Introduction to Regression

Manasi Jayakumar

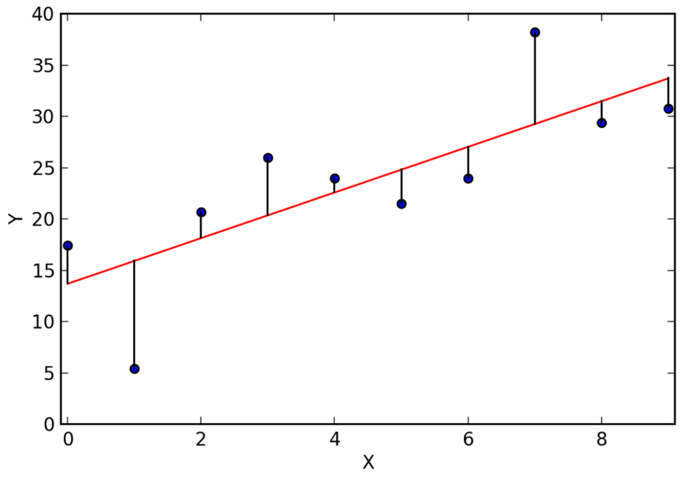
Slides adapted from previous SIPPS workshops by Camille Gasser

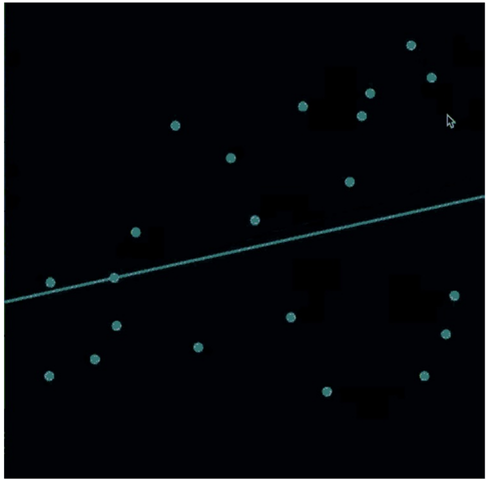
Lesson Plan

- ➔ Review of linear regression models
- ➔ Using lm() function in R

What is a linear regression model?

- Used to quantify the relationship(s) between an **outcome variable** and one (or more!) **predictors**
- What this analysis does, more specifically, is **fit a line** that best minimizes the error between that line and your data points

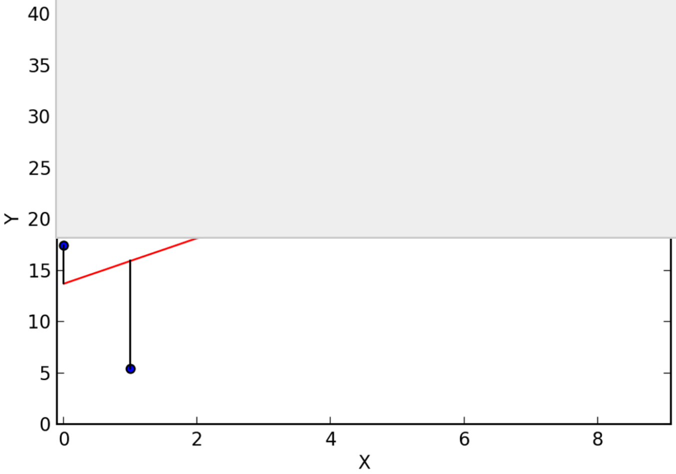


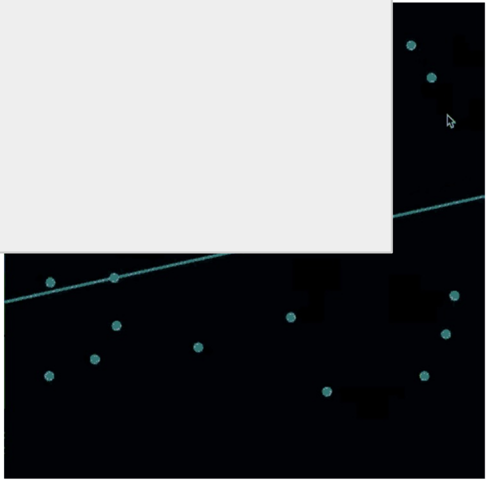


What is a linear regression model?

- Used to quantify the relationship(s) between an **outcome variable** and or
- What t ... best minim

[Interactive Visualization of Linear Regression]



