Lecture 2 – Minimizing Mean Absolute Error

DSC 40A, Fall 2022 @ UC San Diego

Dr. Truong Son Hy, with help from **many others**

Announcements

- Look at the readings linked on the course website!
- First Discussion: Monday, October 3rd 2022
  First Homework Release: Friday September 30th 2022
  First Groupwork Release: Thursday September 29th 2022
  Groupwork Relsease Day: Thursday afternoon
  Groupwork Submission Day: Monday midnight
  Homework Release Day: Friday after lecture
  Homework Submission Day: Friday before
- See Calendar on course website for office hours locations and Zoom links.
  - In-person office hours are now in SDSC. You will get the passcode from the TA and tutors to access the building.

Agenda

1. Recap from Lecture 1 – learning from data.
2. Minimizing mean absolute error.
3. Identifying another choice of error.

Recap from Lecture 1 – learning from data

Last time

- **Question:** How do we turn the problem of learning from data into a math problem?

- **Answer:** Through optimization.

- **Important assumption:** We assume that the data we collected from the past/history is a good representation for the future prediction.

A formula for the mean absolute error

- We have data:

  90,000 94,000 96,000 120,000 160,000

- Suppose our prediction is h.

- The **mean absolute error** of our prediction is:

$$R(h) = \frac{1}{5}\Big(|90{,}000 - h| + |94{,}000 - h| + |96{,}000 - h| + |120{,}000 - h| + |160{,}000 - h|\Big)$$

Many possible predictions

- Last time, we considered four possible **hypotheses** for future salary, and computed the mean absolute error of each.
  - $h_1 = 150{,}000 \implies R(150{,}000) = 42{,}000$
  - $h_2 = 115{,}000 \implies R(115{,}000) = 23{,}000$
  - $h_3 = \text{mean} = 112{,}000 \implies R(112{,}000) = 22{,}400$
  - $h_4 = \text{median} = 96{,}000 \implies R(96{,}000) = 19{,}200$
- Of these four options, the median has the lowest MAE. But is it the **best possible prediction overall**?

A *general* formula for the mean absolute error

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

$$R(h) = \frac{1}{n}\left(|y_1 - h| + |y_2 - h| + \ldots + |y_n - h|\right)$$

- Or, using **summation notation**:

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

The best prediction

- We want the best prediction, h^* (i.e. $R(h^*) = \min_{h>0} R(h)$).
- The smaller $R(h)$, the better h.
- Goal: find h that minimizes $R(h)$.
- Optimization problem (with a constraint $h > 0$):

$$h^* = \operatorname{argmin}_{h>0} R(h)$$

Discussion Question

Can we use calculus to minimize R?

Minimizing mean absolute error

Minimizing with calculus

- Optimization problem:

$$\min_{h>0} R(h)$$

- Calculus: take derivative with respect to h, set equal to zero, solve.

$$\frac{d}{dh} R(h) = 0$$

Minimizing with calculus

Given an **arbitrary** function R, under which conditions the equation

$$\frac{d}{dh}R(h) = 0$$

return to us the solution of the optimization problem

$$\min \ R(h)?$$

Minimizing with calculus

Given an **arbitrary** function R, under which conditions the equation

$$\frac{d}{dh}R(h) = 0$$

return to us the solution of the optimization problem

$$\min \; R(h)?$$

- We are able to compute the derivative or R is differentiable.
- There is a unique global minimum.
- The equation will return to us local minimal and local maximal.

Minimizing with calculus

- Calculus: take derivative with respect to h, set equal to zero, solve.

Given

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

What is $\frac{d}{dh} R(h)$?

Minimizing with calculus

- Calculus: take derivative with respect to h, set equal to zero, solve.

Given

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

What is $\frac{d}{dh} R(h)$?

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

Minimizing with calculus

- Calculus: take derivative with respect to h, set equal to zero, solve.

Given

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

What is $\frac{d}{dh}R(h)$?

