Lecture 5

# Simple Linear Regression

DSC 40A, Fall 2024

## Announcements

- Homework 1 is due **Friday night**.
- Look at the office hours schedule here and plan to start regularly attending!
- Remember to take a look at the supplementary readings linked on the course website.

# Agenda

- 0-1 loss
- Predictin rules using features
- Simple linear regression.
- Minimizing mean squared error for the simple linear model.

## Question 🤔

**Answer at q.dsc40a.com**

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

If the direct link doesn't work, click the "🤔 Lecture Questions" link in the top right corner of dsc40a.com.

## Another example: 0-1 loss

Consider, for example, the **0-1 loss**:

$$L_{0,1}(y_i, h) = \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$$

The corresponding empirical risk is:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{0,1}(y_i, h)$$

## Question 🤔

**Answer at q.dsc40a.com**

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$$

Suppose $y_1, y_2, \ldots, y_n$ are all unique. What is $R_{0,1}(y_1)$?

- A. 0.
- B. $\frac{1}{n}$.
- C. $\frac{n-1}{n}$.
- D. 1.

# Minimizing empirical risk for 0-1 loss

$$R_{0,1}(h) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}0 & y_i = h\\ 1 & y_i \neq h\end{cases}$$

# Summary: Choosing a loss function

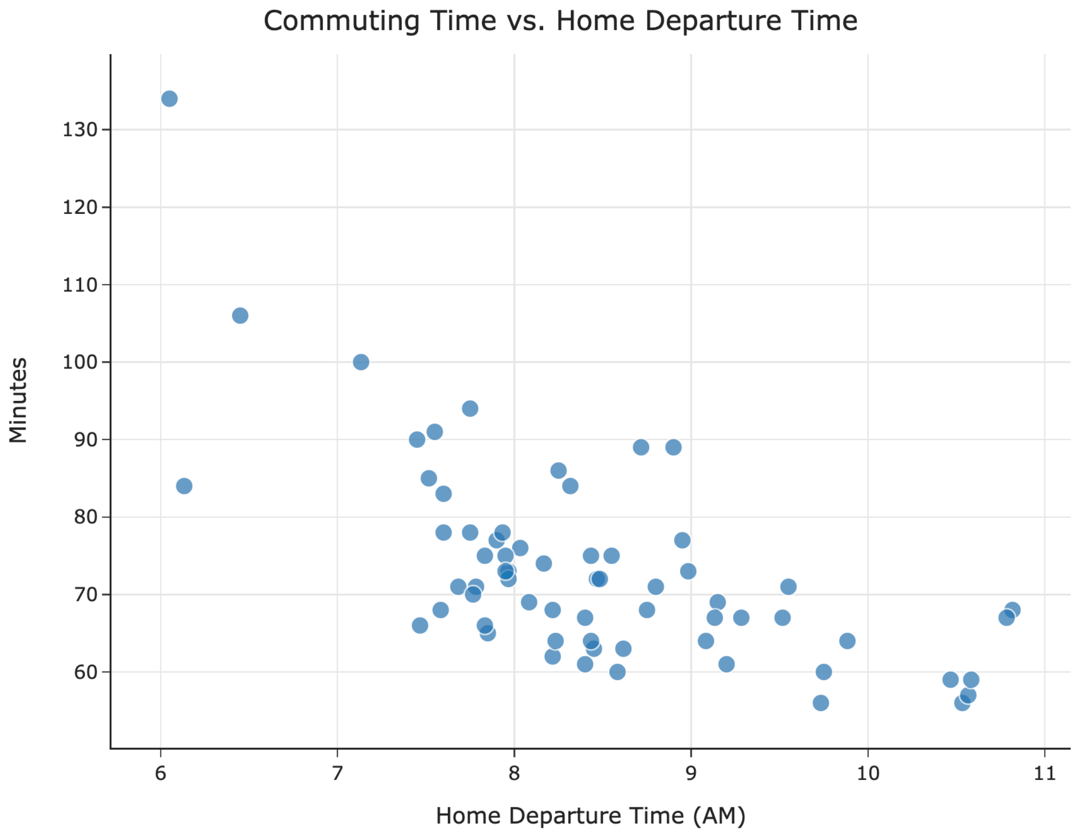
**Key idea**: Different loss functions lead to different best predictions, $h^*$!

| Loss | Minimizer | Always Unique? | Robust to Outliers? | Differentiable? |
|---|---|---|---|---|
| $L_{\text{sq}}$ | mean | yes ✅ | no ❌ | yes ✅ |
| $L_{\text{abs}}$ | median | no ❌ | yes ✅ | no ❌ |
| $L_{\infty}$ | midrange | yes ✅ | no ❌ | no ❌ |
| $L_{0,1}$ | mode | no ❌ | yes ✅ | no ❌ |

The optimal predictions, $h^*$, are all **summary statistics** that measure the **center** of the dataset in different ways.

