**Lecture 3**

# Empirical Risk Minimization – mean absolute error

**DSC 40A, Fall 2025**

# Agenda

- Recap: Mean squared error.

- Another loss function.

- Minimizing mean absolute error.

# Question 🤔

Answer at q.dsc40a.com

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

# 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.

2. Choose a loss function.

3. Minimize average loss to find optimal model parameters.

# Recap: Mean squared error
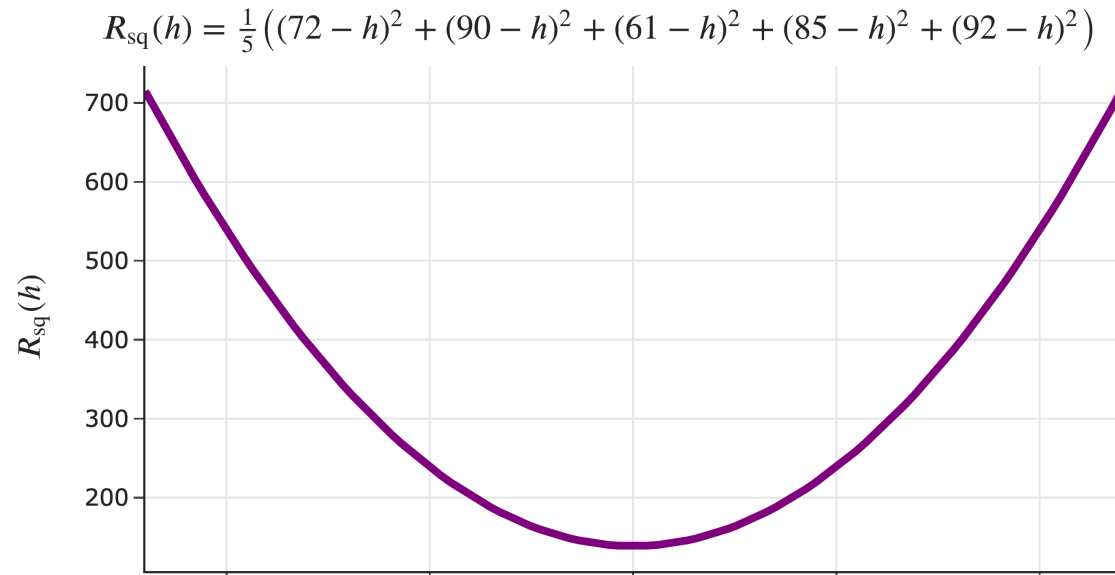
# Minimizing using calculus

We'd like to minimize:

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

In order to minimize $R_{\mathrm{sq}}(h)$, 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 4: Second derivative test

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

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

$$\frac{d^2}{dh^2} R_{sq}(h) = \frac{d}{dh}\left[\frac{2}{n}\sum_{i=1}^{n}(h-y_i)\right] = \frac{2}{n}\sum_{i=1}^{n}\frac{d}{dh}h = \overbrace{\frac{2}{n}\sum_{i=1}^{n}1}^{\substack{n \ \text{times} \\ 1+1+\ldots+1}} = \frac{2}{n}\cdot n = 2 > 0$$

$$\frac{d^2}{dh^2} R_{sq}(h) > 0 \implies h^* \text{ is a minimizer}$$

32

# 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**.

- This answer is always unique!

- We call $h^*$ our **optimal model parameter**, for when we use:
  - the constant model, $H(x) = h$, and
  - the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$.

# Bonus: the mean is easy to compute

```
def mean(numbers):
    total = 0
    for number in numbers:
        total = total + number
    return total / len(numbers)
```

