Lecture 2

Empirical Risk Minimization

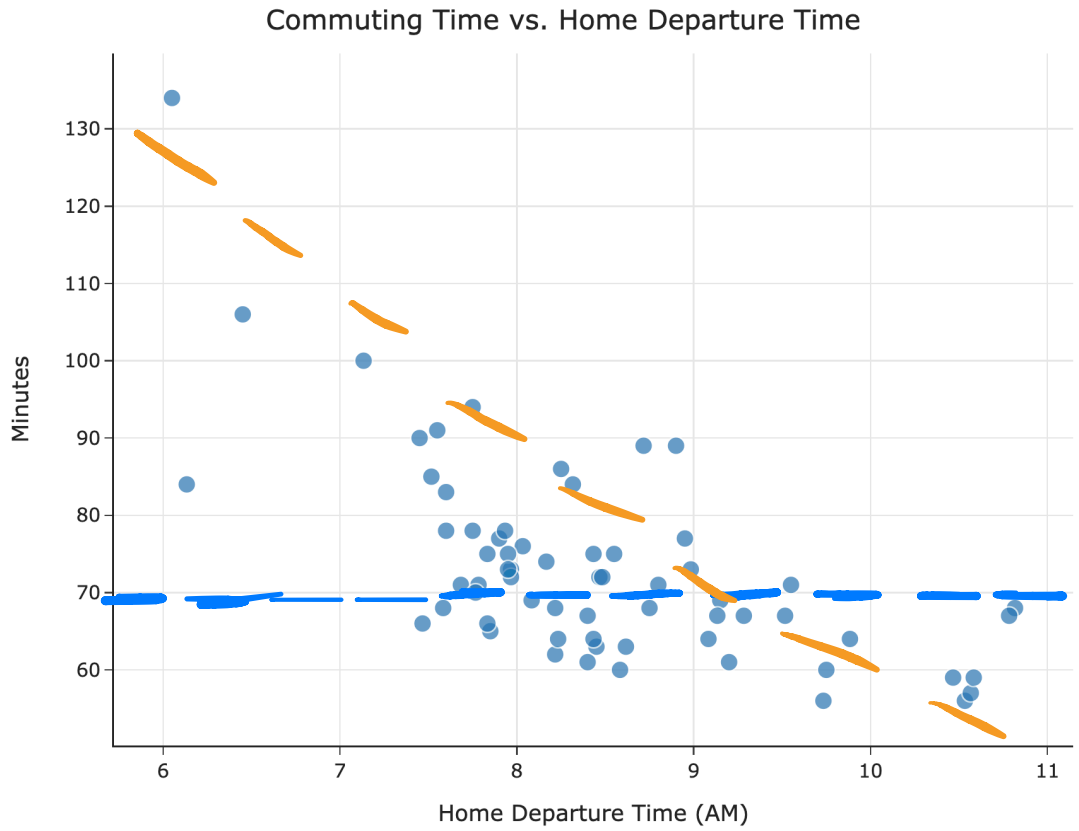
DSC 40A, Summer I 2024

Agenda

- Recap: Mean squared error.
- Minimizing mean squared error.
- Another loss function.
- Minimizing mean absolute error.
- A practice exam problem.

Recap: Mean squared error

Overview



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

Mean squared error

(actual − predicted)^2

- 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:

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

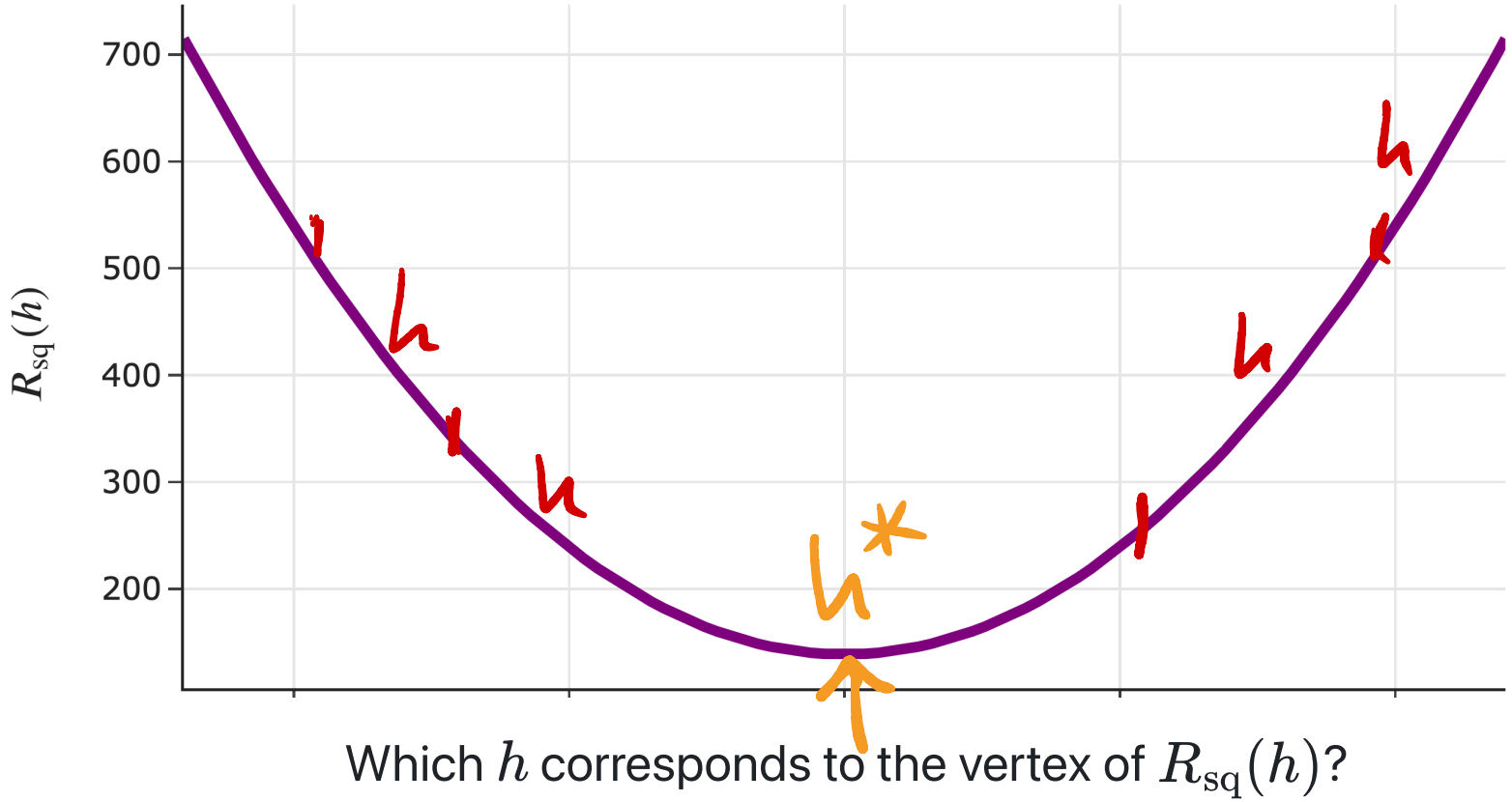
- For example, if we predict $h = 100$, then:

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

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

Visualizing mean squared error

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



Which h corresponds to the vertex of $R_{\text{sq}}(h)$?

The best prediction

- Suppose we collect n commute times, $y_1, y_2, \ldots, y_n$.
- The mean squared error of the prediction h is:

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

- We want the **best** prediction, h^*.
- The smaller $R_{\text{sq}}(h)$ is, the better h is.
- **Goal**: Find the h that minimizes $R_{\text{sq}}(h)$.

  The resulting h will be called h^*.
- **How do we find h^*?** Calculus!? :)

Minimizing mean squared error

Minimizing using calculus

We'd like to minimize:

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

In order to minimize $R_{\text{sq}}(h)$, we:

1. take its derivative with respect to h,
2. set it equal to 0,
3. solve for the resulting h^*, and
4. perform a second derivative test to ensure we found a minimum.

Step 0: The derivative of $(y_i - h)^2$

- Remember from calculus that:
  - if $c(x) = a(x) + b(x)$, then
  - $\frac{d}{dx}c(x) = \frac{d}{dx}a(x) + \frac{d}{dx}b(x)$.
- This is relevant because $R_{\text{sq}}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$ involves the sum of n individual terms, each of which involve h.
- So, to take the derivative of $R_{\text{sq}}(h)$, we'll first need to find the derivative of $(y_i - h)^2$.

$\frac{d}{dh}(y_i - h)^2 =$ $2(y_i - h)\frac{d}{dh}(y_i - h)$

$= 2(y_i - h)(-1) = -2(y_i - h)$

Question 🤔

$$\frac{d}{dh}(y_i - h)^2 = -2(y_i - h)$$

Pause the video and try to answer the question...