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

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

Minimizing with calculus

- Calculus: take derivative with respect to h, set equal to zero, solve.

Given

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

What is $\frac{d}{dh} R(h)$?

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

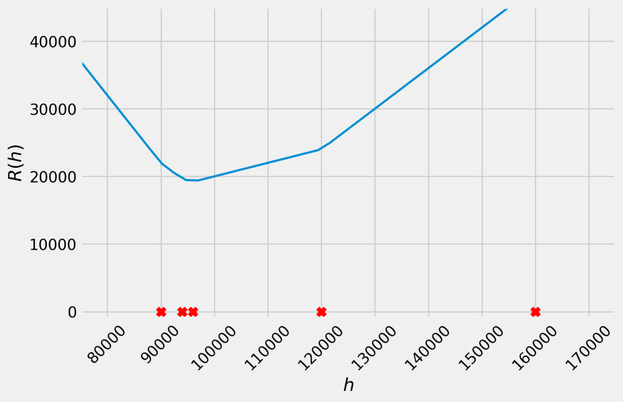
$$\Leftrightarrow \frac{d}{dh} R(h) = \frac{1}{n} \frac{d}{dh} \left(\sum_{i=1}^{n} |y_i - h| \right)$$

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

Uh oh...

- R is **not differentiable**.
- We can't use calculus to minimize it.
- Let's try plotting $R(h)$ instead.

Plotting the mean absolute error



Useful online tool for drawing:

`https://www.desmos.com/calculator`

Discussion Question

A local minimum occurs when the slope goes from ______________. Select all that apply.

A) positive to negative
B) negative to positive
C) positive to zero.
D) negative to zero.

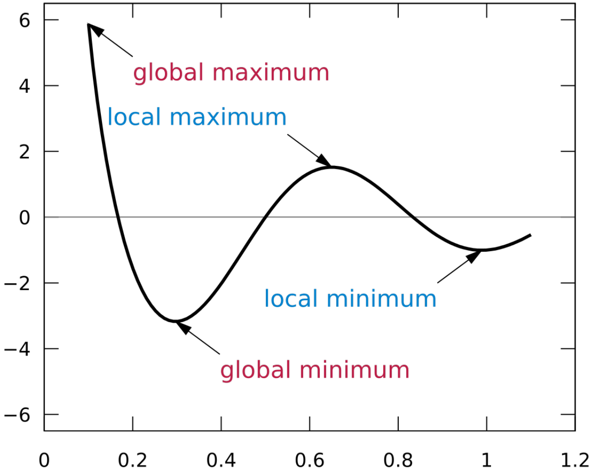
Discussion Question

A local minimum occurs when the slope goes from ________________. Select all that apply.

A) positive to negative
B) negative to positive
C) positive to zero.
D) negative to zero.

Answer: B

What we know from Calculus



Source:
https://en.wikipedia.org/wiki/Maxima_and_minima

What we know from Calculus

The First Derivative Test: Let c be a critical point for a continuous function f

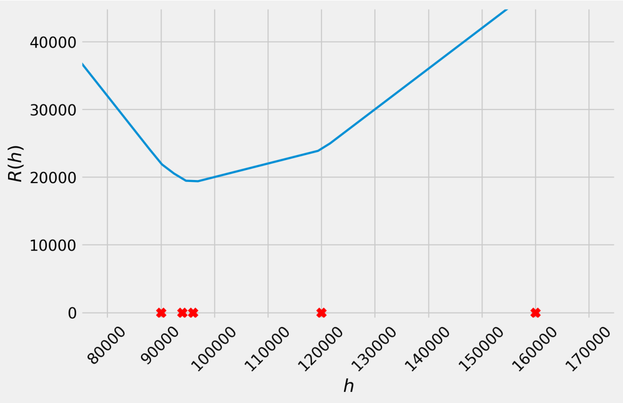
- If $f'(x)$ changes from positive to negative at c, then $f(c)$ is a local maximum.

- If $f'(x)$ changes from negative to positive at c, then $f(c)$ is a local minimum.

