Lecture 1

Introduction to Modeling

DSC 40A, Fall 2024

Agenda

- Introductions.
- What is DSC 40A about?
- Logistics.
- Modeling.
- The constant model.

Introductions

Instructor: Gal Mishne

- Assistant Professor at Halıcıoğlu Data Science Institute since summer 2019.
- Undergrad: EE and Physics at Technion.
- Grad school: EE PhD at Technion.
- Postdoc: Applied math at Yale
- Outside interests: traveling/hiking, cooking, reading, painting.

Course staff

We have 1 TA and 7 tutors, all of whom are excited to help you in discussion and office hours!

Sawyer Robertson

Rebecca (Jiaying) Chen

Zoe Ludena

Brighten Hayama

Utkarsh Lohia

Varun Pabreja

Javier Ponce

Owen Miller

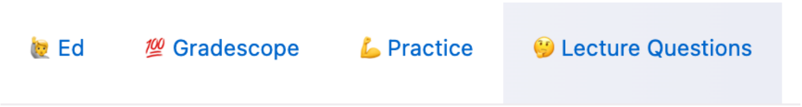
Read more about us at dsc40a.com/staff.

Throughout lecture, ask questions!

- You're always free to ask questions during lecture, and I'll try and stop for them frequently. But still, you may not feel like asking your question out loud.
- You can **type your questions anonymously** at the following link and I'll try and answer them.

q.dsc40a.com

- You'll also use this form to answer questions that I ask you during lecture.
- If the direct link doesn't work, use the 🤔 **Lecture Questions** link in the top right corner of dsc40a.com.



What is DSC 40A about?

Theoretical Foundations of Data Science I

What have you *heard* about DSC 40A?

Here are some responses from the Welcome Survey in the spring quarter.

I've heard the class seeks to uncover a lot of the key concepts of the math behind machine learning, while utilizing a lot of linear algebra. I've heard that the class can be difficult and proof-heavy.

I heard it is conceptual, and therefore, a pretty hard class (to understand conceptually). I also heard it has a lot to do with linear algebra.

That it’s the most awful class in the DSC major, pretty much just pure math/all proofs.

It's a pretty hard class but rewarding in the end.

Why do we need to study theoretical foundations?

Machine learning is about **automatically learning patterns from data**.

Humans are good at understanding handwriting – but how do we get computers to understand handwriting?

Course overview

Part 1: Learning from Data (Weeks 1-5)

- Summary statistics and loss functions; empirical risk minimization.
- Linear regression (including multiple variables); linear algebra.
- Clustering.

Part 2: Probability (Weeks 6-10)

- Set theory and combinatorics; probability fundamentals.
- Conditional probability and independence.
- The Naïve Bayes classifier.

Learning objectives

After this class, you'll...

- understand the basic principles underlying almost every machine learning and data science method.
- be better prepared for the math in upper division: vector calculus, linear algebra, and probability.

What do DSC 80 students have to say about DSC 40A?

Here are some responses from the End-of-Quarter Survey last quarter in DSC 80.

study hardy, pay attention in DSC 40A and start work early :)

40A and Math 18 is super important for this class. Don't wait till the last minute too!

I think DSC40[A] was the most important prerequisite for this class.

Logistics

Getting started

- The course website, dsc40a.com, contains all content. **Read the syllabus carefully!**
    - Click around; you'll find other helpful resources.
- Other sites you'll need to use:
    - Gradescope is where you'll submit all assignments. You'll be automatically added within 24 hours of enrolling.
    - Ed is where all announcements will be made. If you're not enrolled, there's a join link in the syllabus.
    - We aren't using Canvas.
- Make sure to fill out the Welcome Survey ASAP.

Lectures

- Lecture is here, Center Hall 105, MWF 4:00-4:50p.
- Lecture slides will be posted on the course website before class, and annotated slides will be posted after class.
- Lecture will be podcasted.
- **The value of lecture is interaction and discussion, so even though attendance isn't required, it's highly, highly recommended.**

Discussions

- Dicsussion weekly on Mondays, Center Hall 212, 6:00-6:50p.
- Discussion will primarily be used for **groupwork** – that is, working on problems in small groups of size 2-4.
    - You may work in a self-organized group outside of a discussion section for 80% credit, but no matter what, **you cannot work alone**.
- Groupwork worksheets are due to Gradescope on **Mondays at 11:59PM**.
    - Only one group member needs to submit, and should add the rest of the group to the submission.
- **The value of attending is getting support from the TA and tutors**.

Grading

- **Homeworks (40%)**: Due to Gradescope **Fridays at 11:59p**.
  - Graded for correctness. Lowest score is dropped.
- **Groupworks (10%)**: Due to Gradescope on **Mondays at 11:59p**.
  - Graded for effort. Lowest score is dropped.
- **Midterm Exam (20%)**: Monday, November 4th, in class.
- **Final Exam (30%)**: Tuesday, December 10th, 3pm. See the syllabus for the redemption policy.

