

Lecture 3

# Comparing Loss Functions

DSC 40A, Summer 2024

# Announcements

*at midnight*

- Homework 1 is due on **Friday, August 9th**.
  - Before working on it, watch the [Walkthrough Videos](#) on problem solving and using Overleaf.
  - The [Overleaf template](#) has been released.
  - Using the Overleaf template is required for (only) Homework 2.
- Discussion *2 hrs* today, directly after lecture at 2p here.
  - Remember that, in general, groupwork worksheets are released on Tuesday and due Wednesday. *@ midnight* *Solutions will be posted on Ed*
- 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

### Average loss

- Recap: Empirical risk minimization.
- Choosing a loss function.
  - The role of outliers.
- Center and spread.
- Towards linear regression.

**Question** 🤔

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

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

# Recap: Empirical risk minimization

# Goal

We had one goal in Lecture 2: given a dataset of values from the past, **find the best constant prediction** to make.

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

**Key idea:** Different definitions of "best" give us different "best predictions."

## The modeling recipe

In Lecture 2, we made two full passes through our "modeling recipe."

1. Choose a model.

$$H(x) = h$$

2. Choose a loss function.

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

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

3. Minimize average loss to find optimal model parameters.

"mean squared error" ↗  $R_{sq}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2$

$R_{abs}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|$  ↗ "mean absolute err"

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

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

# 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:

choice of loss determines  
our empirical risk

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

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

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

## Empirical risk minimization, in general

Key idea: If  $L(y_i, h)$  is any loss function, the corresponding empirical risk is:

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

averaging

more examples:  
Walkthrough video  
toy pack

**Question** 🤔

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

**What questions do you have?**

$$(|y_1 - h| + |y_2 - h|)^2 = (|y_1 - h|^2 \dots |y_1 - h| \cdot |y_2 - h| \dots (|y_2 - h|^2))$$

Question 🤔

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

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

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

Is the following statement true, for any dataset  $y_1, y_2, \dots, y_n$  and prediction  $h$ ?

*sum then square*       $(R_{\text{abs}}(h))^2 = R_{\text{sq}}(h)$  *square first then sum*

- A. It's true for any  $h$  and any dataset.
- B. It can sometimes be true, but not in general.
- C. It's never true.

*when is*  
 $(a+b)^2 = (a^2 + b^2)$

# Choosing a loss function

## Now what?

$h^*$

- We know that, for the constant model  $H(x) = h$ , the **mean** minimizes mean **squared** error.
- We also know that, for the constant model  $H(x) = h$ , the **median** minimizes mean **absolute** error.
- How does our choice of loss function impact the resulting optimal prediction?

$$61, 72, 85, 90, 92 \quad \bar{x} = 400$$

$$61, 72, 85, 90, 292 \quad \bar{x} = 600$$

## Comparing the mean and median

- Consider our example dataset of 5 commute times.

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

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

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

- Now, the median is **still 85** but the mean is  $\frac{600}{5} = 120$
- Key idea: The mean is quite **sensitive** to outliers.

