STA 235H - Multiple Regression: Overview and Statistical Adjustment

Fall 2023

McCombs School of Business, UT Austin

Today

- Quick multiple regression review
    - How does OLS work?
- What can we say using regressions?
    - Interpreting coefficients

Nothing "Ordinary" about OLS

What do you understand about regressions?

Remembering Regressions

- Linear Regression is a **very useful tool**.
    - Simple supervised learning approach.
    - Many fancy methods are generalizations or extensions of linear regression!
- It's a way to (partially) describe a **data generating process (DGP)**.

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$$

Essential Parts of a Regression

Y

Outcome Variable

Response Variable

Dependent Variable

Thing you want to explain or predict

Essential Parts of a Regression

Y	X
Outcome Variable	Explanatory Variable
Response Variable	Predictor Variable
Dependent Variable	Independent Variable
Thing you want to explain or predict	*Thing you use to explain or predict Y*

Identify the variables

A study examines the effect of smoking on lung cancer

Identify the variables

A study examines the effect of smoking on lung cancer

Fantasy football fanatics predict the performance of a player based on past performance, health status, and characteristics of the opposite team

Identify the variables

A study examines the effect of smoking on lung cancer

Fantasy football fanatics predict the performance of a player based on past performance, health status, and characteristics of the opposite team

You want to see if taking more AP classes in high school improves college grades

Identify the variables

A study examines the effect of smoking on lung cancer

Fantasy football fanatics predict the performance of a player based on past performance, health status, and characteristics of the opposite team

You want to see if taking more AP classes in high school improves college grades

Netflix uses your past viewing history, the day of the week, and the time of the day to guess which show you want to watch next

Two Purposes of Regression

Prediction

Forecast the future

Focus is on Y

Netflix trying to guess your next show

Explanation

Explain the effect of X on Y

Focus is on X

Netflix looking at the effect of time of the day on show selection

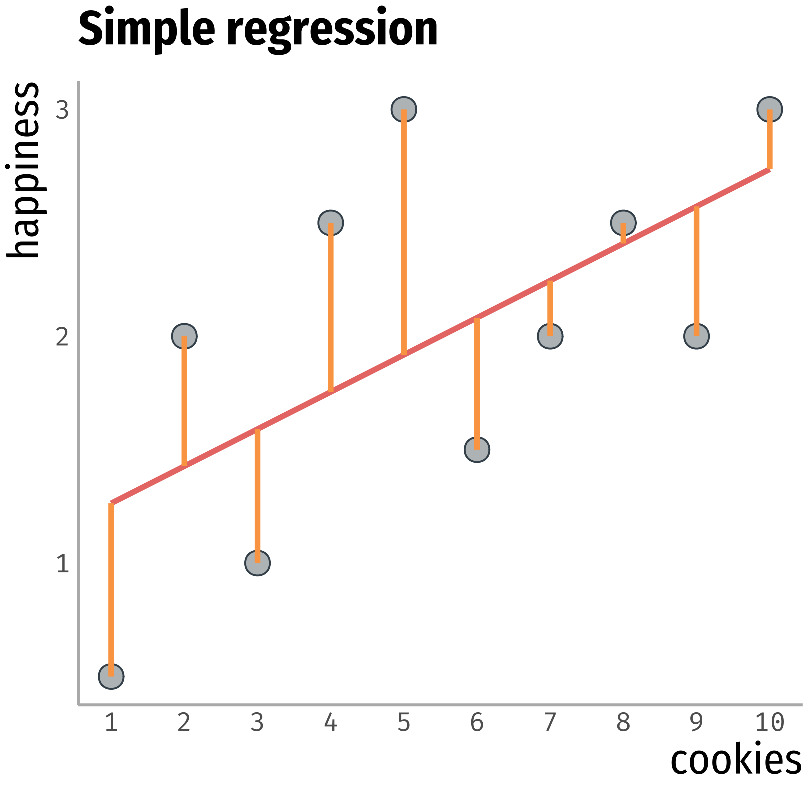
What do we want to estimate in a regression?