Support

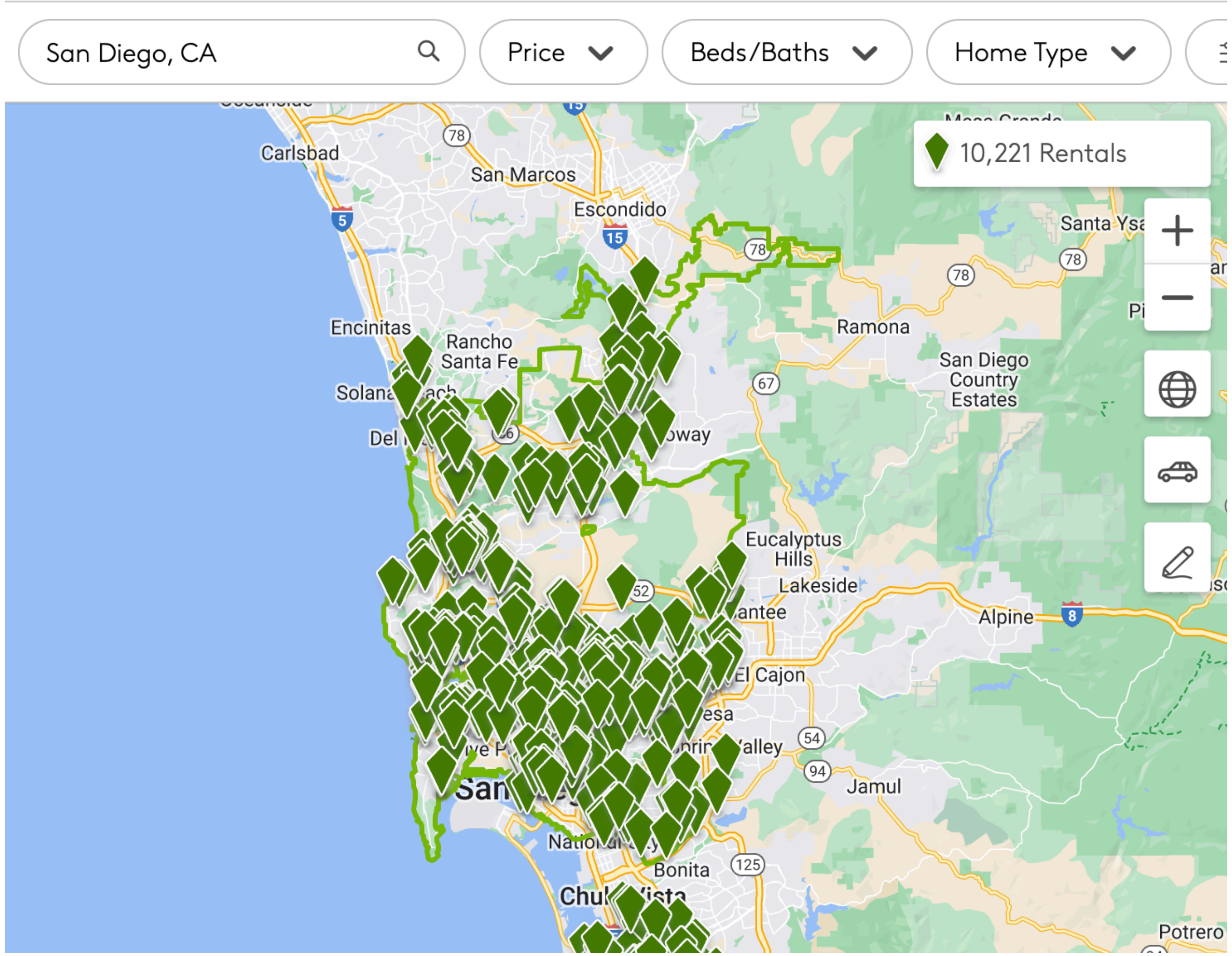
We know this is a challenging class, and we're here to help:

- **Office hours**: In-person in HDSI 155. Plan to attend at least once a week for homework help.
- **Ed**: Use it! We're here to help you. Post conceptual questions publicly – just don't post answers to homework questions.

A bunch of new-ish things to improve the student experience:

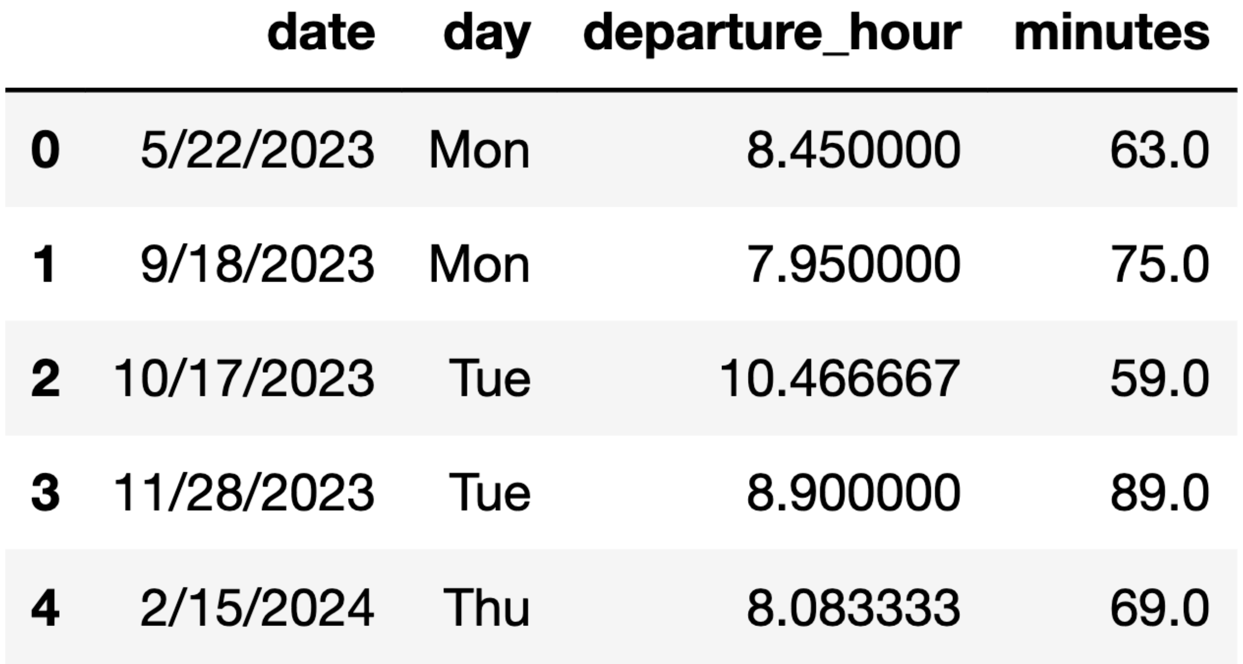
- practice.dsc40a.com to give you access to practice exam problems, categorized by topic.
- Walkthrough videos to show you our thought process when answering questions.
- More time reviewing linear algebra.

Modeling



San Diego, CA

You might be starting to look for off-campus apartments, none of which are affordable.

