Lecture 12

Loss Functions and Simple Linear Regression

EECS 398: Practical Data Science, Winter 2025

practicaldsc.org • github.com/practicaldsc/wn25 • 📣 See latest announcements **here on Ed**

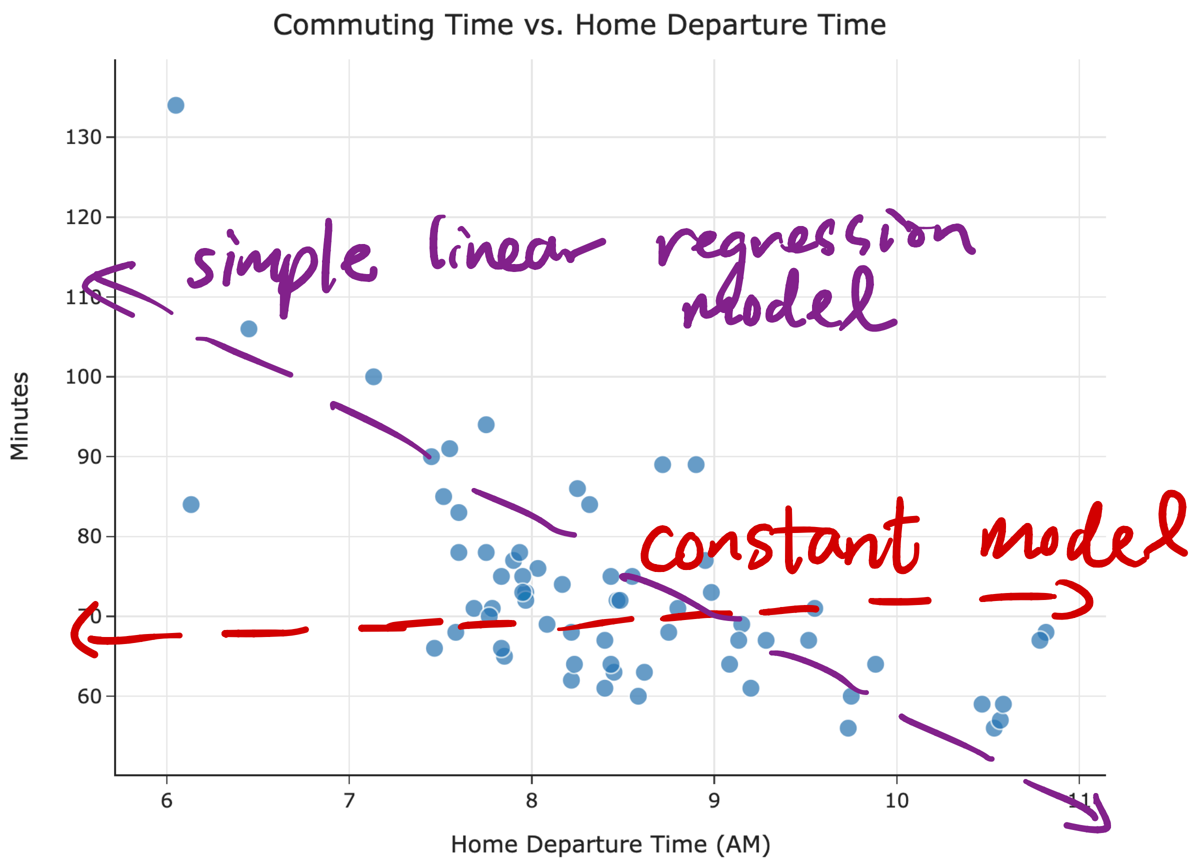
Agenda 📆

- Recap: Models and loss functions.
- Another loss function.
- Towards simple linear regression.
- Minimizing mean squared error for the simple linear model.
- Correlation.
- Interpreting the formulas.

There are several important videos for Lectures 11 and 12; they are all in this YouTube playlist.

Recap: Models and loss functions

Overview



- We started by introducing the idea of a hypothesis function, $H(x_i)$.
- We looked at two possible models:
  - The constant model, $H(x_i) = h$.
  - The simple linear regression model, $H(x_i) = w_0 + w_1 x_i$.
- We decided to find the **best constant prediction** to use for predicting commute times, in minutes.

squared loss function : $(\text{actual} - \text{predicted})^2$

Recap: Mean squared error

- Let's suppose we have just a smaller dataset of just five historical commute times in minutes.

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 92$$

- The **mean squared error** of the constant prediction h is:

$$R_{\text{sq}}(h) = \frac{1}{5}\left((72 - h)^2 + (90 - h)^2 + (61 - h)^2 + (85 - h)^2 + (92 - h)^2\right)$$

- For example, if we predict $h = 100$, then:

$$R_{\text{sq}}(100) = \frac{1}{5}\left((72 - 100)^2 + (90 - 100)^2 + (61 - 100)^2 + (85 - 100)^2 + (92 - 100)^2\right)$$
$$= \boxed{538.8}$$

- **We can pick any h as a prediction, but the smaller $R_{\text{sq}}(h)$ is, the better h is!**

The mean minimizes mean squared error!

- The problem we set out to solve was, find the h^* that minimizes:

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

took derivative, set to 0, solved.

- The answer is:

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

- The **best constant prediction**, in terms of mean squared error, is always the **mean**.
- We call h^* our **optimal model parameter**, for when we use:
  - the constant model, $H(x_i) = h$, and
  - the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$.
- Review the derivation steps from Lecture 11's slides, and watch the video we posted.

The modeling recipe

- We've implicitly introduced a three-step process for finding optimal model parameters (like h^*) that we can use for making predictions:

    1. Choose a model.

       constant model: $H(x_i) = h$

    2. Choose a loss function.

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

    3. Minimize average loss to find optimal model parameters.

       $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2 \implies h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$

- Most modern machine learning methods today, including neural networks, follow this recipe, and we'll see it repeatedly this semester!

Question 🤔

Answer at practicaldsc.org/q

What questions do you have?

Another loss function

Another loss function

- We started by computing the **error** for each of our predictions, but ran into the issue that some errors were positive and some were negative.

$$e_i = y_i - H(x_i)$$

→ actual − predicted

- The solution was to **square** the errors, so that all are non-negative. The resulting loss function is called **squared loss**.

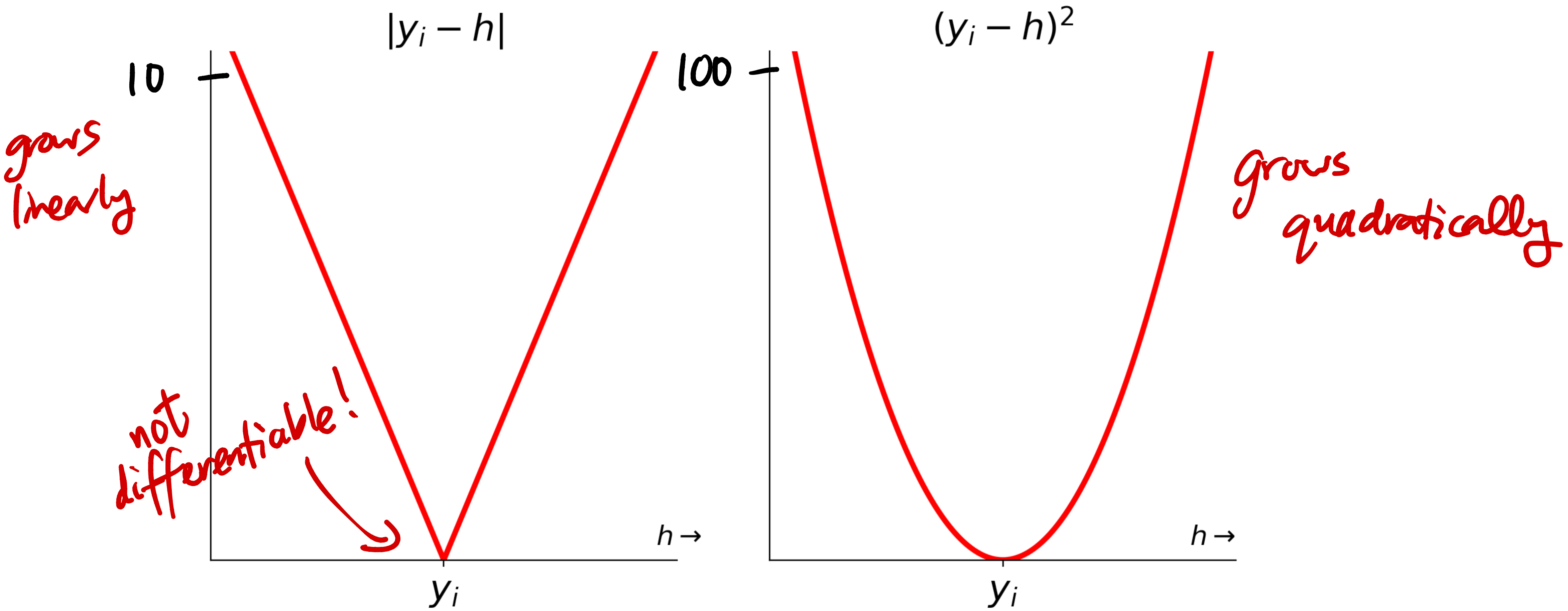
$$L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$$

- Another loss function, which also measures how far $H(x_i)$ is from y_i, is **absolute loss**.