- If $f'(x)$ does not change sign at c, then $f(c)$ is neither a local maximum or minimum.

Note: Critical points are the solutions of equation $f'(x) = 0$.

Goal



- Find where slope of R goes from negative to non-negative.
- Want a formula for the slope of R at h.

Sums of linear functions

- Let

$$f_1(x) = 3x + 7 \qquad f_2(x) = 5x - 4 \qquad f_3(x) = -2x - 8$$

- What is the slope of $f(x) = f_1(x) + f_2(x) + f_3(x)$?

Sums of linear functions

- Let

$$f_1(x) = 3x + 7 \qquad f_2(x) = 5x - 4 \qquad f_3(x) = -2x - 8$$

- What is the slope of $f(x) = f_1(x) + f_2(x) + f_3(x)$?

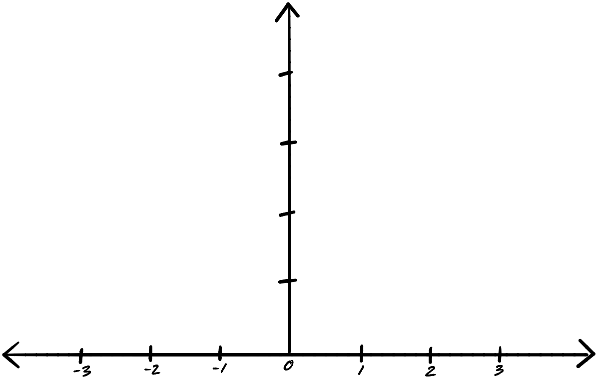
We can do it analytically:

$$f(x) = (3x + 7) + (5x - 4) + (-2x - 8) = 6x - 5$$

So the slope is 6. Because in this case, we don't have the absolute value.

Absolute value functions

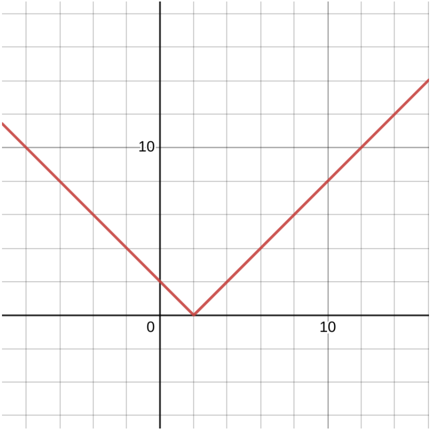
Recall, $f(x) = |x - a|$ is an absolute value function centered at $x = a$.



First, start with $f(x) = |x|$ and then shift the plot by a units to the right.

Absolute value functions

Recall, $f(x) = |x - a|$ is an absolute value function centered at $x = a$.



