# EECS 245 Fall 2025
# Math for ML

Lecture 2: Models and Loss Functions

→ Read Ch. 1.2 and 1.3

→ Announcements on Ed

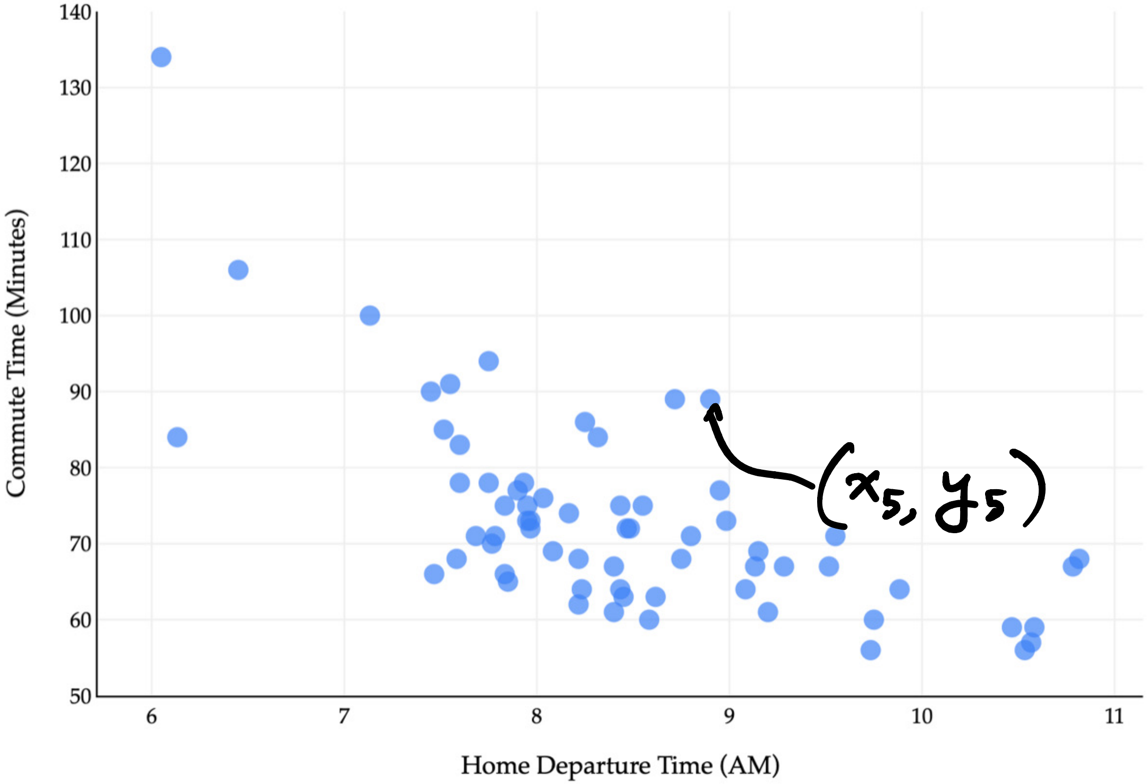
# Agenda

1. Hypothesis functions and parameters
2. Loss functions
3. Finding optimal model parameters using calculus
4. (Time permitting) "Empirical risk minimization" and another loss function

Ch. 1.2

Ch. 1.3

Mostly done, but still adding some content to the end



140
130
120
110
100
90
80
70
60
50
Commute Time (Minutes)
6
7
8
9
10
11
Home Departure Time (AM)
$(x_5, y_5)$