$$\frac{400}{5} + \frac{200}{5} = 120$$

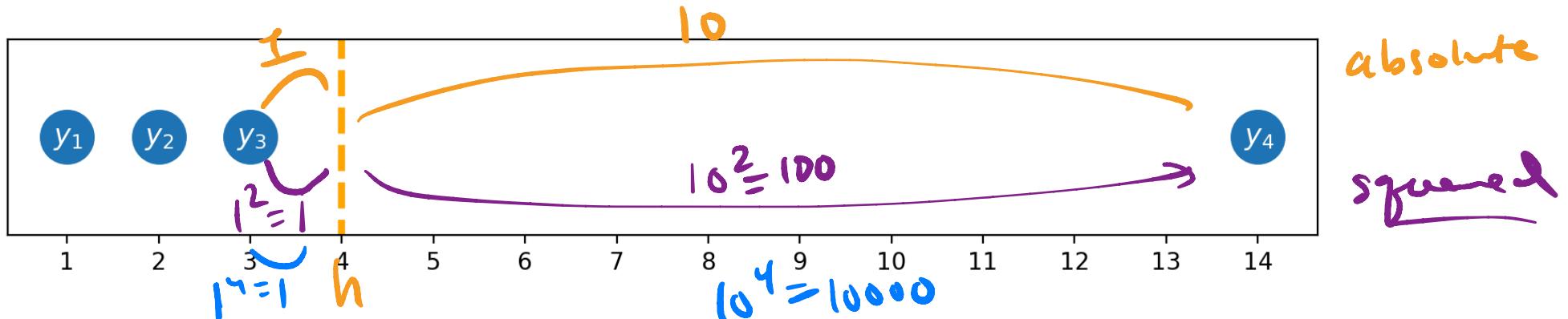
↑ old mean      ↑ extra

$$(y_i - h)^4$$

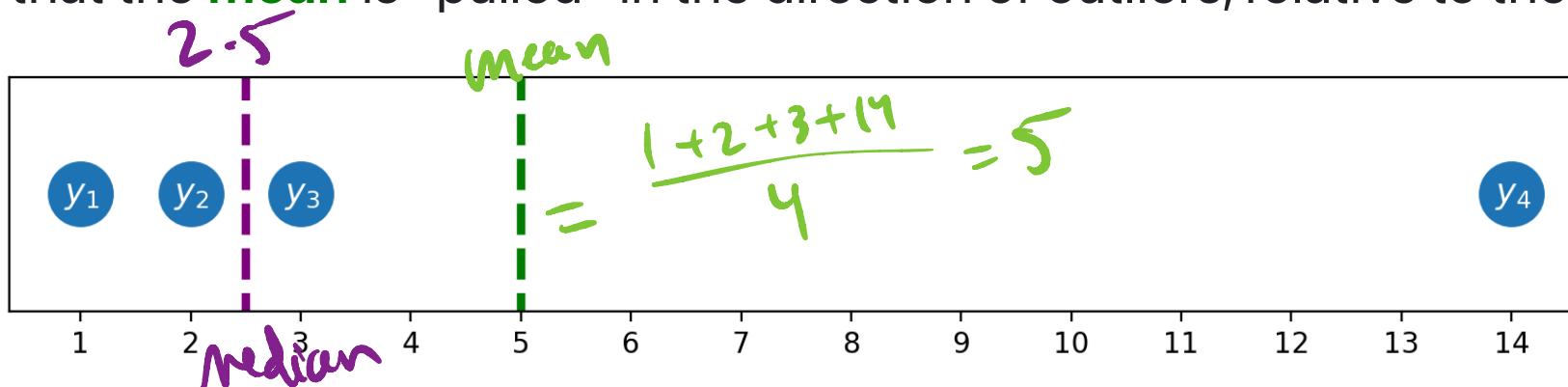
quartic mean

## Outliers

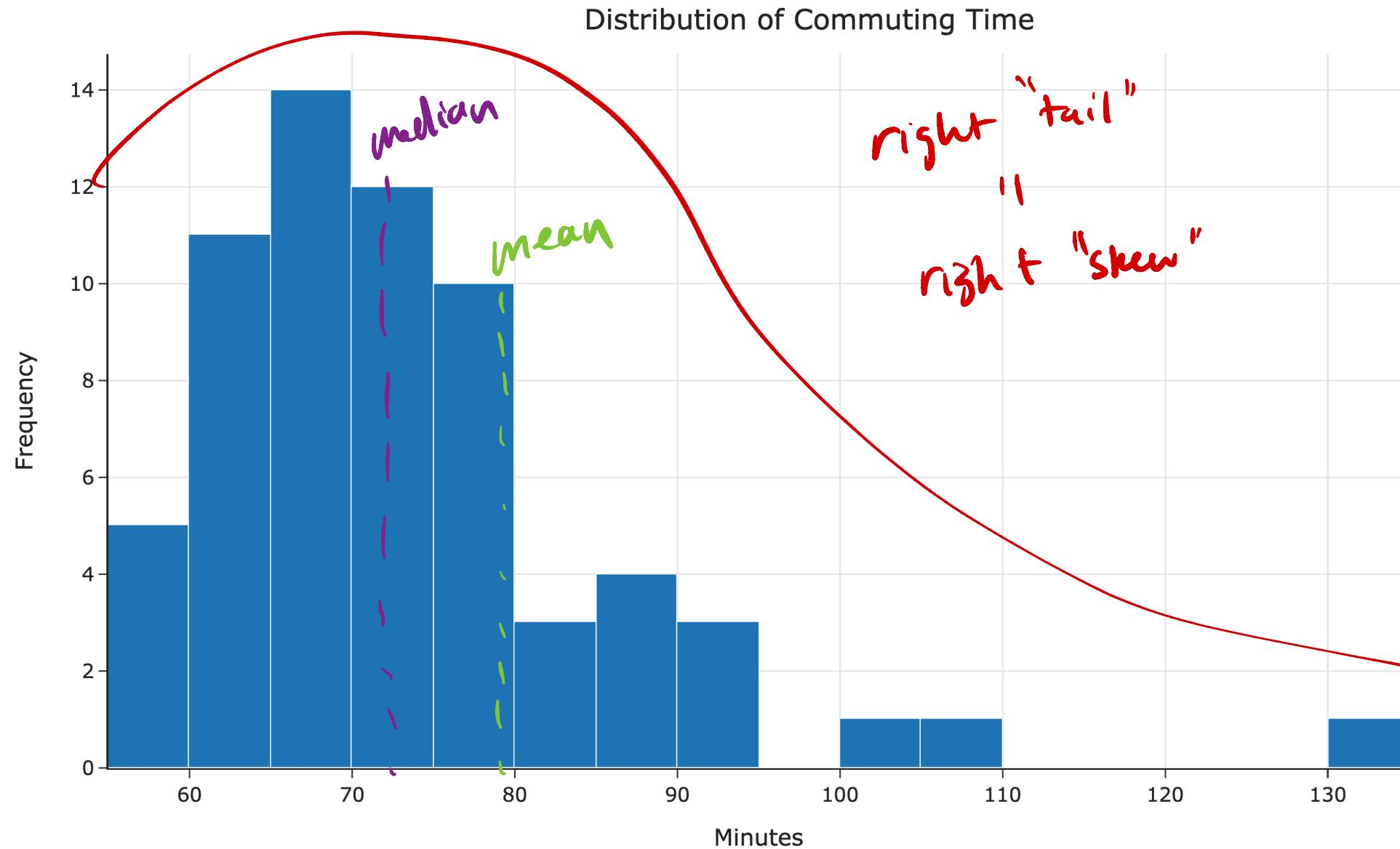
Below,  $|y_4 - h|$  is 10 times as big as  $|y_3 - h|$ , but  $(y_4 - h)^2$  is 100 times  $(y_3 - h)^2$ .