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

(actual − predicted)²

Absolute loss vs. squared loss



Mean absolute error

- Suppose we collect n commute times, $y_1, y_2, \ldots, y_n$.
- The **average absolute loss**, or **mean absolute error (MAE)**, of the prediction h is:

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

- We'd like to find the best constant prediction, h^*, by finding the h that minimizes **mean absolute error** (a new objective function).
- Any guesses? median!

The median minimizes mean absolute error!

- It turns out that the constant prediction h^* that minimizes mean absolute error,

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

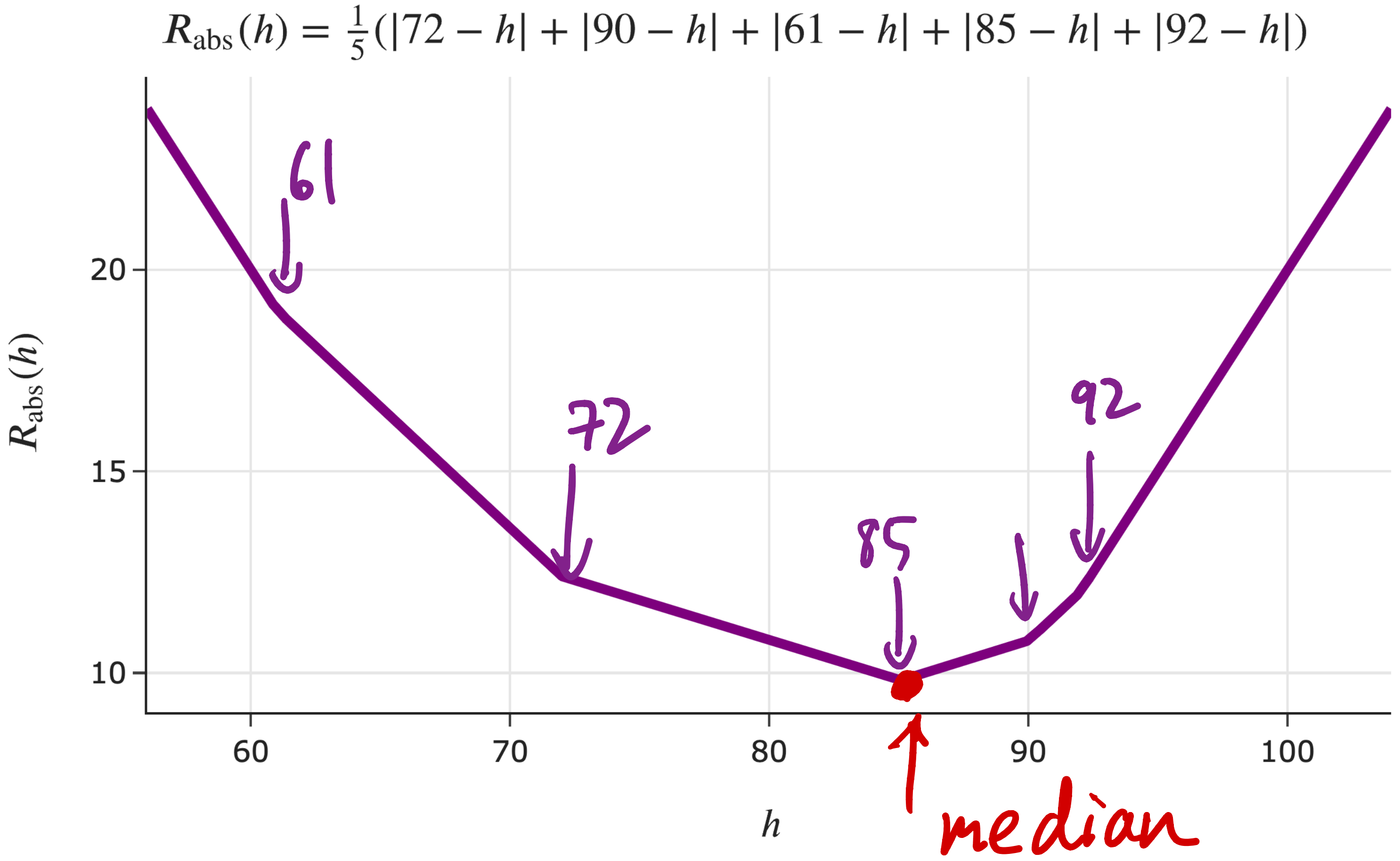
  is:

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

- We won't prove this in lecture, but this extra video walks through it.
  Watch it!
- To make a bit more sense of this result, let's graph $R_{\text{abs}}(h)$.

Visualizing mean absolute error

$R_{abs}(h)$ is a piecewise linear function!