- When we run a regression we have an outcome Y and explanatory variables or covariates X.
- **We want to estimate the β's**
- One important distinction:
    - β's are the **population parameters** we want to estimate.
    - $\hat{\beta}$ are the **estimates** of those parameters.

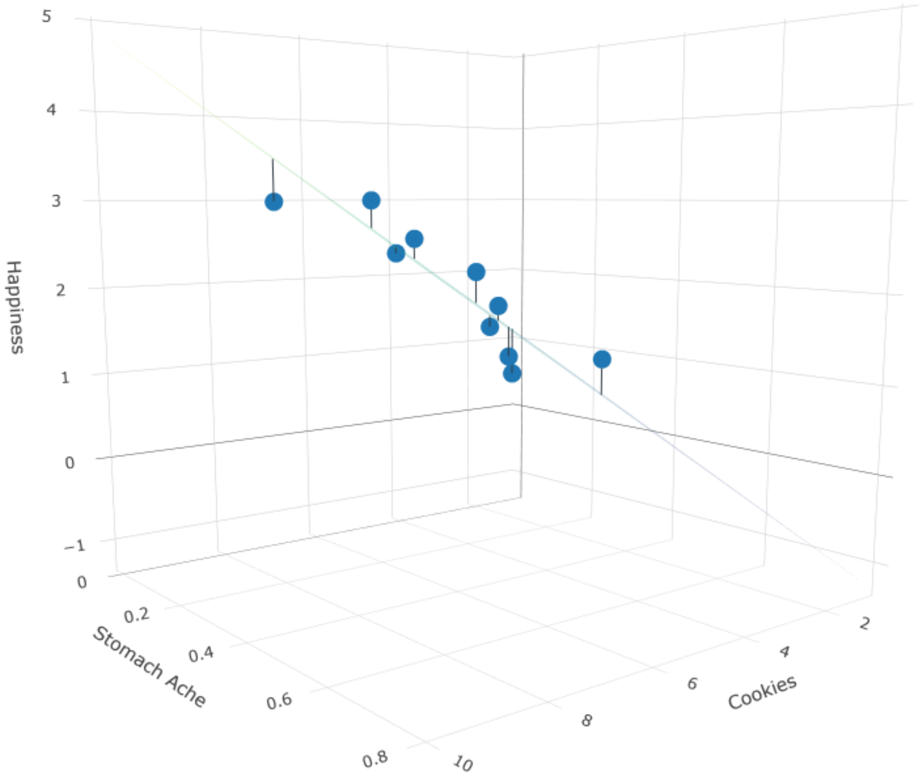
How do we estimate the coefficients in a regression ?