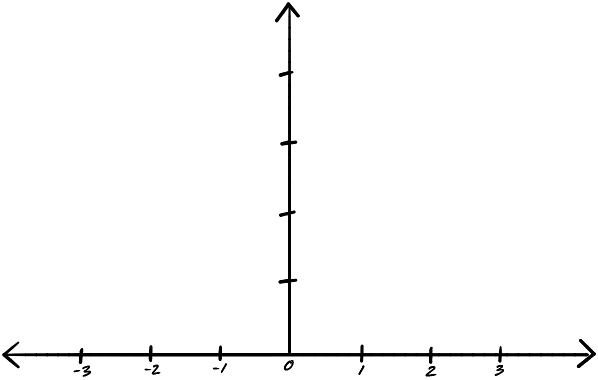
$a = 2$

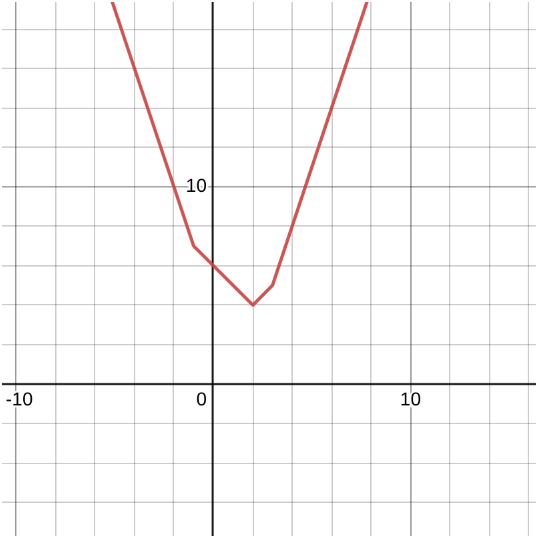
Sums of absolute values

- Let

$$f_1(x) = |x - 2| \qquad f_2(x) = |x + 1| \qquad f_3(x) = |x - 3|$$

- What is the slope of $f(x) = f_1(x) + f_2(x) + f_3(x)$?





https://www.desmos.com/calculator

The slope of the mean absolute error

$R(h)$ is a sum of absolute value functions (times $\frac{1}{n}$):

$$R(h) = \frac{1}{n}\left(|h - y_1| + |h - y_2| + \ldots + |h - y_n|\right)$$

The slope of the mean absolute error

$R(h)$ is a sum of absolute value functions (times $\frac{1}{n}$):

$$R(h) = \frac{1}{n}\left(|h - y_1| + |h - y_2| + \ldots + |h - y_n|\right)$$

We have:

$$R(h) = \frac{1}{n}\left(\sum_{i : y_i < h} |h - y_i| + \sum_{i : y_i > h} |h - y_i|\right)$$

The slope of the mean absolute error

$R(h)$ is a sum of absolute value functions (times $\frac{1}{n}$):

$$R(h) = \frac{1}{n}\left(|h - y_1| + |h - y_2| + \ldots + |h - y_n|\right)$$

We have:

$$R(h) = \frac{1}{n}\left(\sum_{i:y_i<h} |h - y_i| + \sum_{i:y_i>h} |h - y_i|\right)$$

$$\Leftrightarrow R(h) = \frac{1}{n}\left(\sum_{i:y_i<h} (h - y_i) + \sum_{y_i>h} (y_i - h)\right)$$

The slope of the mean absolute error

$R(h)$ is a sum of absolute value functions (times $\frac{1}{n}$):

$$R(h) = \frac{1}{n}\left(|h - y_1| + |h - y_2| + \ldots + |h - y_n|\right)$$

We have:

$$R(h) = \frac{1}{n}\left(\sum_{i:y_i<h} |h - y_i| + \sum_{i:y_i>h} |h - y_i|\right)$$

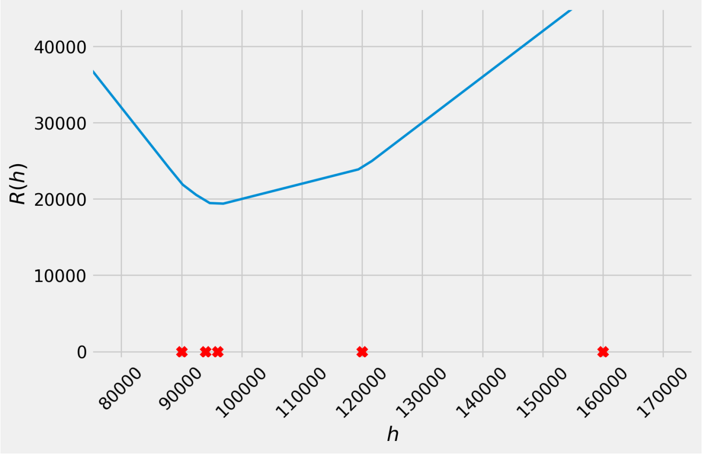
$$\Leftrightarrow R(h) = \frac{1}{n}\left(\sum_{i:y_i<h} (h - y_i) + \sum_{y_i>h} (y_i - h)\right)$$

$$\Leftrightarrow R(h) = \frac{1}{n}\left(\sum_{i:y_i<h} 1 - \sum_{i:y_i>h} 1\right)h + \text{constant}$$

The slope of the mean absolute error

The slope of R at h is:

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



Where the slope's sign changes

The slope of R at h is:

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

Discussion Question

Suppose that n is odd. At what value of h does the slope of R go from negative to non-negative?