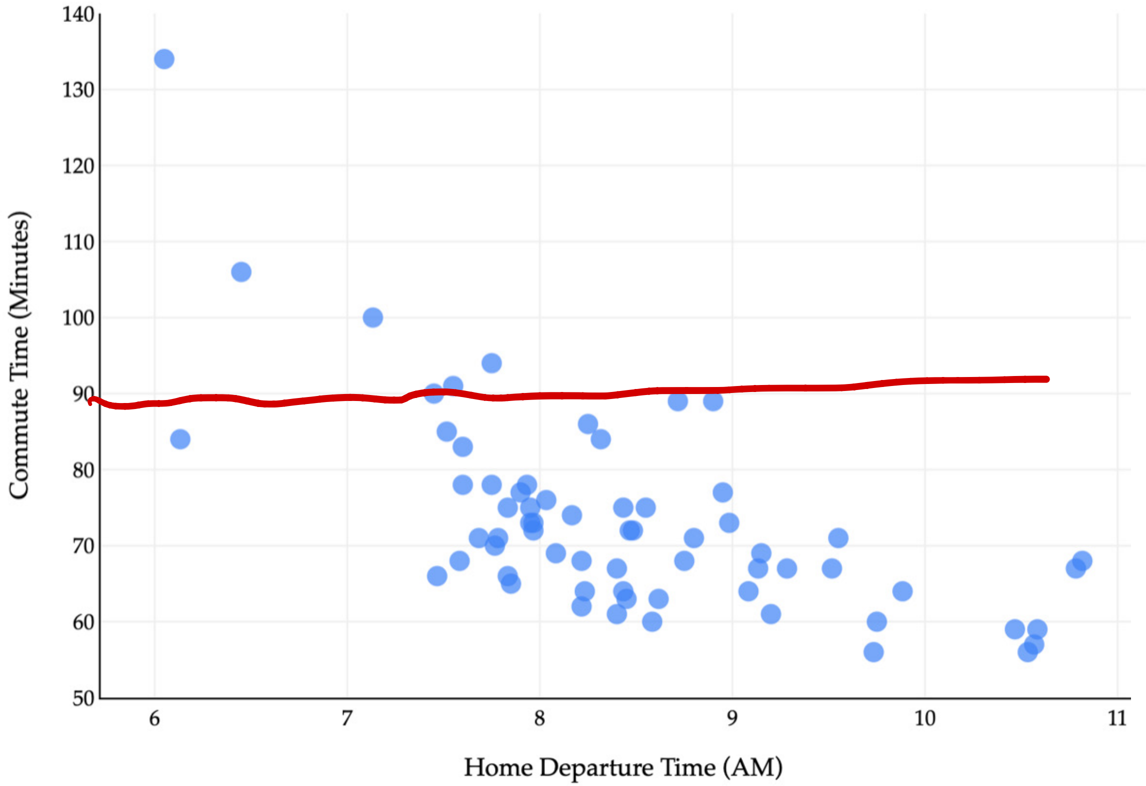
Hypothesis functions, $h$, take in features and return predictions

input/independent variable (here, only feature is departure time)

e.g.

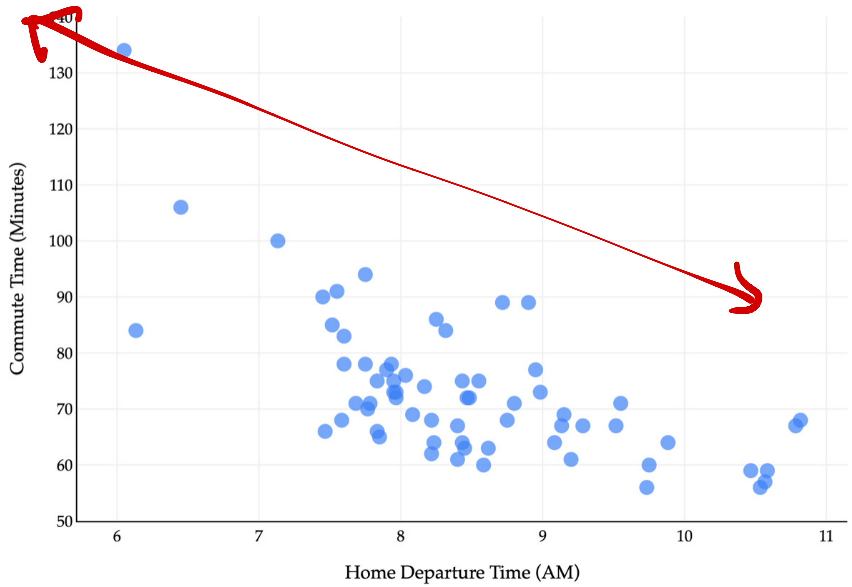
$$h(x_i) = 90$$

"constant model"



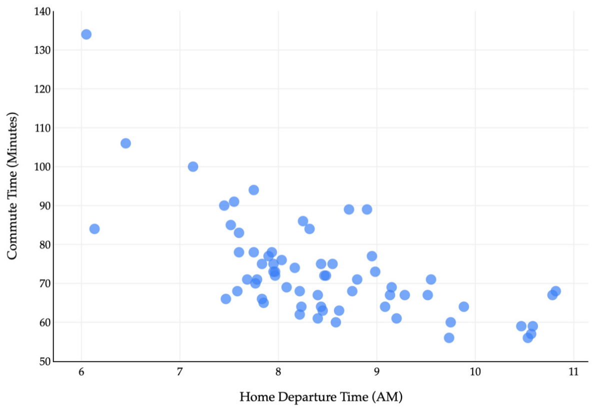

$$h(x_i) = 100 - 3x_i$$

"simple linear regression"



Parameters, $w$

Constant model: $h(x_i) = \underbrace{w}_{\text{the one parameter for the constant model}}$




