Lecture 4

Comparing Loss Functions

DSC 40A, Fall 2024

Announcements

- Homework 1 will be relased by tomorrow and will be due on **Friday, October 11th**.
  - 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).
- Remember that in, general, groupwork worksheets are released on Sunday and due Monday.
- 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: Empirical risk minimization.
- Choosing a loss function.
  - The role of outliers.
- Center and spread.
- Towards linear regression.

Question 🤔

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

Recap: Empirical risk minimization

Goal

We had one goal in Lectures 2 and 3: given a dataset of values from the past, **find the best constant prediction** to make.

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

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

mean median

both the "best", under diff conditions.

The modeling recipe

In Lectures 2 and 3, we made two full passes through our "modeling recipe."

1. Choose a model.

$$H(x) = h$$

actual

prediction

2. Choose a loss function.

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

3. Minimize average loss to find optimal model parameters.

$$h* = \text{mean}(y_1, \ldots, y_n) \qquad h* = \text{median}(y_1, \ldots, y_n)$$

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

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:

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

- 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:

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

Question 🤔

Answer at q.dsc40a.com

What questions do you have?

Question 🤔

Answer at q.dsc40a.com

$$[R_{abs}(h)]^2 = \left(\frac{1}{n}(|y_1 - h| + |y_2 - h| + \ldots |y_n - h|)\right)^2$$

$$= \frac{1}{n^2}\left((y_1 - h)^2 + |y_1 - h| \cdot |y_2 - h| + \ldots\right) \neq R_{sq}(h)$$

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

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

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

$$\left(R_{\text{abs}}(h)\right)^2 = R_{\text{sq}}(h)$$

- ~~A. It's true for any h and any dataset.~~
- B. It's true for at least one h for any dataset, but not in general.
- ~~C. It's never true.~~

Choosing a loss function

Now what?

- 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?**

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?

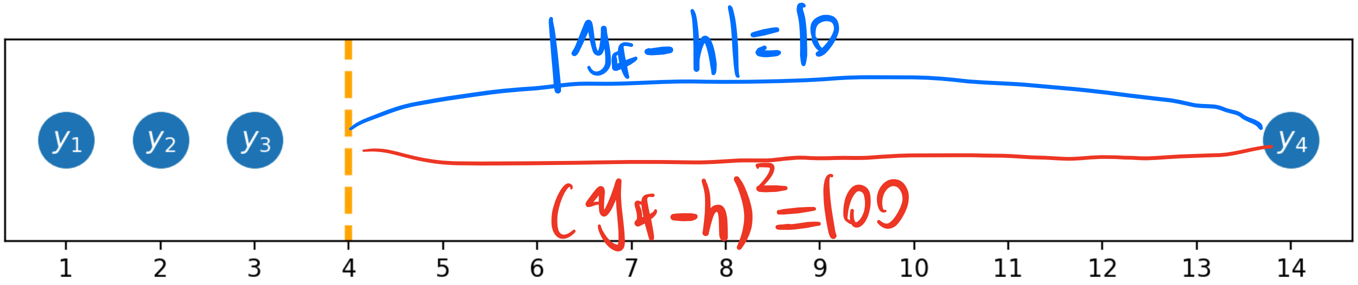
$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 292$$

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

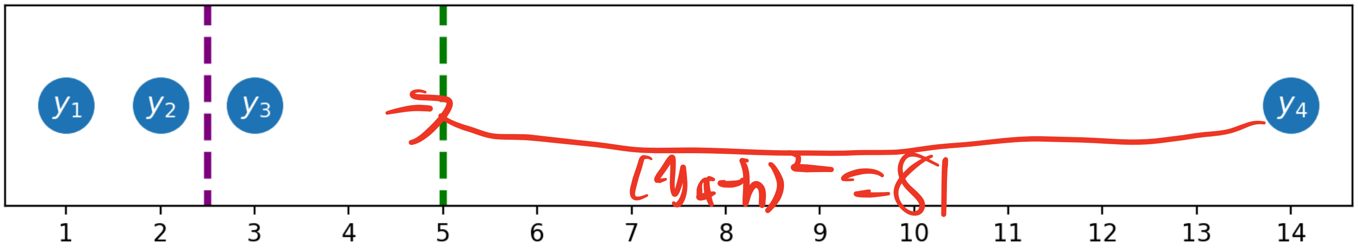
5 data point, 200 added to the dataset

Outliers

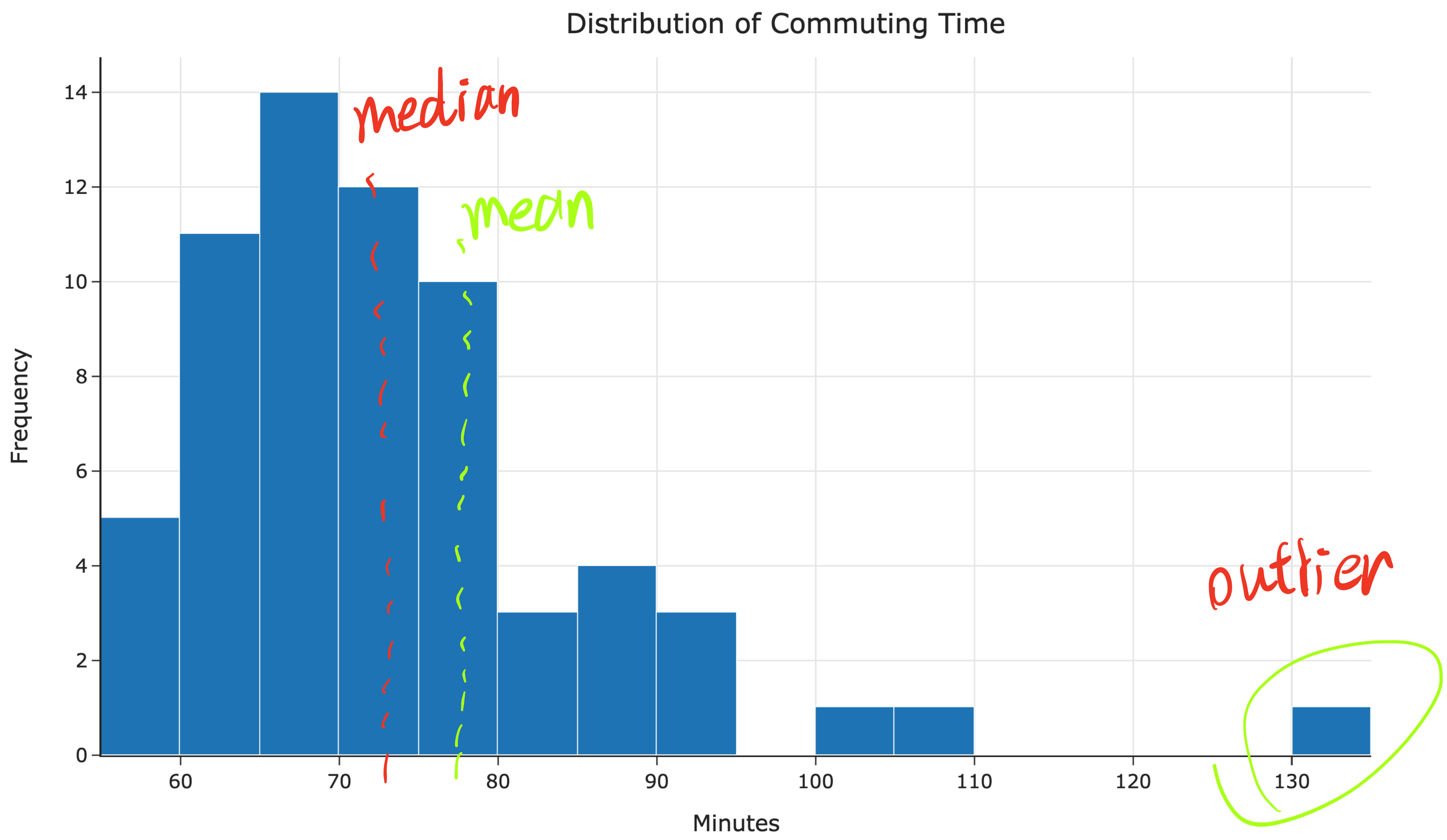
Below, $|y_4 - h|$ is 10 times as big as $|y_3 - h|$, but $(y_4 - h)^2$ is 100 times $(y_3 - h)^2$.



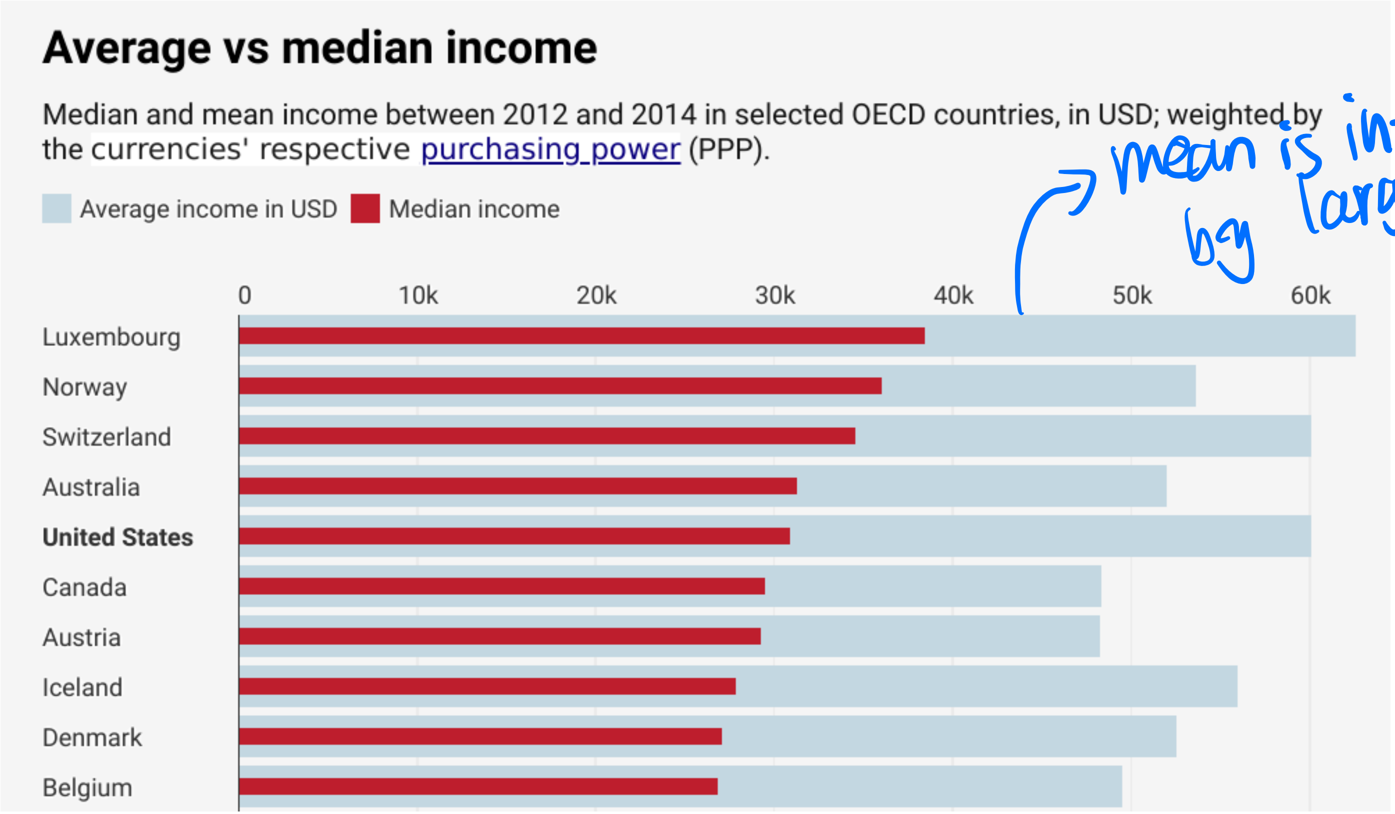
The result is that the **mean** is "pulled" in the direction of outliers, relative to the **median**.



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


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

Example: Income inequality

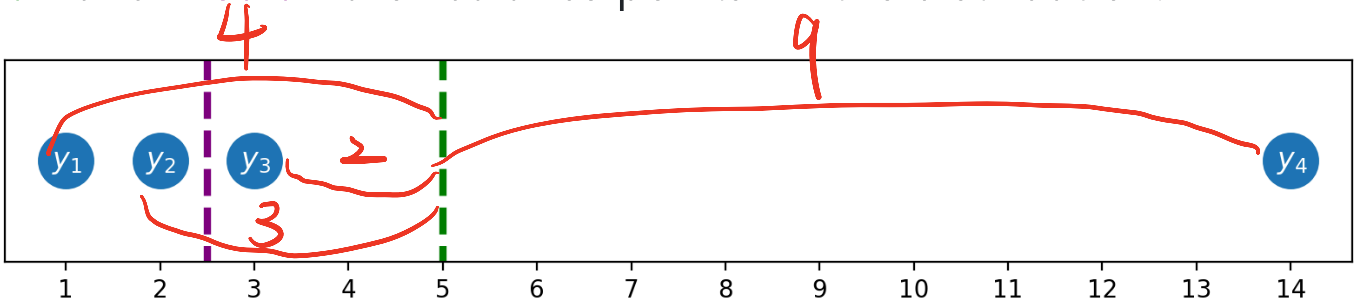


mean is influenced by large outlier

For mean: sum of distance below = sum of distance above

Balance points

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



- The mean is the point where $\sum_{i=1}^{n} (y_i - h) = 0$.

number of

- The median is the point where $\# \, (y_i < h) = \# \, (y_i > h)$.

2 points to the left of median
2 points to the right of median

Why stop at squared loss?

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

if using odd exponent need absolute value otherwise loss will be negative

Generalized L_p loss

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

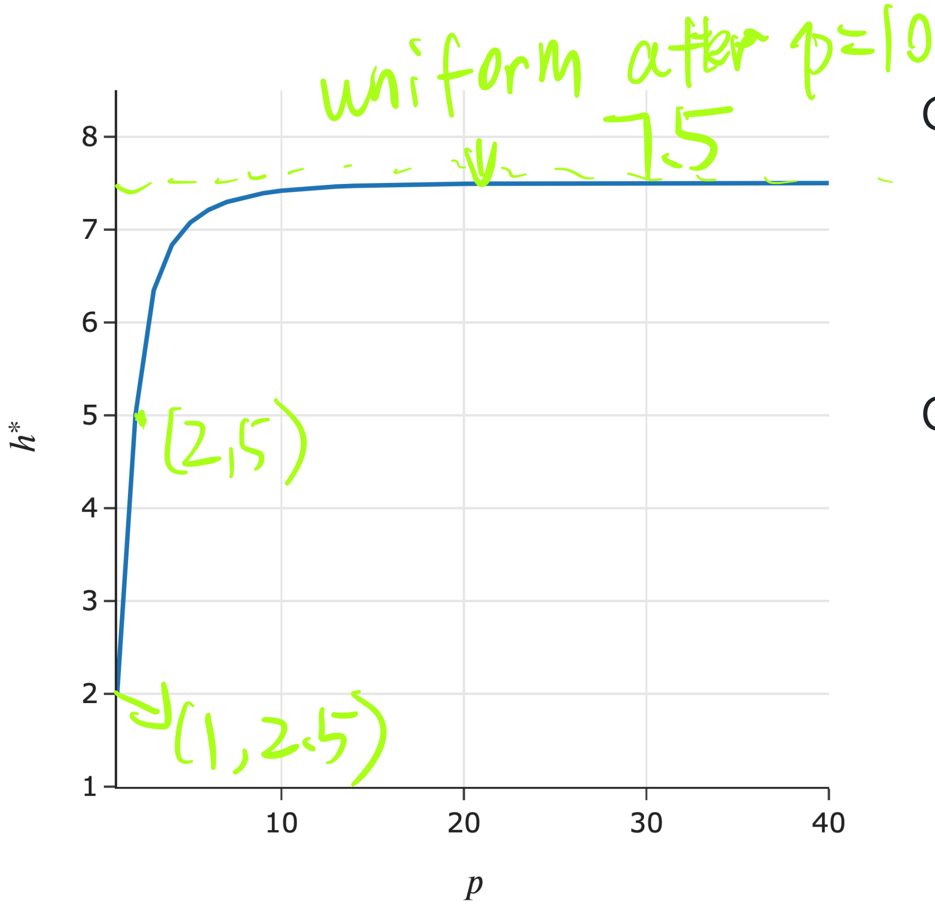
$$L_p(y_i, h) = |y_i - h|^p$$

The corresponding empirical risk is:

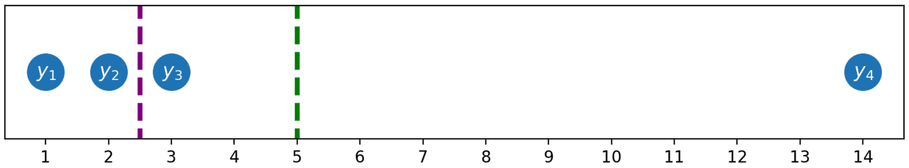
$$R_p(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|^p$$

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

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



Consider the dataset $1, 2, 3, 14$:



On the left:

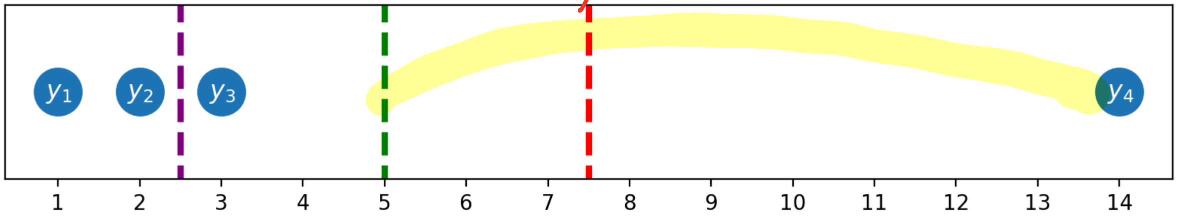
- The x-axis is p.
- The y-axis is h^*, the optimal constant prediction for L_p loss:

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

$\frac{14+1}{2} = 7.5$

The *midrange* minimizes average L_∞ loss!

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



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

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

$\Rightarrow$ proportion of points NOT equal to h

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

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

$R_{0,1}(y_1)$ = proportion of points $\neq y_i$

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

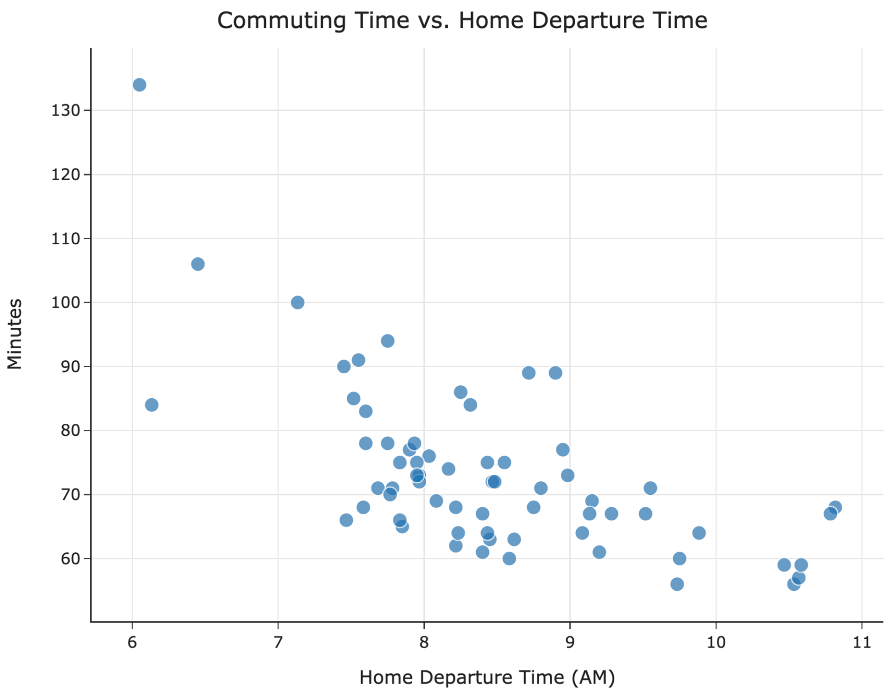
Key idea: Different loss functions lead to different best predictions, h^*!

Loss	Minimizer	Always Unique?	Robust to Outliers?	Differentiable?
L_{sq}	mean	yes ✅	no ❌	yes ✅
L_{abs}	median	no ❌	yes ✅	no ❌
L_{∞}	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.

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**, $H(x) = w_0 + w_1 x$.
- 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.