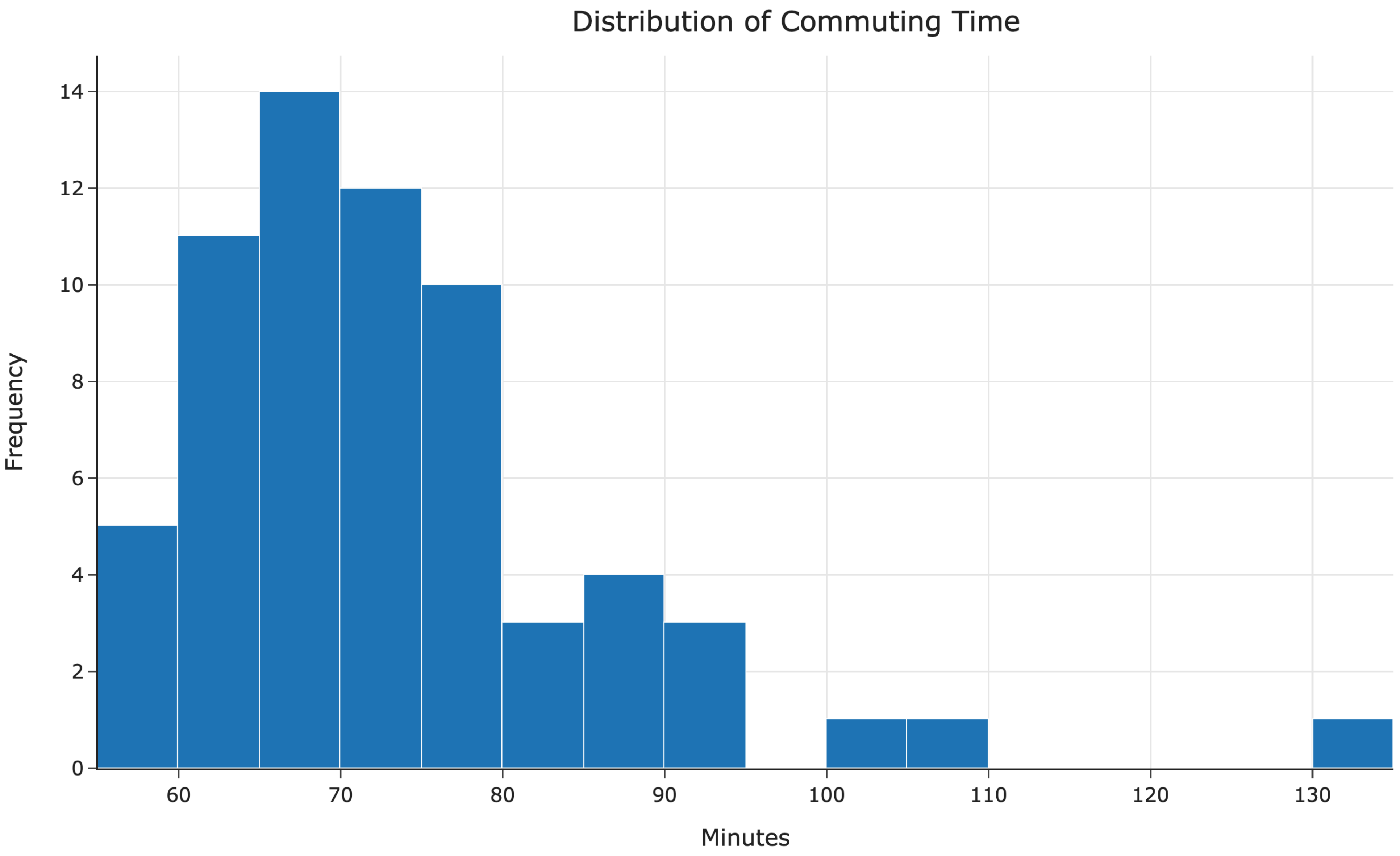


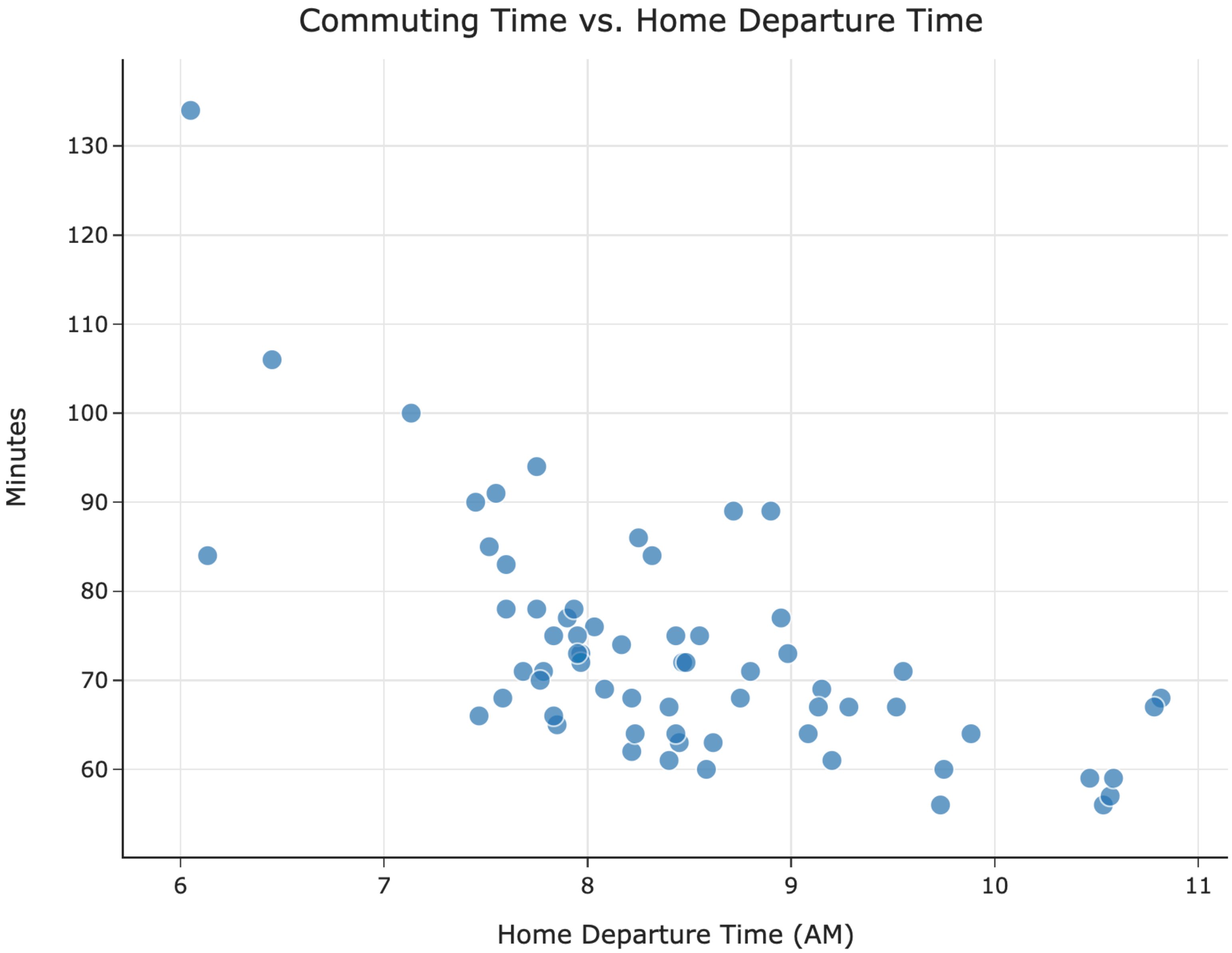
	date	day	departure_hour	minutes
0	5/22/2023	Mon	8.450000	63.0
1	9/18/2023	Mon	7.950000	75.0
2	10/17/2023	Tue	10.466667	59.0
3	11/28/2023	Tue	8.900000	89.0
4	2/15/2024	Thu	8.083333	69.0

