

Lecture 2

Empirical Risk Minimization

DSC 40A, Fall 2025

Announcements

- Remember, there is no Canvas: all information is at dsc40a.com.
- Please fill out the [Welcome Survey](#) if you haven't already.
- Look at the office hours schedule [here](#) and plan to start regularly attending!
(more to be added)
- There are now readings linked on the course website for the next few weeks – read them for supplementary explanations.
 - They cover the same ideas, but in a different order and with different examples.
- Discussion after lecture today

Agenda

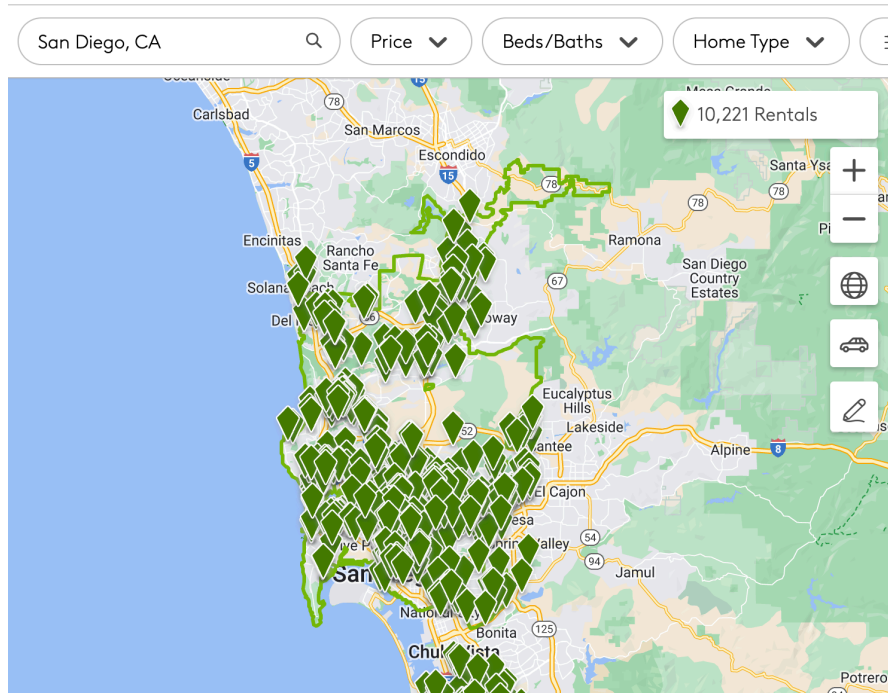
- The constant model.
- Mean squared error.
- Minimizing mean squared error.

Question 🤔

Answer at q.dsc40a.com

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

Modeling



	date	day	departure_hour	minutes
0	5/22/2023	Mon	8.450000	63.0
1	9/18/2023	Mon	7.950000	75.0
2	10/17/2023	Tue	10.466667	59.0
3	11/28/2023	Tue	8.900000	89.0
4	2/15/2024	Thu	8.083333	69.0

Due to housing costs, you are living in Orange County and commuting to campus every day.

Goal: Predict your commute time.
That is, predict how long it'll take to get to school.

How can we do this? Learning from data

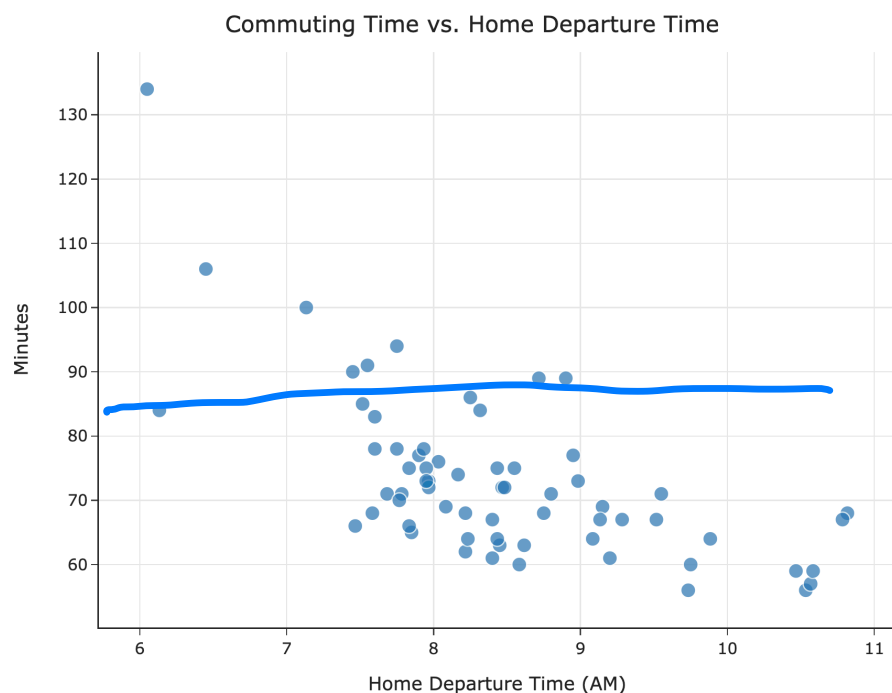
What will we need to assume? Future commute times are similar to past commute times.