Simple linear: $h(x_i) = w_0 + w_1 x_i$

slope

intercept

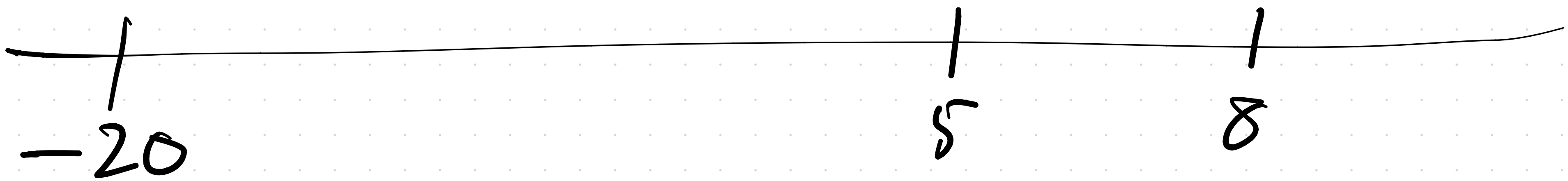
Question! how do we find the best parameters?

"Error"

$$e_i = \underbrace{y_i}_{\text{actual}} - \underbrace{h(x_i)}_{\text{predicted}}$$

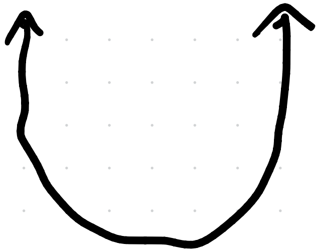
e.g. $y_i = 80$

1) if $h(x_i) = 75 \longrightarrow e_i = 80 - 75 = 5$
2) if $h(x_i) = 72 \longrightarrow e_i = 80 - 72 = 8$
3) if $h(x_i) = 100 \longrightarrow e_i = 80 - 100 = -20$



**Loss functions**: describe the quality of a prediction for a single data point

① Squared loss:



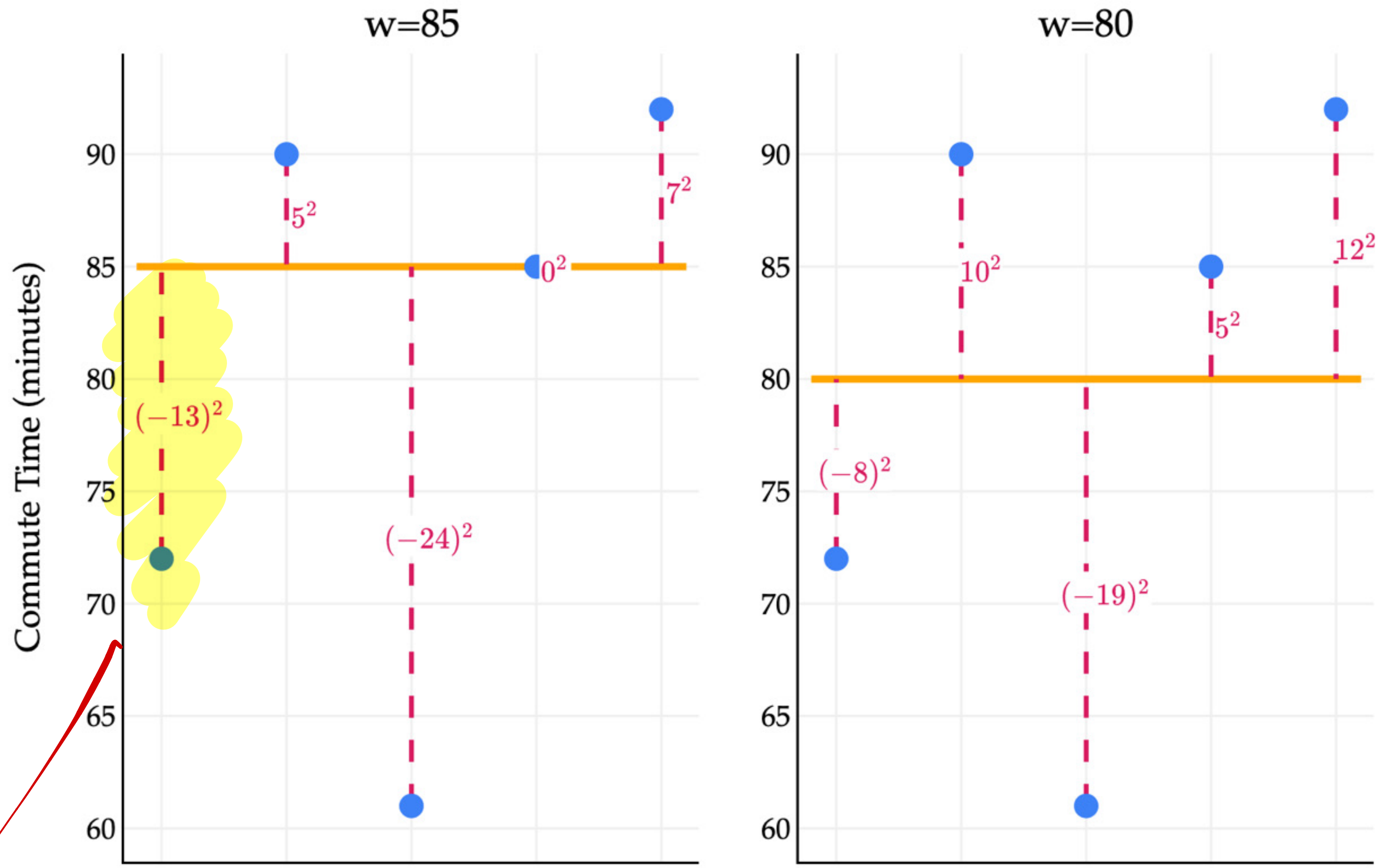
$$L_{sq}(y_i, h(x_i)) = \underbrace{(y_i - h(x_i))^2}_{(\text{actual}-\text{predicted})^2}$$