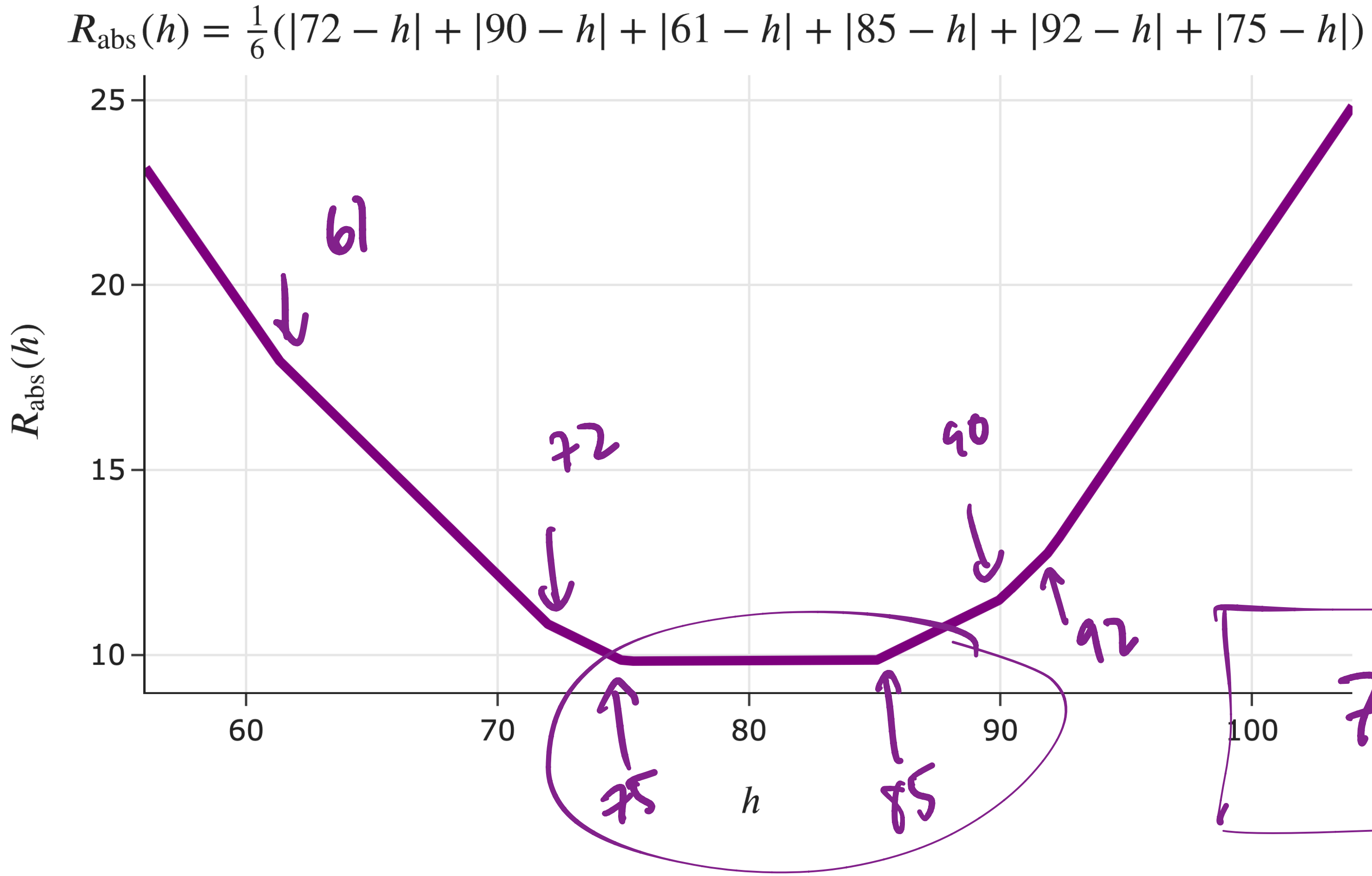


- Consider, again, our example dataset of five commute times.

$$72, 90, 61, 85, 92$$

- Where are the "bends" in the graph of $R_{abs}(h)$ – that is, where does its slope change?

Visualizing mean absolute error, with an even number of points



- What if we add a sixth data point?

$$72, 90, 61, 85, 92, 75$$

- Is there a unique h^*?

no!

$75 \leq h^* \leq 85$

The median minimizes mean absolute error!

- The new problem we set out to solve was, find the h^* that minimizes:

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

- The answer is:

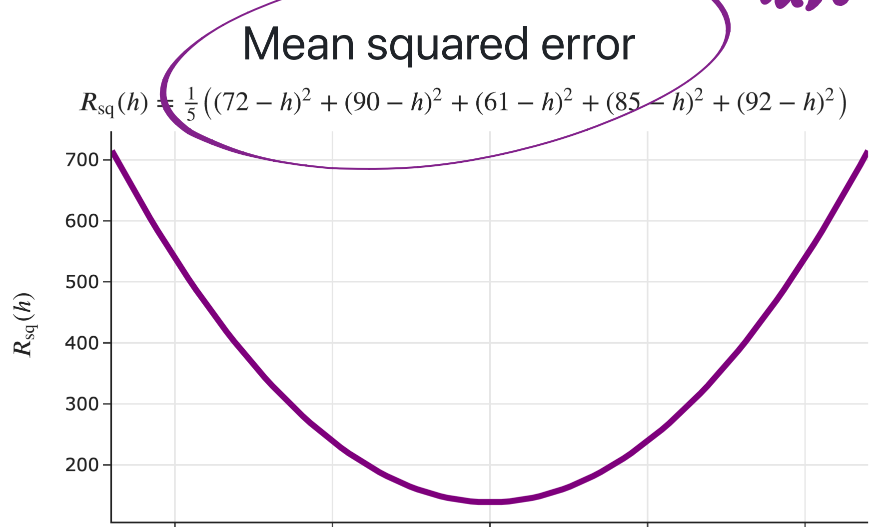
$$h^* = \text{Median}(y_1, y_2, \ldots, y_n)$$

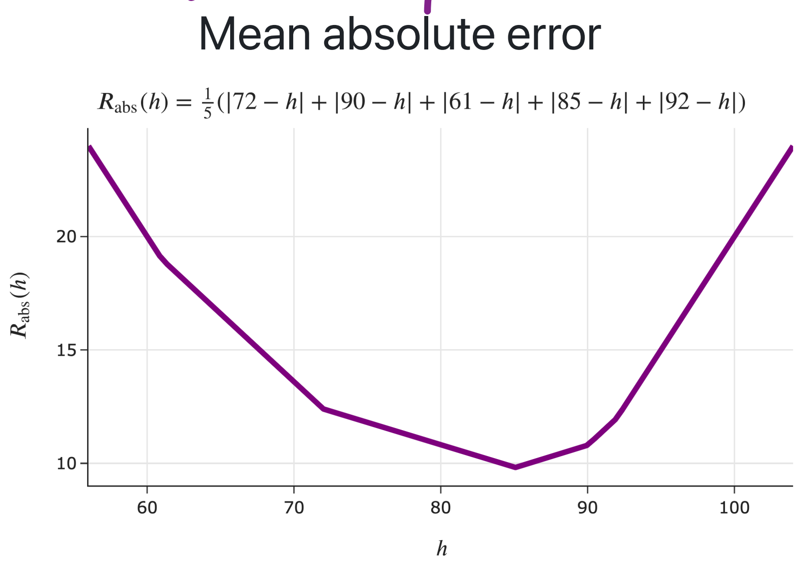
- The **best constant prediction**, in terms of mean absolute error, is always the **median**.
  - When n is odd, this answer is unique.
  - When n is even, any number between the middle two data points (when sorted) also minimizes mean absolute error.
  - When n is even, define the median to be the mean of the middle two data points.

Choosing a loss function

- For the constant model $H(x_i) = h$, the **mean** minimizes mean **squared** error.
- For the constant model $H(x_i) = h$, the **median** minimizes mean **absolute** error.
- In practice, squared loss is the more common choice, as the resulting objective function is more easily **differentiable**.

also: connections to probability





- But how does our choice of loss function impact the resulting optimal prediction?

Comparing the mean and median

- Consider our example dataset of 5 commute times.

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 92$$

- As of now, the **median is 85** and the **mean is 80**.
- What if we add 200 to the largest commute time, 92?

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 292$$

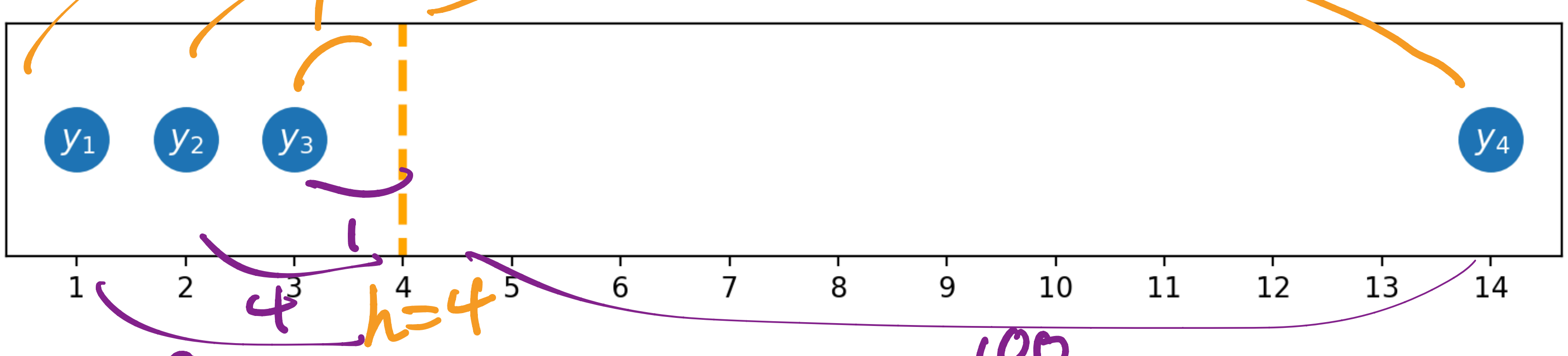
- Now, the median is still 85 but the mean is 120 !
- **Key idea**: The mean is quite **sensitive** to outliers.

  But why?

