

# STA 235H - Multiple Regression: Overview and Analysis

Fall 2021

McCombs School of Business, UT Austin

# Before we start...

## Feedback/Questions on the JITT:

- Students not very familiar with STA 301 topics. E.g.:
  - Took it a year ago.
  - Online classes did not facilitate learning.
  - Didn't take STA 301.

**No need to remember everything from STA 301**

**Intuition about regression + R basics**

**Check the course website for resources!**

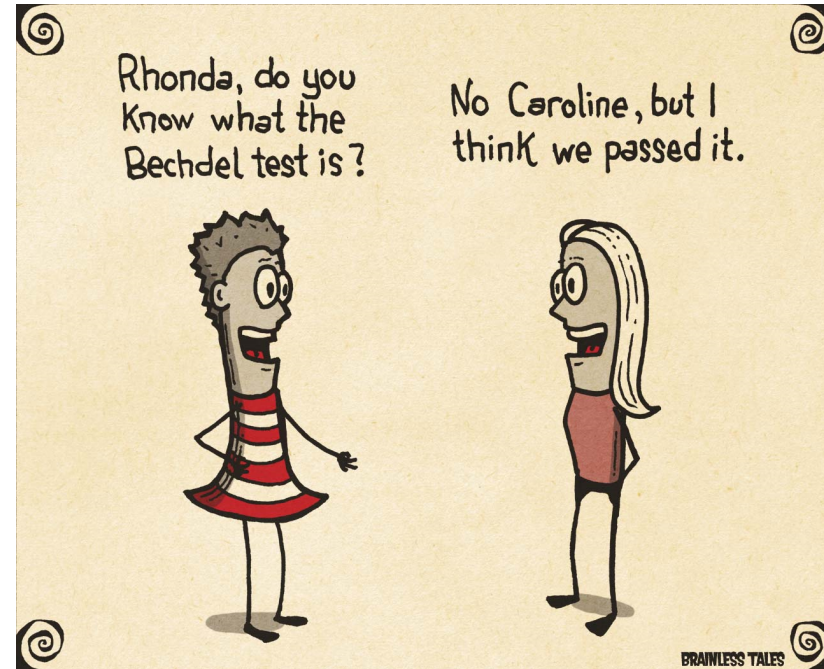
# Today

- Quick **multiple regression** review
  - Continuation from last week (outliers, comparing effect sizes)
- **Interpreting regressions:**
  - Interaction models
- **Potential issues in regressions:**
  - Multicollinearity
  - Heteroskedasticity