② $$L_{abs}(y_i, h(x_i)) = |y_i - h(x_i)|$$

loss functions have tradeoffs;

start with squared loss

ex: $y_1 = 72,\ y_2 = 90,\ y_3 = 61,\ y_4 = 85,\ y_5 = 92$




$$L_{sq}(72, 85) = (72 - 85)^2 = (-13)^2 = 169$$

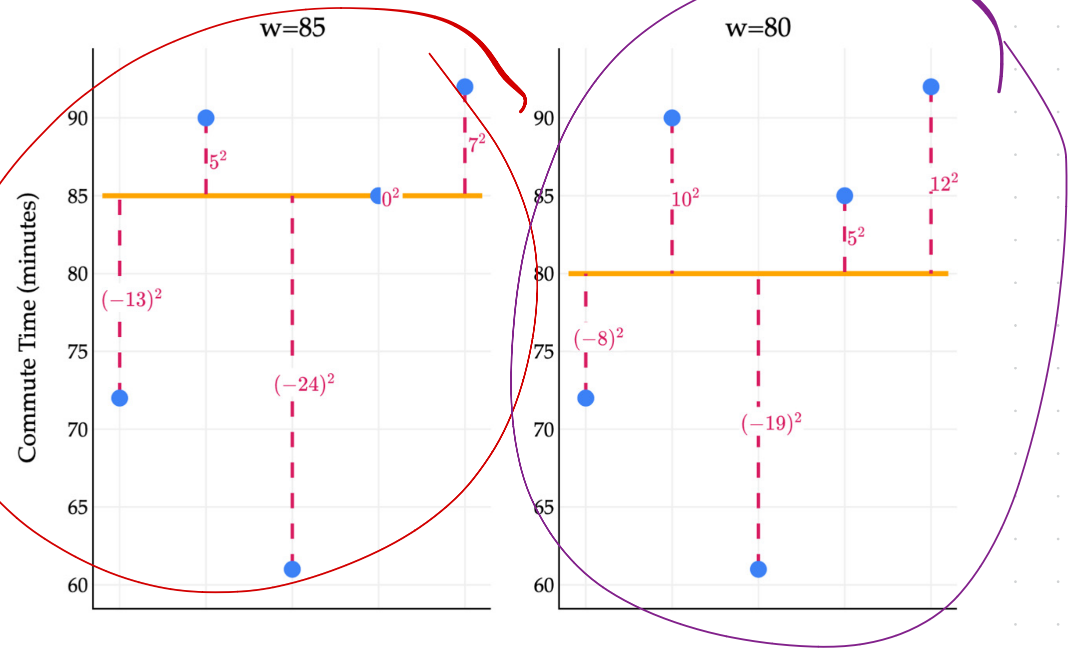
Average squared loss will give me one number that measures the quality of $w=85$