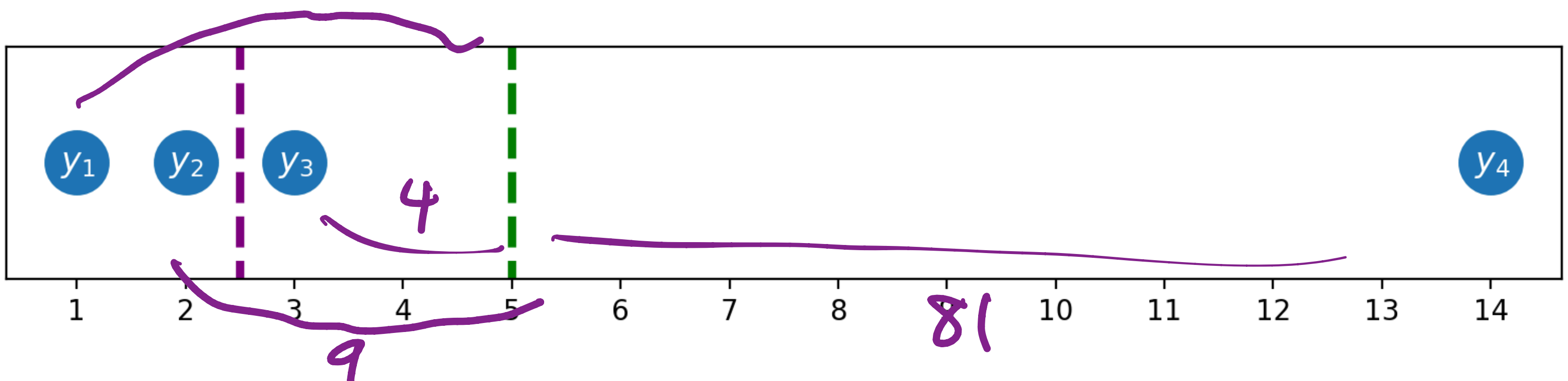
$80 + \frac{200}{5}$ } 40

Outliers

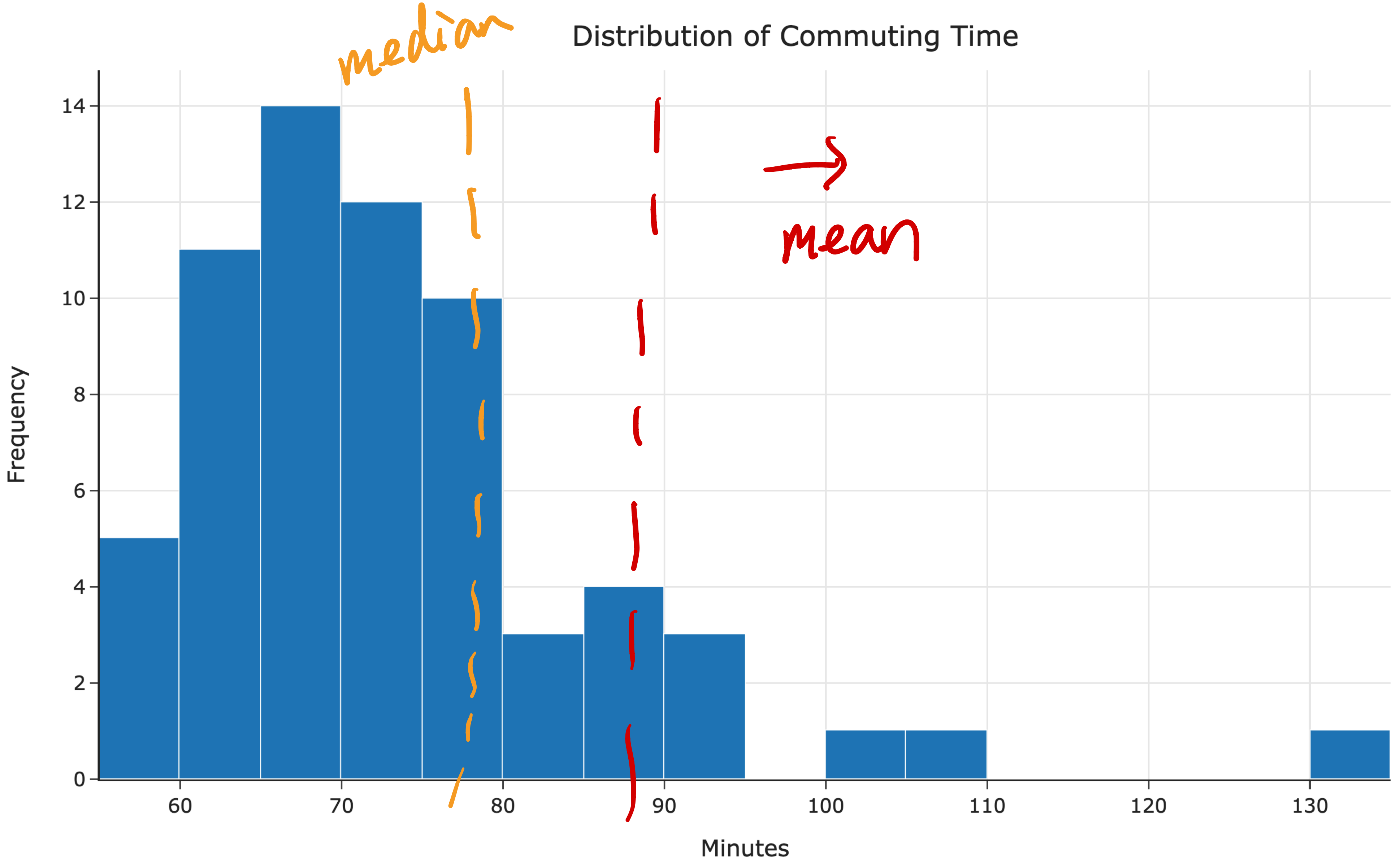
- Below, $|y_4 - h|$ is 10 times as big as $|y_3 - h|$, but $(y_4 - h)^2$ is 100 times $(y_3 - h)^2$.



- The result is that the **mean** is "pulled" in the direction of outliers, relative to the **median**.

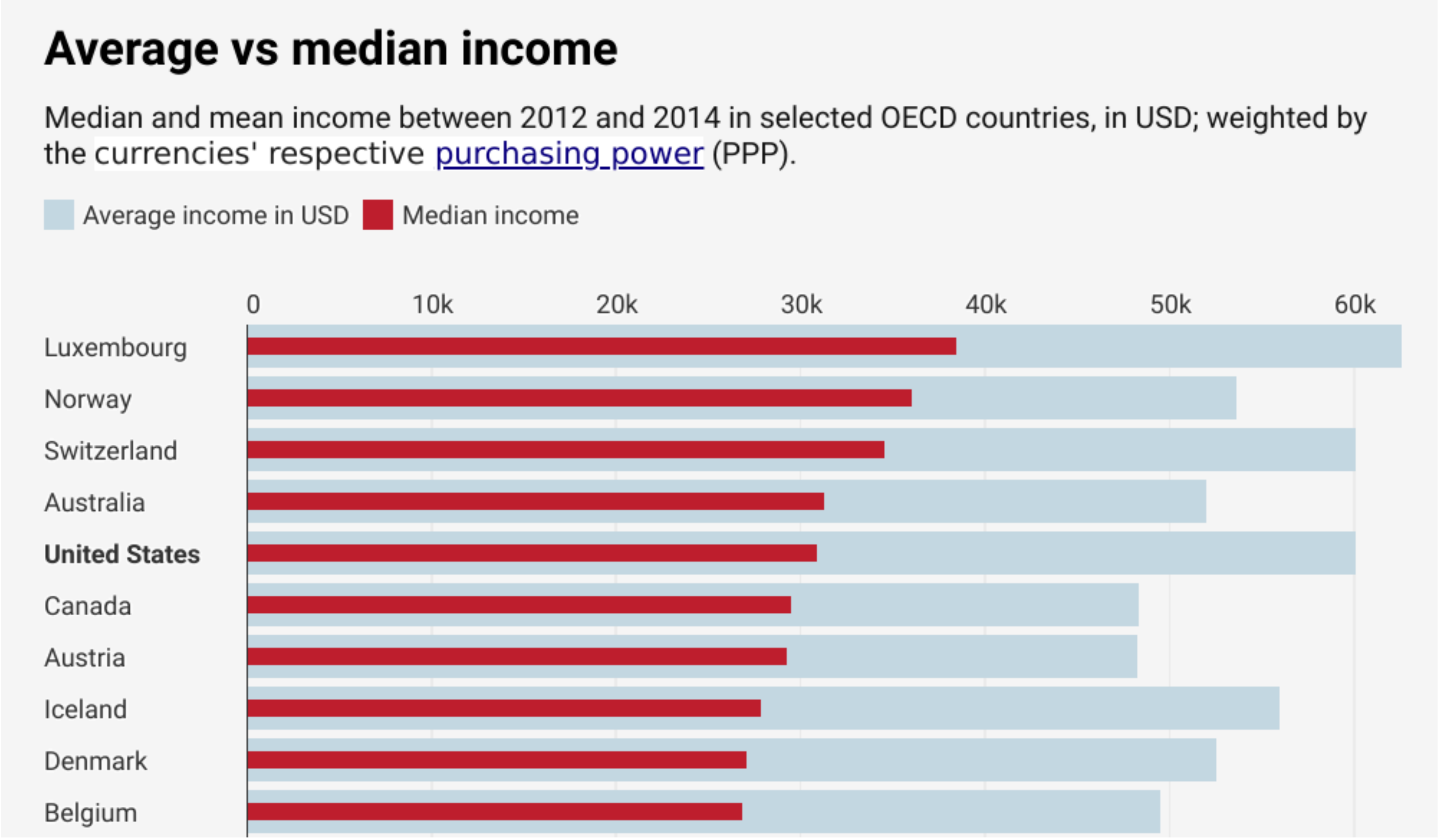


- As a result, we say the **median** – and absolute loss more generally – is **robust**.