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



As a result, we say the **median** is **robust** to outliers. But the **mean** was easier to solve for.



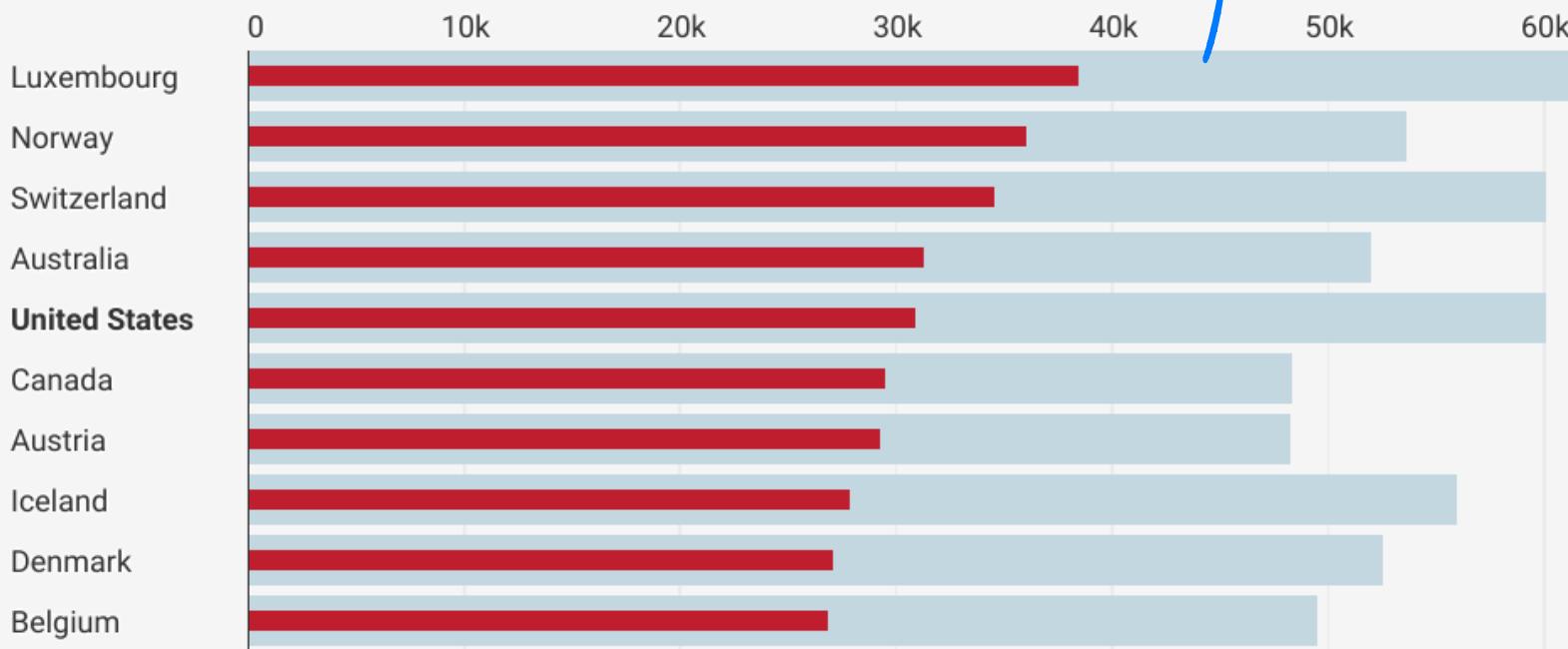
# Example: Income inequality

mean is influenced by  
large outliers

## Average vs median income

Median and mean income between 2012 and 2014 in selected OECD countries, in USD; weighted by the currencies' respective purchasing power (PPP).