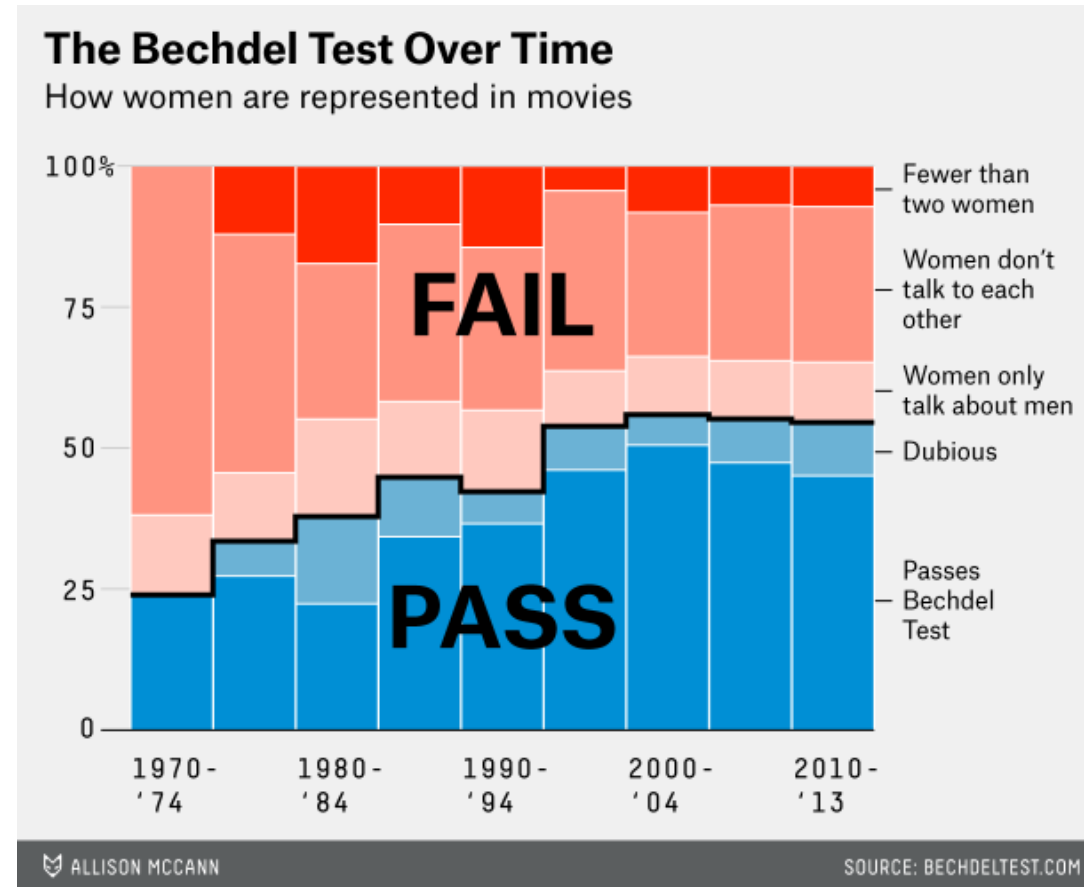


# Let's introduce an example: The Bechdel Test

- **Three criteria:**
  1. At least two named women
  2. Who talk to each other
  3. About something besides a man



# Do movies pass the test?



# 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 would fit?

# 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 would fit?

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

Is this right?





**What should we do before we ran any model?**

**Inspect your data!**

# vtable() can be of help

```
library(tidyverse)
library(vtable)

rawData <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/C1")

# Select movies post 1990
rawData <- rawData %>% filter(Year>1989)

# Create return on Investment (ROI) measures
# Passes Bechdel test:
rawData <- rawData %>% mutate(ROI = Revenue/Budget, # Total ROI
                             pass_bechdel = ifelse(rating==3, "PASS", "FAIL"))

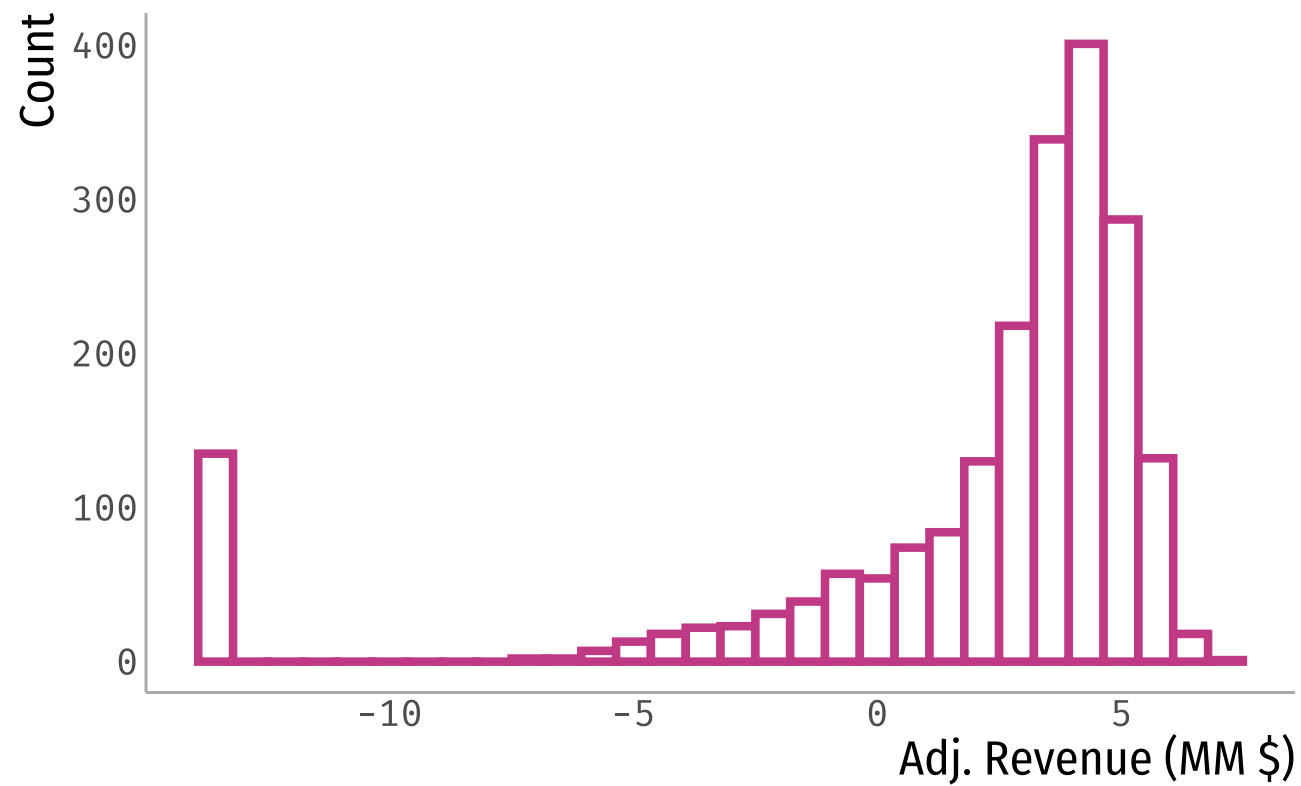
vtable(rawData)
```

