# DSC 40A

## Theoretical Foundations of Data Science I

## Last Time

- To predict future salary:
  - Gather salaries $y_1, y_2, \dots, y_n$.
  - Find a prediction $h^*$ which minimizes the **mean error**:

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

our choice

  - We saw that $R(h)$ is minimized by Median$(y_1, \dots, y_n)$.
- We turned learning into a math problem and solved it.

## Two things we don't like

1. **Minimizing** the mean error wasn't so easy. ← not diff

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

## In This Video

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

## Recommended Reading

Course Notes: Chapter 1, Section 1

# The mean 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 error is not differentiable

- We can't compute $\frac{d}{dh}|y_i - h|$.
- Remember: $\underbrace{|y_i - h|}$ 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**?

**Question**

Which of these would work?

$= (y_i - h)^2$

derivative

$2(y_i - h) \cdot -1$

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

b) $|y_i - h|^2$

c) $\cancel{|y_i - h|^3}$

d) $\cancel{\cos}(y_i - h)$

## 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 error, measures how far $h$ is from $y$.

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

$$\frac{d}{dh}(y - h)^2 = 2(y-h) \cdot -1$$

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

$$= 2(h-y)$$

## The Mean Squared Error

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

| salary | 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: $1.0652 \times 10^{10}$
**mean squared error**: $2.13 \times 10^{9}$

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

## The Mean Squared Error

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

| salary | 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: $3.47 \times 10^9$
**mean squared error**: $6.95 \times 10^8$

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

## The New Idea

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

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

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

## 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.

**Question**

Which of these is $dR_{\text{sq}}/dh$?

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

b) $0$

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

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

## Solution

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

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

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

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

## Set to zero and solve for minimizer

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

constant

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

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

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

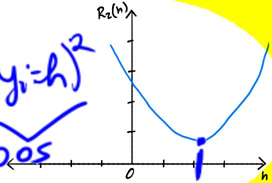
$$\boxed{h=\frac{1}{n}\sum_{i=1}^{n}y_i}\rightarrow$$

average of the data set

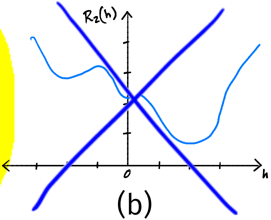
## Question

Suppose $y_1, \ldots, y_n$ are salaries. Which plot could be $R_{\text{sq}}(h)$?

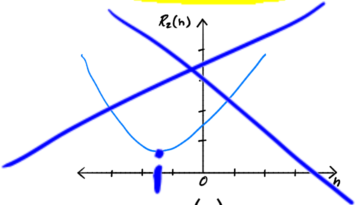
$\frac{1}{n}\sum_{i=1}^{n}\underbrace{(y_i-h)^2}_{\text{pos}}$



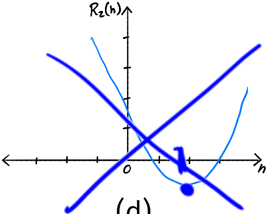

(a)




(b)




(c)




(d)

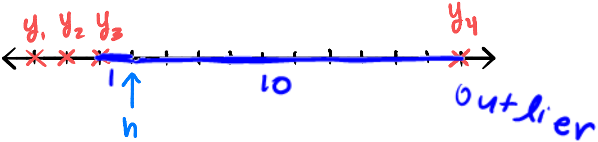
## Bonus: the mean is easy to compute

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

- Time complexity: $\Theta(n)$
- Median by sorting: $\Theta(n \log n)$
- But there's a $\Theta(n)$ way to find median: `quickselect`.
- DSC 40B.

# Outliers

- The mean is quite **sensitive** to outliers.




- $|y_4 - h|$ is 10 times as big as $|y_3 - h|$. error $|y_i - h|$

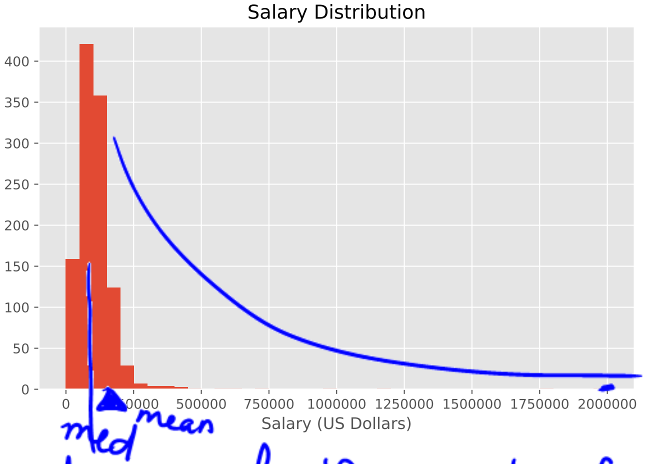
- But $(y_4 - h)^2$ is 100 times as big as $(y_3 - h)^2$.

- Squared error can be dominated by outliers.

## Example: Data Science Salaries

- Data set of 1121 self-reported data science salaries in the United States from the 2018 StackOverflow survey.
- Median = \$100,000
- Mean = \$111,032
- Max = \$2,000,000 outlier
- Min = \$52
- 95th Percentile: \$200,000

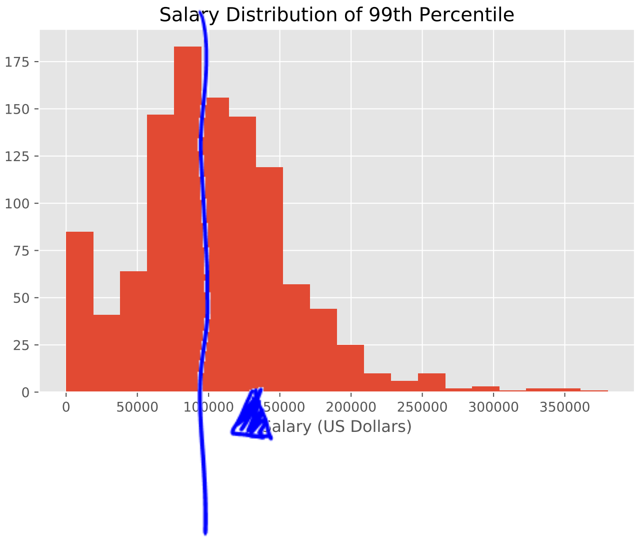
# Example: Data Science Salaries




median = halfway pt of area

mean = balance point

# Example: Data Science Salaries



# Example: Income Inequality

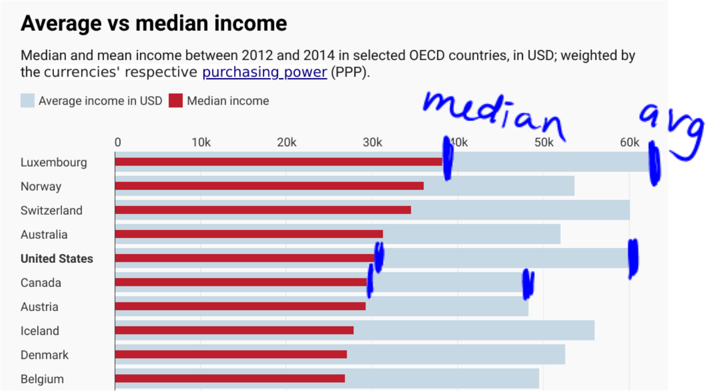
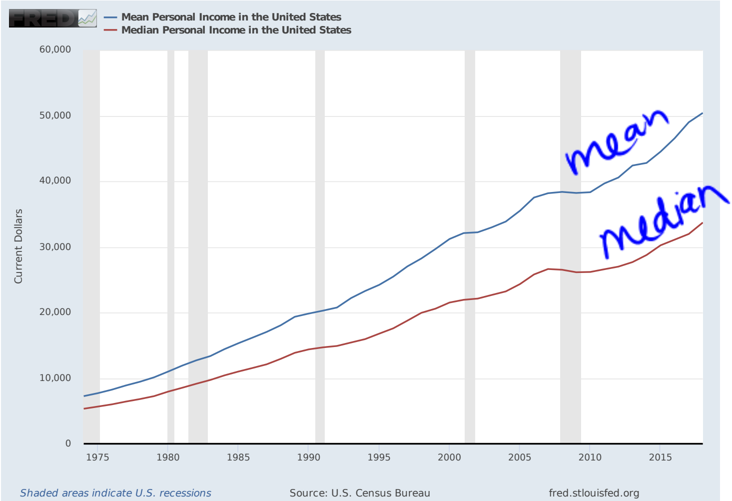



Chart: Lisa Charlotte Rost, Datawrapper

# Example: Income Inequality



## Summary: The Mean Minimizes the Mean Squared Error

- Our problem was: find $h^*$ which minimizes the mean squared error, $\boxed{R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2}$.

- The answer is: $\underline{\text{Mean}(y_1, \ldots, y_n)}$.

- Using mean squared error biases the prediction towards outliers.

- **Next time:** We consider both the mean error and the mean squared error as part of a more general framework.