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

Which of the following is $\frac{d}{dh}R_{\text{sq}}(h)$?

- A. 0
- B. $\sum_{i=1}^{n} y_i$
- C. $\frac{1}{n}\sum_{i=1}^{n}(y_i - h)$
- D. $\frac{2}{n}\sum_{i=1}^{n}(y_i - h)$
- ✓ E. $-\frac{2}{n}\sum_{i=1}^{n}(y_i - h)$

if $c(x) = K \cdot a(x)$

Constant

$$\frac{d}{dx}c(x) = K\frac{d}{dx}a(x)$$

Step 1: The derivative of $R_{\text{sq}}(h)$

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

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

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

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

Steps 2 and 3: Set to 0 and solve for the minimizer, h^*

$$\frac{d}{dh} R_{sq}(h) = \frac{-2}{n} \sum_{i=1}^{n} (y_i - h) = 0$$

$$\sum_{i=1}^{n} (y_i - h) = 0$$

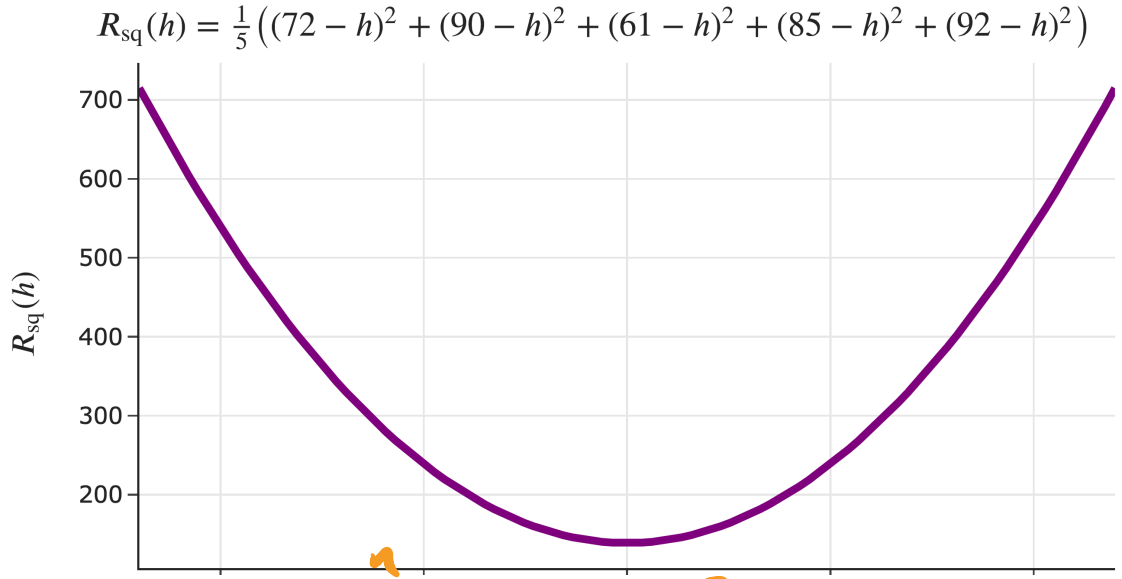
$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} h = 0$$

$$\sum_{i=1}^{n} y_i - nh = 0$$

$$\sum_{i=1}^{n} h = (h + h + \cdots + h) = nh \quad (n \text{ times})$$

$$h^* = \frac{\sum_{i=1}^{n} y_i}{n} = \text{mean}(y_1, y_2, \ldots y_n)$$

Step 4: Second derivative test



We already saw that $R_{\text{sq}}(h)$ is **convex**, i.e. that it opens upwards, so the h^* we found must be a minimum, not a maximum.

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

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

$$\sum_{i=1}^{n}(-1) \quad (-1)+(-1)+\cdots+(-1) = -n$$

$$\frac{d^2}{dh^2}R_{sq}(h) = \frac{-2}{n}\sum_{i=1}^{n}(-1) = \frac{2}{n}(n) = 2 > 0 \qquad R_{sq}(h)$$

The mean minimizes mean squared error!

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

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

- 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, $H(x) = h$, and
  - the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$.

Aside: Notation

Another way of writing

$$h^* \text{ is the value of } h \text{ that minimizes } \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2$$

is

$$h^* = \underset{h}{\text{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2 \right)$$

h^* is the solution to an **optimization problem**.

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.