**Look at the data**

# Look at the data

What can you say about this variable?

# Logarithms to the rescue?



# What to do with outliers?

## 1. Check them!

- Make sure there's no coding error; try to understand what's happening there.

## 2a. If they are wrongly coded:

- You can remove them, always adding a note of why you did so. Issues with the analysis will come from sample selection.

## 2b. If they are correctly coded:

- Run analysis both with and without outliers (don't just drop them!). E.g. Results do not depend exclusively on a few observations.

# Let's analyze some models

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

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   17.0321     0.0808 210.9100     0
## bechdel_test  -0.4418     0.1079  -4.0954     0
```

- How do you interpret these results?
- What are the units for the dependent variable?



## A side note on log-transformed variables...

$$\log(y) = \hat{\beta}_0 + \hat{\beta}_1 x$$

# A side note on log-transformed variables...

$$\log(y) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\log(y_1) - \log(y_0) = \hat{\beta}_0 + \hat{\beta}_1(x + 1) - (\hat{\beta}_0 + \hat{\beta}_1 x)$$

$$\log\left(\frac{y_1}{y_0}\right) = \hat{\beta}_1$$

$$\log\left(1 + \frac{y_1 - y_0}{y_0}\right) = \hat{\beta}_1$$

## A side note on log-transformed variables...

$$\log(y) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\log(y_1) - \log(y_0) = \hat{\beta}_0 + \hat{\beta}_1(x + 1) - (\hat{\beta}_0 + \hat{\beta}_1 x)$$

$$\log\left(\frac{y_1}{y_0}\right) = \hat{\beta}_1$$

$$\log\left(1 + \frac{y_1 - y_0}{y_0}\right) = \hat{\beta}_1$$

$$\rightarrow \frac{\Delta y}{y} = \exp(\hat{\beta}_1) - 1$$

# Let's analyze some models

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

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   17.0321     0.0808  210.9100      0
## bechdel_test  -0.4418     0.1079   -4.0954      0
```

- $(e^{\beta} - 1) \cdot 100 \rightarrow$  A movie that passes the Bechdel test is associated with a 36% decrease in Revenue

**Negative effect of including more women?**

# What gives?

## FiveThirtyEight

Politics Sports Science Podcasts Video

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(log(Adj_Revenue) ~ bechdel_test + log(Adj_Budget) + Metascore + imdb, data = bechdel)
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	1.3798	0.5126	2.6921	0.0072
## bechdel_test	0.2275	0.0665	3.4229	0.0006
## log(Adj_Budget)	0.8594	0.0256	33.6160	0.0000
## Metascore	0.1012	0.0293	3.4512	0.0006
## imdb	0.0864	0.0517	1.6716	0.0948

**Positive and significant!**

# Comparing effect sizes

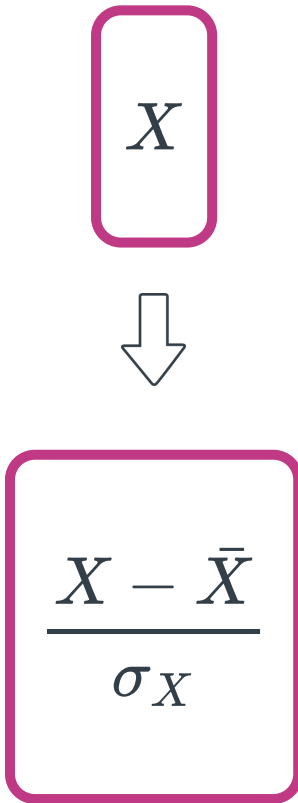
- Another investor says that it's better to bring in a better actor because it will increase ratings.
- **How do you compare effect sizes?**
  - How does one more point on IMDB compare to passing/failing the Bechdel test?