Distribution of Commuting Time
median
mean
Frequency
Minutes
0
2
4
6
8
10
12
14
60
70
80
90
100
110
120
130

Example: Income inequality



Summary: Choosing a loss function

- **Key idea**: Different loss functions lead to different best predictions, h^*!

Loss	Minimizer	Always Unique?	Robust to Outliers?	Differentiable?
$L_{\text{sq}}(y_i, h) = (y_i - h)^2$	mean	yes ✅	no ❌	yes ✅
$L_{\text{abs}}(y_i, h) = \|y_i - h\|$	median	no ❌	yes ✅	no ❌
$L_{0,1}(y_i, h) = \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$	mode	no ❌	yes ✅	no ❌
$L_\infty(y_i, h)$ See HW 6.	???	yes ✅	no ❌	no ❌

- The optimal predictions, h^*, are all **summary statistics** that measure the **center** of the dataset in different ways.

Question 🤔

Answer at practicaldsc.org/q

What questions do you have?

The modeling recipe

- We've now made two full passes through our modeling recipe.
    1. Choose a model.
    2. Choose a loss function.
    3. Minimize average loss to find optimal model parameters.

constant model: $H(x_i) = h$

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

$\Downarrow$ $\qquad$ $\Downarrow$

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

$\Downarrow$ $\qquad$ $\Downarrow$

$h^* = \text{Mean}$ $\qquad$ $h^* = \text{Median}$

Empirical risk minimization

- The formal name for the process of minimizing average loss is **empirical risk minimization**; another name for "average loss" is **empirical risk**.
- When we use the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$, the corresponding empirical risk is mean squared error:

$$R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2 \implies h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$$

- When we use the absolute loss function, $L_{\text{abs}}(y_i, h) = |y_i - h|$, the corresponding empirical risk is mean absolute error:

$$R_{\text{abs}}(h) = \frac{1}{n}\sum_{i=1}^{n}|y_i - h| \implies h^* = \text{Median}(y_1, y_2, \ldots, y_n)$$

Empirical risk minimization, in general

- **Key idea**: If L is **any** loss function, and H is any hypothesis function, the corresponding empirical risk is:

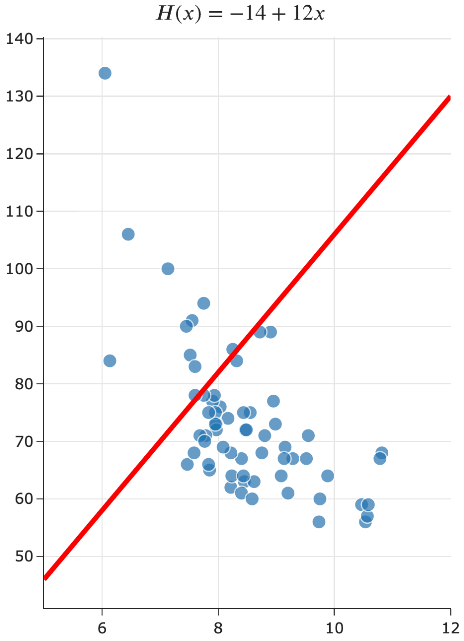
$$R(H) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, H(x_i))$$

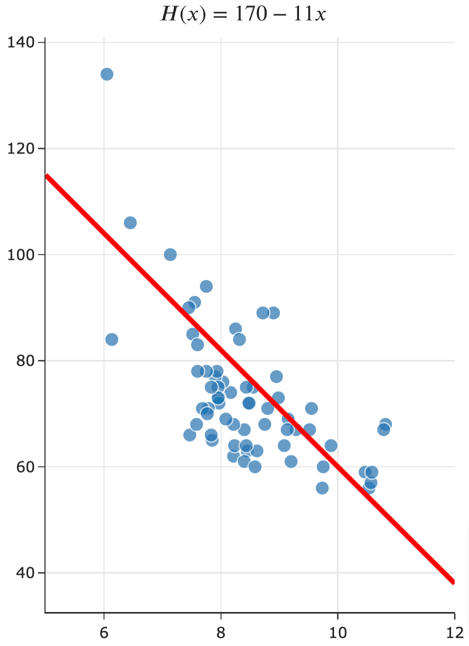
- In Homework 6 and tomorrow's discussion, there are several questions where:
    - You are given a new loss function L.
    - You have to find the optimal parameter h^* for the constant model $H(x_i) = h$.

Towards simple linear regression

Recap: Hypothesis functions and parameters

- A hypothesis function, H, takes in an x_i as input and returns a predicted y_i.
- **Parameters** define the relationship between the input and output of a hypothesis function.
- **Example**: The simple linear regression model, $H(x_i) = w_0 + w_1 x_i$ has two parameters: w_0 and w_1.





The modeling recipe

1. Choose a model.

$H(x_i) = w_0 + w_1 x_i$ simple linear regression model

x_i : departure hour

y_i : commute time

2. Choose a loss function.

$R_{sq}(y_i, H(x_i)) = (y_i - H(x_i))^2$ squared loss

3. Minimize average loss to find optimal model parameters.

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

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

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:

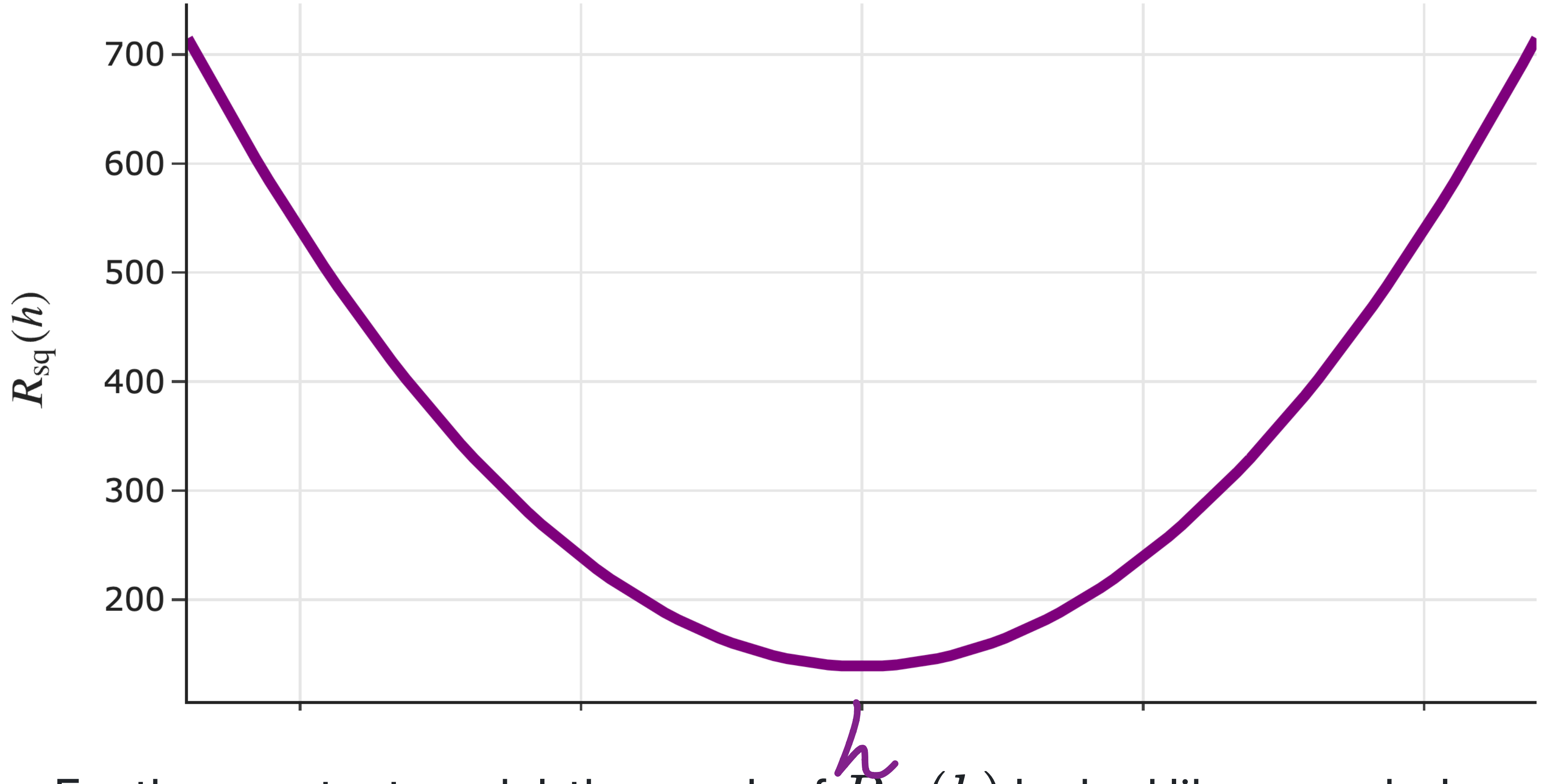
$$R_{\text{sq}}(H) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - H(x_i)\right)^2$$