# Predictions with features

# Towards simple linear regression



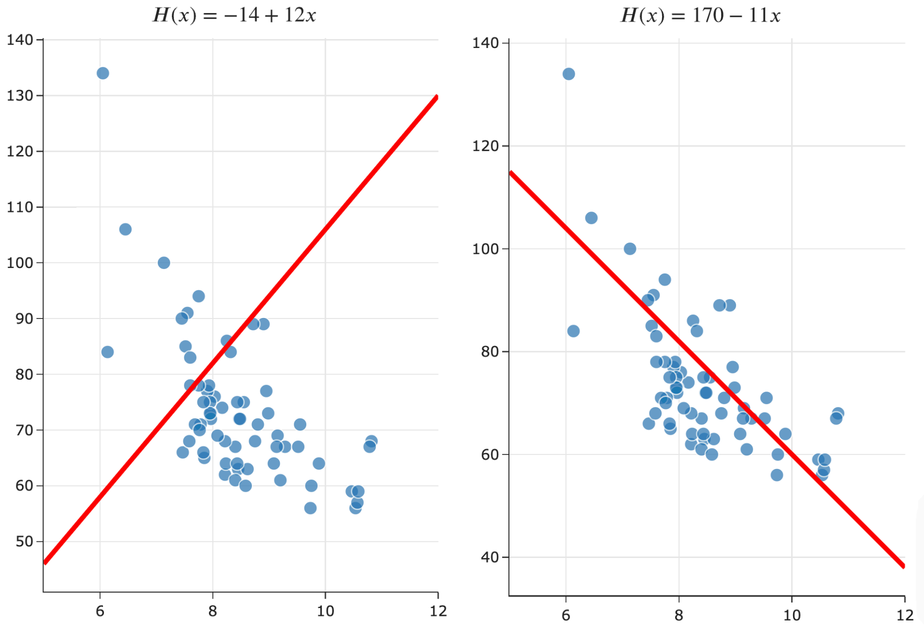

- In Lecture 1, we introduced the idea of a hypothesis function, $H(x)$.
- We've focused on finding the best **constant model**, $H(x) = h$.
- Now that we understand the modeling recipe, we can apply it to find the best **simple linear regression model**, $H(x) = w_0 + w_1 x$.
- This will allow us to make predictions that aren't all the same for every data point.

# Recap: Hypothesis functions and parameters

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

**Parameters** define the relationship between the input and output of a hypothesis function.

The simple linear regression model, $H(x) = w_0 + w_1 x$, has two parameters: $w_0$ and $w_1$.



# The modeling recipe

1. Choose a model.

2. Choose a loss function.

3. Minimize average loss to find optimal model parameters.

## Features

A **feature** is an attribute of the data – a piece of information.

- **Numerical**: maximum allowed speed, time of departure
- **Categorical**: day of week
- **Boolean**: was there a car accident on the road?

Think of features as columns in a DataFrame (i.e. table).

| Departure time | Day of week | Accident on route | Commute time |
|---|---|---|---|
| 7:05 | Monday | yes | 101 |
| 8:03 | Tuesday | no | 87 |
| 10:20 | Wednesday | yes | 79 |
| 8:30 | Thursday | no | 76 |

# Modeling

- We believe that commute time is a function of departure time.
- I.e., there is a function $H$ so that:

  commute time $\approx H$(departure time)
- $H$ is called a **hypothesis function** or **prediction rule**.
- Our goal: find a good prediction rule $H$.

## Possible Hypothesis Functions

- $H_1$(departure time) = 90 - 10 ·(departure time-7)
- $H_2$(departure time) = 90 - (departure time-8)$^2$
- $H_3$(departure time) = 20 + 6·departure time

These are all valid prediction rules.