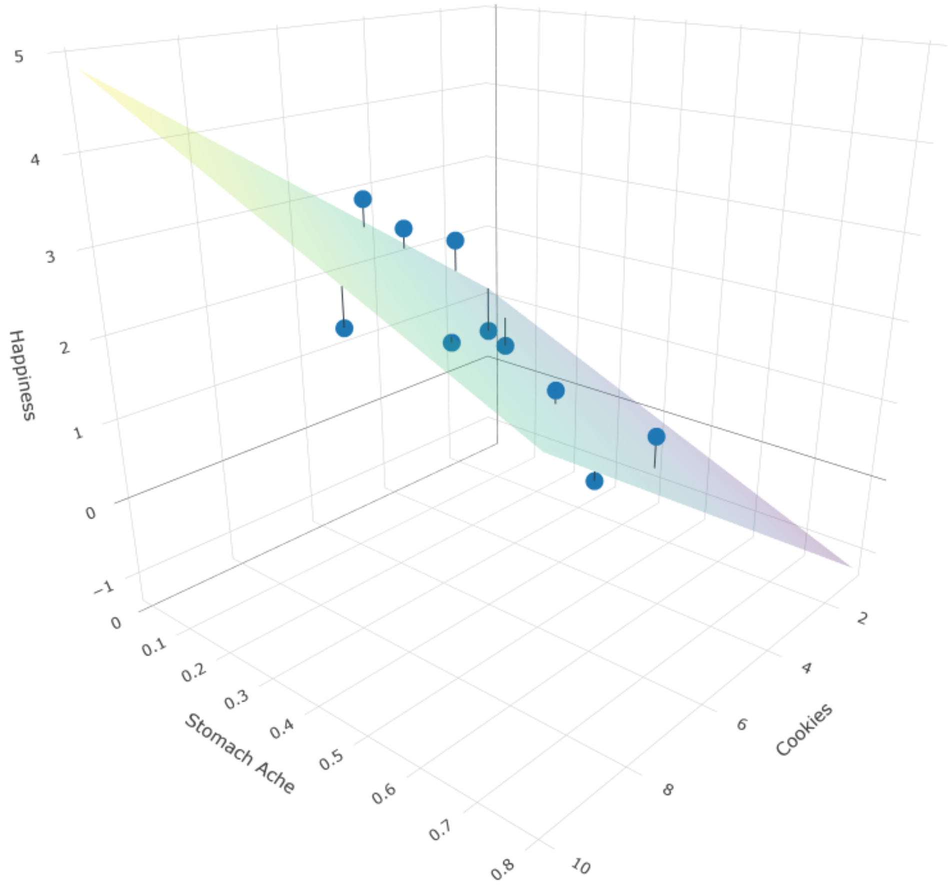
- Ordinary Least Squares is the most popular way.

$$\min_{\beta} \sum [Y_i - (\sum_{j=1}^{p} \beta_j X_{ij})]^2$$



How do we estimate the coefficients in a regression ? (cont.)