A) h = mean of $y_1, \dots, y_n$
B) h = median of $y_1, \dots, y_n$
C) h = mode of $y_1, \dots, y_n$

Where the slope's sign changes

The slope of R at h is:

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

Discussion Question

Suppose that n is odd. At what value of h does the slope of R go from negative to non-negative?

A) h = mean of $y_1, \dots, y_n$
B) h = median of $y_1, \dots, y_n$
C) h = mode of $y_1, \dots, y_n$

Answer: B

The median minimizes mean absolute error, when n is odd

- Our problem was: find h^* which minimizes the mean absolute error, $R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$.

- We just determined that when n is odd, the answer is $\text{Median}(y_1, \ldots, y_n)$. This is because the median has an equal number of points to the left of it and to the right of it.

- But wait — what if n is **even**?

Discussion Question

Consider again our example dataset of 5 salaries.

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 120{,}000 \quad 160{,}000$$

Suppose we collect a 6th salary, so that our data is now

$$90{,}000 \quad 94{,}000 \quad 96{,}000 \quad 108{,}000 \quad 120{,}000 \quad 160{,}000$$

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

A) 96,000 only
B) 108,000 only
C) 102,000 only
D) Any value between 96,000 and 108,000, inclusive

Discussion Question

Consider again our example dataset of 5 salaries.

90,000   94,000   96,000   120,000   160,000

Suppose we collect a 6th salary, so that our data is now

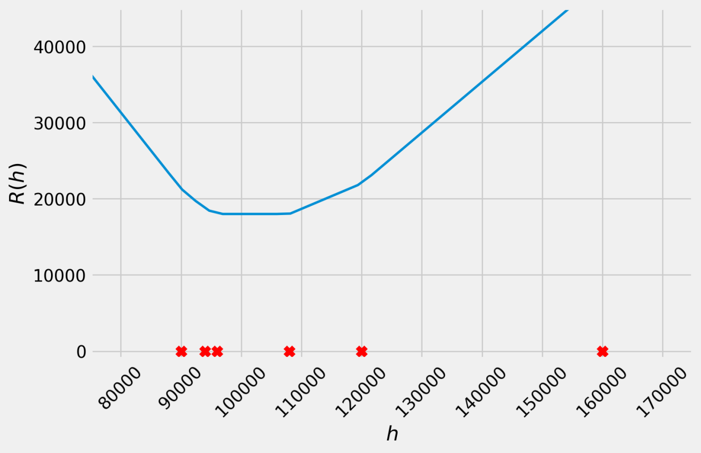
90,000   94,000   96,000   108,000   120,000   160,000

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

A) 96,000 only
B) 108,000 only
C) 102,000 only
D) Any value between 96,000 and 108,000, inclusive

Answer: D

Plotting the mean absolute error, with an even number of data points



- What do you notice?

The median minimizes mean absolute error

- Our problem was: find h^* which minimizes the mean absolute error, $R(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$.

- **Regardless of if n is odd or even**, the answer is $h^* = \text{Median}(y_1, \dots, y_n)$. The **best prediction**, in terms of mean absolute error, is the **median**.
  - When n is odd, this answer is unique.
  - When n is even, any number between the middle two data points also minimizes mean absolute error.
  - We define the median of an even number of data points to be the mean of the middle two data points.

Identifying another type of error

Two things we don't like

1. **Minimizing** the mean absolute error wasn't so easy.

2. Actually **computing** the median isn't so easy, either.

- **Question**: Is there another way to measure the quality of a prediction that avoids these problems?