- Since linear hypothesis functions are of the form $H(x_i) = w_0 + w_1 x_i$, we can re-write R_{sq} as a function of w_0 and w_1:

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

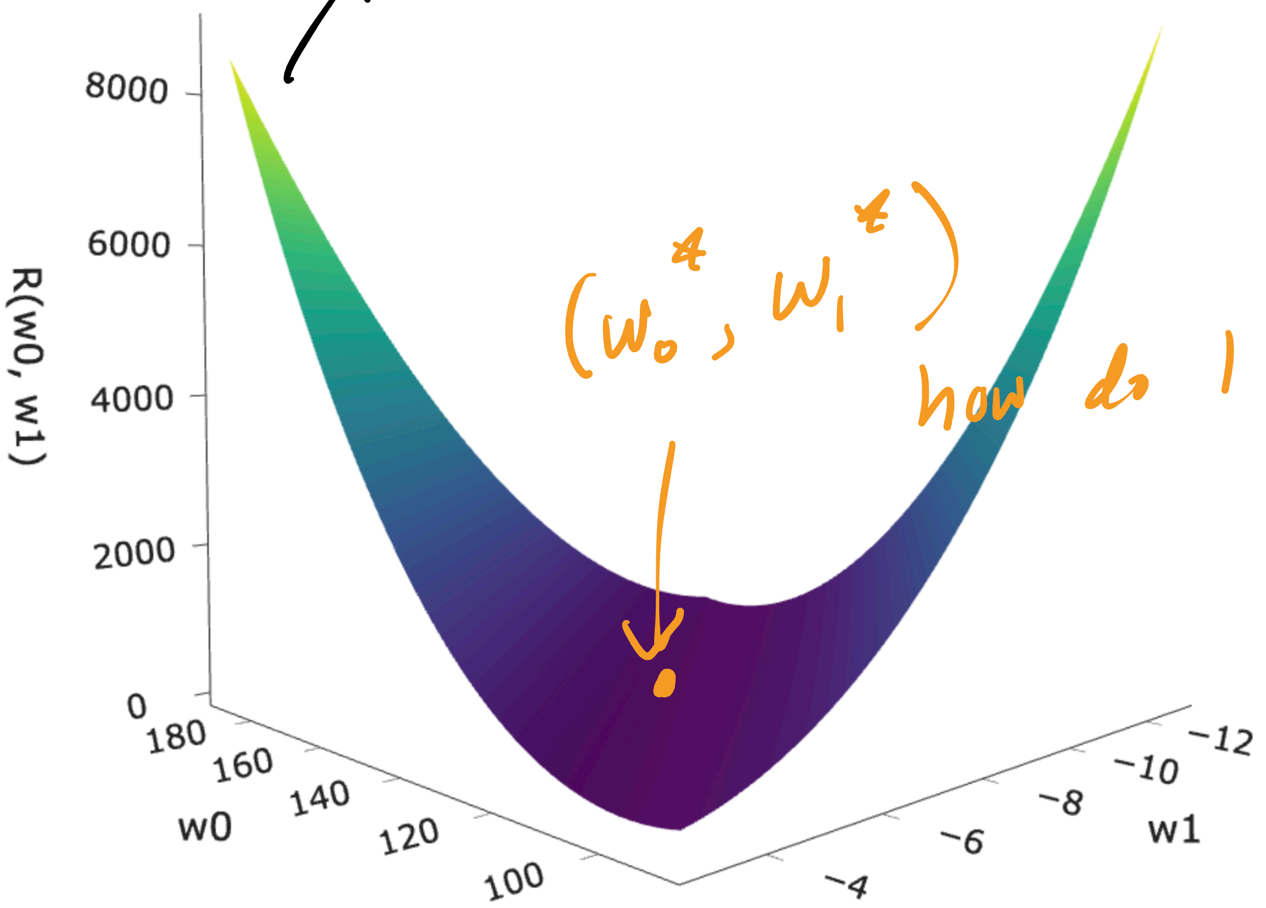
- **How do we find the parameters w_0^* and w_1^* that minimize $R_{\text{sq}}(w_0, w_1)$?**

$$R_{\text{sq}}(h) = \frac{1}{5}\left((72-h)^2 + (90-h)^2 + (61-h)^2 + (85-h)^2 + (92-h)^2\right)$$



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

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



The graph of $R_{\text{sq}}(w_0, w_1)$ for the simple linear regression model is 3 dimensional **bowl**, and is called a **loss surface**.

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:

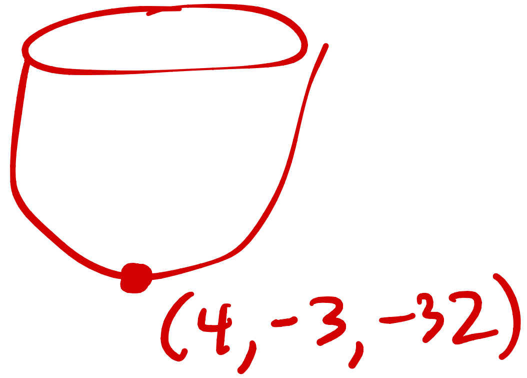
$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

- R_{sq} is a function of two variables: w_0 and w_1, and is a bowl-like shape in 3D.
- To minimize a function of multiple variables:
  - Take partial derivatives with respect to each variable.
  - Set all partial derivatives to 0 and 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).
- To save time, we won't do the derivation live in class, but you are responsible for it! Here's a video of me walking through it, and the slides will be annotated with it.

Example

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

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



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

This is the "partial derivative" of f with respect to x; it treats y as a constant.

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

To solve for where the function is minimized/maximized, set ALL partial derivatives to 0 and solve.

$2x - 8 = 0$ ① → solve and get

$2y + 6 = 0$ ② → $x = 4, y = -3$ as the minimizing input

$(z = f(4, -3) = -32)$

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.

unlike on the last slide, BOTH partial derivatives will involve BOTH variables, w_0 and w_1

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

chain rule!

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

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

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

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

again, chain rule!

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

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

Strategy

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

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

this says that the sum of (actual y_i - predicted y_i) should be 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^* ← optimal intercept!

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

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

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

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

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

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

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

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

mean of y!

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

optimal intercept!

optimal intercept is defined in terms of optimal slope!

Solving for w_1^*

optimal slope

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

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

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

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

grouping like terms

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

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

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

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

formula for optimal slope!

Least squares solutions

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

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

optimal slope

optimal intercept

where:

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

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

An equivalent formula for w_1^*

another equivalent formula!