Average income in USD Median income

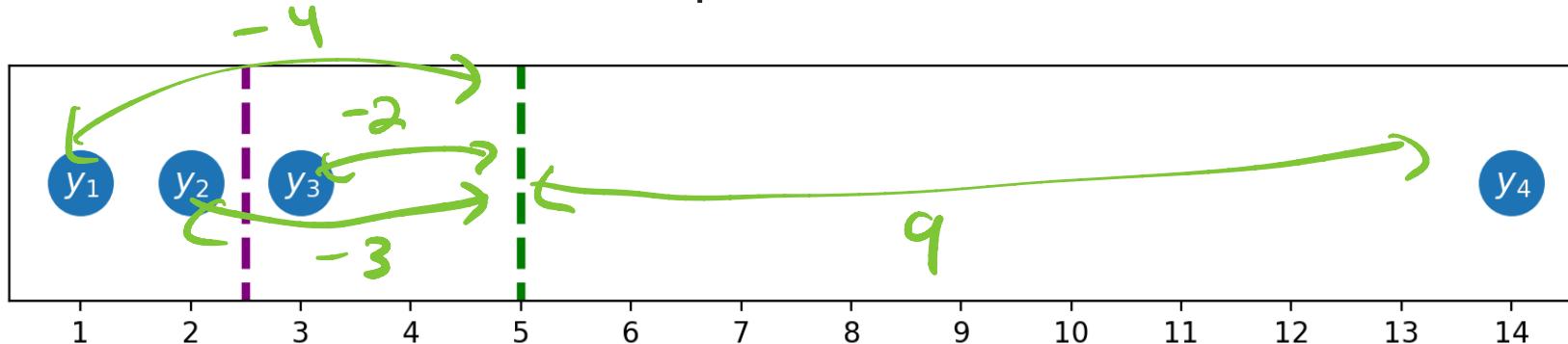


Mean: sum of distances below = sum. dist. above

## Balance points

$$\sum_{y_i < h} |y_i - h| = \sum_{y_i > h} |y_i - h|$$

Both the **mean** and **median** are "balance points" in the distribution.



- The **mean** is the point where  $\sum_{i=1}^n (y_i - h) = 0$ .
  - This appears in Homework 1!
- The **median** is the point where  $\#(y_i < h) = \#(y_i > h)$ .

Here : 2 pts left of median  $\{y_1, y_2\}$

2 points right of median  $\{y_3, y_4\}$

$$\frac{1}{n} \left[ \#(y_i < h) - \#(y_i > h) \right]$$

Why stop at squared loss?

Mean squared error →  
 $(-x)^3 < 0$  neg →

Empirical Risk, $R(h)$	Derivative of Empirical Risk, $\frac{d}{dh} R(h)$	Minimizer
$\frac{1}{n} \sum_{i=1}^n  y_i - h $	$\frac{1}{n} \left( \sum_{y_i < h} 1 - \sum_{y_i > h} 1 \right)$	median
$\frac{1}{n} \sum_{i=1}^n (y_i - h)^2$	$\frac{-2}{n} \left[ \sum_{i=1}^n (y_i - h) \right]$ set = 0 to find mean	mean
$\frac{1}{n} \sum_{i=1}^n  y_i - h ^3$		???
$\frac{1}{n} \sum_{i=1}^n (y_i - h)^4$		???
$\frac{1}{n} \sum_{i=1}^n (y_i - h)^{100}$		???
...	...	...

## Generalized $L_p$ loss

For any  $p \geq 1$ , define the  $L_p$  loss as follows:

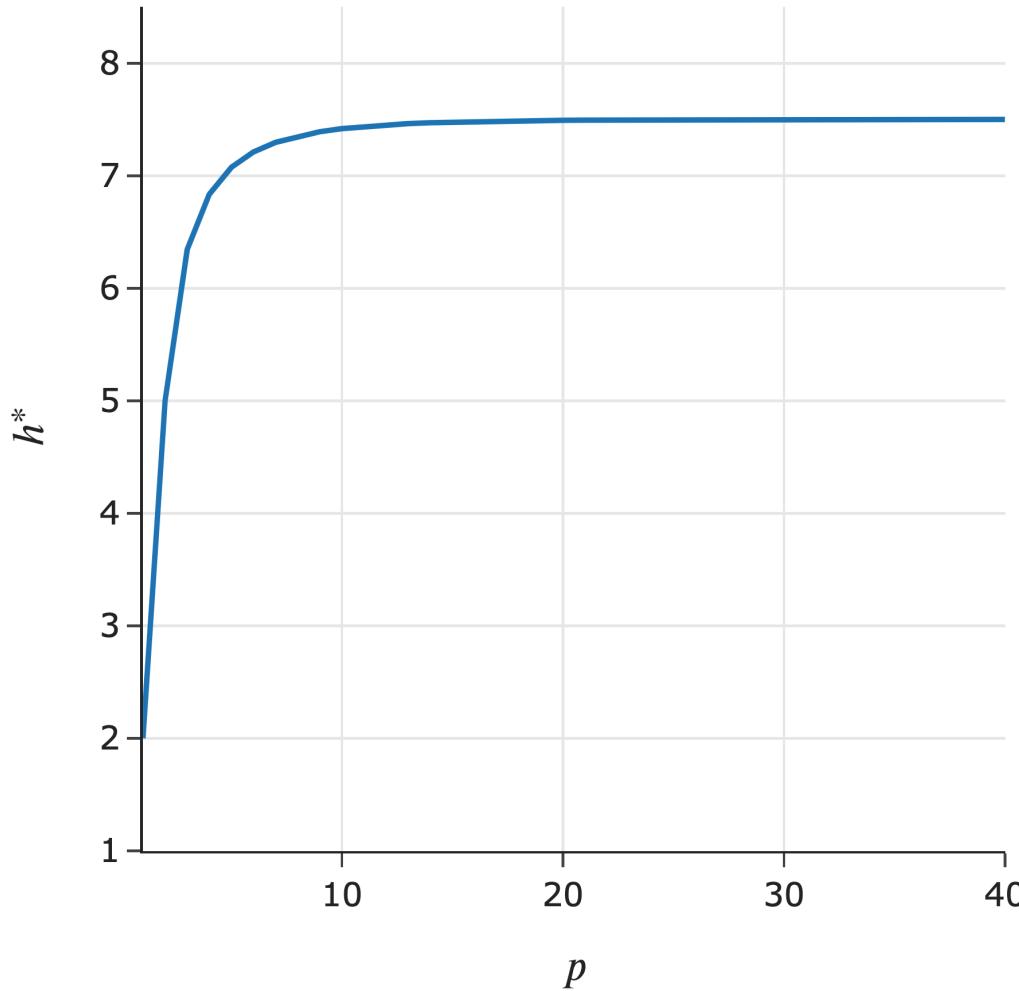
$$L_p(y_i, h) = |y_i - h|^p$$

The corresponding empirical risk is:

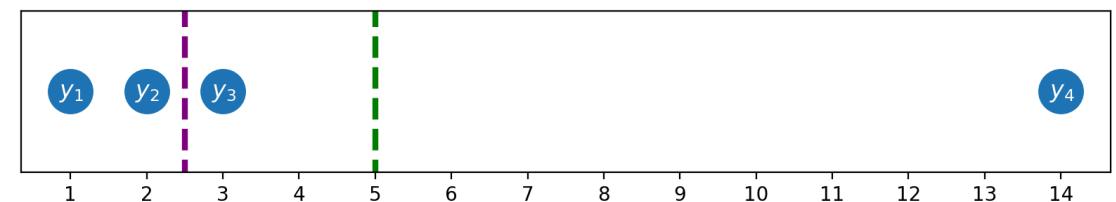
$$R_p(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|^p$$

- When  $p = 1$ ,  $h^* = \text{Median}(y_1, y_2, \dots, y_n)$ .
- When  $p = 2$ ,  $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$ .
- What about when  $p = 3$ ?
- What about when  $p \rightarrow \infty$ ?

What value does  $h^*$  approach, as  $p \rightarrow \infty$ ?



Consider the dataset 1, 2, 3, 14:



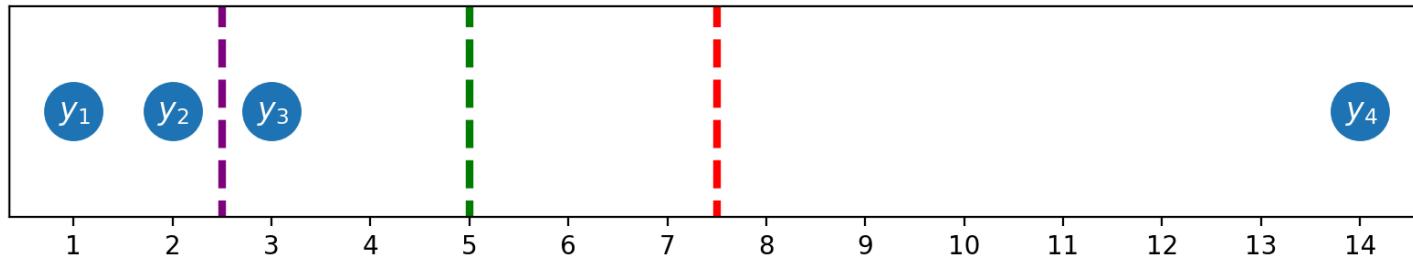
On the left:

- The  $x$ -axis is  $p$ .
- The  $y$ -axis is  $h^*$ , the optimal constant prediction for  $L_p$  loss:

$$h^* = \operatorname{argmin}_h \frac{1}{n} \sum_{i=1}^n |y_i - h|^p$$

## The *midrange* minimizes average $L_\infty$ loss!

On the previous slide, we saw that as  $p \rightarrow \infty$ , the minimizer of mean  $L_p$  loss approached the **midpoint of the minimum and maximum values in the dataset**, or the **midrange**.