- Time complexity $\Theta(n)$

# 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

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

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

# 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.

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

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

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

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

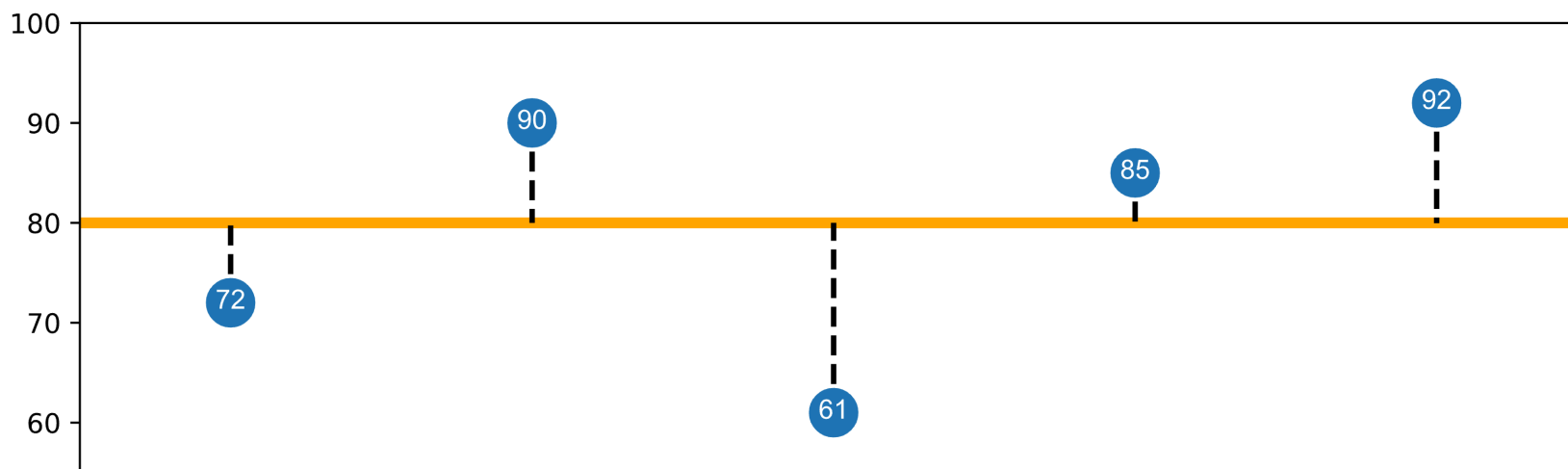$$L_{\mathrm{abs}}(y_i, H(x_i)) = |y_i - H(x_i)|$$

# Squared loss vs. absolute loss

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

- Squared loss: $L_{\mathrm{sq}}(y_i, h) = (y_i - h)^2$.
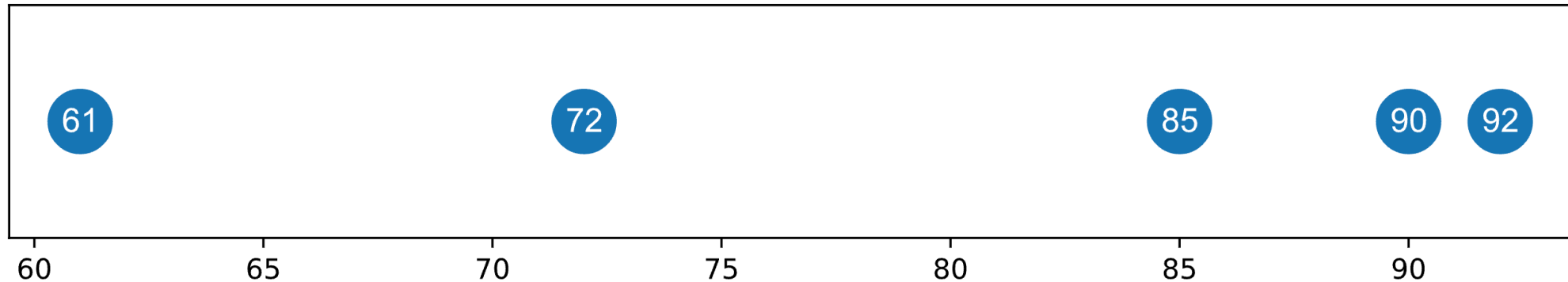- Absolute loss: $L_{\mathrm{abs}}(y_i, h) = |y_i - h|$.

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 **<u>average</u> absolute loss**, or **<u>mean</u> absolute error (MAE)**, of the prediction $h$ is:

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

- We'd like to find the best prediction, $h^*$.

- Previously, when using squared loss we used calculus to find the optimal model parameter $h^*$ that minimized $R_{\mathrm{sq}}$.