- Claim:

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

"the sum of deviations is 0"

- Proof: Start with the fact that $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$

Then, on the numerator, starting from the left side:

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

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

numerator ✓ same logic for denom

Least squares solutions

"least squares" → line with LEAST mean SQUARED error

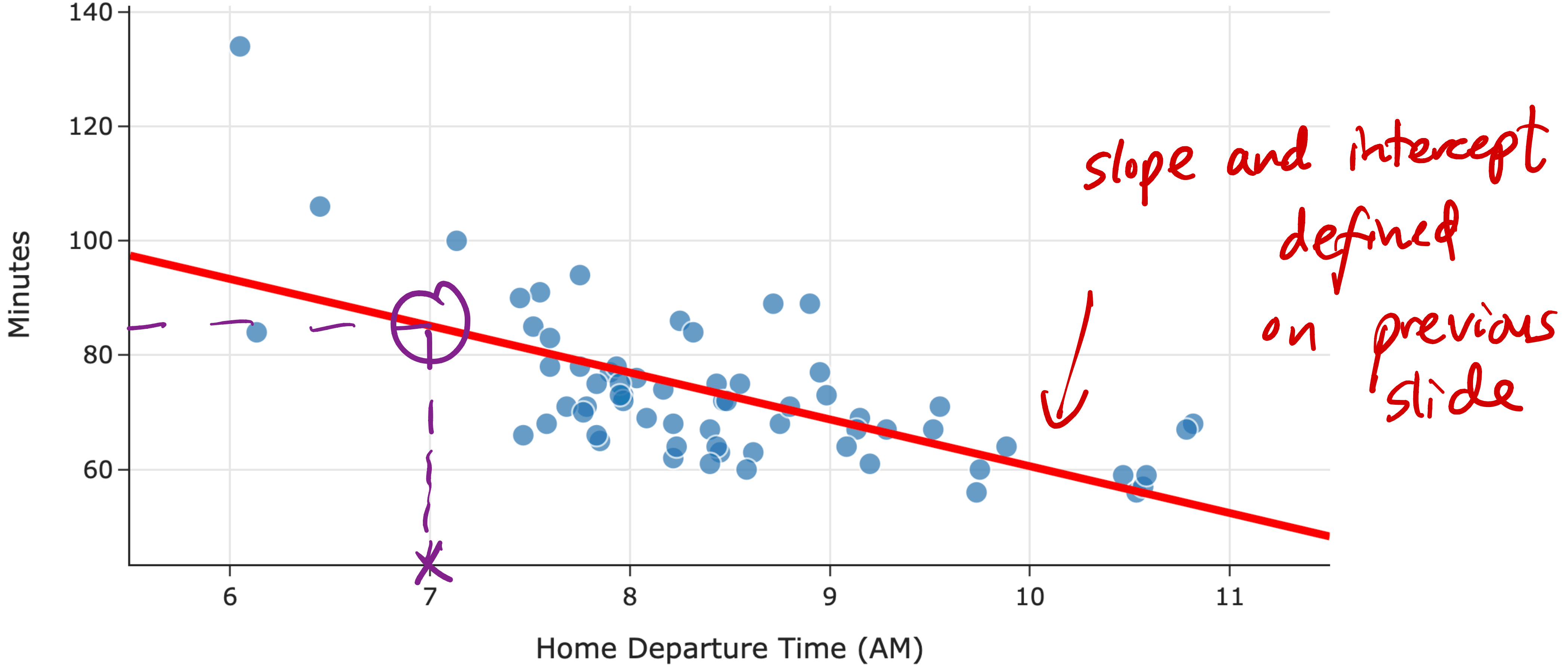
- The **least squares solutions** for the intercept w_0 and slope w_1 are:

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

"line of best fit"

- 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**." "training the model"
- To make predictions about the future, we use $\boxed{H^*(x) = w_0^* + w_1^* x}$.