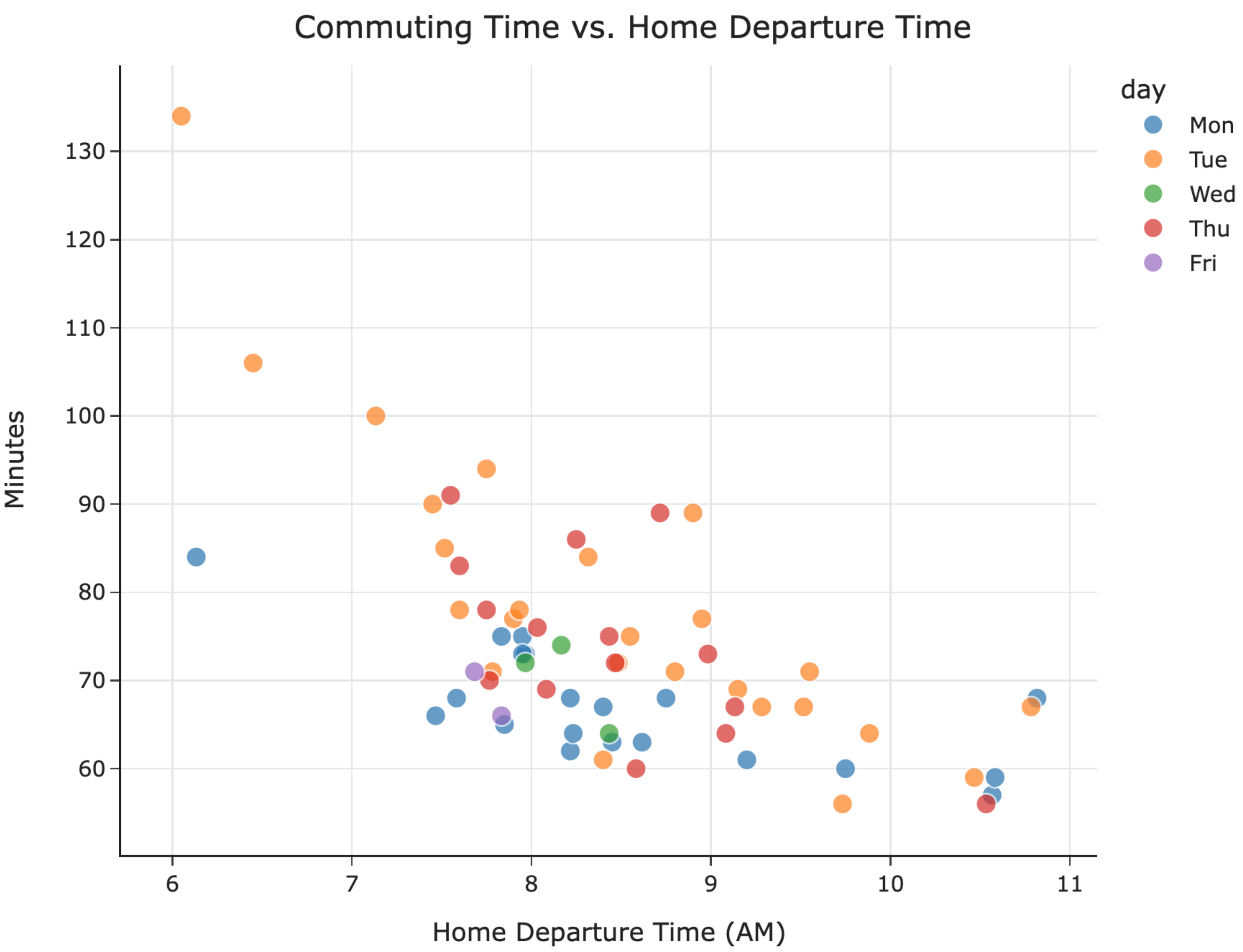
...

You decide to live with your parents in Orange County and commute.
You keep track of how long it takes you to get to school each day.

This is a real dataset, collected by Joseph Hearn! However, he lived in the Seattle area, not San Diego.