# Standardized Partial Coefficients

- **Main idea:** Transform everything to the same scale (standard deviations)



The diagram illustrates the transformation of a variable  $X$  into its standardized form. At the top, the variable  $X$  is enclosed in a vertical pink rounded rectangle. A white downward-pointing arrow is positioned below this rectangle. At the bottom, the standardized formula  $\frac{X - \bar{X}}{\sigma_X}$  is enclosed in a wider pink rounded rectangle.

$$X$$

↓

$$\frac{X - \bar{X}}{\sigma_X}$$

- Will this change our estimates? How?

# Transform the data

```
scale2 <- function(x, na.rm = FALSE) (x - mean(x, na.rm = na.rm)) / sd(x, na.rm)
```

```
bechdel_std <- bechdel %>% select(log_Adj_Revenue, log_Adj_Budget,  
                                bechdel_test, Metascore, imdb) %>%  
  mutate_at(vars(log_Adj_Budget, bechdel_test, Metascore, imdb),  
            ~scale2(., na.rm=TRUE))
```

```
lm(log_Adj_Revenue ~ bechdel_test + log_Adj_Budget + Metascore + imdb,  
   data = bechdel_std)
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	17.5882	0.0328	537.0138	0.0000
## bechdel_test	0.1129	0.0330	3.4229	0.0006
## log_Adj_Budget	1.1122	0.0331	33.6160	0.0000
## Metascore	0.1676	0.0486	3.4512	0.0006
## imdb	0.0812	0.0486	1.6716	0.0948

- What are the units on bechdel\_test now? Does it make sense?

# Transform the data

```
scale2 <- function(x, na.rm = FALSE) (x - mean(x, na.rm = na.rm)) / sd(x, na.rm)

bechdel_std <- bechdel %>% select(log_Adj_Revenue, log_Adj_Budget,
                                bechdel_test, Metascore, imdb) %>%
  mutate_at(vars(log_Adj_Budget, bechdel_test, Metascore, imdb),
             ~scale2(., na.rm=TRUE))

lm(log_Adj_Revenue~bechdel_test+log_Adj_Budget + Metascore + imdb,
   data = bechdel_std)
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	17.5882	0.0328	537.0138	0.0000
## bechdel_test	0.1129	0.0330	3.4229	0.0006
## log_Adj_Budget	1.1122	0.0331	33.6160	0.0000
## Metascore	0.1676	0.0486	3.4512	0.0006
## imdb	0.0812	0.0486	1.6716	0.0948

- What are the units on bechdel\_test now? Does it make sense?

# Transform the data

```
scale2 <- function(x, na.rm = FALSE) (x - mean(x, na.rm = na.rm)) / sd(x, na.rm)

bechdel_std <- bechdel %>% select(log_Adj_Revenue, log_Adj_Budget,
                                bechdel_test, Metascore, imdb) %>%
  mutate_at(vars(log_Adj_Budget, bechdel_test, Metascore, imdb), ~scale2(., na.rm=TRUE))

lm(log_Adj_Revenue ~ bechdel_test + log_Adj_Budget + Metascore + imdb,
   data = bechdel_std)
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	17.5882	0.0328	537.0138	0.0000
## bechdel_test	0.1129	0.0330	3.4229	0.0006
## log_Adj_Budget	1.1122	0.0331	33.6160	0.0000
## Metascore	0.1676	0.0486	3.4512	0.0006
## imdb	0.0812	0.0486	1.6716	0.0948

- What are the units on bechdel\_test now? Does it make sense?

# Main takeaway points

- Data can tell different stories depending on how you handle it.
  - Does that mean that we can get data to say **anything**?

# Main takeaway points

- Data can tell different stories depending on how you handle it.
  - Does that mean that we can get data to say **anything**?

**"If you torture the data long enough, it will confess to anything"**

- Assumptions and measures matter.

# Main takeaway points

- Data can tell different stories depending on how you handle it.
  - Does that mean that we can get data to say **anything**?

**"If you torture the data long enough, it will confess to anything"**

- Assumptions and measures matter.
- **Plot your data!**

# 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 Openness in Data Journalism". *Github Repository*