^2$

We now know how to find the optimal slope and intercept for linear hypothesis functions.

$H(x_i) = w_0 + w_1 x_i + w_2 x_i^2$

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!