- As  $p \rightarrow \infty$ ,  $R_p(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|^p$  minimizes the "worst case" distance from any data point". (Read more [here](#)).
- If your measure of "good" is "not far from any one data point", then the midrange is the best prediction.

## Another example: 0-1 loss

Consider, for example, the **0-1 loss**:

$$L_{0,1}(y_i, h) = \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$$

The corresponding empirical risk is:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^n L_{0,1}(y_i, h)$$

## Question 🤔

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

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

Suppose  $y_1, y_2, \dots, y_n$  are all unique. What is  $R_{0,1}(y_1)$ ?

- A. 0.
- B.  $\frac{1}{n}$ .
- C.  $\frac{n-1}{n}$ .
- D. 1.

## Minimizing empirical risk for 0-1 loss

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

## Summary: Choosing a loss function

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

Loss	Minimizer	Always Unique?	Robust to Outliers?	Differentiable?
$L_{\text{sq}}$	mean	yes	no	yes
$L_{\text{abs}}$	median	no	yes	no
$L_{\infty}$	midrange	yes	no	no
$L_{0,1}$	mode	no	yes	no

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

# Center and spread

## What does it mean?

- The general form of empirical risk, for any loss function  $L(y_i, h)$ , is:

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

- As we just saw, the input  $h^*$  that minimizes  $R(h)$  is some measure of the **center** of the dataset.
  - Examples include the mean ( $L_{\text{sq}}$ ), median ( $L_{\text{abs}}$ ), and mode ( $L_{0,1}$ ).
- The minimum output,  $R(h^*)$ , represents some measure of the **spread**, or variation, in the dataset.

## Squared loss

- The empirical risk for squared loss, i.e. mean squared error, is:

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

- $R_{\text{sq}}(h)$  is minimized when  $h^* = \text{Mean}(y_1, y_2, \dots, y_n)$ .
- Therefore, the minimum value of  $R_{\text{sq}}(h)$  is:

$$\begin{aligned} R_{\text{sq}}(h^*) &= R_{\text{sq}}(\text{Mean}(y_1, y_2, \dots, y_n)) \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \text{Mean}(y_1, y_2, \dots, y_n))^2 \end{aligned}$$