Distribution of Commuting Time
Frequency
0
2
4
6
8
10
12
14
60
70
80
90
100
110
120
130
Minutes


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)


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)
day
● Mon
● Tue
● Wed
● Thu
● Fri

Goal: Predict your commute time.

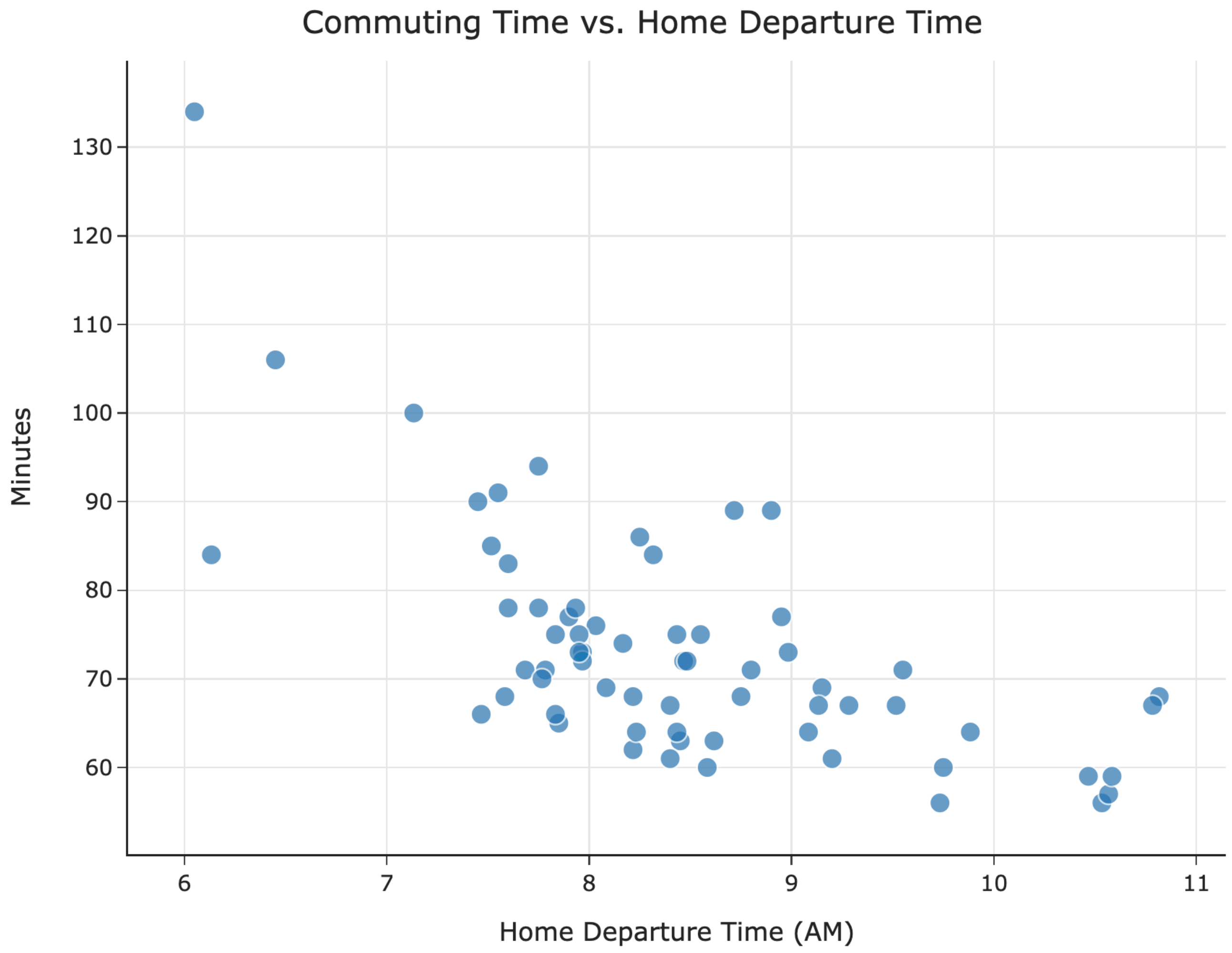
That is, predict how long it'll take to get to school.

How can we do this?

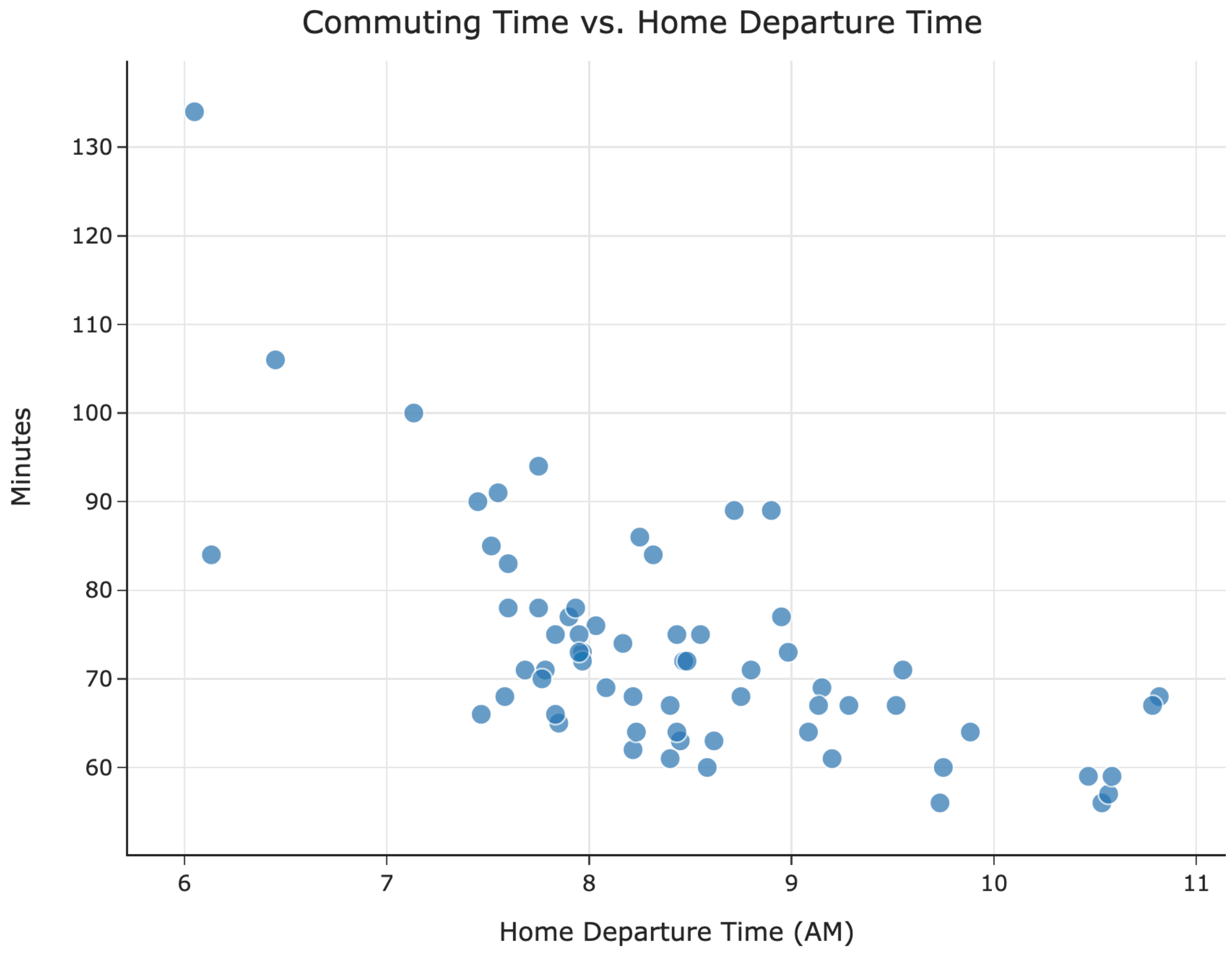
What will we need to assume?