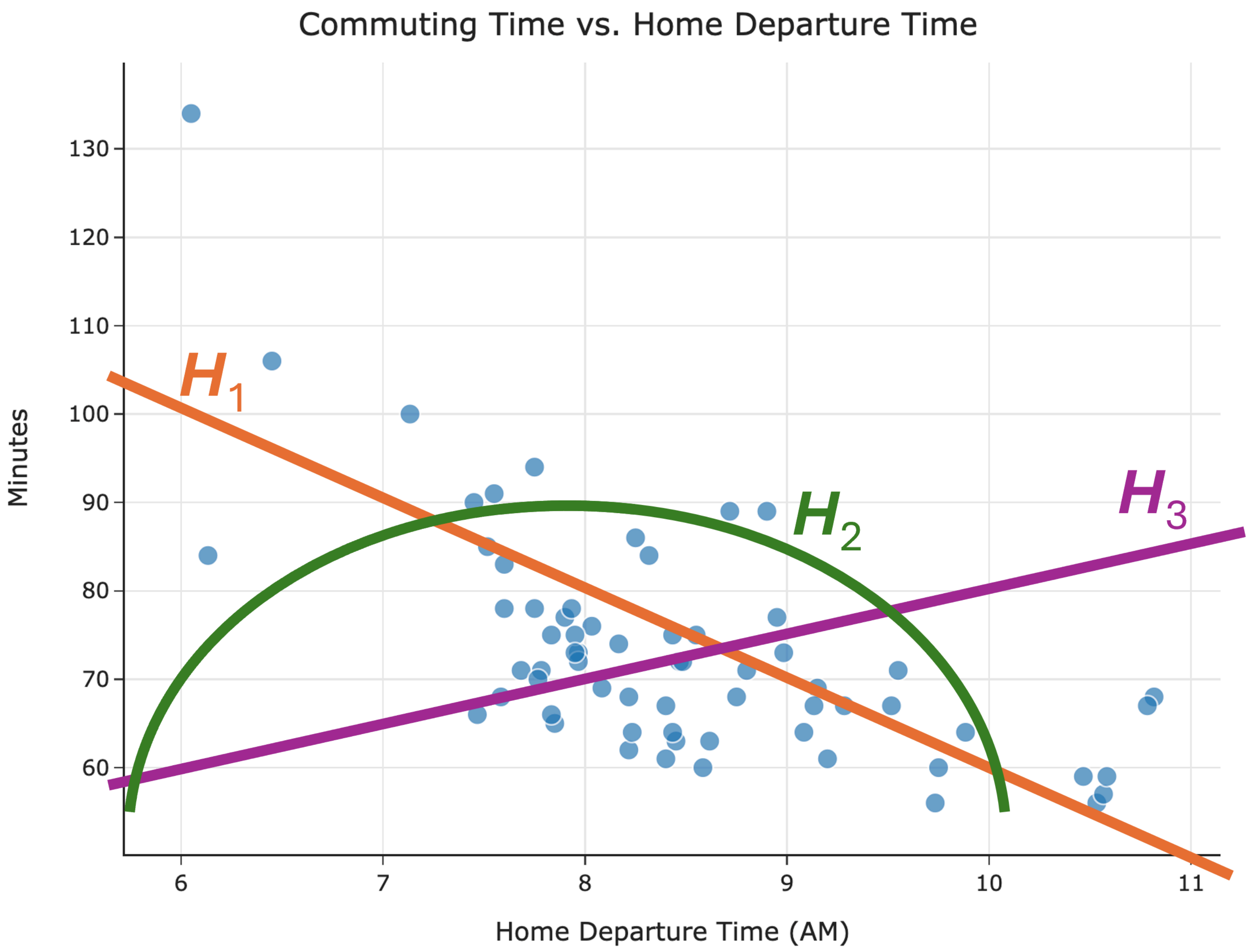
Some are better than others.

## Comparing predictions

- How do we know which hyppthesis is best: $H_1,\ H_2,\ H_3$?
- We gather data from $n$ days of commute. Let $x_i$ be departure time, $y_i$ be commute time:

$$\begin{array}{lcl} (\text{departure time}_1\ ,\ \text{commute time}_1) & & (x_1, y_1) \\ (\text{departure time}_2\ ,\ \text{commute time}_2) & & (x_2, y_2) \\ \ldots & \rightarrow & \\ (\text{departure time}_n\ ,\ \text{commute time}_n) & & (x_n, y_n) \end{array}$$

- See which rule works better on data.