The constant model

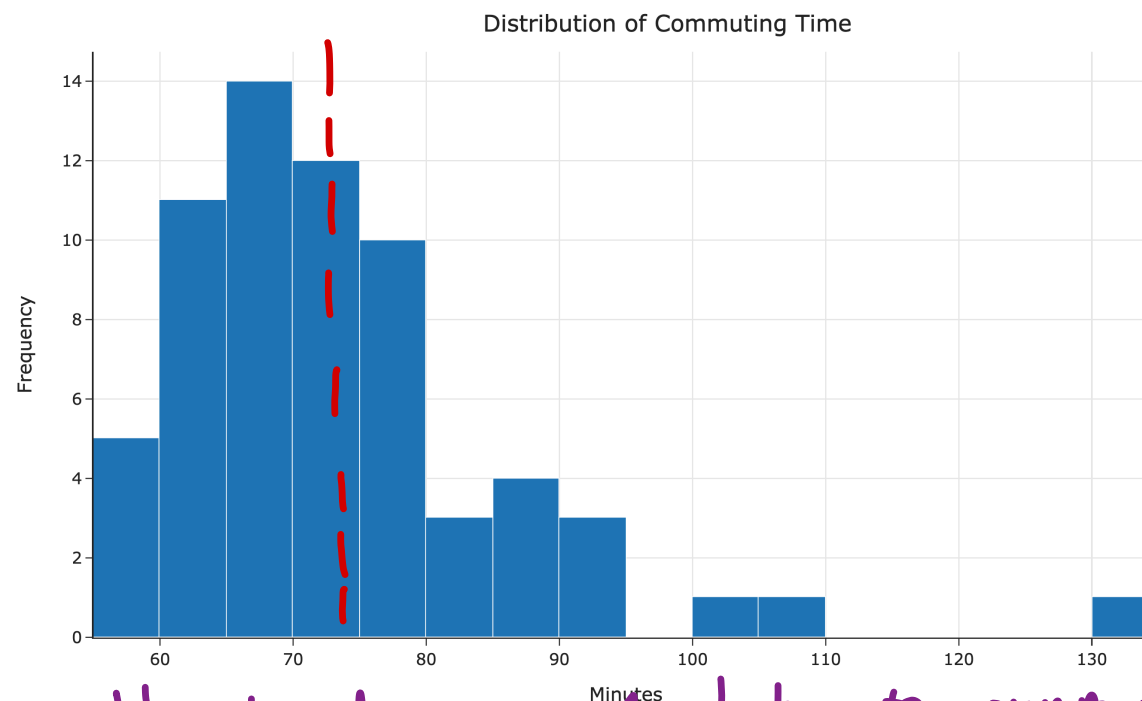
The constant model

A hypothesis function, H , takes in an x as input and returns a predicted y .

The constant model, $H(x) = h$, has one parameter: h .



Want to find best horizontal line =



Where to draw a vertical line to summarize the histogram?

A concrete example

Let's suppose we have just a smaller dataset of just five historical commute times in minutes.

$$y_1 = 72$$

$$y_2 = 90$$

$$y_3 = 61$$

$$y_4 = 85$$

$$y_5 = 92$$

Given this data, can you come up with a prediction for your future commute time?

How?