For $w=85 \rightarrow \dfrac{(-13)^2+5^2+(-24)^2+0^2+7^2}{5} = 163.8$

$w=80 \rightarrow \dfrac{(-8)^2+10^2+\dots+12^2}{5}$

$= 138.8$

$138.8 < 163.8$, so $w=80$ better than $w=85$



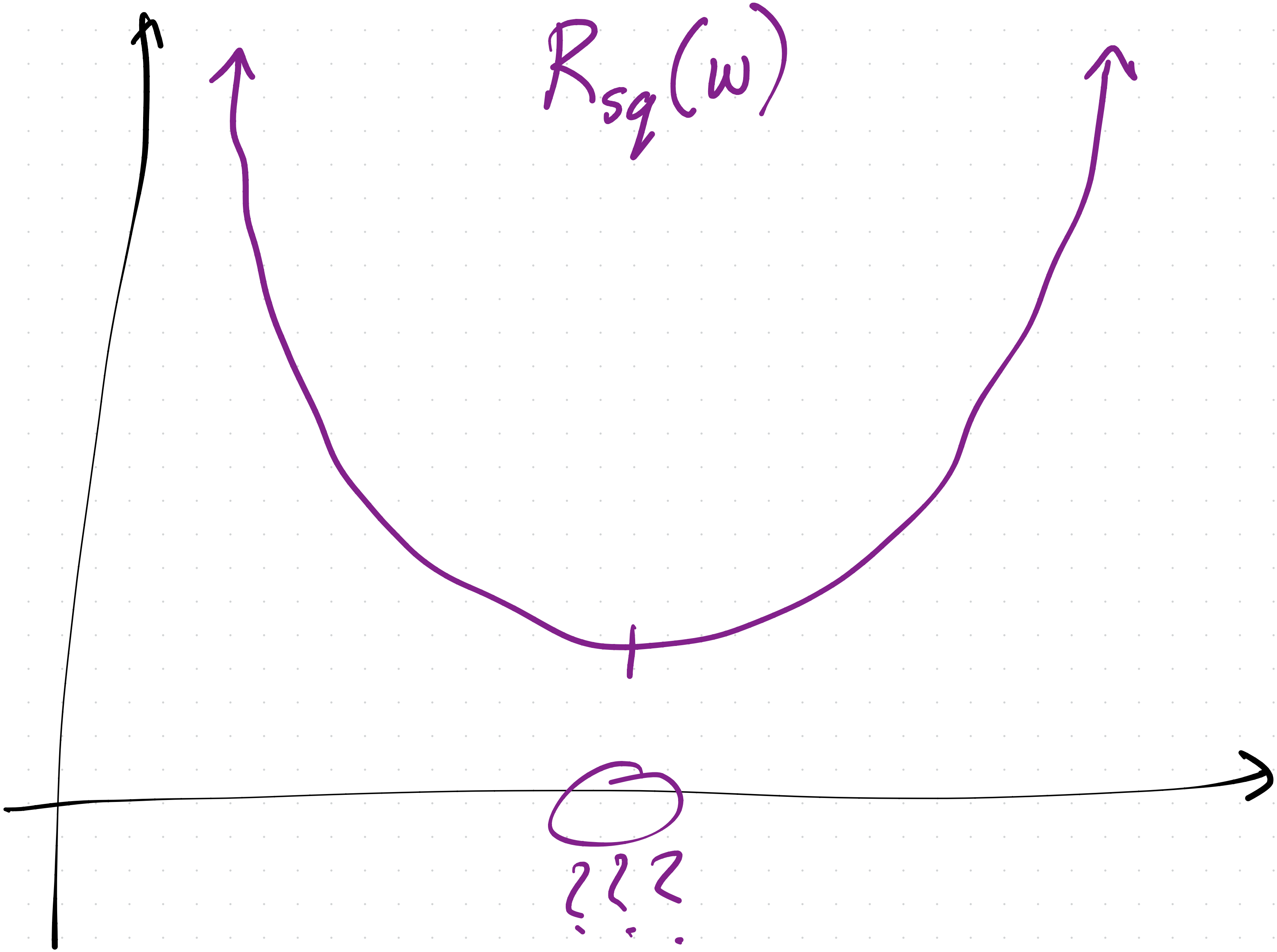
$y_1 = 72,\ y_2 = 90,\ y_3 = 61,\ y_4 = 85,\ y_5 = 92$

For any constant prediction $w$, average sq'd loss:

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

$L$: loss for a single data point

$R$: average loss across data set



$R_{sq}(w)$
???