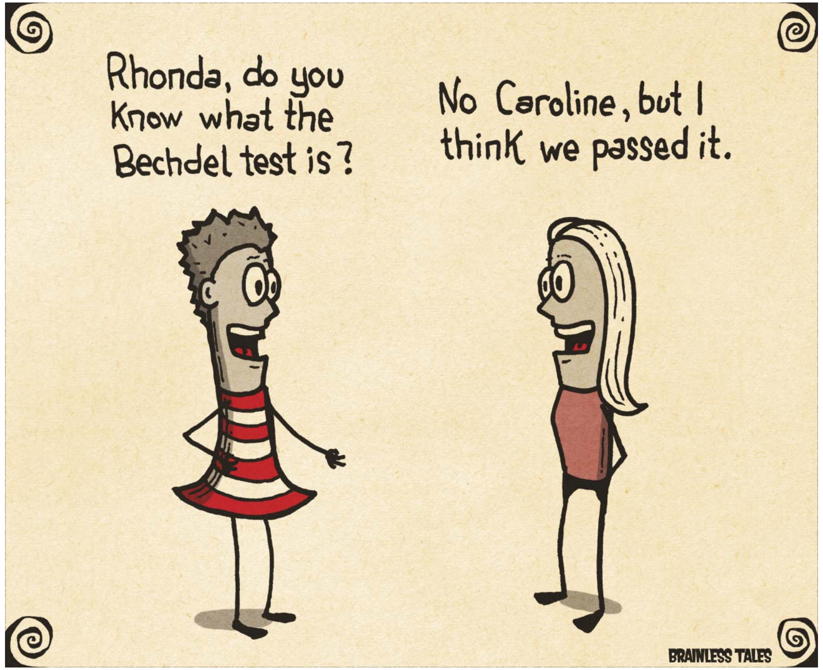


Let's get into some data

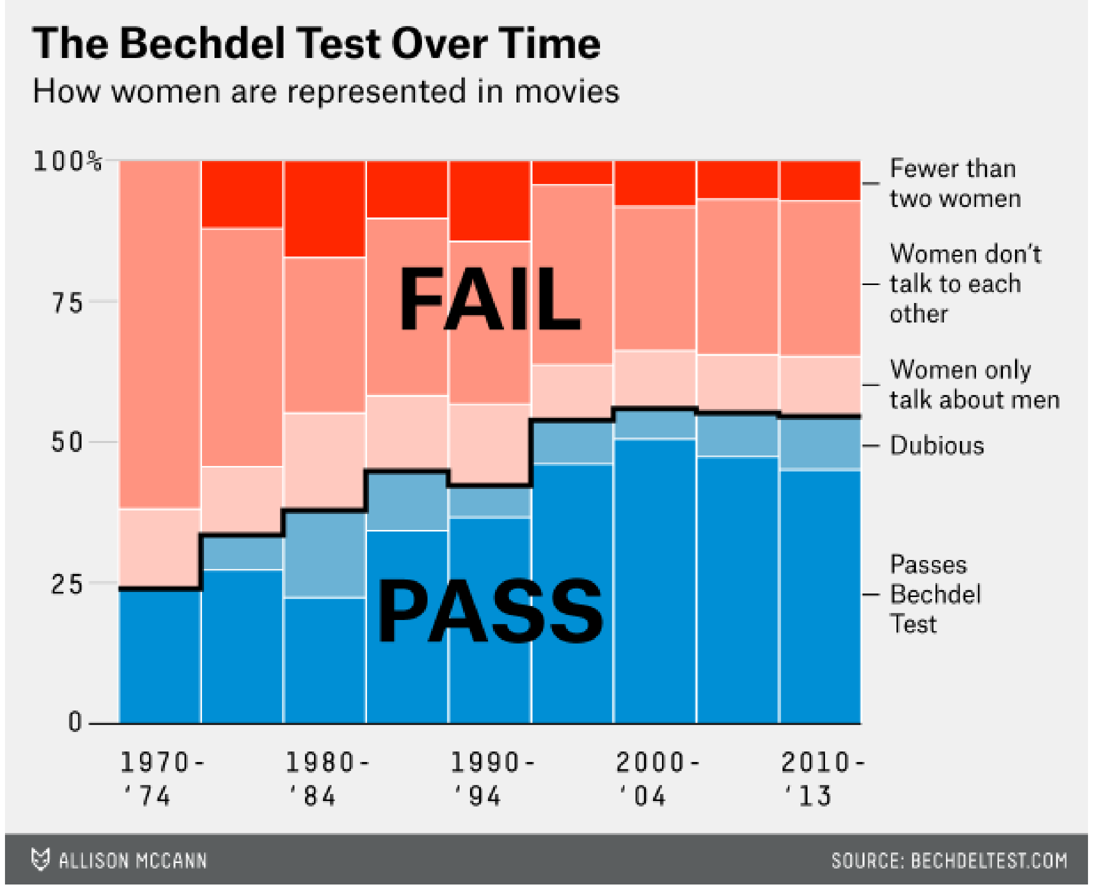
Let's introduce an example: The Bechdel Test

- Three criteria:

  1. At least two named women
  2. Who talk to each other
  3. About something besides a man


Rhonda, do you Know what the Bechdel test is?
No Caroline, but I think we passed it.
BRAINLESS TALES

Do movies pass the test?



The Bechdel Test Over Time

How women are represented in movies

ALLISON MCCANN

SOURCE: BECHDELTEST.COM

Is it convenient for my movie to pass the Bechdel test?

- I'm a profit-maximizing investor and want to know whether it's in my best interest to switch a male for a female character.
    - What is the simplest model you could fit?

$$Revenue = \alpha + \beta Bechdel + \varepsilon$$

Let's analyze some models

- We have some data and code on the course website
- Dataset from fivethirtyeight.com:
    - Focus on 1990 onward

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Year	2087	2004.963	6.755	1990	1999	2011	2014
Adj_Revenue	2087	66.254	92.07	0	4.36	86.936	968.41
Adj_Budget	1369	61.498	57.784	0.02	19.3	88.47	470.839
Metascore	1755	5.663	1.66	1.1	4.5	6.8	9.7
imdbRating	2085	6.546	0.979	1.5	6	7.2	9.3
bechdel_test	2087	0.571	0.495	0	0	1	1

Let's analyze some models

```
summary(lm(Adj_Revenue ~ bechdel_test, data = bechdel))
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.4553     3.0641 24.9521        0
## bechdel_test -17.8616     4.0544 -4.4055        0
```

- How do you interpret these results?

Let's analyze some models

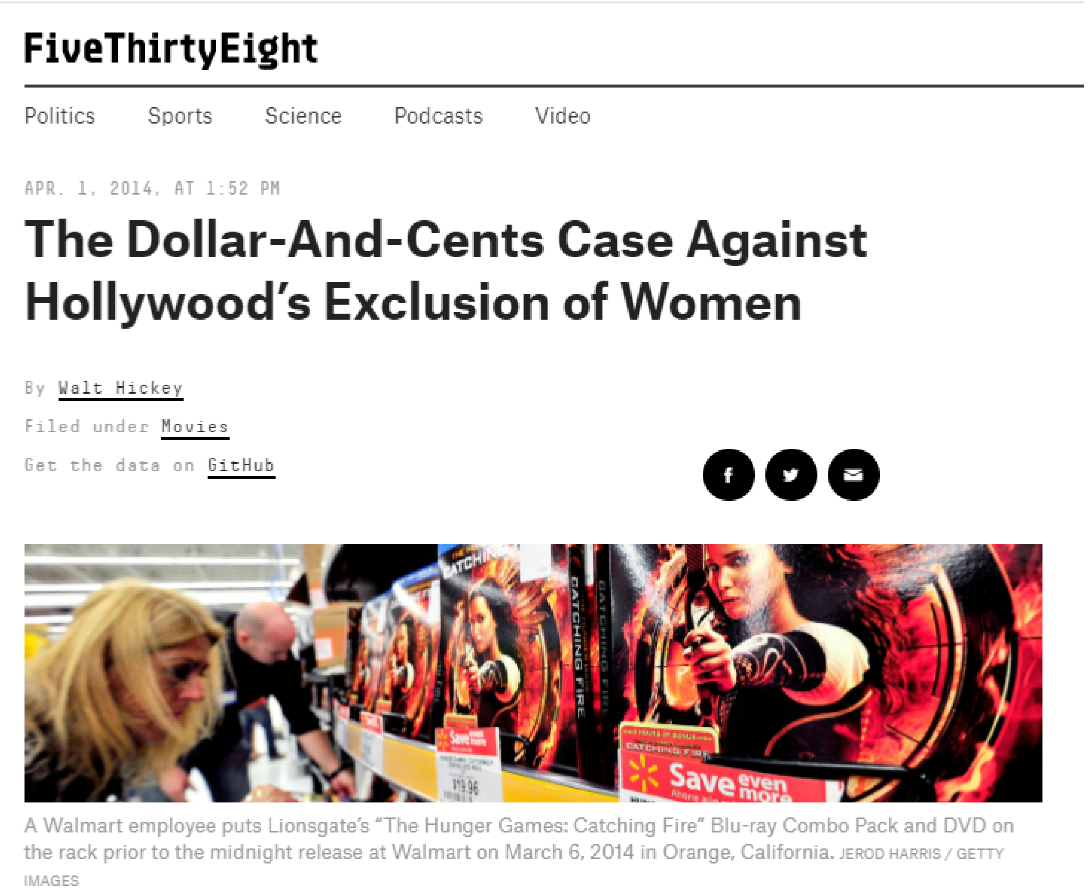
```
summary(lm(Adj_Revenue ~ bechdel_test, data = bechdel))
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.4553     3.0641 24.9521        0
## bechdel_test -17.8616     4.0544 -4.4055        0
```

- $\hat{\beta}_0$ is the average adjusted revenue (in millions of dollars) for movies that do not pass the Bechdel test.
- On average, movies that pass the Bechdel test have an adjusted revenue that is $|\hat{\beta}_1|$ million dollars less than a movie that doesn't pass the Bechdel test.

Negative effect of including more women?

What gives?



FiveThirtyEight

APR. 1, 2014, AT 1:52 PM

The Dollar-And-Cents Case Against Hollywood's Exclusion of Women

By Walt Hickey

Filed under Movies

Get the data on GitHub

A Walmart employee puts Lionsgate's "The Hunger Games: Catching Fire" Blu-ray Combo Pack and DVD on the rack prior to the midnight release at Walmart on March 6, 2014 in Orange, California. JEROD HARRIS / GETTY IMAGES

More variables

- Bechdel test could be capturing the effect of other variables:
    - What type of movies are the ones that pass the test?
    - What is their budget?

More variables

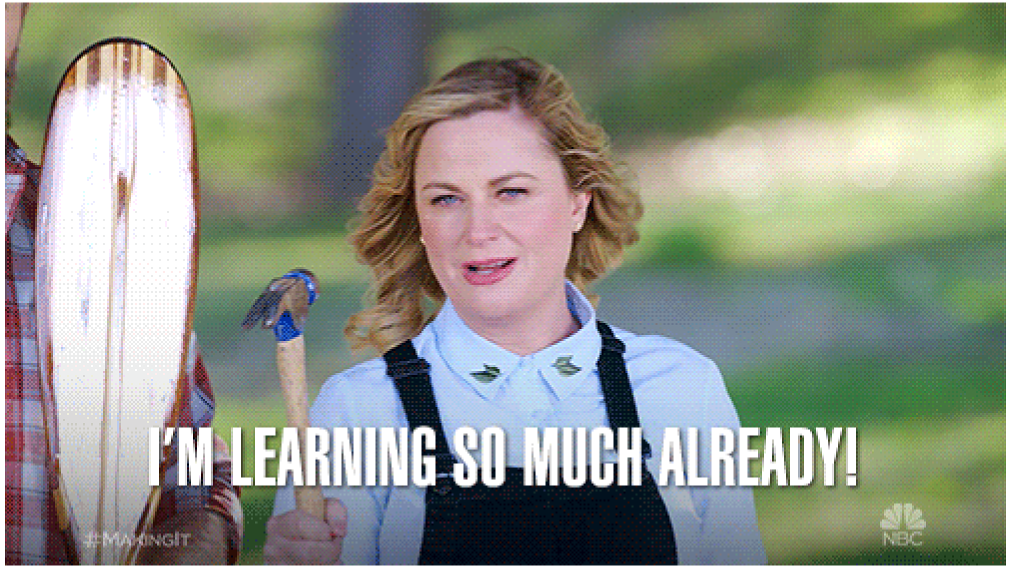
```
lm(Adj_Revenue ~ bechdel_test + Adj_Budget + Metascore + imdbRating, data=bechdel)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -127.0710    17.0563 -7.4501   0.0000
## bechdel_test     11.0009     4.3786  2.5124   0.0121
## Adj_Budget        1.1192     0.0367 30.4866   0.0000
## Metascore         7.0254     1.9058  3.6864   0.0002
## imdbRating       15.4631     3.3914  4.5595   0.0000
```

Positive and significant!

- How do we interpret the relevant coefficient now?

Main takeaway points



- Regressions are super useful…
  - But you need to know how to interpret them.
- Be sure not to overstate your claims!
- Remember the magic words for interpretation

Next class

- Continue with multiple regression models:
  - Interactions and how to interpret them
- "Nonlinear" models

References

- Heiss, A. (2020). "Course: Program Evaluation for Public Service". *Slides for Regression and Inference*.
- Ismay, C. & A. Kim. (2021). “Statistical Inference via Data Science”. Chapter 10.
- Keegan, B. (2018). "The Need for Openess in Data Journalism". *Github Repository*