Average: 80

Median: 85

mode

max commute: 92
min : 61

$$\text{max} - \text{min} = 31$$

Some common approaches

- The mean:

$$\frac{1}{5}(72 + 90 + 61 + 85 + 92) = \boxed{80}$$

- The median:

$$61 \quad 72 \quad \boxed{85} \quad 90 \quad 92$$

- Both of these are familiar **summary statistics** – they summarize a collection of numbers with a single number.
- But which one is better? Is there a "best" prediction we can make?

The cost of making predictions

A **loss function** quantifies how bad a prediction is for a single data point.

- If our prediction is **close** to the actual value, we should have **low** loss.
- If our prediction is **far** from the actual value, we should have **high** loss.

A good starting point is error, which is the difference between **actual** and **predicted** values.

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

↖ actual commute time (known)

↘ predicted commute time (unknown)

Suppose my commute **actually** takes 80 minutes.

- If I predict 75 minutes: **error**: $80 - 75 = 5$
- If I predict 72 minutes: **error**: $80 - 72 = 8$
- If I predict 100 minutes: **error**: $80 - 100 = -20$
negative ↗

$$H(x_i) = h \text{ for all } x_i$$

ideas:
- absolute value
- square

Squared loss

One loss function is squared loss, L_{sq} , which computes $(\text{actual} - \text{predicted})^2$. error

$$L_{\text{sq}}(\textcolor{blue}{y_i}, \textcolor{orange}{H(x_i)}) = (\textcolor{blue}{y_i} - \textcolor{orange}{H(x_i)})^2$$

knownprediction/hypothesis..

Note that for the constant model, $H(x_i) = h$, so we can simplify this to:

$$L_{\text{sq}}(\textcolor{blue}{y_i}, \textcolor{orange}{h}) = (\textcolor{blue}{y_i} - \textcolor{orange}{h})^2$$

Note: $(y_i - h)^2 = (h - y_i)^2$

Squared loss is not the only loss function that exists! Soon, we'll learn about absolute loss. main pro: differentiable! V ← cusp absolute: not diff.

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

A concrete example, revisited

Consider again our smaller dataset of just five historical commute times in minutes. Suppose we predict the median, $h = 85$. What is the squared loss of 85 for each data point?

$$y_1 = 72$$

$$y_2 = 90$$

$$y_3 = 61$$

$$y_4 = 85$$

$$y_5 = 92$$

$$(y_1 - h)^2 = (72 - 85)^2 = 169$$

$$(y_2 - h)^2 = (90 - 85)^2 = 25$$

$$(61 - 85)^2 = 576$$

$$(85 - 85)^2 = 0$$

$$(92 - 85)^2 = 49$$

Goal

Find a single number that describes the loss for the prediction for the whole dataset

Averaging squared losses

We'd like a single number that describes the quality of our predictions across our entire dataset. One way to compute this is as the **average of the squared losses**.

- For the median, $h = 85$:

$$\frac{1}{5} ((72 - 85)^2 + (90 - 85)^2 + (61 - 85)^2 + (85 - 85)^2 + (92 - 85)^2) = \boxed{163.8}$$

- For the mean, $h = 80$:

$$\frac{1}{5} ((72 - 80)^2 + (90 - 80)^2 + (61 - 80)^2 + (85 - 80)^2 + (92 - 80)^2) = \boxed{138.8}$$

low loss is good!

Which prediction is better? Could there be an even better prediction?

80 is a better prediction than 85

L : loss for an individual point
 R : risk, avg loss over all points

Mean squared error

- Another term for average squared loss is mean squared error (MSE).
- The mean squared error on our smaller dataset for any prediction h is of the form:

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

R stands for "risk", as in "empirical risk." We'll see this term again soon.

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

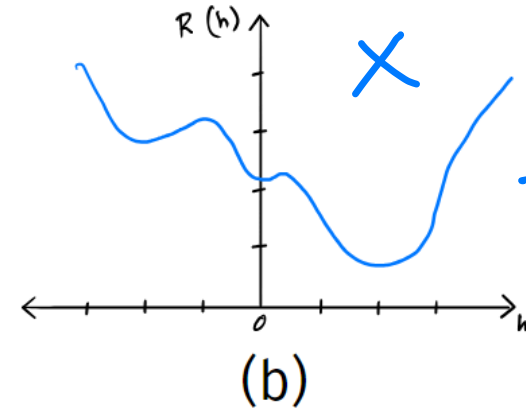
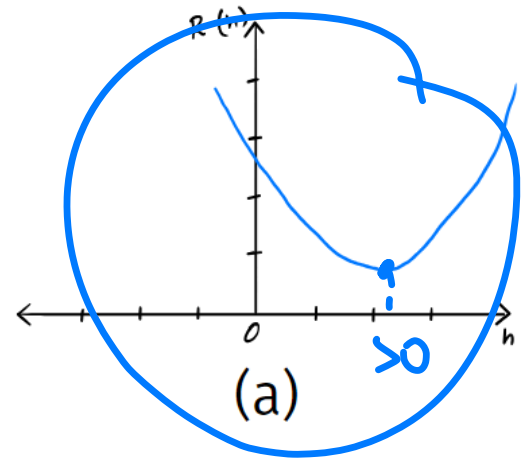
$$\begin{aligned} R_{\text{sq}}(100) &= \frac{1}{5} ((72 - 100)^2 + (90 - 100)^2 + (61 - 100)^2 + (85 - 100)^2 + (92 - 100)^2) \\ &= \boxed{538.8} > R_{\text{sq}}(95) > R_{\text{sq}}(90) > ? \end{aligned}$$

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

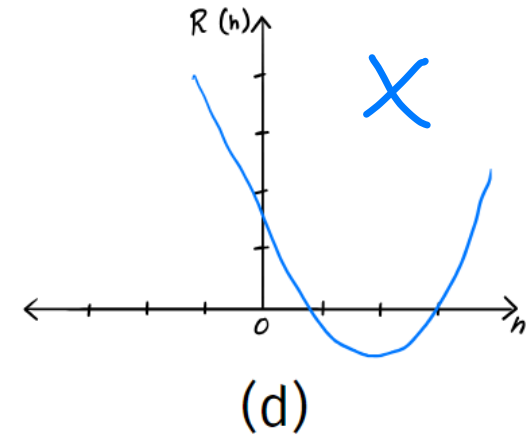
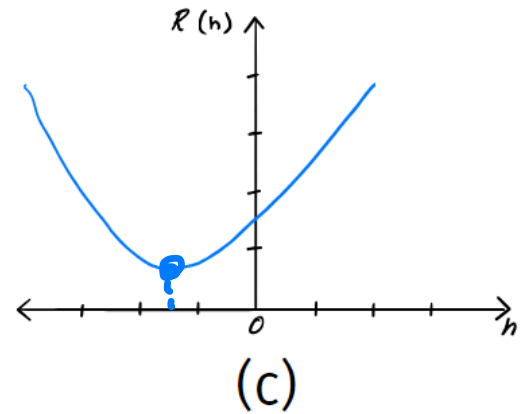
Goal: h with smallest risk

$$R_{sq} = \frac{1}{n} ((y_1 - h)^2 + (y_2 - h)^2 + \dots + (y_n - h)^2) > 0$$

Question 🤔 Answer at q.dsc40a.com



→ not parabolic



→ $R_{sq} < 0$

best can't be
neg. commute
time ←

Suppose y_1, \dots, y_n are commute times - which plot could be $R_{sq}(h)$?