Given $y_1, y_2, \ldots, y_n$ (all numbers),

Goal is to find the best constant prediction, $w$,

by minimizing $R_{sq}$!

$$R_{sq}(w) = \frac{1}{n}\left((y_1 - w)^2 + (y_2 - w)^2 + \cdots + (y_n - w)^2\right)$$

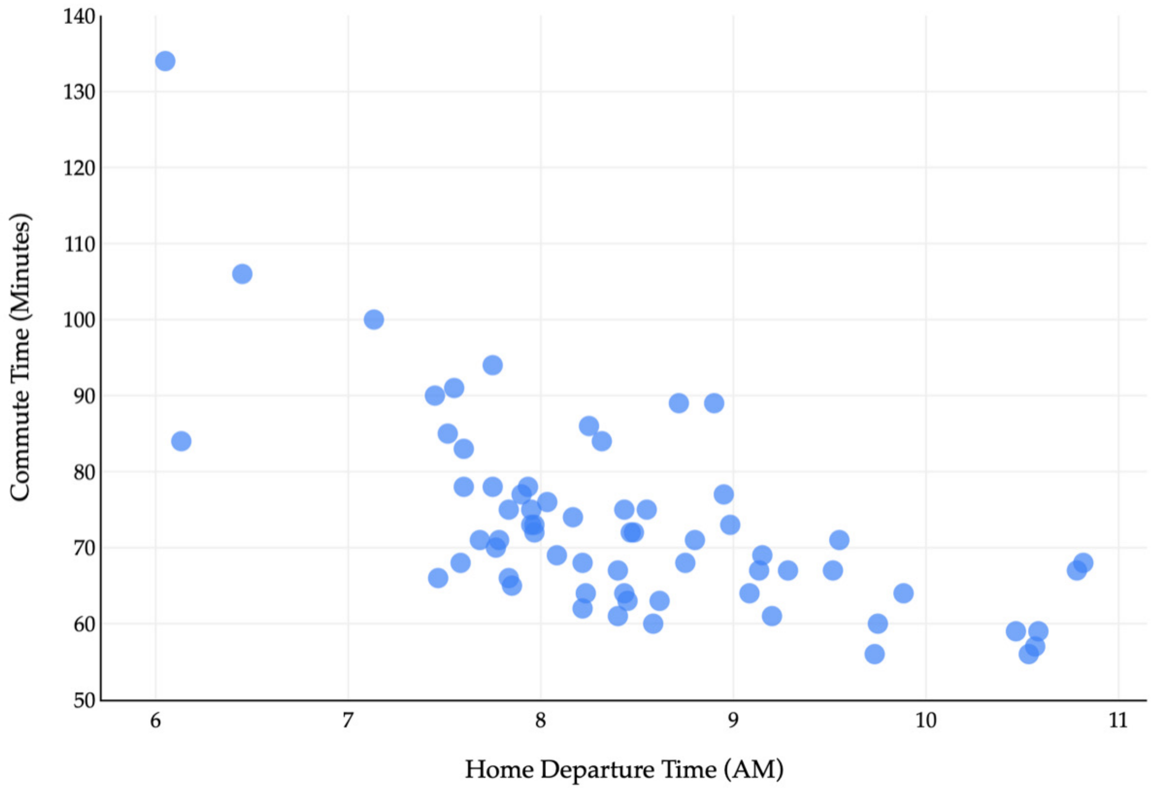
$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - w)^2$$

Tip: Think of $y_i$'s as constants, this is just a function of $w$!

"mean squared error"

// same thing

"average squared loss"



140
130
120
110
100
90
80
70
60
50
Commute Time (Minutes)
6
7
8
9
10
11
Home Departure Time (AM)

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

function of $w$ only!

1. Take derivative w.r.t $w$
2. set to 0
3. Second derivative test

Step 1

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

$$\frac{dR}{dw} = \frac{1}{n}\frac{d}{dw}\sum_{i=1}^{n}(y_i - w)^2$$

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

from next slide

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

$$= \underbrace{\boxed{-\frac{2}{n}\sum_{i=1}^{n}(y_i - w)}}_{\frac{dR}{dw}}$$

$$f(x) + g(x) + h(x)$$

$$\downarrow$$

$$f'(x) + g'(x) + \ldots$$

Aside: what is

$$\frac{d}{dw}(y_i - w)^2$$

$$= 2(y_i - w)^1 * \underbrace{\frac{d(y_i - w)}{dw}}_{\text{chain rule}} \quad -1$$

$$= -2(y_i - w)$$

$$= 2(w - y_i)$$

Step 2

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

$\downarrow$ multiply BS by $\frac{-n}{2}$

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

$$\sum_{i=1}^{n} y_i - \underbrace{\sum_{i=1}^{n} w}_{w+w+\dots+w} = 0$$

$*$ = best/optimal

$$\sum_{i=1}^{n} y_i - nw = 0 \rightarrow \boxed{w^* = \frac{\sum_{i=1}^{n} y_i}{n}}$$

Aside

$$\sum_{i=1}^{5} 2 = \underbrace{2+2+2+2+2}_{5 \text{ times}} = 2 \cdot 5$$

step 3 second derivative test

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

$$\downarrow$$

$$\frac{d^2R}{dw^2} = -\frac{2}{n}\sum_{i=1}^{n}\frac{d}{dw}(y_i - w) \quad = -1$$