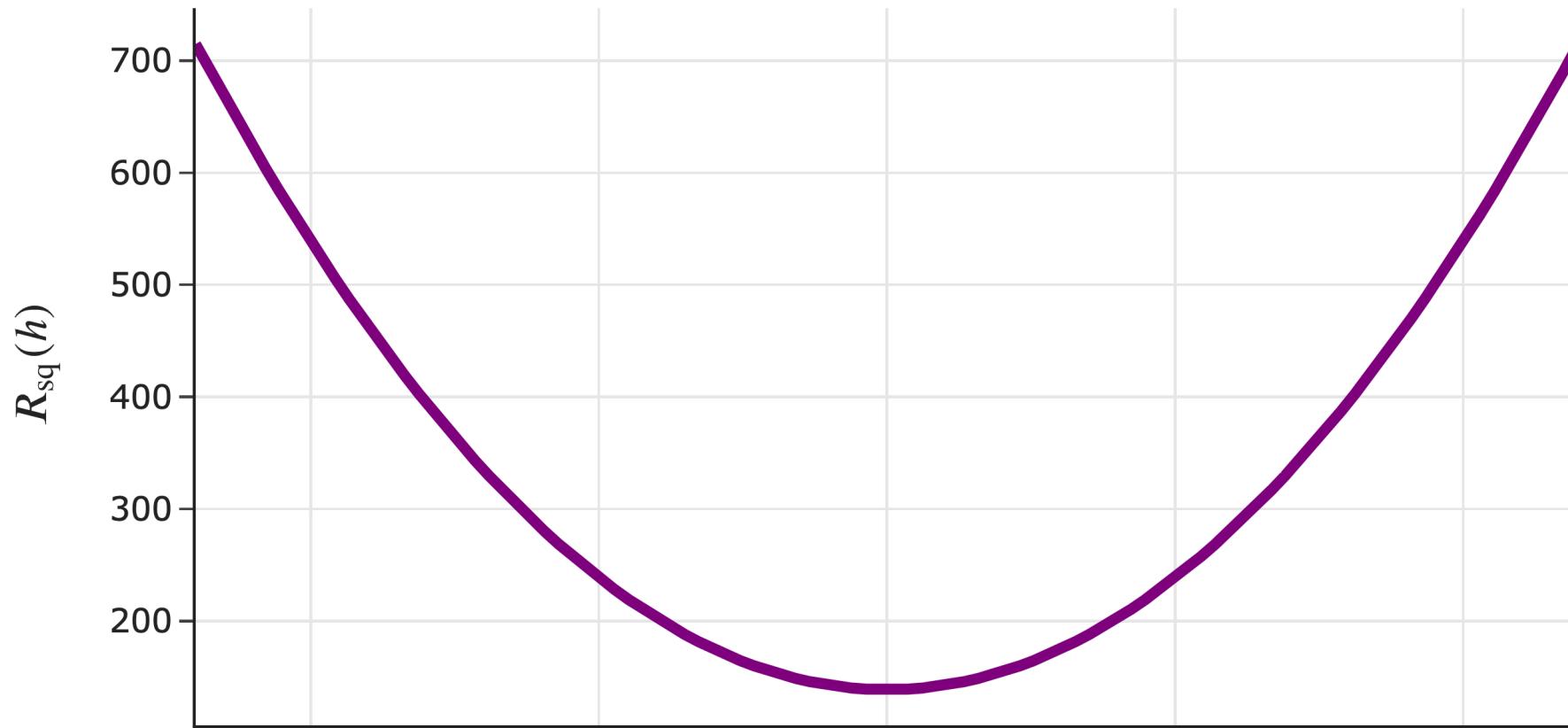
# Variance

- The minimum value of  $R_{\text{sq}}(h)$  is the mean squared deviation from the mean, more commonly known as the **variance**.

$$\text{Variance}(y_1, y_2, \dots, y_n) = \frac{1}{n} \sum_{i=1}^n (y_i - \text{Mean}(y_1, y_2, \dots, y_n))^2$$

- It measures the squared distance of each data point from the mean, on average.
- Its square root is called the **standard deviation**.

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



## Absolute loss

- The empirical risk for absolute loss, i.e. mean absolute error, is:

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

- $R_{\text{abs}}(h)$  is minimized when  $h^* = \text{Median}(y_1, y_2, \dots, y_n)$ .
- Therefore, the minimum value of  $R_{\text{abs}}(h)$  is:

$$\begin{aligned} R_{\text{abs}}(h^*) &= \frac{1}{n} \sum_{i=1}^n |y_i - h| \\ &= R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - \text{Median}(y_1, y_2, \dots, y_n)| \end{aligned}$$

## Mean absolute deviation from the median

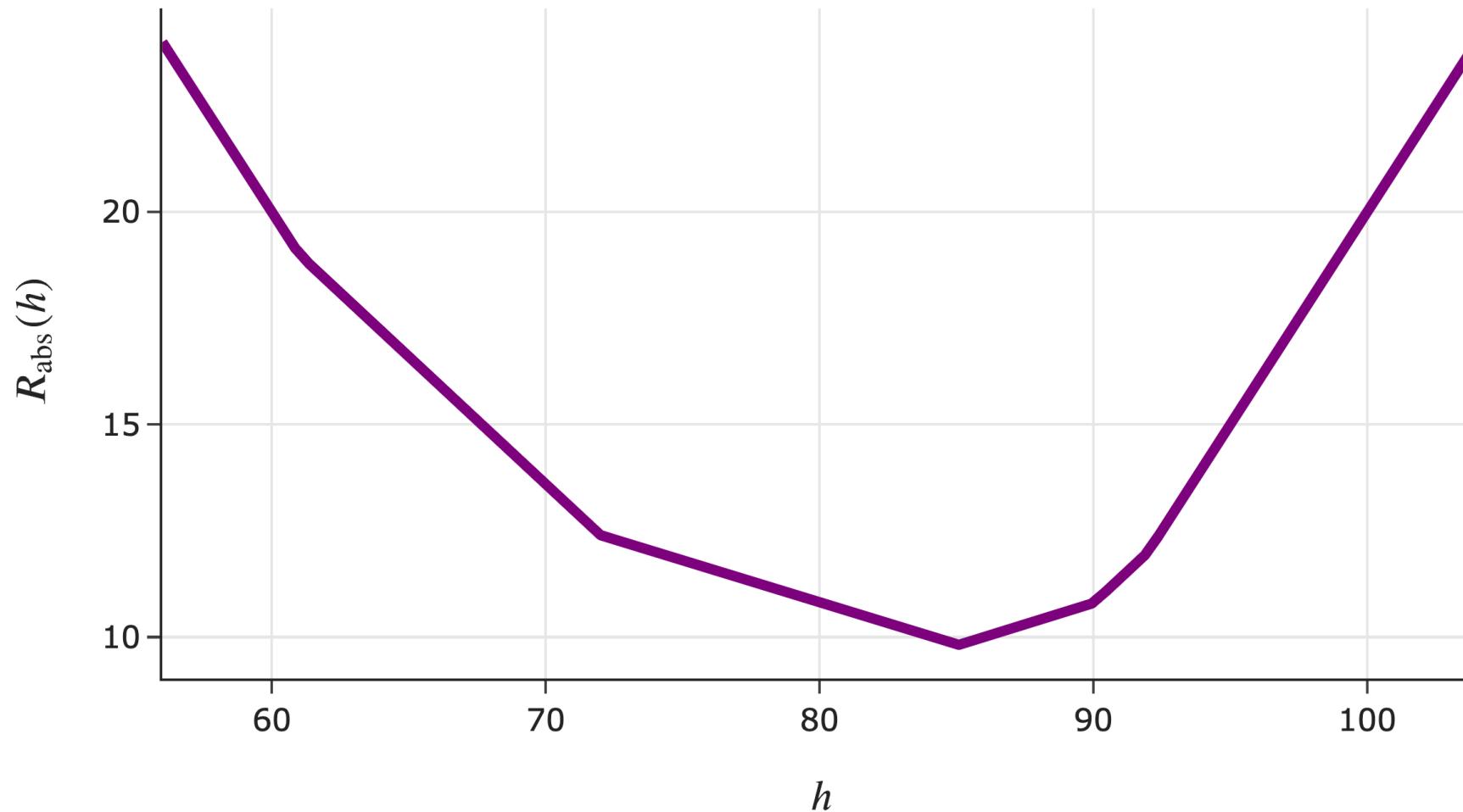
- The minimum value of  $R_{\text{abs}}(h)$  is the **mean absolute deviation from the median**.