$H(x) = h$

Regression $H(x) = w_0 + w_1 x$

2. Choose a loss function.

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

$|y_i - h|$

3. Minimize average loss to find optimal model parameters.

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

Question 🤔

Take a moment to pause and reflect...

If you have any questions post online to our class forms/Q&A site.

Course staff will answer them ASAP!

Another loss function

Another loss function

- Last lecture, 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.

$$e_i = y_i - H(x_i) \rightarrow h$$

- 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**.

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

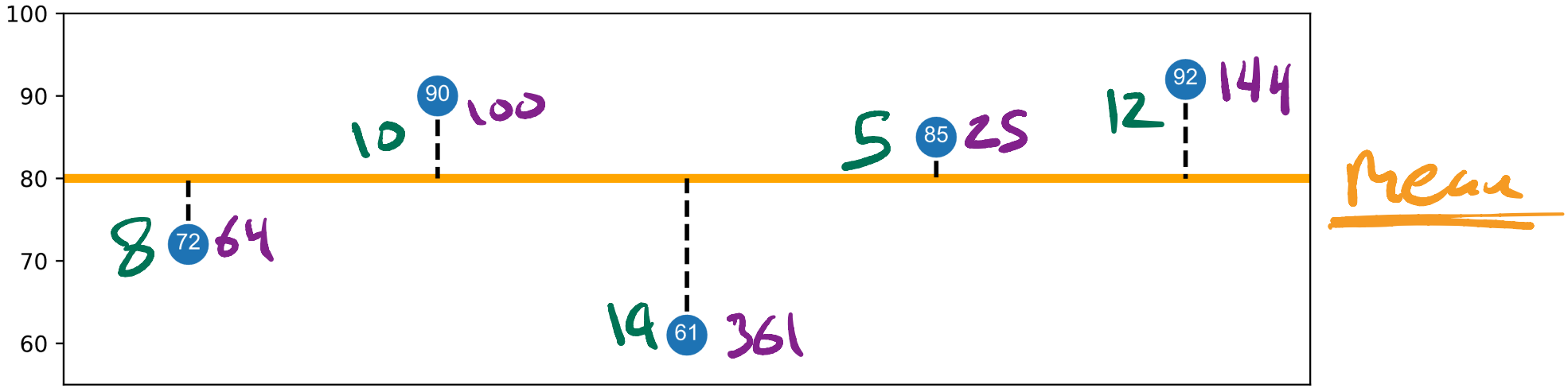
Squared loss vs. absolute loss

For the constant model, $H(x_i) = h$, so we can simplify our loss functions as follows:

- Squared loss: $L_{\text{sq}}(y_i, h) = (y_i - h)^2$. → 80, the mean of ys minimizes
- Absolute loss: $L_{\text{abs}}(y_i, h) = |y_i - h|$.

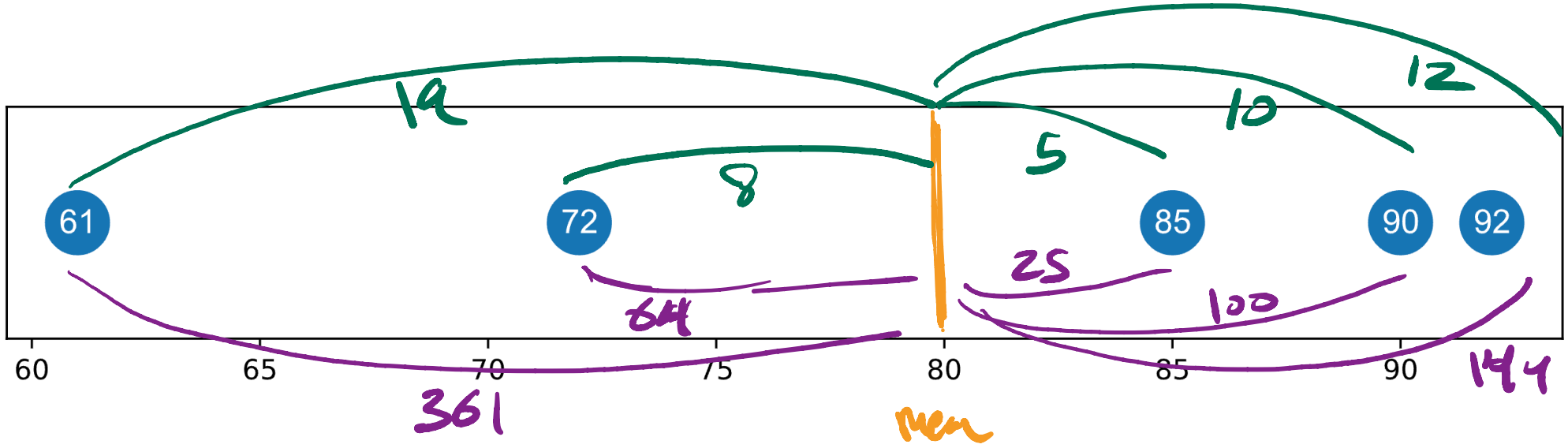
Consider, again, our example dataset of five commute times and the prediction $h = 80$.