- Can we use calculus to minimize $R_{\mathrm{abs}}(h)$, too?

# Minimizing mean absolute error
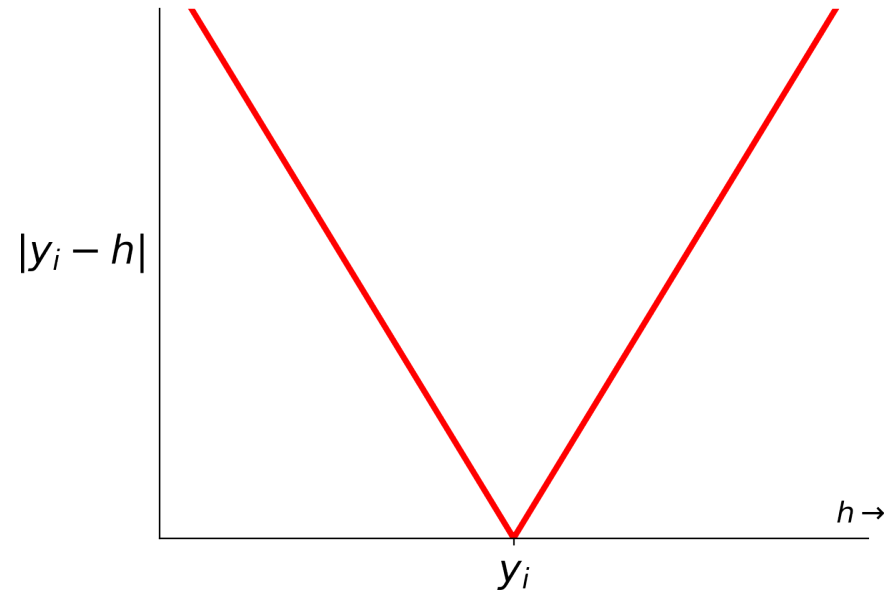
# Minimizing using calculus, again

We'd like to minimize:

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

In order to minimize $R_{\mathrm{abs}}(h)$, 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.
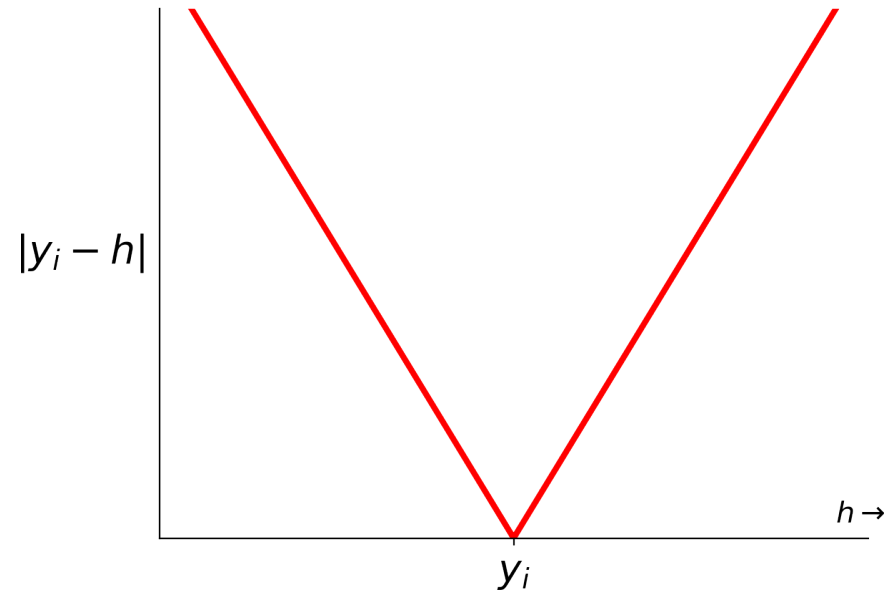
# Step 0: The derivative of $|y_i - h|$

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 $h$:

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

$|y_i - h|$

$h \rightarrow$

$y_i$

18

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

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

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

# Question 🤔

Answer at

The slope of $R_{\text{abs}}$ at $h$ is

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

Suppose that the number of points $n$ is odd. At what value of $h$ does the slope change from negative to positive?

- A) $h$ = mean of $\{y_1, \ldots, y_n\}$
- B) $h$ = median of $\{y_1, \ldots, y_n\}$
- C) $h$ = mode of $\{y_1, \ldots, y_n\}$

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

# The median minimizes mean absolute error!

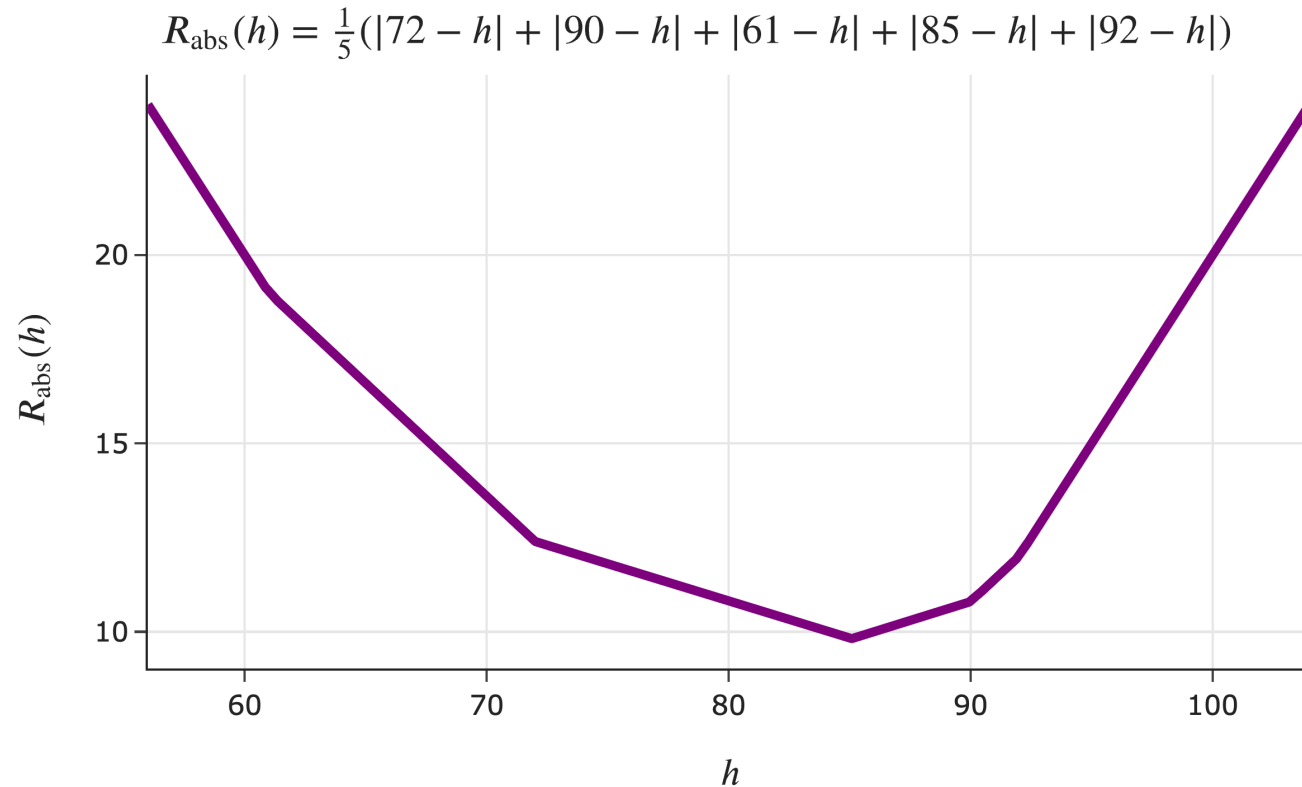- The new problem we set out to solve was, find the $h^*$ that minimizes:

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

- 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

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



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

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

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

# Question 🤔

Consider, again, our example dataset of five commute times.
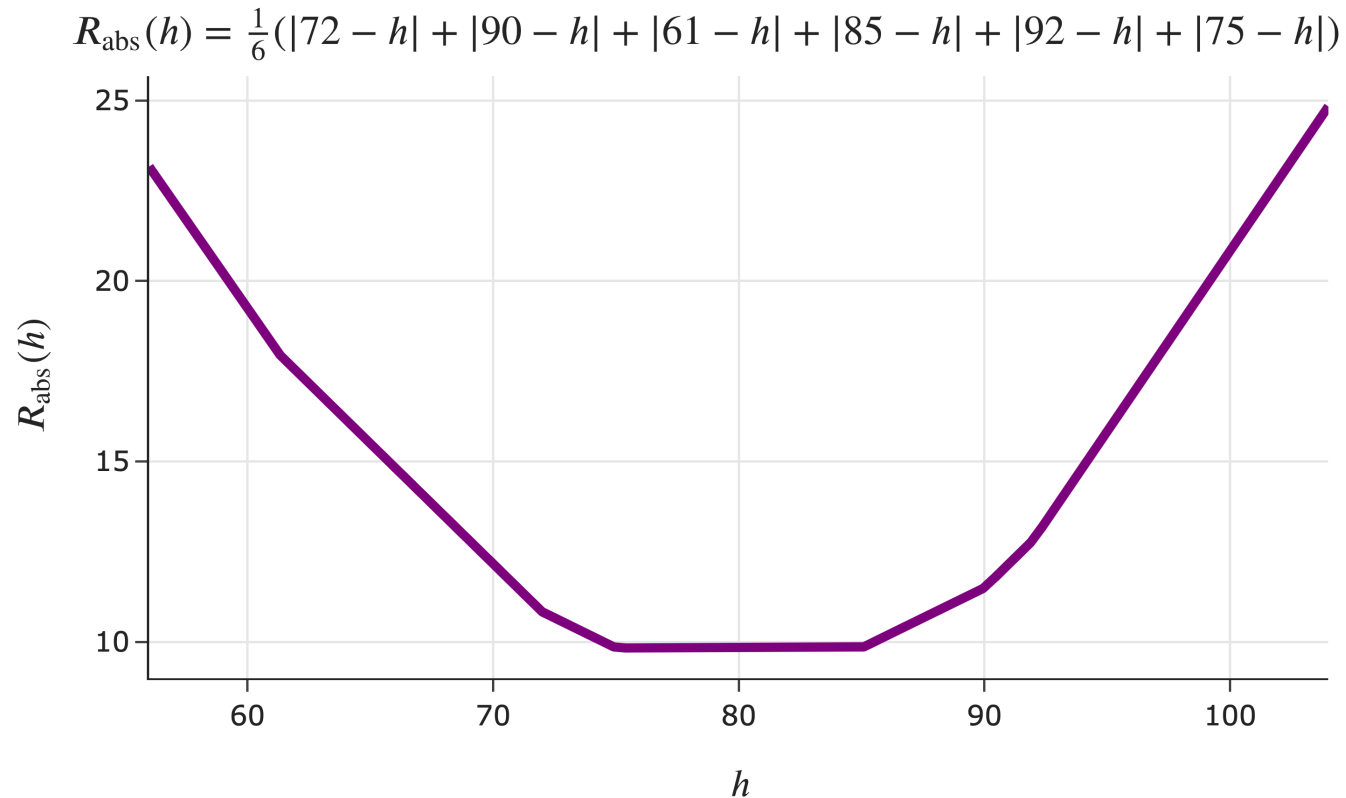
$$72, 90, 61, 85, 92$$

Suppose we add a sixth point so that our data is now

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

Which of the following correctly describes the $h*$ that minimizes mean absolute error for our new dataset?

- A) 85 only
- B) 75 only
- C) 80 only
- D) Any value between 75 and 85 inclusive

26

# Visualizing mean absolute error, with an even number of points

$$R_{\text{abs}}(h) = \frac{1}{6}(|72 - h| + |90 - h| + |61 - h| + |85 - h| + |92 - h| + |75 - h|)$$



What if we add a sixth data point?

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

Is there a unique $h^*$?

# The median minimizes mean absolute error!

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

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

- The answer is:

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

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.


2. Choose a loss function.


3. Minimize average loss to find optimal model parameters.

# 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, $L_{\mathrm{sq}}(y_i, h) = (y_i - h)^2$, the corresponding empirical risk is mean squared error:

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

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

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

# 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?

# 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 $R_{\text{abs}}(h^*)$?