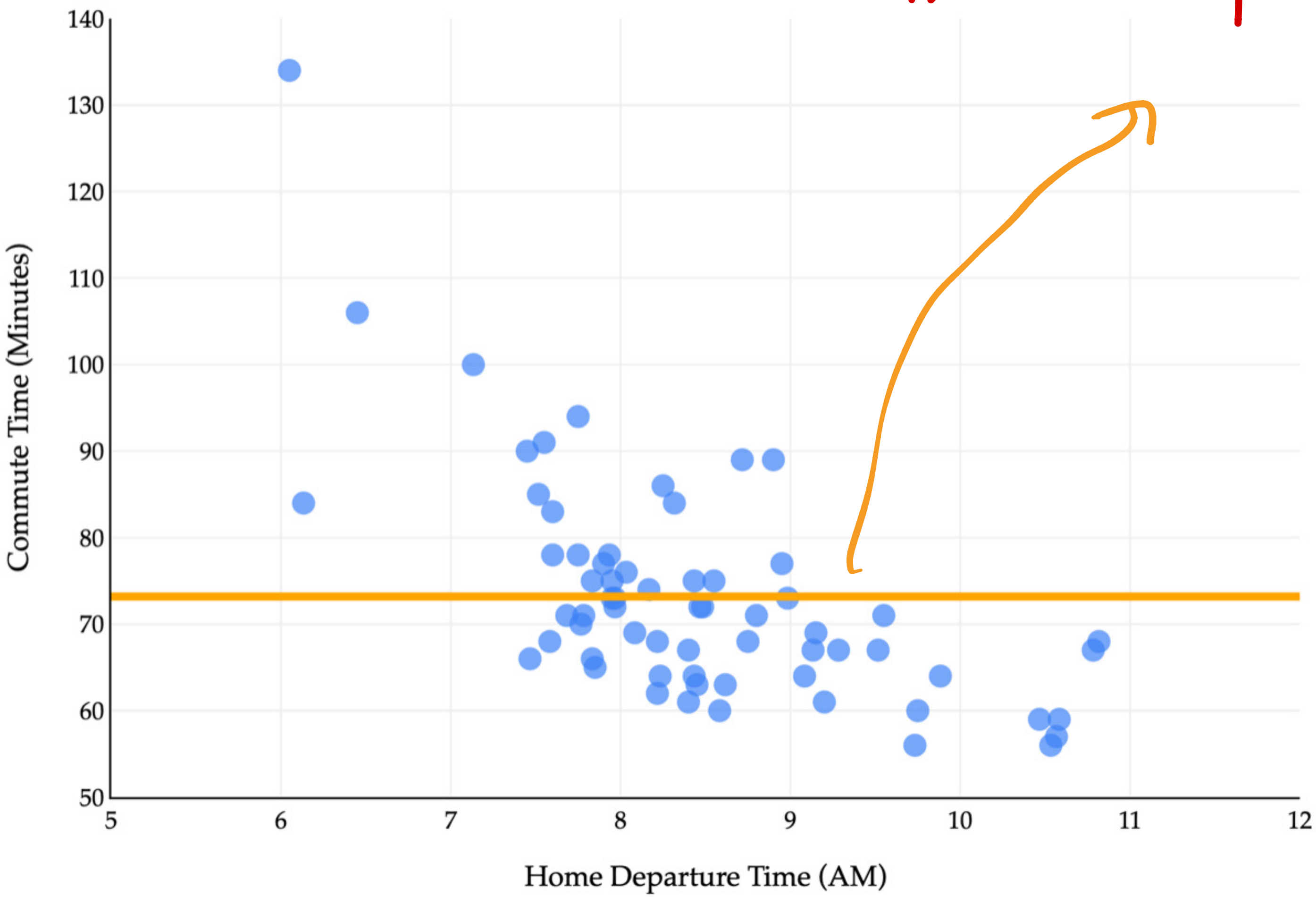
$$= -\frac{2}{n}\underbrace{\sum_{i=1}^{n}(-1)}_{(-n)}$$

$$= \left(-\frac{2}{\cancel{n}}\right)(-\cancel{n}) = \boxed{2} > 0$$

so, $R$ is concave up everywhere $\rightarrow w^*$ is minimum

$$w^* = \text{Mean}(y_1, \ldots, y_n) = \bar{y}$$

is the "optimal constant prediction", when minimizing average squared loss




$w^*$ "optimal model parameter"

$$h(x_i) = w$$

end of Ch. 1.2

start ch. 1.3

"Three-step modeling recipe"