$$\text{MAD from the median}(y_1, y_2, \dots, y_n) = \frac{1}{n} \sum_{i=1}^n |y_i - \text{Median}(y_1, y_2, \dots, y_n)|$$

- It measures how far each data point is from the median, on average.
- **Example:** What's the MAD from the median in the dataset 2, 3, 3, 4, 5?

## Mean absolute deviation from the median

$$R_{\text{abs}}(h) = \frac{1}{5}(|72 - h| + |90 - h| + |61 - h| + |85 - h| + |92 - h|)$$



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

## A poor way to measure spread

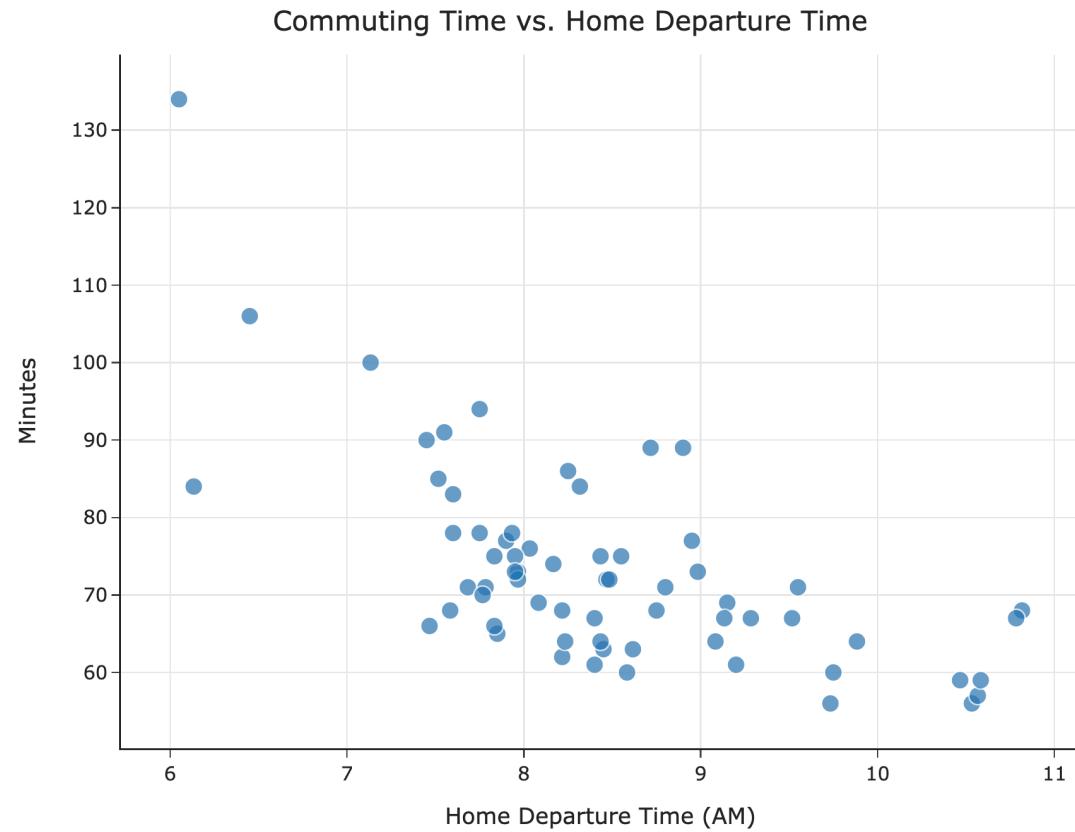
- The minimum value of  $R_{0,1}(h)$  is the proportion of data points not equal to the mode.
- A higher value means less of the data is clustered at the mode.
- Just as the mode is a very basic way of measuring the center of the data,  $R_{0,1}(h^*)$  is a very basic and uninformative way of measuring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

# What's next?

# 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.
2. Choose a loss function.
3. Minimize average loss to find optimal model parameters.