Commuting Time vs. Home Departure Time
Minutes
130
120
110
100
90
80
70
60
6
7
8
9
10
11
Home Departure Time (AM)
$H_1$
$H_2$
$H_3$

## Quantifying the performance of a model

- Reminder: one loss function, which measures how far $H(x_i)$ is from $y_i$, is **absolute loss**.

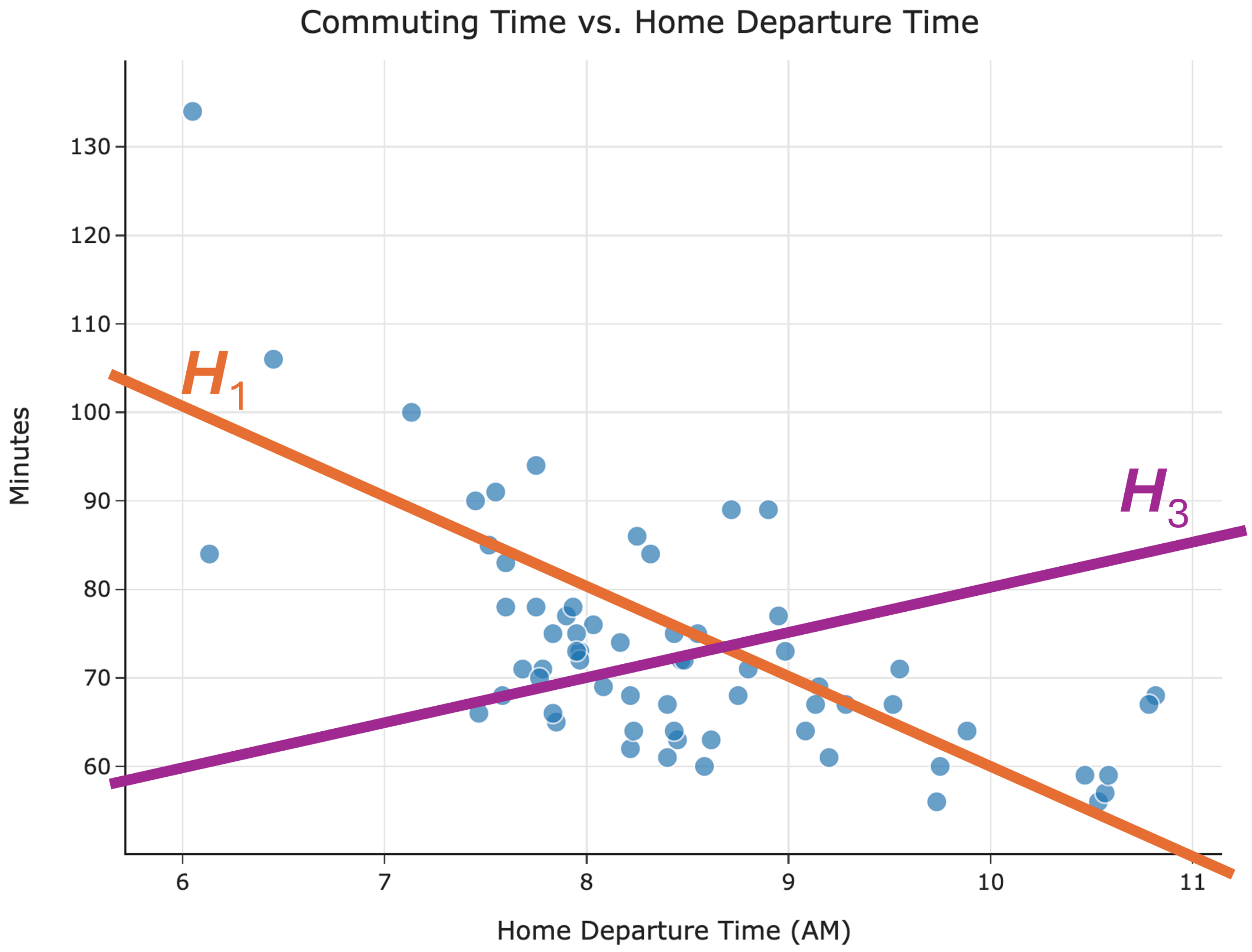
- The mean absolute error of $H(x)$ is

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

- We want the **best** prediction, $H^*(x)$.

- The smaller $R_{\text{abs}}(h)$ is, the better the hypothesis.