① Choose a model

$$h(x_i) = w$$

"constant model"

② Choose a loss function

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

"squared loss"

③ Minimize average loss to find optimal parameters

$$R_{sq}(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - w)^2 \xrightarrow[\text{calc.}]{} \boxed{w^* = \bar{y}}$$

① Constant model $h(x_i) = w$

② Absolute loss

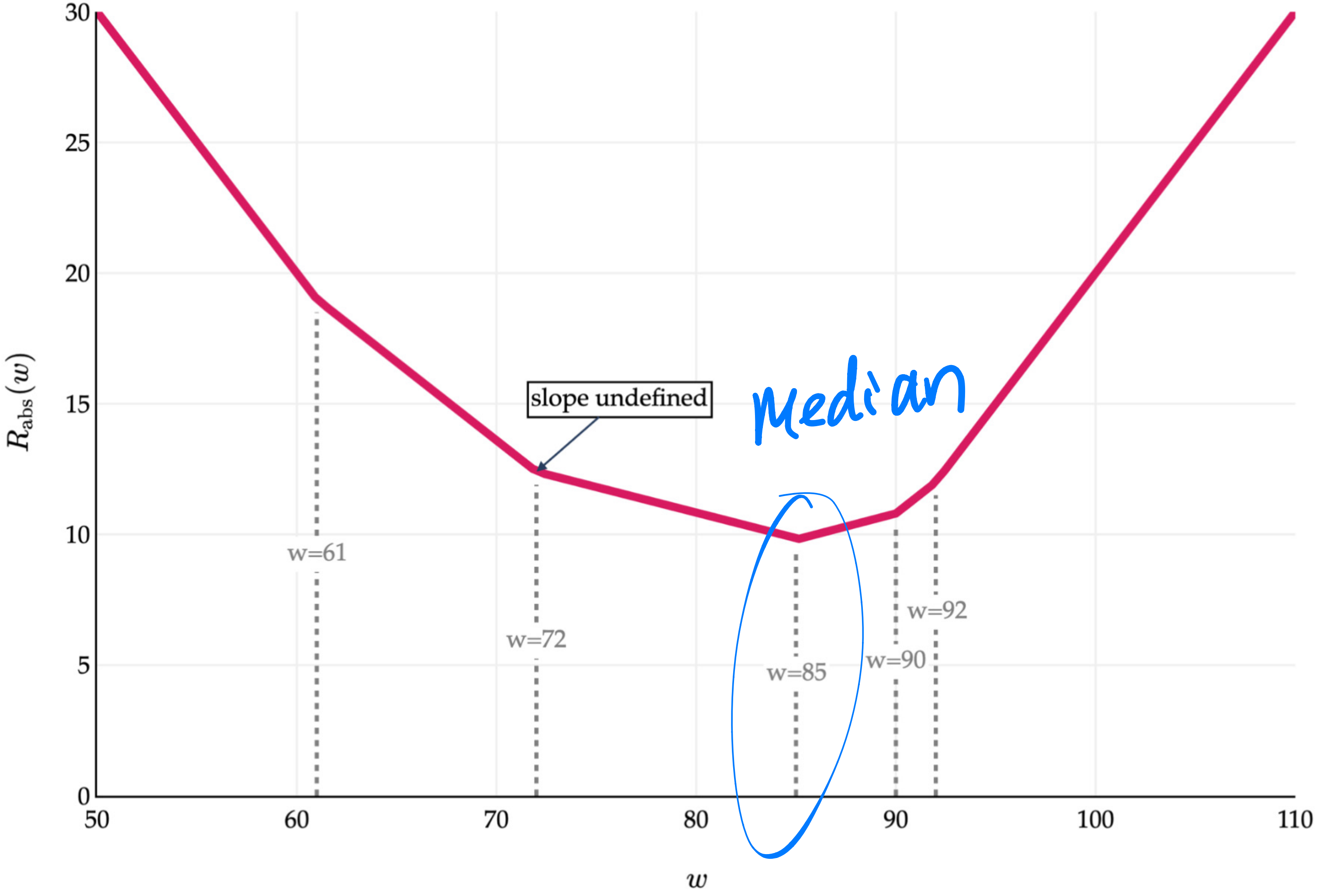
③

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

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

"mean absolute error"

$$R_{\text{abs}}(w) = \frac{1}{5}(|72 - w| + |90 - w| + |61 - w| + |85 - w| + |92 - w|)$$



even # of points

$$R_{\text{abs}}(w) = \frac{1}{6}(|72 - w| + |90 - w| + |61 - w| + |85 - w| + |92 - w| + |78 - w|)$$

And its graph is: