Lectures 5-7

Simple Linear Regression

DSC 40A, Fall 2025

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
- Prediction 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

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

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

$\parallel$

$R_{0,1}(h = y_1)$

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

Proportion of all pts different from y_1

If $h \neq y_i$ for any y_i:

$$R_{0,1}(h) = \frac{1}{n} \sum_{i=1}^{n} 1 = \frac{1}{n}(\underbrace{1 + 1 + \ldots + 1}_{n \text{ times}}) = \frac{n}{n} = 1$$

Minimizing empirical risk for 0-1 loss

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

= proportion of pts not equal to h

$R_{0,1}$ is minimized for the value that is most frequent (appears the most) in the data

$h^* = \text{Mode}(y_1, y_2, \ldots, y_n)$
isn't necessarily unique (Ex: $\{1, 2, 3, 14\}$
$\{1, 2, 2, 3, 3, 5\}$)

Summary: Choosing a loss function

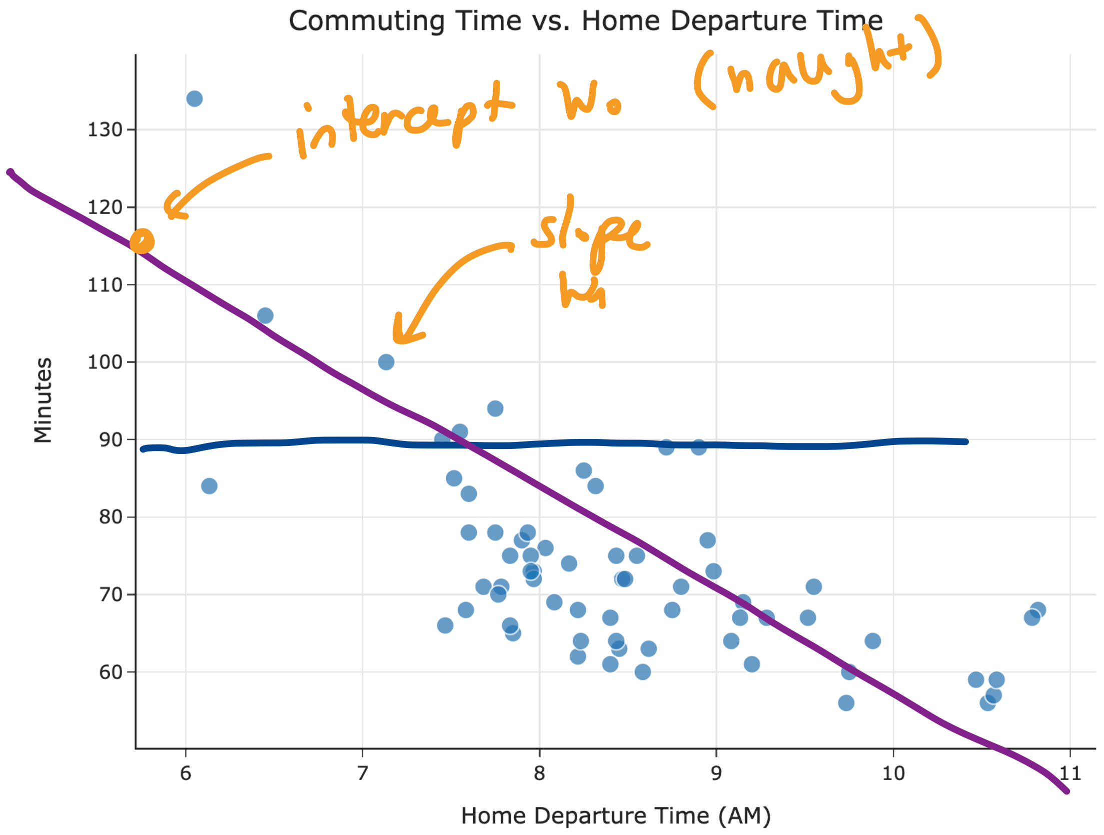
Key idea: Different loss functions lead to different best predictions, h^*!

Loss	Minimizer	Always Unique?	Robust to Outliers?	Differentiable?
L_{sq}	mean	yes ✅	no ❌	yes ✅
L_{abs}	median	no ❌	yes ✅	no ❌
L_{∞}	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.

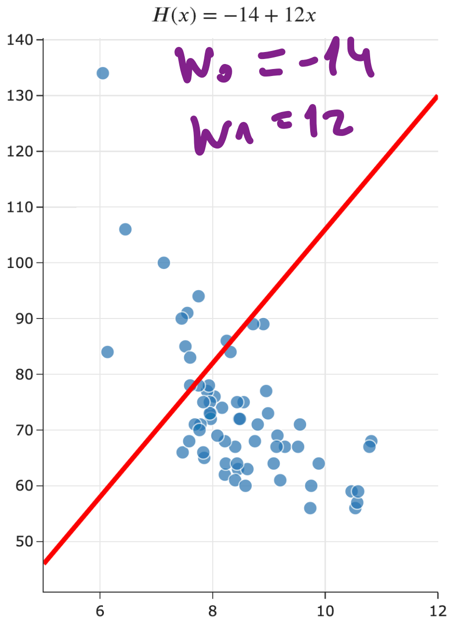
$H\left(\text{departure time}\right) \longrightarrow$ predicted commute time

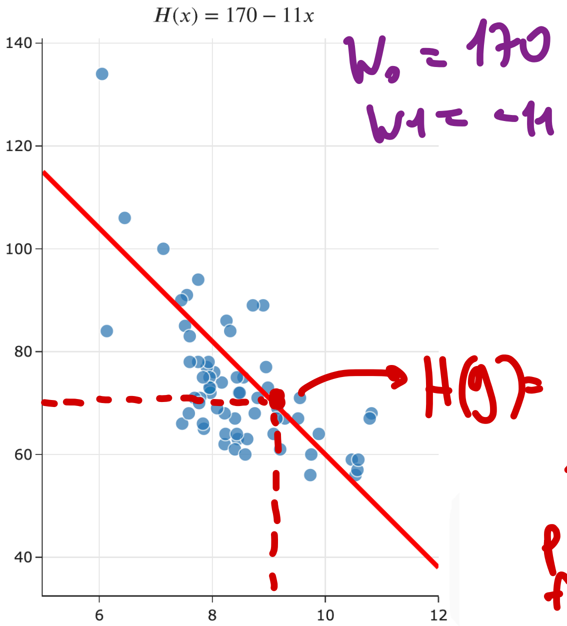
Recap: Hypothesis functions and parameters

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.





$H(9) = 170 - 11 \cdot 9 =$
$170 - 99 = 71$
predicted commute for leaving at 9am

We want to find
best slope w_1^*
best intercept w_0^*

The modeling recipe

1. Choose a model.

Before : Constant $H(x) = h$

Now : SLR $H(x) = w_0 + w_1 x$

2. Choose a loss function.

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

actual

prediction

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

3. Minimize average loss to find optimal model parameters.

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

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

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).

x_i y_i

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

Variables

- The features, x, that we base our predictions on are called predictor variables.
- The quantity, y, that we're trying to predict based on these features is called the response variable, dependent variable or target.
- We are trying to predict our commute time as a function of departure time.

y x

H

Modeling

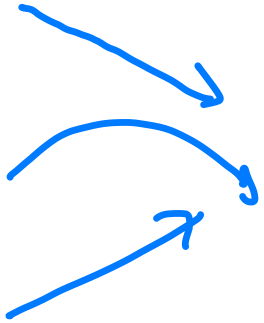
- We believe that commute time is a function of departure time.
- I.e., there is a function H so that:

  commute time $\approx H$(departure time)

  y_i x_i
- 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.

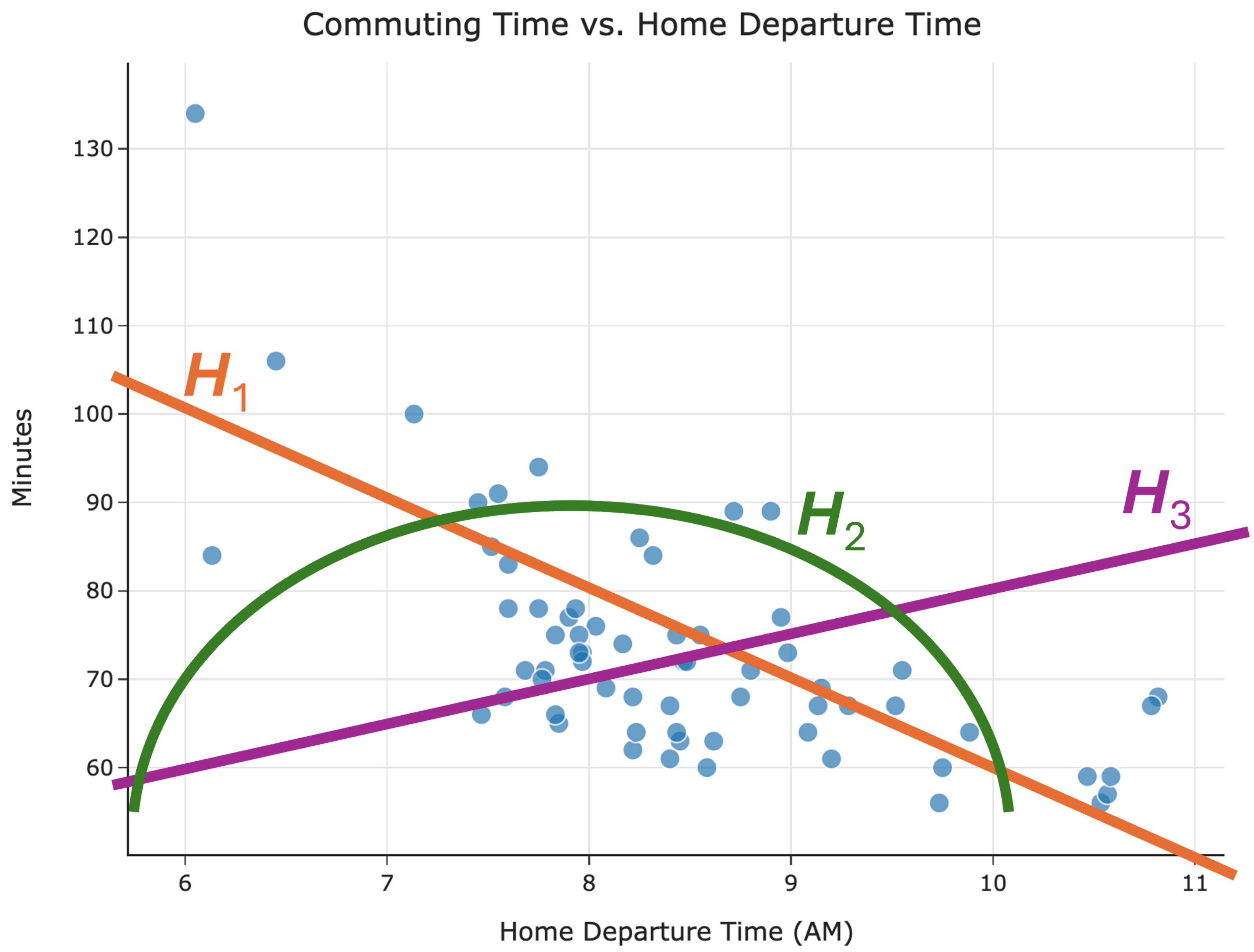
Some are better than others.

Comparing predictions

- How do we know which is best: $H_1,\ H_2,\ H_3$?
- We gather data from n days of commute. Let xi be experience, yi be salary:

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

- See which rule works better on data.



H_1 seems to be the best
→ how to quantify?

How to find optimal solution?

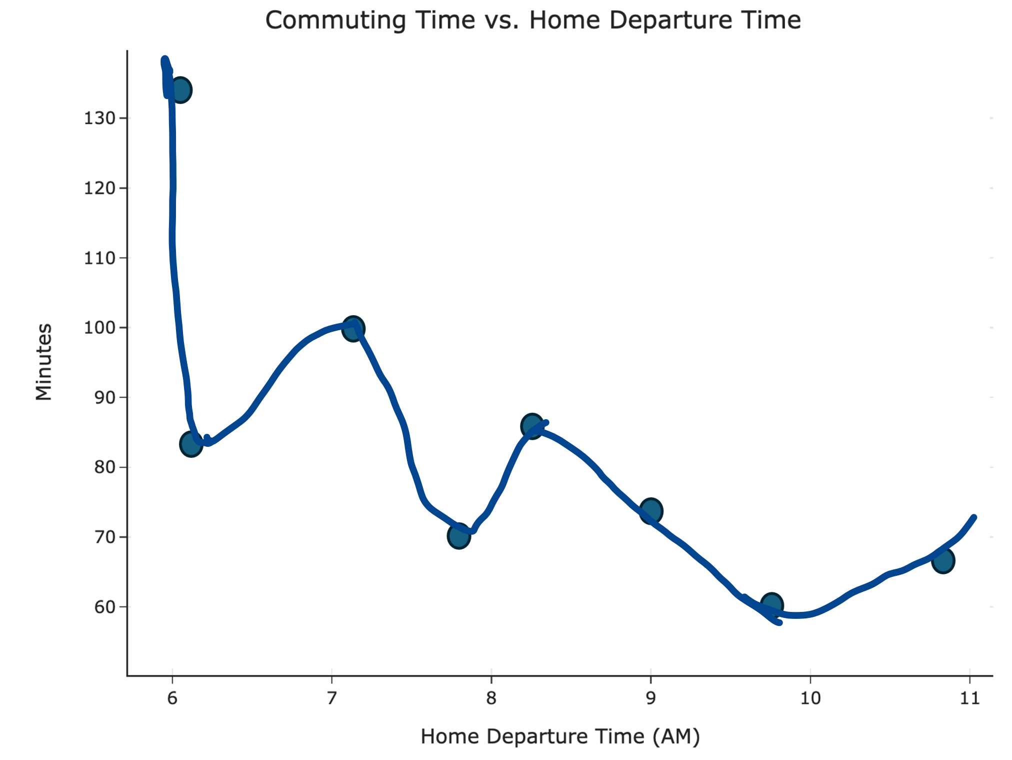
$$\ell_{abs}(y_i, H(x_i)) = |y_i - H(x_i)|$$

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

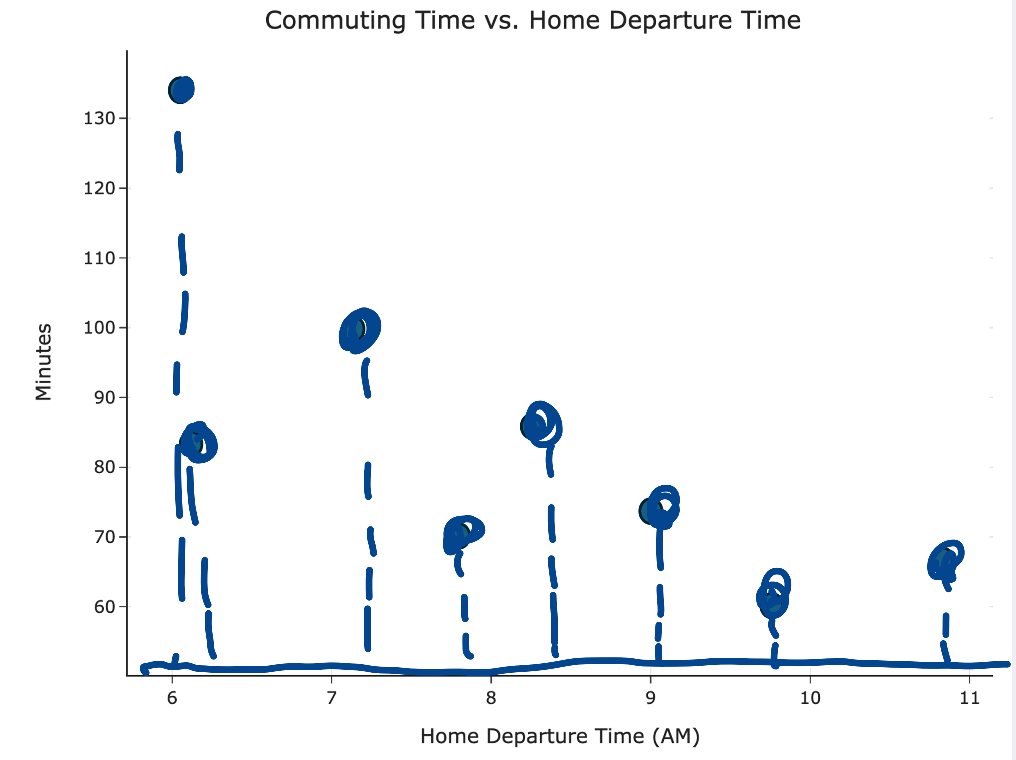
$$R_{abs}(H) = \frac{1}{n} \sum |y_i - H(x_i)|$$



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



Solution

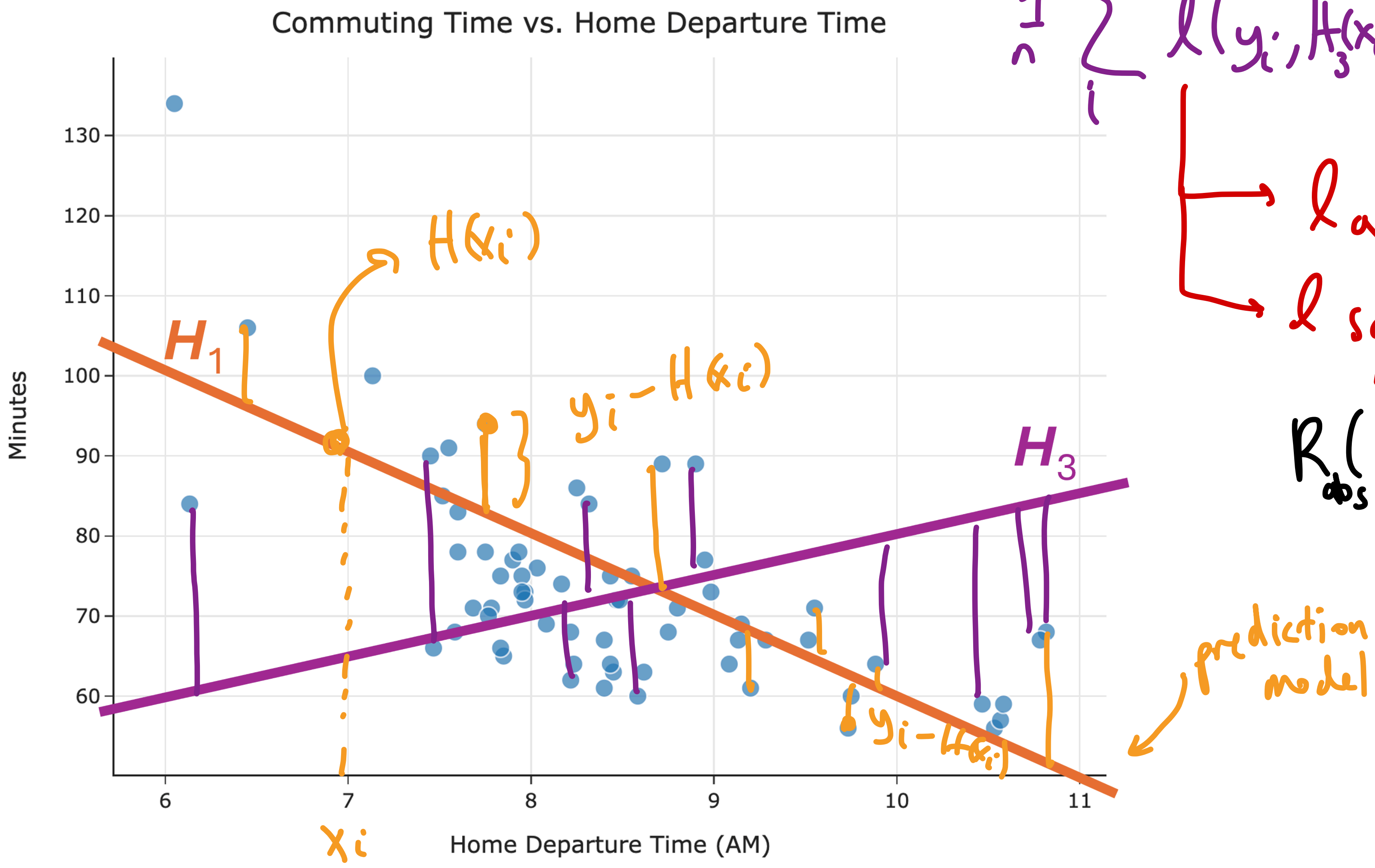
- Don't allow H to be just any function.
- Require that it has a certain form.
- Examples:
    - Linear: $H(x) = w_0 + w_1 x$. ← this week
    - Quadratic: $H(x) = w_0 + w_1 x_1 + w_2 x^2$. ← nonlinear, in a few weeks
    - Exponential: $H(x) = w_0 e^{w_1 x}$.
    - Constant: $H(x) = w_0$. ← last week

Goal: linear relationship between x_i and y_i

Find the best linear model

$\Rightarrow$ finding w_0^*, w_1^*

Comparing predictions



$R(H_3)$ $R(H_1)$

$\frac{1}{n} \sum_i \ell(y_i, H_3(x_i)) > \frac{1}{n} \sum \ell(y_i, H_1(x_i))$

ℓ_{abs}

ℓ_{sq}

$R_{abs}(H)$ averages the lengths of the error lines

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:

$w_0 + w_1 x_i$

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

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

func. of two parameters

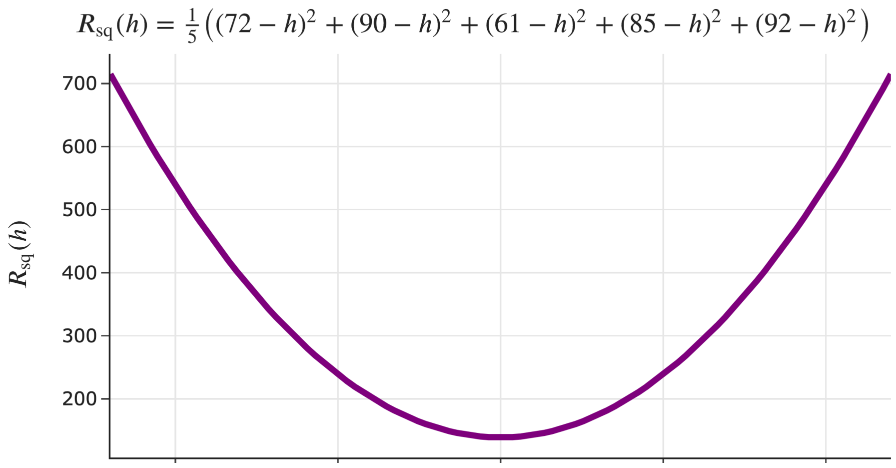
$$\boxed{R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2}$$

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

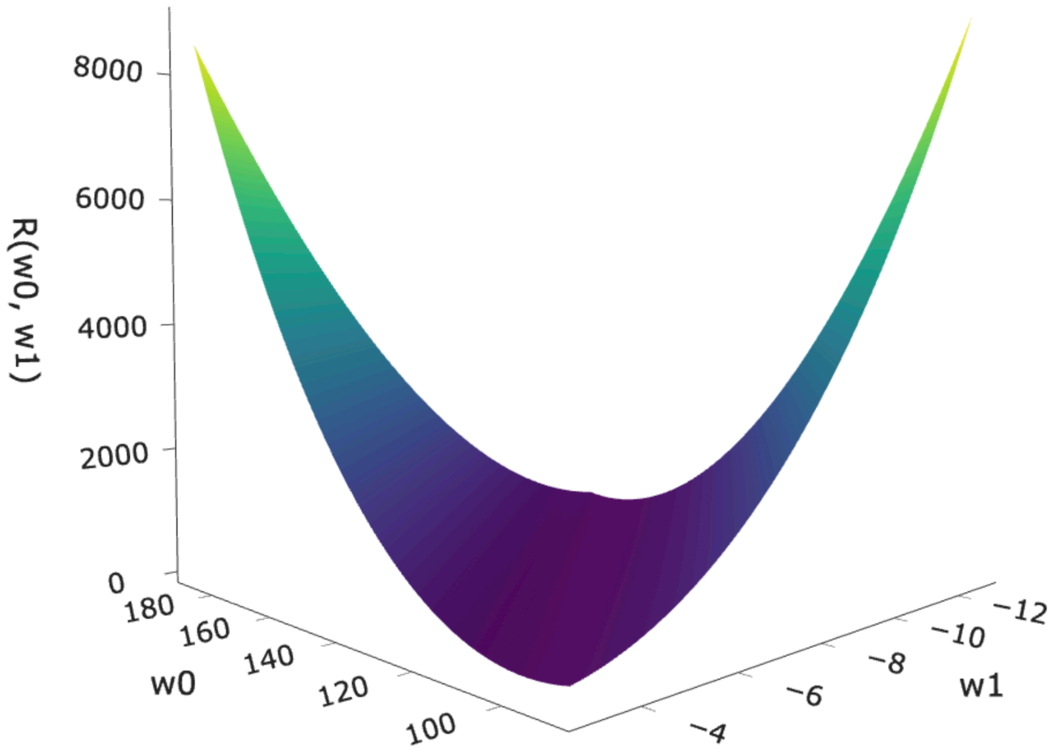
Multivariate calculus

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 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_{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).

Example

Find the point (x, y, z) at which the following function is minimized.

$$f(x, y) = x^2 - 8x + y^2 + 6y - 7$$