The mean absolute error is not differentiable

- We can't compute $\frac{d}{dh}|y_i - h|$.
- Remember: $|y_i - h|$ measures how far h is from y_i.
- Is there something besides $|y_i - h|$ which:
    1. Measures how far h is from y_i, *and*
    2. is **differentiable**?

The mean absolute error is not differentiable

- We can't compute $\frac{d}{dh}|y_i - h|$.
- Remember: $|y_i - h|$ measures how far h is from y_i.
- Is there something besides $|y_i - h|$ which:
  1. Measures how far h is from y_i, *and*
  2. is **differentiable**?

Discussion Question

Which of these would work?

a) $e^{|y_i - h|}$

b) $|y_i - h|^2$

c) $|y_i - h|^3$

d) $\cos(y_i - h)$

The squared error

- Let h be a prediction and y be the right answer. The **squared error** is:

$$|y - h|^2 = (y - h)^2$$

- Like absolute error, measures how far h is from y.

- But unlike absolute error, the squared error is **differentiable**:

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

The squared error

Reminder that:

$$\frac{d}{dx}x^n = n \cdot x^{n-1}$$

Thus:

$$\frac{d}{dx}x^2 = 2 \cdot x$$

Reminder about the derivative of composite function:

$$(f \circ g)' = \frac{d}{dx}[f(g(x))] = f'(g(x)) \cdot g'(x)$$

Therefore:

The squared error

Reminder that:

$$\frac{d}{dx}x^n = n \cdot x^{n-1}$$

Thus:

$$\frac{d}{dx}x^2 = 2 \cdot x$$

Reminder about the derivative of composite function:

$$(f \circ g)' = \frac{d}{dx}[f(g(x))] = f'(g(x)) \cdot g'(x)$$

Therefore:

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

$$= 2 \cdot (y-h) \cdot \left(\frac{dy}{dh} - \frac{dh}{dh}\right) = 2 \cdot (y-h) \cdot (-1) = 2(h-y)$$

The **mean squared error**

- Suppose we predicted a future salary of $h_1 = 150{,}000$ *before* collecting data.

salary	absolute error of h_1	squared error of h_1
90,000	60,000	$(60{,}000)^2$
94,000	56,000	$(56{,}000)^2$
96,000	54,000	$(54{,}000)^2$
120,000	30,000	$(30{,}000)^2$
160,000	10,000	$(10{,}000)^2$

total squared error: 1.0652×10^{10}
mean squared error: 2.13×10^{9}

- A good prediction is one with small **mean squared error**.

The mean squared error

- Now suppose we had predicted $h_2 = 115{,}000$.

salary	absolute error of h_2	squared error of h_2
90,000	25,000	$(25{,}000)^2$
94,000	21,000	$(21{,}000)^2$
96,000	19,000	$(19{,}000)^2$
120,000	5,000	$(5{,}000)^2$
160,000	45,000	$(45{,}000)^2$

total squared error: 3.47×10^9
mean squared error: 6.95×10^8

- A good prediction is one with small **mean squared error**.

The new idea

- Make prediction by minimizing the **mean squared error**:

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

- Strategy: Take derivative, set to zero, solve for minimizer.

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

Discussion Question

Which of these is dR_{sq}/dh?

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

B) 0

C) $\sum_{i=1}^{n} y_i$

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

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

Discussion Question

Which of these is dR_{sq}/dh?

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

B) 0

C) $\sum_{i=1}^{n} y_i$

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

Answer: D

The new idea

- Make prediction by minimizing the **mean squared error**:

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

- Strategy: Take derivative, set to zero, solve for minimizer.

We have:

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

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

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

Summary

Summary

- Our first problem was: find h^* which minimizes the mean absolute error, $R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$.
  - The answer is: $\text{Median}(y_1, \dots, y_n)$.
  - The **best prediction**, in terms of mean absolute error, is the **median**.
- We then started to consider another type of error, squared error, that is differentiable and hence is easier to minimize.
- **Next time:** We will finish determining the value of h^* that minimizes mean squared error, and see how it compares to the median.