# Mean absolute error



# Finding the best hypothesis $H(x)$

- Goal: out of all functions$\mathbb{R} \to \mathbb{R}$, find the function $H$ with the smallest mean absolute error.
- That is, $H^*$ should be the function that minimizes

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

## Finding the best hypothesis $H(x)$

- Goal: out of all functions $\mathbb{R} \to \mathbb{R}$, find the function $H$ with the smallest mean absolute error.
- That is, $H^*$ should be the function that minimizes

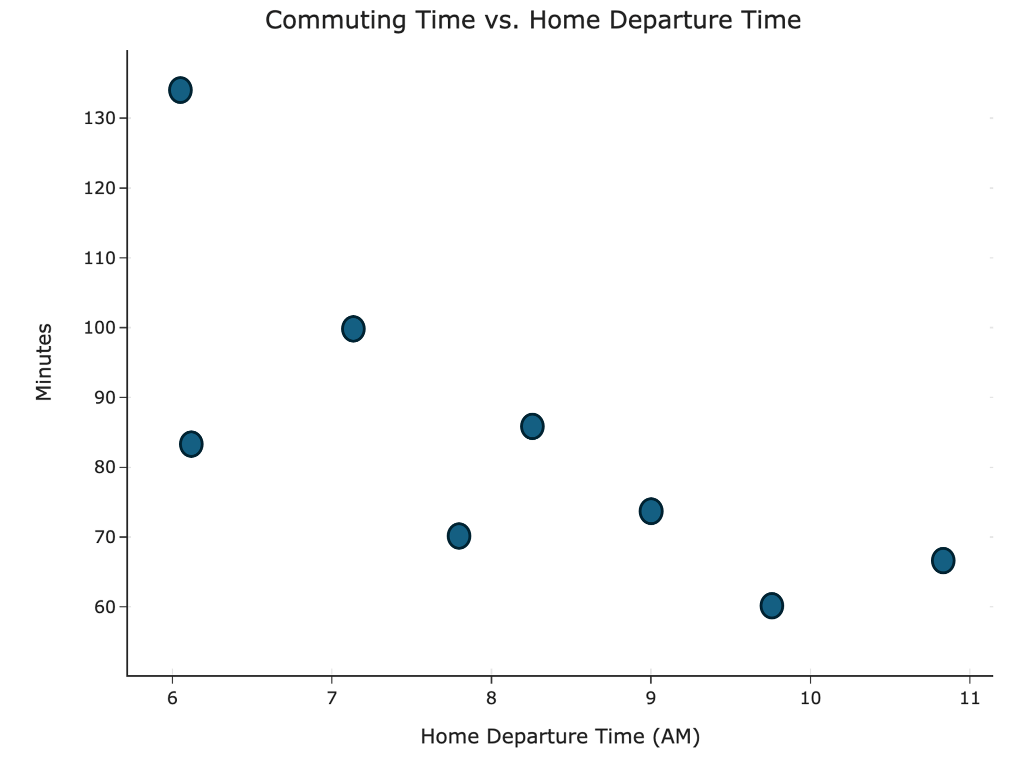
$$R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - H(x_i))|$$

- **There are two problems with this.**

# Question 🤔 Answer at q.dsc40a.com

Given the data below, is there a prediction rule H which has zero mean absolute error?