A **model** is a set of assumptions about how data were generated.

Possible models



Notation



x: "input", "independent variable", or "feature"

y: "response", "dependent variable", or "target"

We use x to predict y.

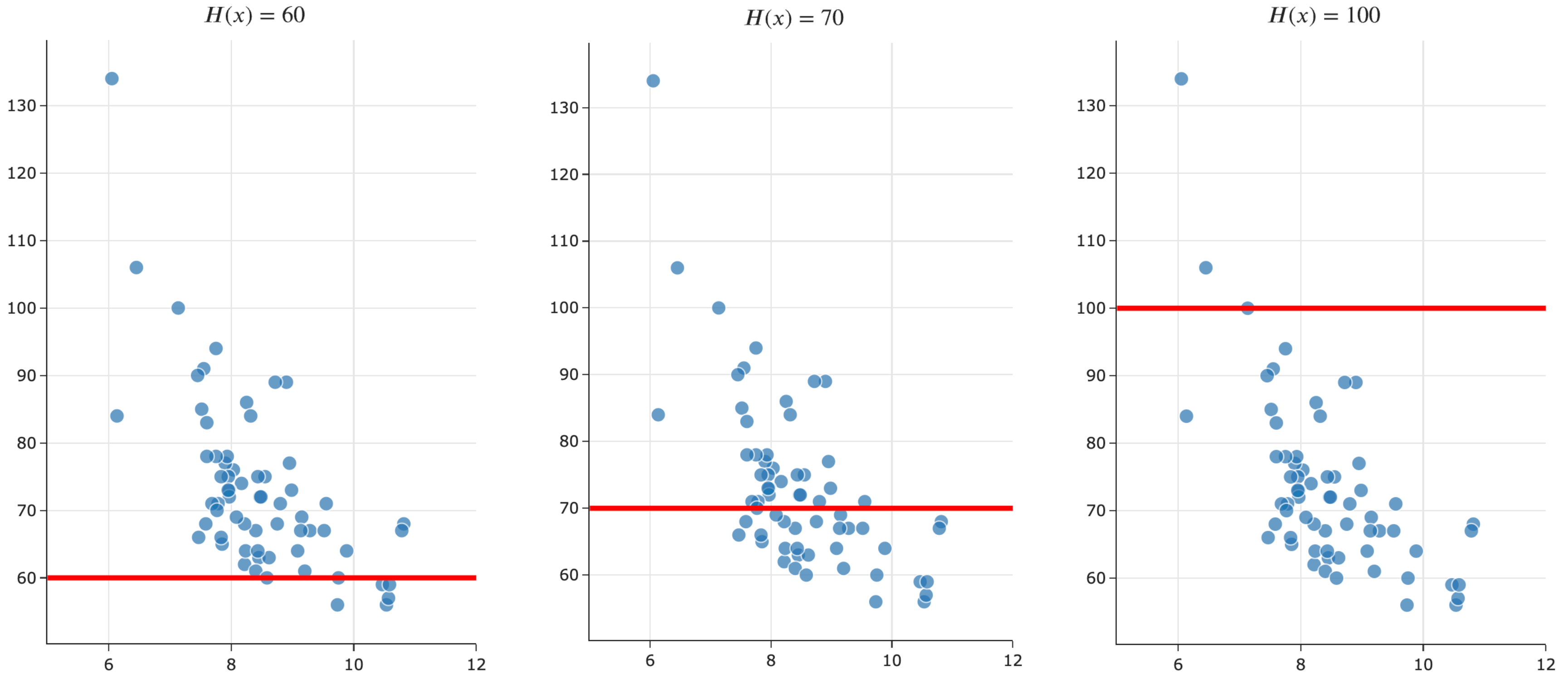
The ith observation is denoted (x_i, y_i).