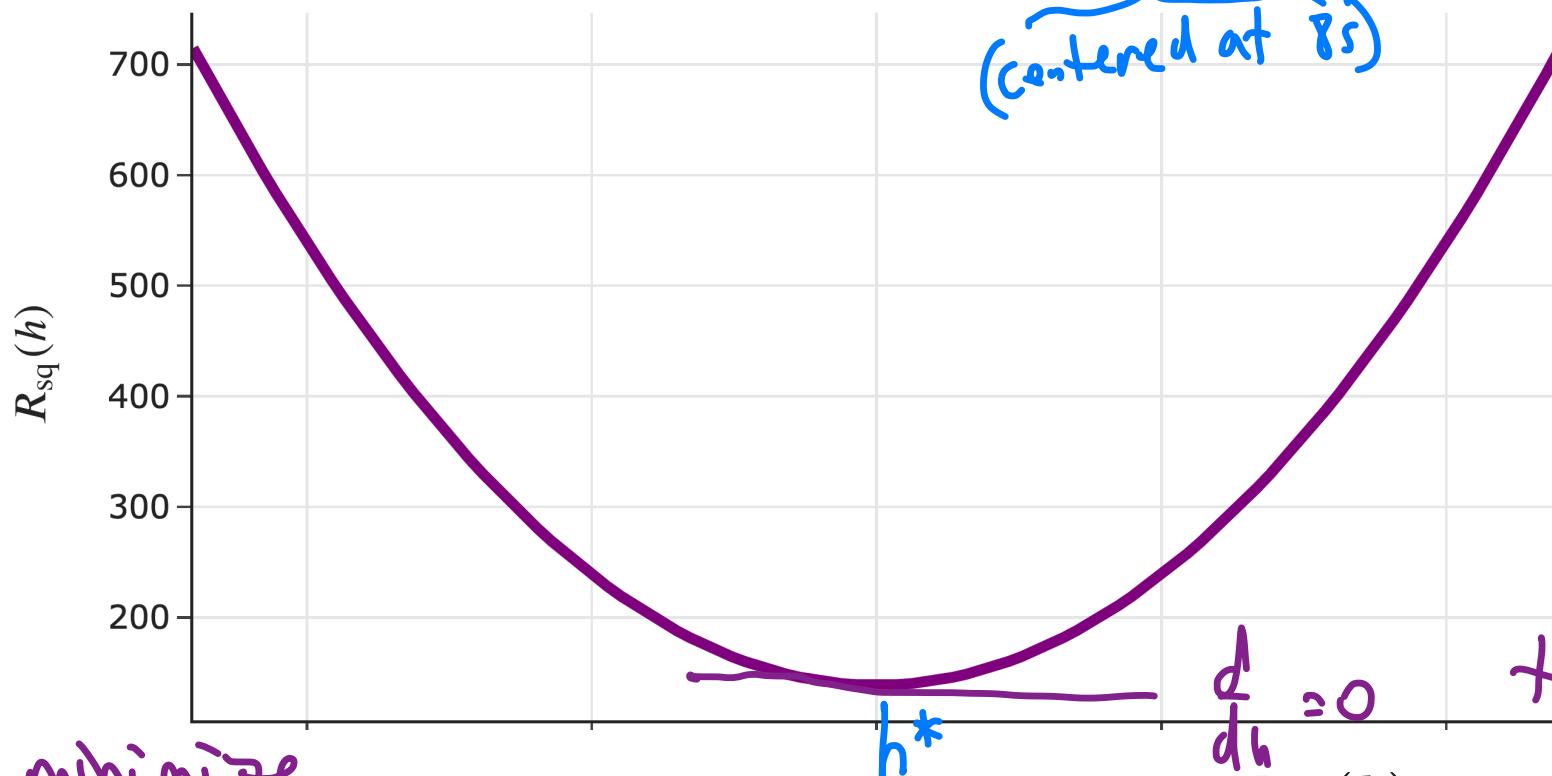
Visualizing mean squared error

each loss function is quadratic

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

(centered at 85)

$\cup_{72} + \cup_{90} + \cup_{61}$
 $+ \cup_{85} + \cup_{92}$



Want to minimize
MSE and find h^*

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

tangent is
flat

Mean squared error, in general

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

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

averaging (pointing to $\frac{1}{n}$)

sum of squared losses (under the bracket)

- Or, using summation notation:

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

equivalent (pointing from the summation notation to the expanded form above)

The best prediction

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

Handwritten annotations:
A blue arrow points from y_i to the word "known".
An orange arrow points from h to the word "unknown".

- 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^* ?

Using 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:

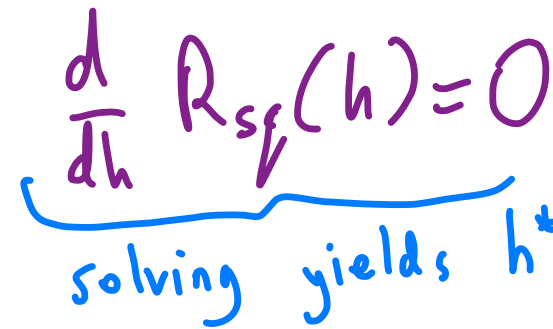
Minimizing using calculus

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.