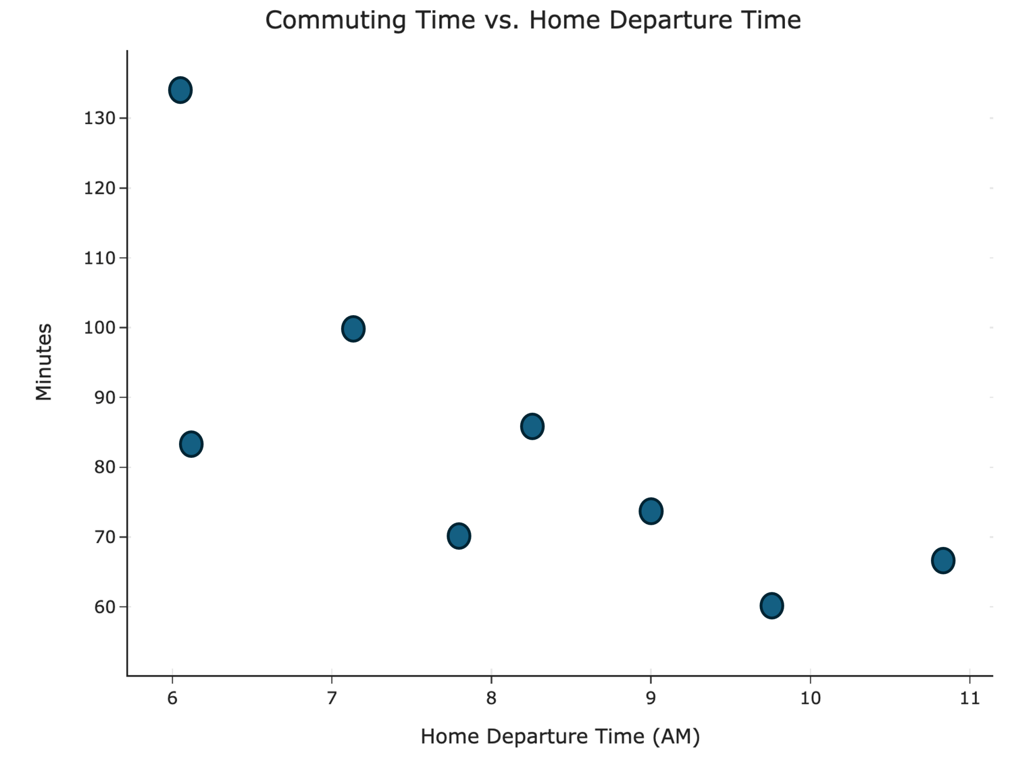
- A. yes
- B. no



## Question 🤔 Answer at q.dsc40a.com

Given the data below, is there a prediction rule H which has zero mean absolute error?

- A. yes
- B. no



## Problem

- We can make mean squared error very small, even zero!
- But the function will be weird.
- This is called **overfitting**.
- Remember our real goal: make good predictions on data **we haven't seen**.

## Solution

- Don't allow $H$ to be just any function.
- Require that it has a certain form.
- Examples:
  - Linear: $H(x) = w_0 + w_1 x$.
  - Quadratic: $H(x) = w_0 + w_1 x_1 + w_2 x^2$.
  - Exponential: $H(x) = w_0 e^{w_1 x}$.
  - Constant: $H(x) = w_0$.

## Finding the best linear model

- **Goal:** Out of all **linear** functions $\mathbb{R} \to \mathbb{R}$, find the function $H^*$ with the smallest mean squared error.
  - Linear functions are of the form $H(x) = w_0 + w_1 x$.
  - They are defined by a slope ($w_1$) and intercept ($w_0$).
- That is, $H^*$ should be the linear function that minimizes

$$R_{abs}(H) = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - H(x_i) \right|$$

## Finding the best linear model

- **Goal:** Out of all **linear** functions $\mathbb{R} \to \mathbb{R}$, find the function $H^*$ with the smallest mean squared error.
  - Linear functions are of the form $H(x) = w_0 + w_1 x$.
  - They are defined by a slope ($w_1$) and intercept ($w_0$).
- That is, $H^*$ should be the linear function that minimizes

$$R_{abs}(H) = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - H(x_i) \right|$$

- 
  - **There is still a problem with this.**

## Problem #2

It is hard to minimize the mean absolute error:

$$R_{abs}(H) = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - H(x_i) \right|$$

- Not differentiable!
- What can we do?

## Minimizing mean squared error for the simple linear model

- We'll choose squared loss, since it's the easiest to minimize.
- Our goal, then, is to find the linear hypothesis function $H^*(x)$ that minimizes empirical risk:

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - H(x_i)\right)^2$$

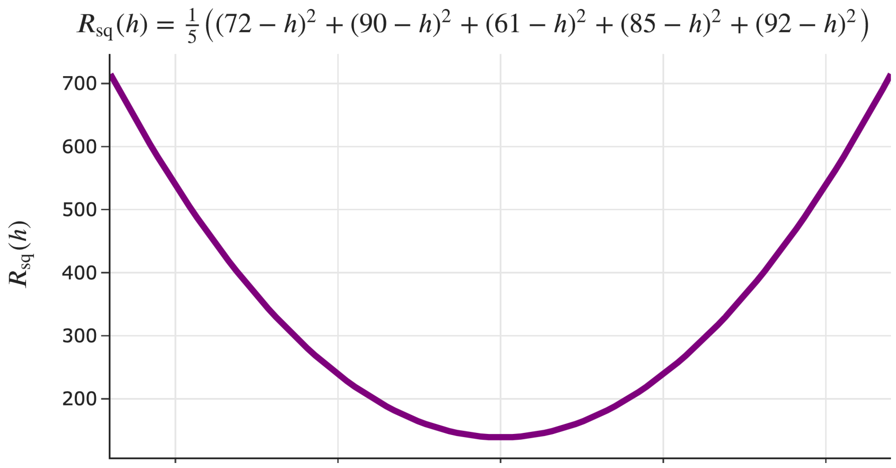
- Since linear hypothesis functions are of the form $H(x) = w_0 + w_1 x$, we can re-write $R_{\text{sq}}$ as a function of $w_0$ and $w_1$:

$$\boxed{R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i)\right)^2}$$

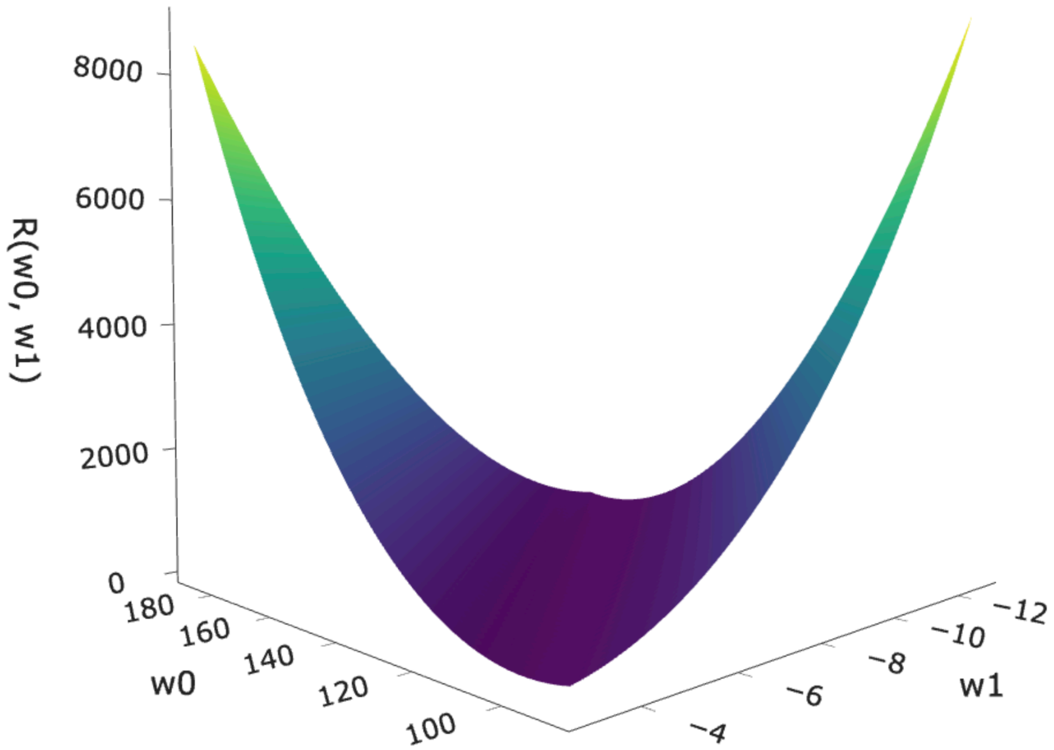
- **How do we find the parameters $w_0^*$ and $w_1^*$ that minimize $R_{\text{sq}}(w_0, w_1)$?**

# Loss surface

For the constant model, the graph of $R_{\text{sq}}(h)$ looked like a parabola.




What does the graph of $R_{\text{sq}}(w_0, w_1)$ look like for the simple linear regression model?



# Minimizing mean squared error for the simple linear model

## Minimizing multivariate functions

- Our goal is to find the parameters $w_0^*$ and $w_1^*$ that minimize mean squared error:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (w_0 + w_1 x_i)\right)^2$$

- $R_{\text{sq}}$ is a function of two variables: $w_0$ and $w_1$.
- To minimize a function of multiple variables:
  - Take partial derivatives with respect to each variable.
  - Set all partial derivatives to 0.
  - Solve the resulting system of equations.
  - Ensure that you've found a minimum, rather than a maximum or saddle point (using the second derivative test for multivariate functions).