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



Squared loss vs. absolute loss

- When we use squared loss, h^* is the point at which the average squared loss is minimized.
- When we use absolute loss, h^* is the point at which the average absolute loss is minimized.



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 prediction, h^*.
- Previously, we used calculus to find the optimal model parameter h^* that minimized R_{sq} – that is, when using squared loss.
- Can we use calculus to minimize $R_{\text{abs}}(h)$, too?

Minimizing mean absolute error

Minimizing using calculus, again

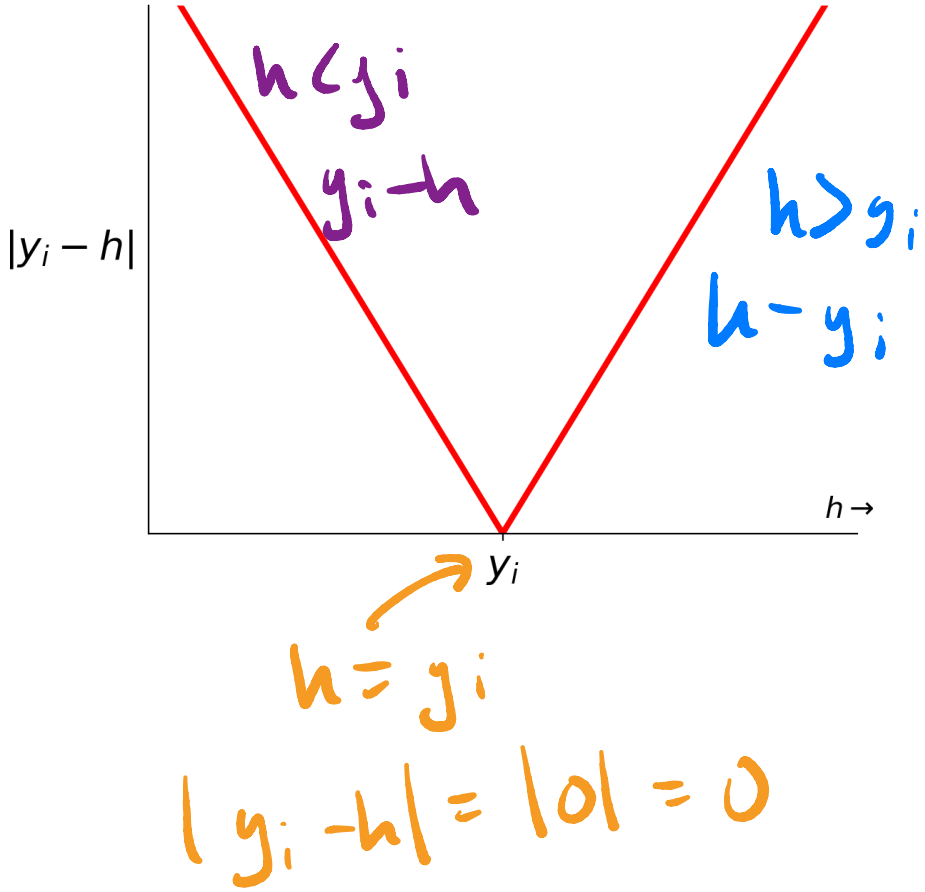
We'd like to minimize:

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

In order to minimize $R_{\text{abs}}(h)$, we:

1. take its derivative with respect to h,
2. set it equal to 0,
3. solve for the resulting h^*, and
4. perform a second derivative test to ensure we found a minimum.

Step 0: The derivative of $|y_i - h|$



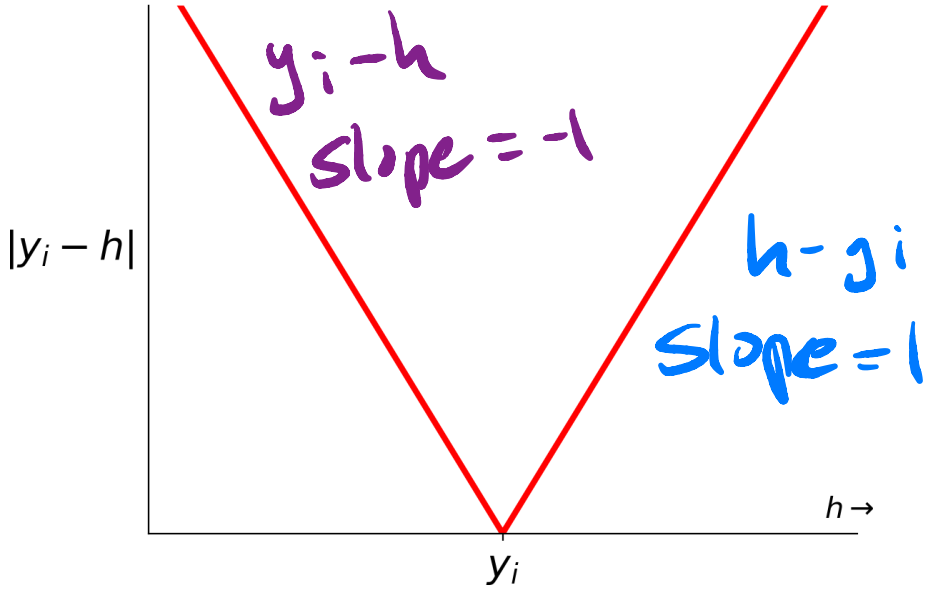
Remember that $|x|$ is a **piecewise linear** function of x:

$$|x| = \begin{cases} x & x > 0 \\ 0 & x = 0 \\ -x & x < 0 \end{cases}$$

So, $|y_i - h|$ is also a piecewise linear function of h:

$$|y_i - h| = \begin{cases} y_i - h & h < y_i \\ 0 & y_i = h \\ h - y_i & h > y_i \end{cases}$$

Step 0: The "derivative" of $|y_i - h|$



$$|y_i - h| = \begin{cases} y_i - h & h < y_i \\ 0 & y_i = h \\ h - y_i & h > y_i \end{cases}$$

What is $\frac{d}{dh}|y_i - h|$?

$$\frac{d}{dh}|y_i - h| = \begin{cases} -1 & h < y_i \\ \text{undefined} & y_i = h \\ 1 & h > y_i \end{cases}$$

Step 1: The "derivative" of $R_{\text{abs}}(h)$

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

$$\frac{d}{dh}|y_i - h| \begin{cases} -1 & h < y_i \\ +1 & h > y_i \end{cases}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{d}{dh}|y_i - h| \qquad \begin{matrix} +1 & h > y_i \\ -1 & h < y_i \end{matrix}$$

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

slope of MAE

$h = 81$

81 72 | 85 90 92

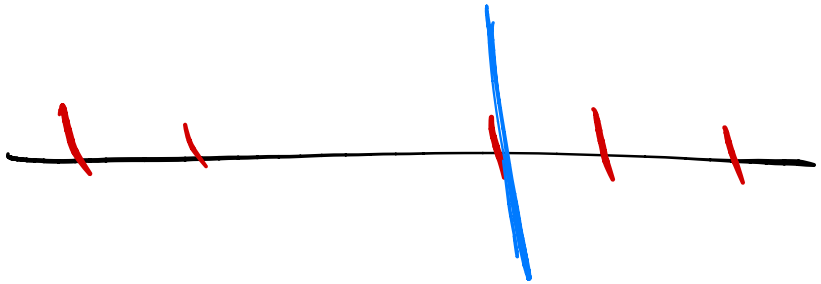
$$\frac{d}{dh} R_{\text{abs}}(81) = \frac{2-3}{5}$$

$$= -\frac{1}{5}$$

Steps 2 and 3: Set to 0 and solve for the minimizer, h^*

$$\frac{d}{dh} R_{cbs}(h) = \frac{1}{n}\left[\#(h > y_i) - \#(h < y_i)\right] = 0$$

$$= \#(h > y_i) - \#(h < y_i) = 0$$



$$= \#(h > y_i) = \#(h < y_i)$$

Median = Points to the Left of h = Points to the Right of h

The median minimizes mean absolute error!

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

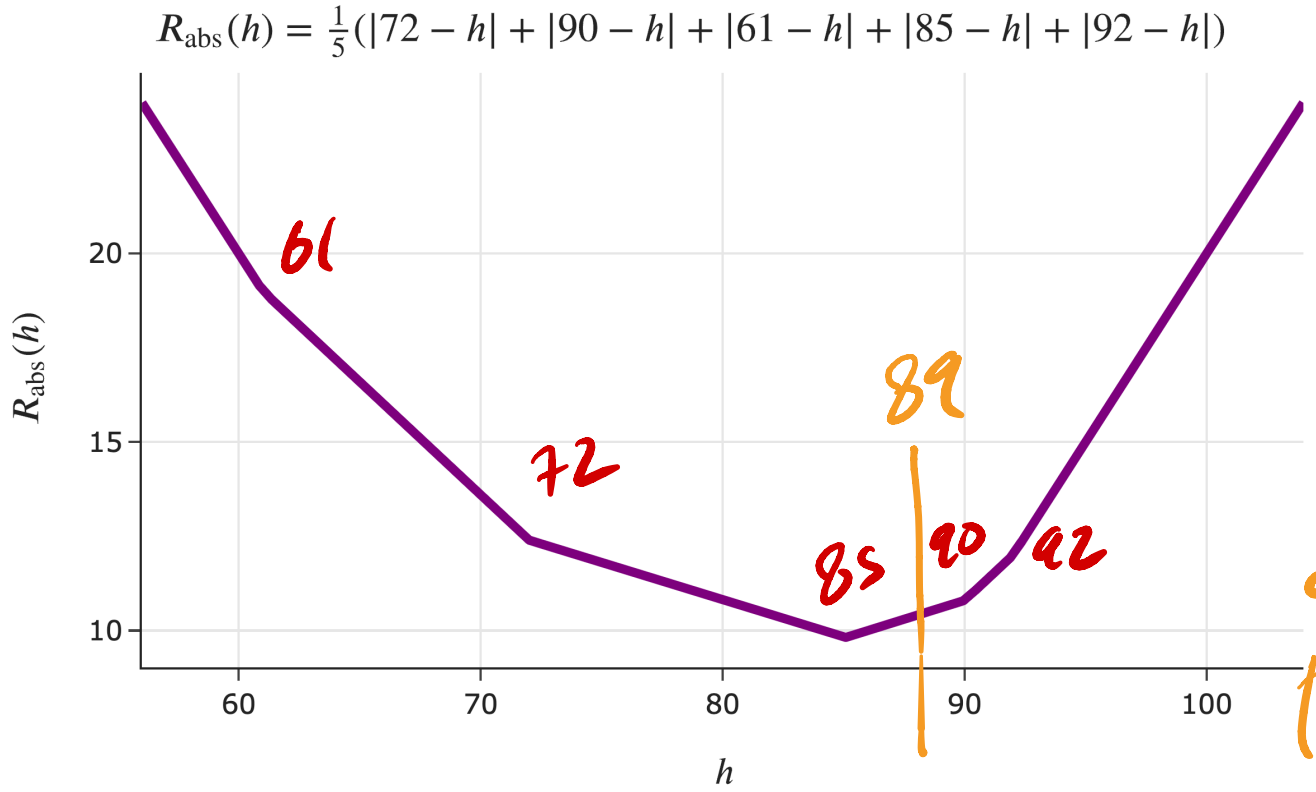
$$R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

- The answer is:

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

- This is because the median has an equal number of data points to the left of it and to the right of it.
- To make a bit more sense of this result, let's graph $R_{\text{abs}}(h)$.

Visualizing mean absolute error



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

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

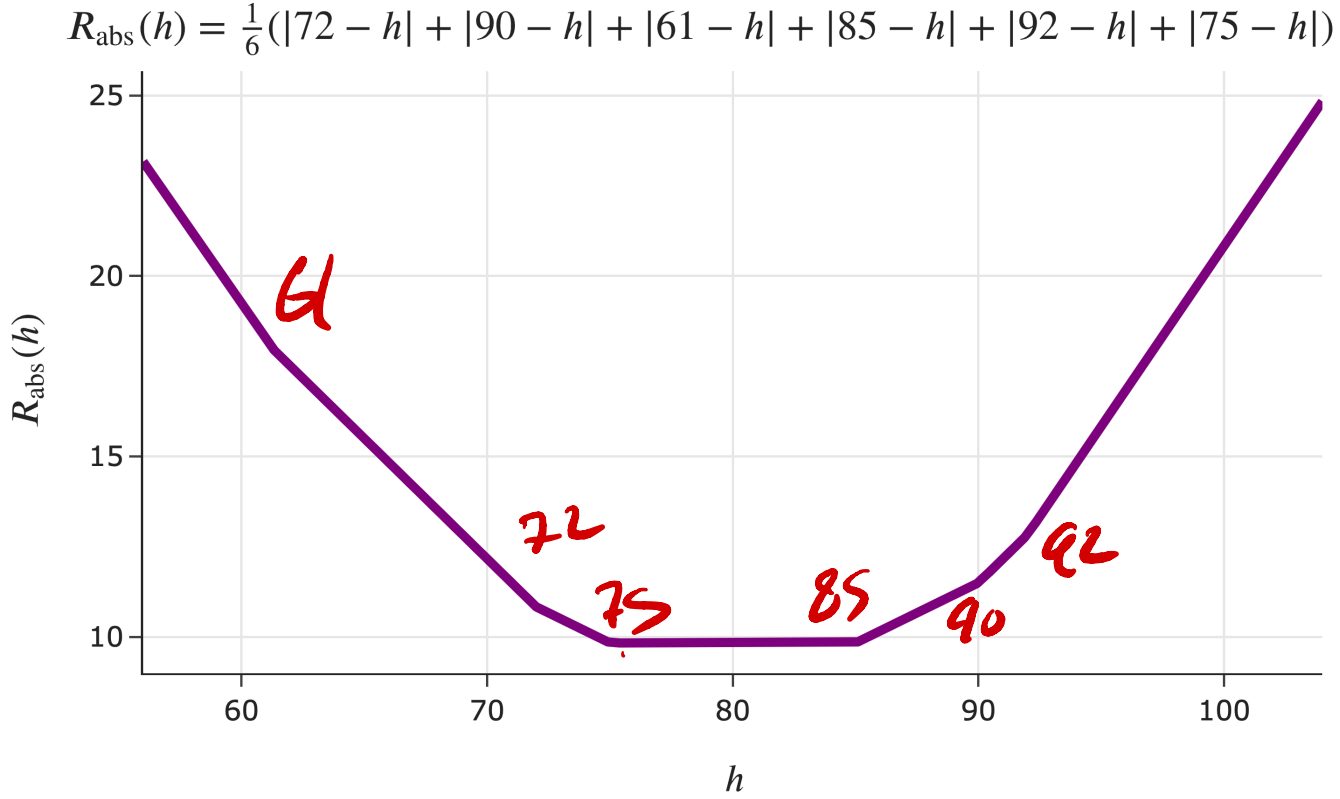
Where are the "bends" in the graph of $R_{\text{abs}}(h)$ – that is, where does its slope change?

$\frac{d}{dh} R_{\text{abs}}(h) = \frac{1}{n}[\text{left} - \text{right}]$

$R_{\text{abs}}(89) = \frac{1}{5}[3 - 2]$

$R_{\text{abs}}(89) = \frac{1}{5}$

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^*?

$75 \leq h^* \leq 85$

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.

The modeling recipe, again

We've now made two full passes through our "modeling recipe."

1. Choose a model.

$H(x) = h$

2. Choose a loss function.

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

3. Minimize average loss to find optimal model parameters.

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

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:

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

- 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 🤔

Take a moment to pause and reflect...

If you have any questions post online to our class forms/Q&A site.

Course staff will answer them ASAP!

Summary, next time

- $h^* = \text{Mean}(y_1, y_2, \ldots, y_n)$ minimizes mean squared error, $R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^{n} (y_i - h)^2$.
- $h^* = \text{Median}(y_1, y_2, \ldots, y_n)$ minimizes mean absolute error, $R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$.
- $R_{\text{sq}}(h)$ and $R_{\text{abs}}(h)$ are examples of **empirical risk** – that is, average loss.
- **Next time**: What's the relationship between the mean and median? What is the significance of $R_{\text{sq}}(h^*)$ and $R_{\text{abs}}(h^*)$?

A practice exam problem

An exam problem? Already?

- Homework 1 has been or will be released soon!
- In it, you'll be asked to *show* or *prove* that various facts hold true – but you may have never done this before!
- To help you practice, we'll walk through an old exam problem together.

Define the extreme mean (EM) of a dataset to be the average of its largest and smallest values. Let $f(x) = -3x + 4$.

Show that for any dataset $x_1 \leq x_2 \leq \ldots \leq x_n$,

$$\mathrm{EM}(f(x_1), f(x_2), \ldots, f(x_n)) = f(\mathrm{EM}(x_1, x_2, \ldots, x_n))$$