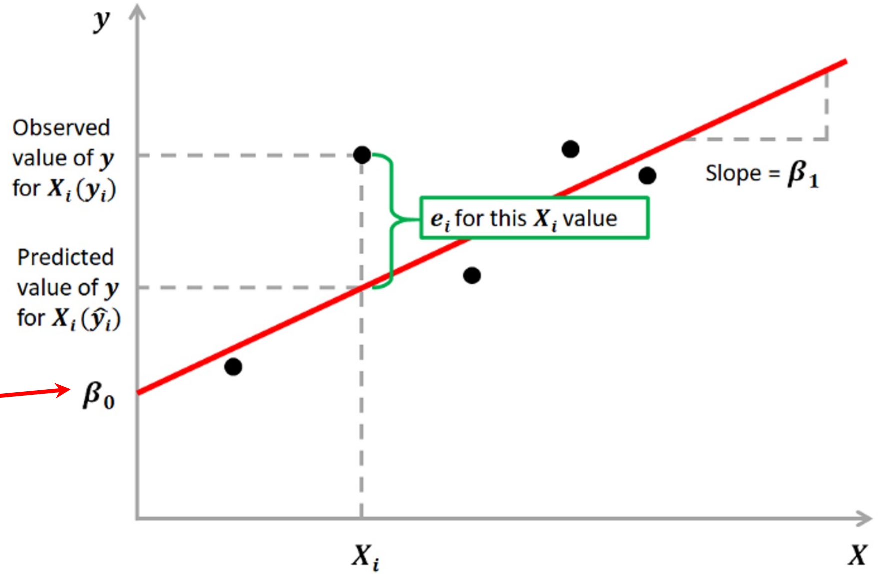


Slope and Intercept

- When we fit a line to the data, we get an **intercept** and at least one **slope.**

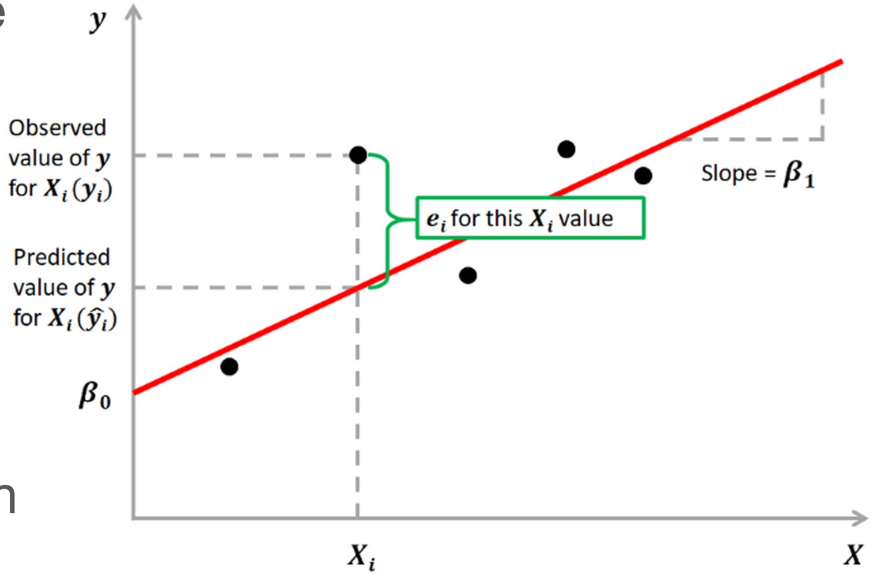
- **Intercept:**

  Value of the outcome variable (Y) when all predictors (Xs) are 0.



Slope and Intercept

- **Slope** values (also called **betas**) are what we're most interested in:
  - Change in outcome variable (Y) for every unit change in the predictor variable (X)
  - Rise / Run
- if a slope associated with a given X variable is significantly different than zero, we can conclude that the value of X is meaningfully related to the value of Y



Example of a linear regression model

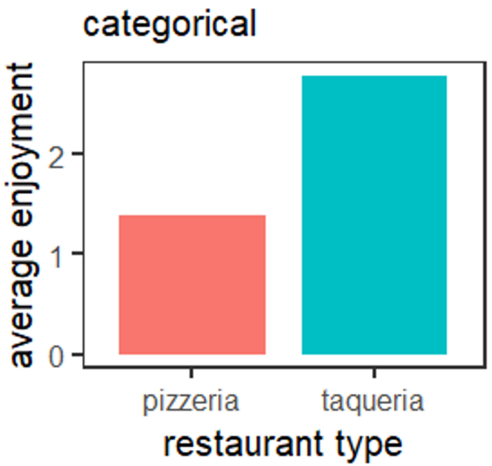
- Let's say you are interested in which variables affect how much you'll enjoy a particular take-out restaurant
- You could hypothesize that your enjoyment will depend on (at least) two things:
    - The average Yelp rating
    - Whether the restaurant is a pizzeria or a taqueria

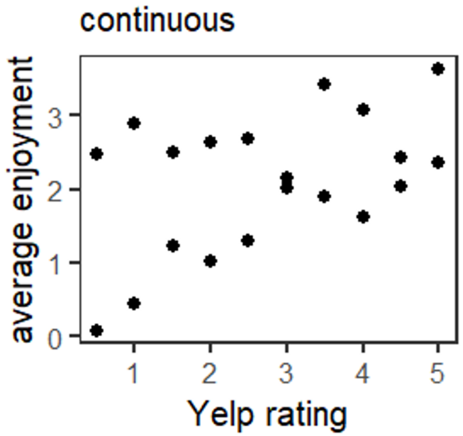
enjoyment ~ (yelp rating) + (type of restaurant)

if both slopes are significant, we can say that both the yelp rating and the type of restaurant are significantly associated with food enjoyment

Types of variables

- regression models are inherently flexible, and allow you to quantify many different kinds of variables & relationships
- in our toy model, for example, we are looking at two different types of moderators/predictors: continuous (the yelp rating) & categorical (the type of restaurant)





Introducing lm() function in R

- In R, lm() is a function that allows you to run linear models
- Using it requires two main arguments: 1) the dataframe you want to work with, and 2) the equation of the model you want to run
- Equations follow this format:
  Y ~ X1 [+ X2 + X3 + …]
- So we might run:
  lm(data = mydata, enjoyment ~ yelp_rating + restaurant_type)