$\frac{d}{dh} R_{\text{sq}}(h) = 0$
solving yields h^*

Calculus reminders

- 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 .
- Remember from calculus that:
 - $\frac{d}{dx}x^n = nx^{n-1}$
 - $\frac{d}{dx}(f(g(x))) = f'(g(x)) \cdot g'(x)$

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

- To take the derivative of $R_{\text{sq}}(h)$, we'll first need to find the derivative of $(y_i - h)^2$.

$$\begin{aligned} \frac{d}{dh} (y_i - h)^2 &= 2(y_i - h) \underbrace{\frac{d}{dh} (y_i - h)}_{\text{chain rule}} \\ &= 2(y_i - h) \cdot (-1) \\ &= -2(y_i - h) = 2(h - y_i) \end{aligned}$$

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

Question 🤔

Answer at q.dsc40a.com

$$\begin{aligned}\frac{d}{dh} (y_i - h)^2 &= 2(h - y_i) \\ &= -2(y_i - h)\end{aligned}$$

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

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

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

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

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

reminder:

if $c(x) = k a(x)$

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

then: $\frac{d}{dx} c(x) = k \cdot \frac{d}{dx} a(x)$

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

pull the constant

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

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

/ multiply by $\frac{n}{2}$

$$\underbrace{h+h+\dots+h}_{n \text{ times}} \\ = nh \\ = h \sum_{i=1}^n 1$$

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

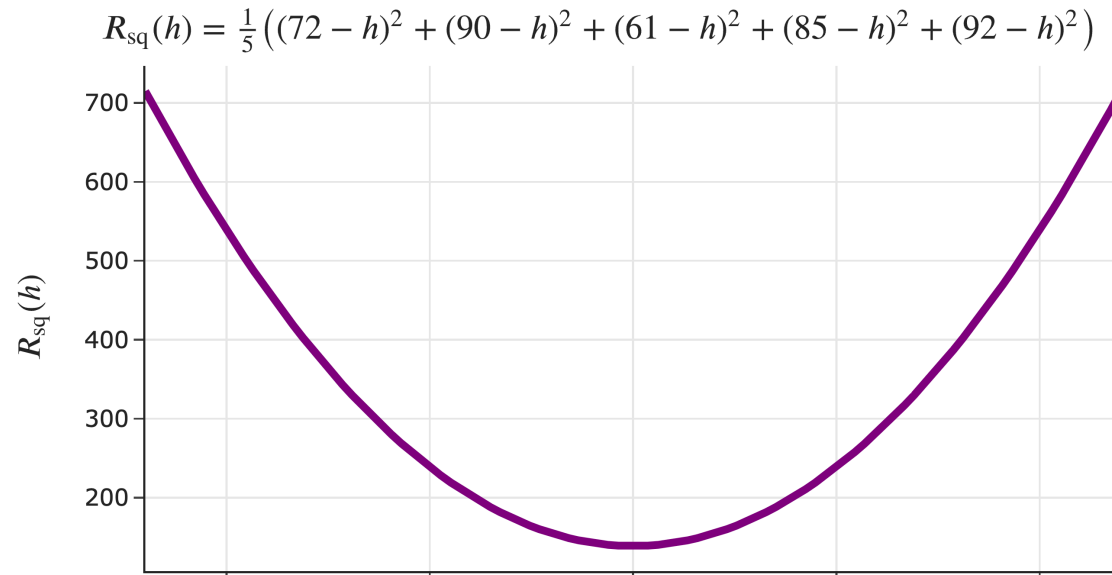
$$\sum_{i=1}^n h - \sum_{i=1}^n y_i = 0$$

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

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

$$\boxed{h^* = \frac{1}{n} \sum_{i=1}^n y_i} \\ = \frac{y_1 + y_2 + \dots + y_n}{n} \\ = \text{mean}\{y_1, \dots, y_n\}$$

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.

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:

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

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

is

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

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.
2. Choose a loss function.
3. Minimize average loss to find optimal model parameters.

Question 🤔

Answer at q.dsc40a.com

What questions do you have?

Summary

- We started with the abstract problem:
 - Given historical commute times, predict your future commute time.
- We've turned it into a formal optimization problem:
 - Find the prediction h^* that has the smallest mean squared error $R_{\text{sq}}(h)$ on the data.
- Implicitly, we introduced a three-step modeling process that we'll keep revisiting:
 - i. Choose a model.
 - ii. Choose a loss function.
 - iii. Minimize average loss, R .
- $R_{\text{sq}}(h)$ is an example of **empirical risk** (average loss) minimized by the mean.