Predicted Commute Time = 142.25 - 8.19 * Departure Hour

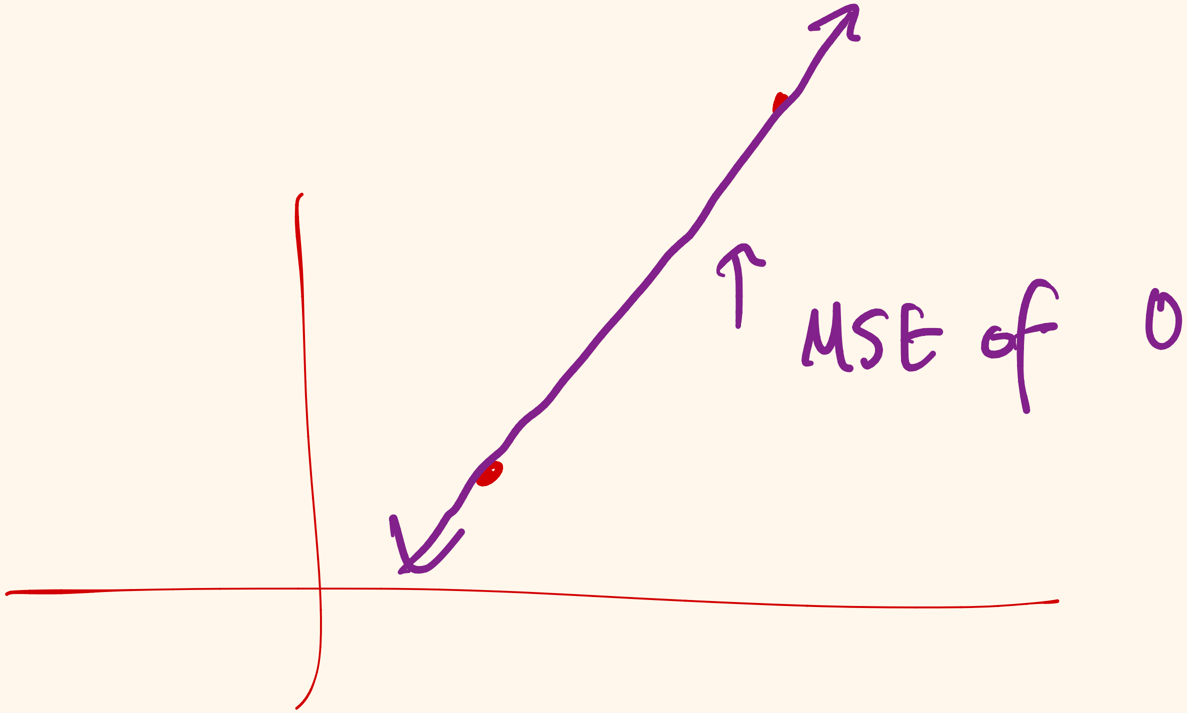


Question 🤔

Answer at practicaldsc.org/q

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. $w_0^* = -5, w_1^* = 5$

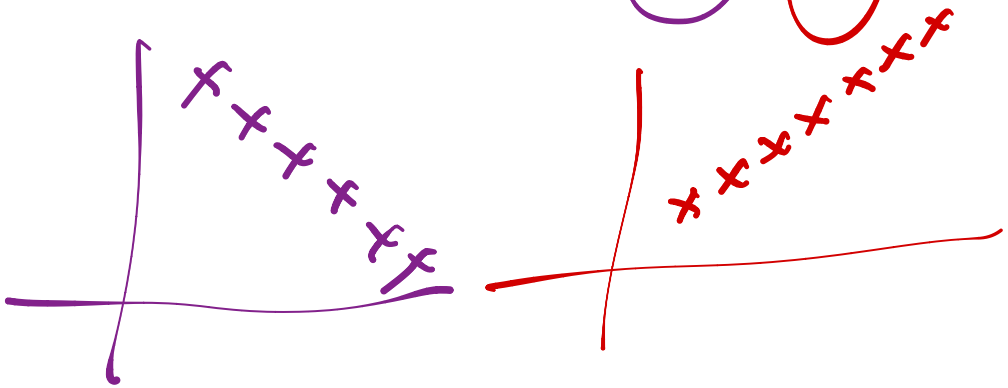


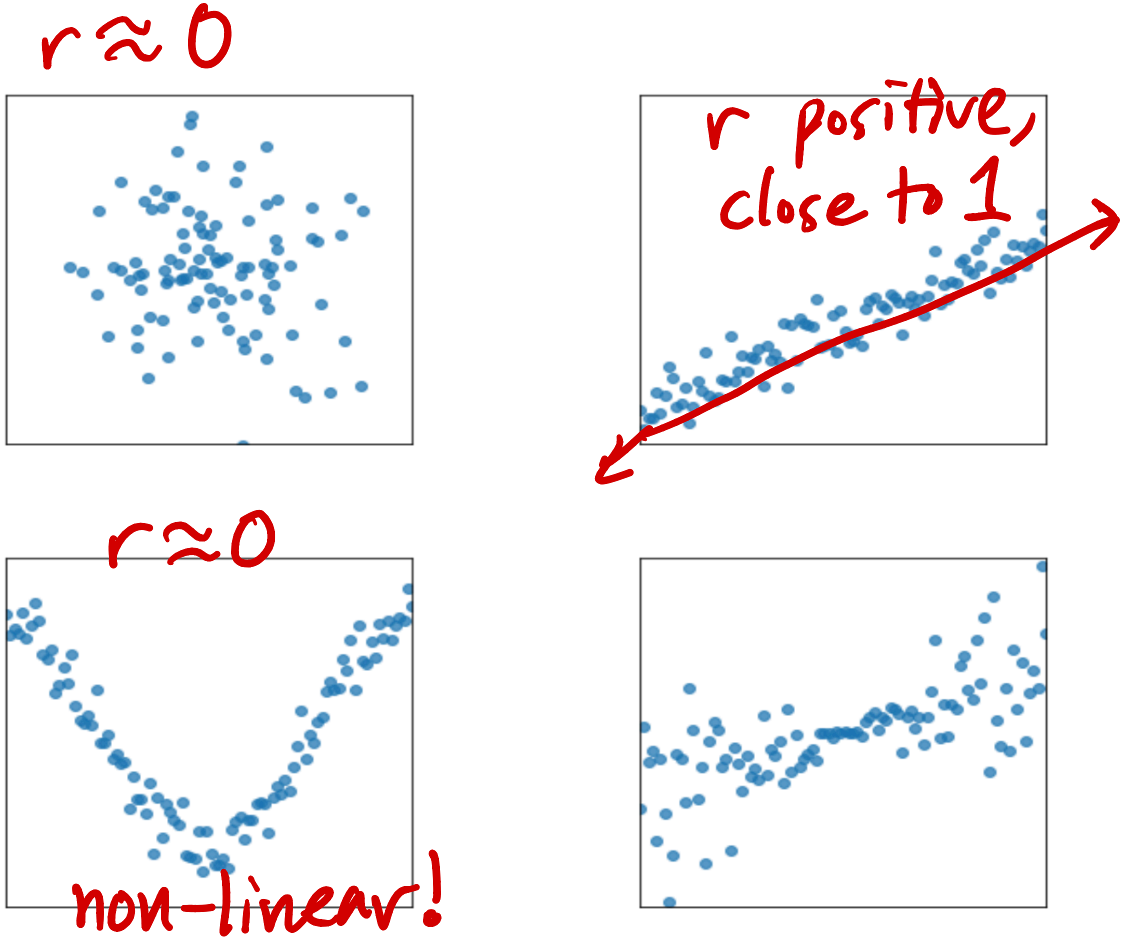
Correlation

Correlation ≠ causation

Quantifying patterns in scatter plots

- The **correlation coefficient**, r, 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.





"Pearson's correlation"

The correlation coefficient

- The correlation coefficient, r, is defined as the **average of the product of x and y, when both are *standardized***.
- Let σ_x be the standard deviation of the x_is, and $\bar{x}$ be the mean of the x_is.
- x_i standardized is $\frac{x_i - \bar{x}}{\sigma_x}$. subtract the mean, then divide by SD — "z-scores!"
- The correlation coefficient, then, is:

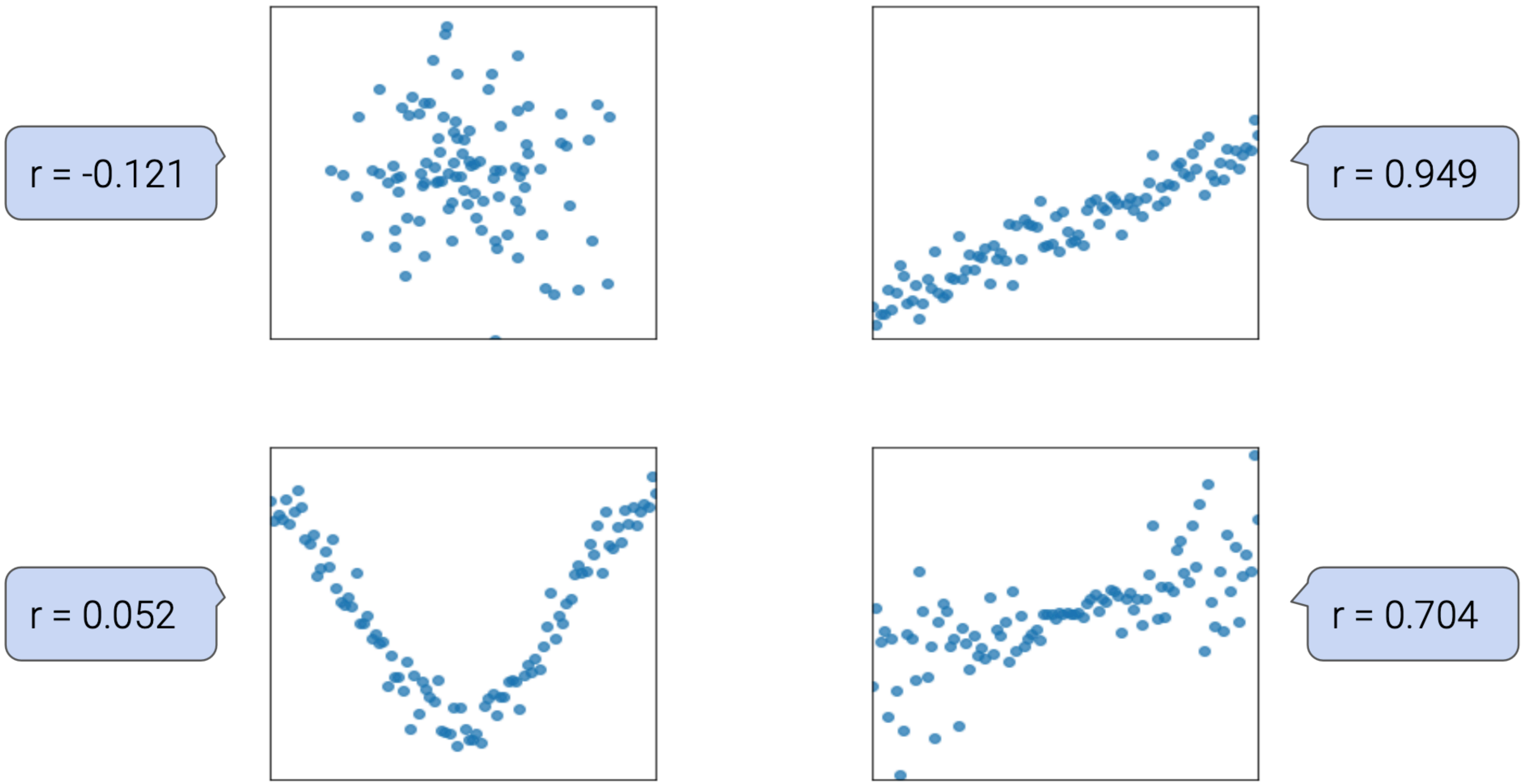
$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right) \times \left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

average

z-scored / standardized x_i

standardized y_i

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!

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

Note that w_1^* has the same sign as r!

- 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^*:

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

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

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

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

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

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

Remember,

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

so

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

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

Recap: Simple linear regression

- **Goal**: Use the modeling recipe to find the "best" simple linear hypothesis function.

1. **Model**: $H(x_i) = w_0 + w_1 x_i$.
2. **Loss function**: $L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2$.
3. **Minimize empirical risk**: $R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$.

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

- The resulting line, $H^*(x) = w_0^* + w_1^* x$, is the line that minimizes mean squared error.

It's often called the **(least squares) regression line**, and the **optimal linear predictor**.