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

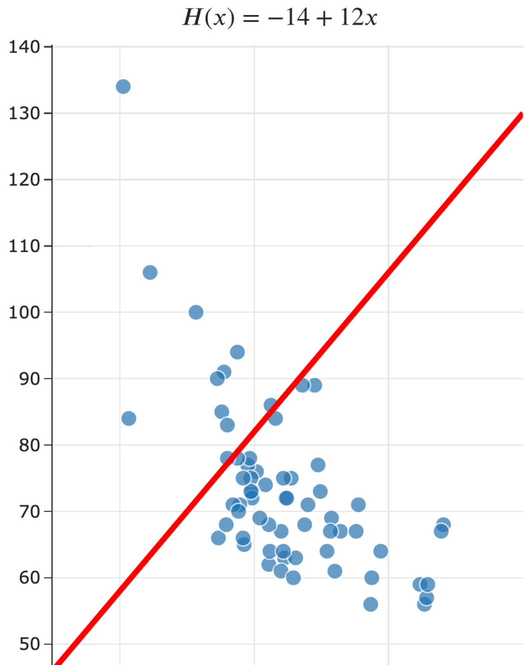


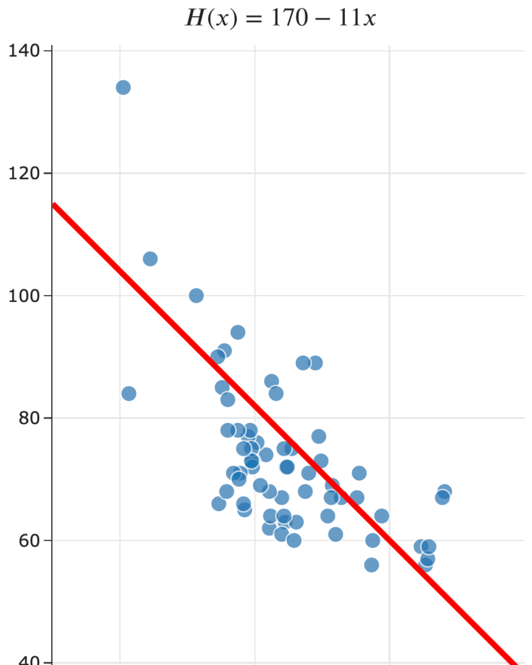
Hypothesis functions and parameters

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.

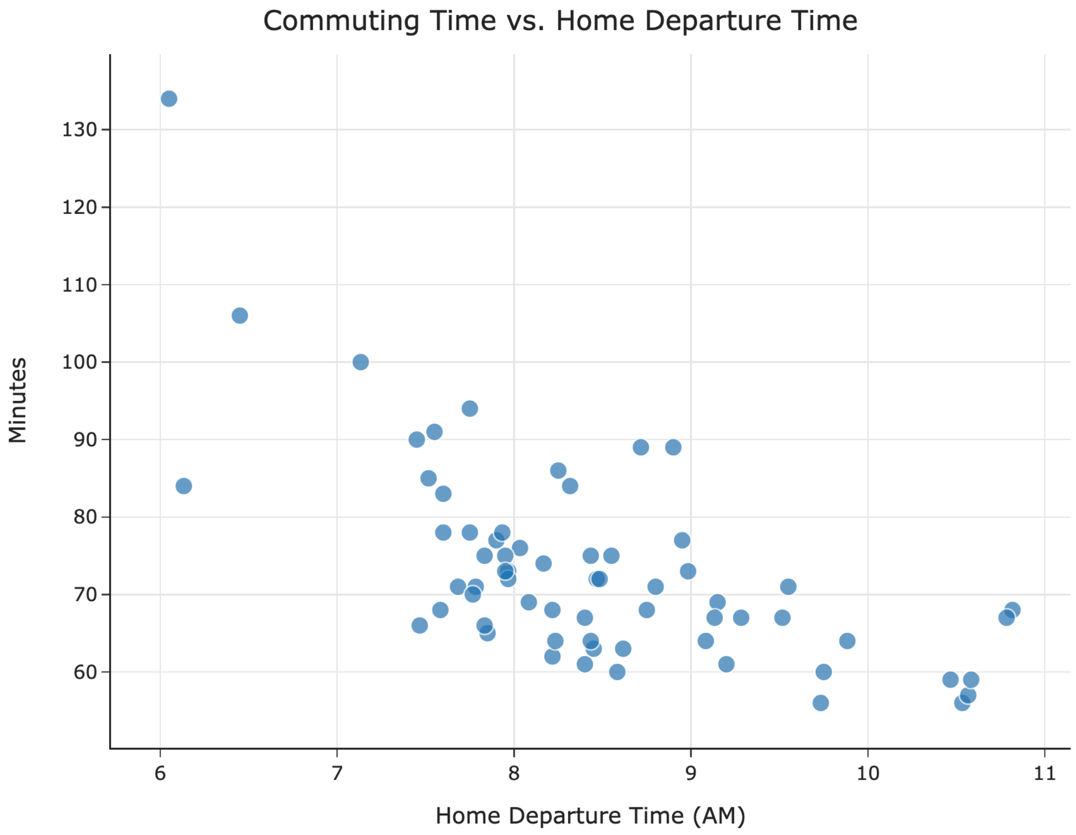
The simple linear regression model, $H(x) = w_0 + w_1 x$, has two parameters: w_0 and w_1.

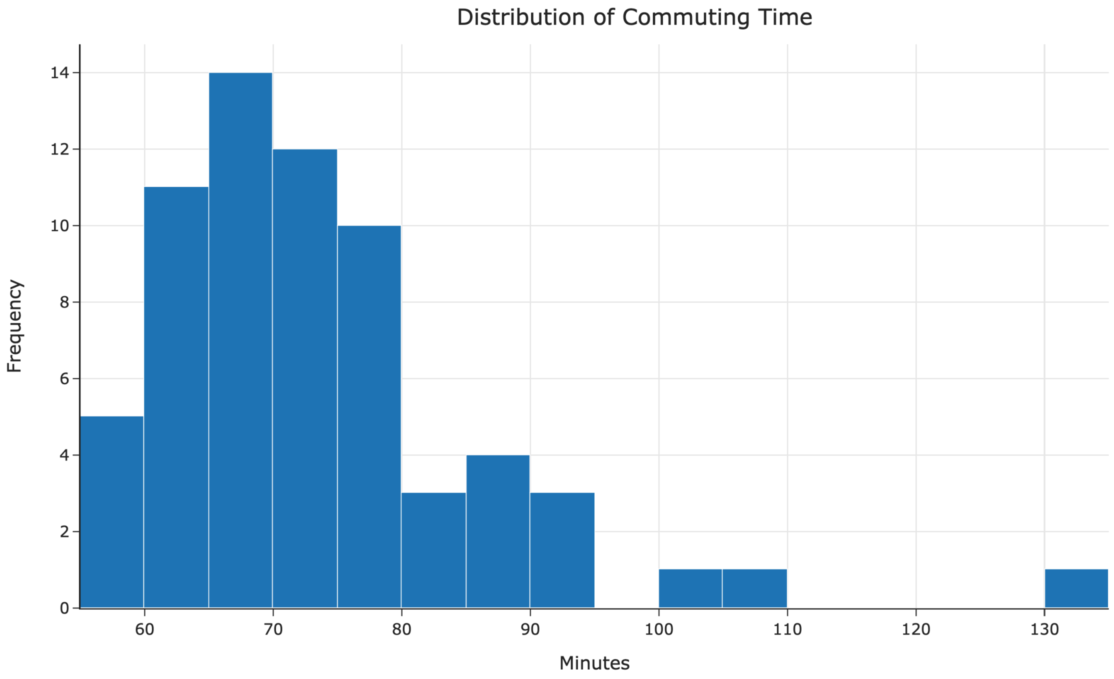




The constant model

The constant model





A concrete example

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

$$\begin{aligned} y_1 &= 72 \\ y_2 &= 90 \\ y_3 &= 61 \\ y_4 &= 85 \\ y_5 &= 92 \end{aligned}$$

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

Some common approaches

- The **mean**:

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

- The **median**:

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

- 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.

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

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

- If I predict 75 minutes:
- If I predict 72 minutes:
- If I predict 100 minutes:

Squared loss

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

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

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

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

Squared loss is not the only loss function that exists! Soon, we'll learn about absolute loss.

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?

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

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, $h = 85$:

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

- For the mean, $h = 80$:

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

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

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:

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

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

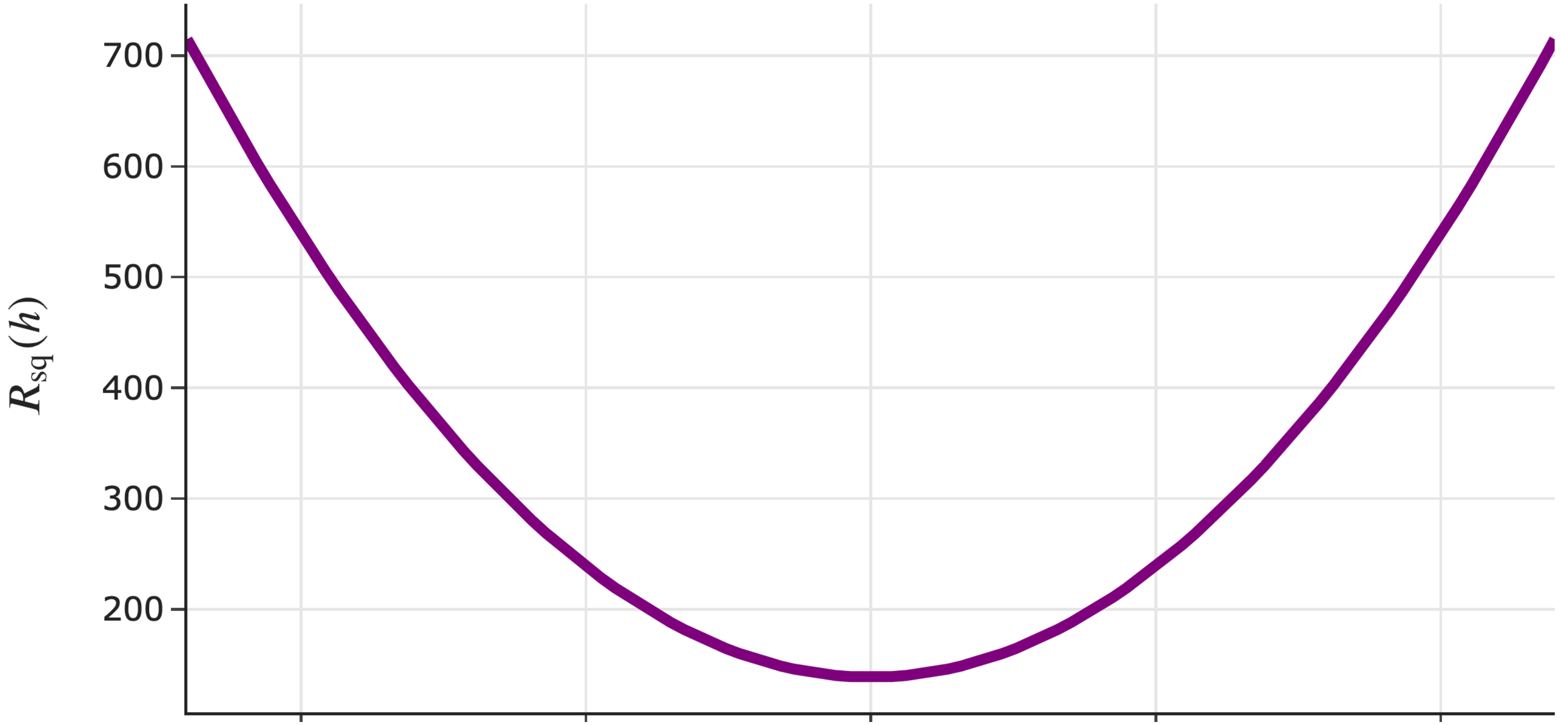
- For example, if we predict $h = 100$, then:

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

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

Visualizing mean squared error

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



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

Mean squared error, in general

- Suppose we collect n commute times, $y_1, y_2, \ldots, y_n$.
- The mean squared error of the prediction h is:
- Or, using **summation notation**:

The best prediction

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

- 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^*?**

Summary, next time

- 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.
- **Next time**: We'll solve this optimization problem by hand.