STA 235 - Multiple Regression

Spring 2021

McCombs School of Business, UT Austin

Today

- Quick **multiple regression** review
  - How does OLS work?
- Comparing **effect sizes**
- **Uncertainty quantification** in regression

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)**.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- What are β's?

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)**.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- What are β's?
  - β's are the **population parameters** we want to estimate.
  - $\hat{\beta}$ are the **estimates** of those parameters.

Shoutout: Slides based on Andrew Heiss' "Program Evaluation for Public Service" course.

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

Outcome Variable

Response Variable

Dependent Variable

Thing you want to explain or predict

X

Explanatory Variable

Predictor Variable

Independent Variable

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

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

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

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

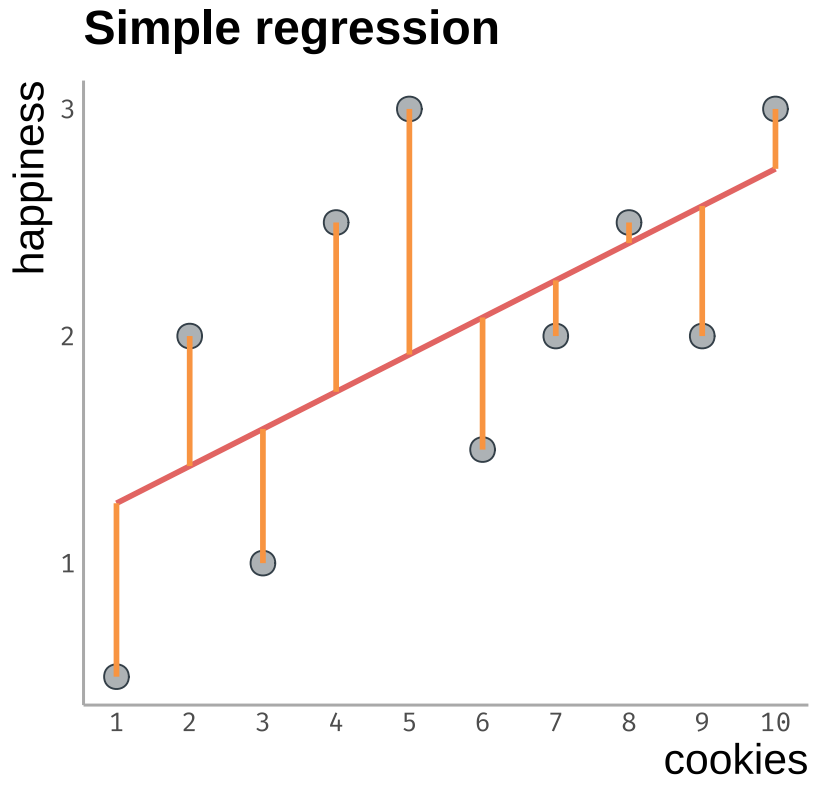
Focus is on X

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

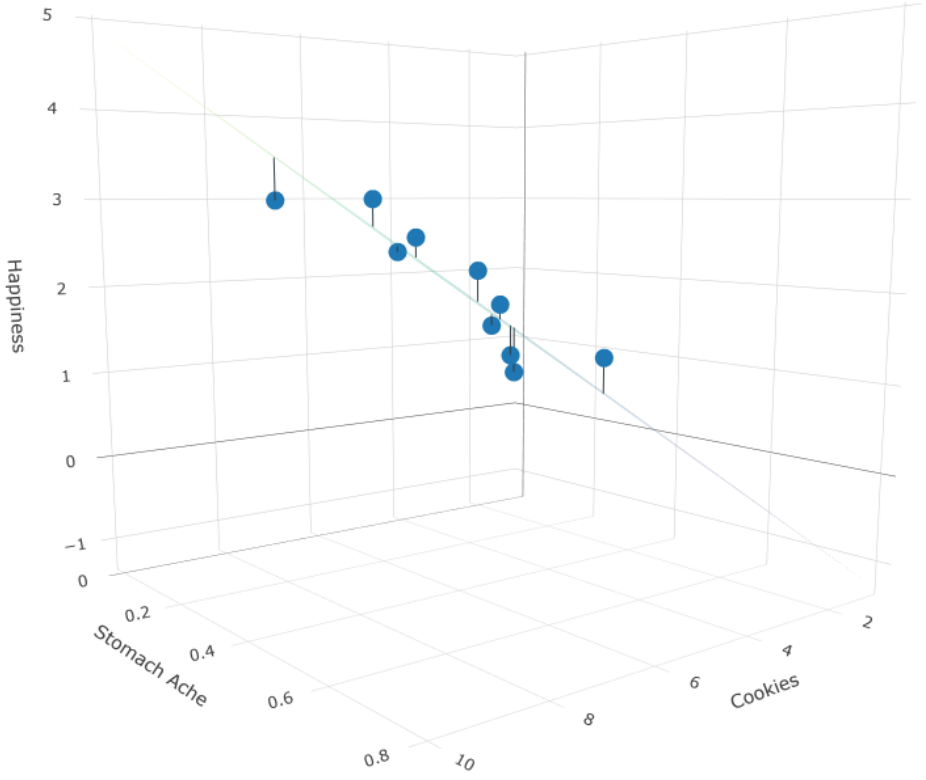
How do we estimate the coefficients in a regression ?

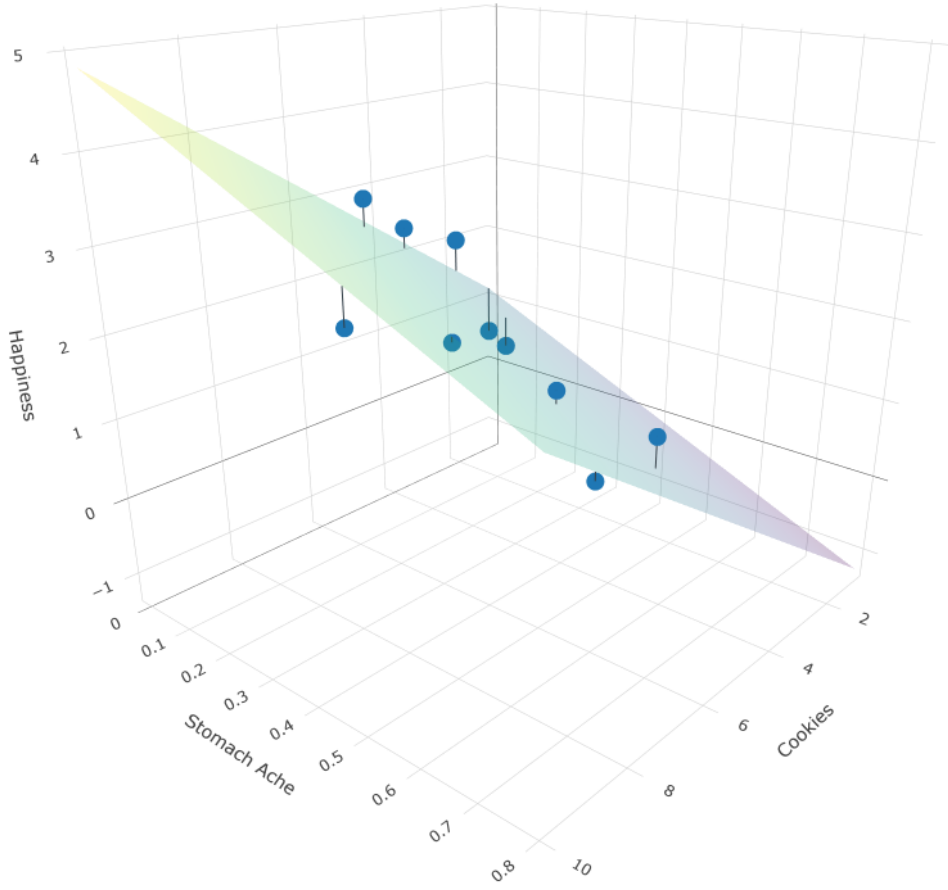
- **Ordinary Least Squares** is the most popular way.

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



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

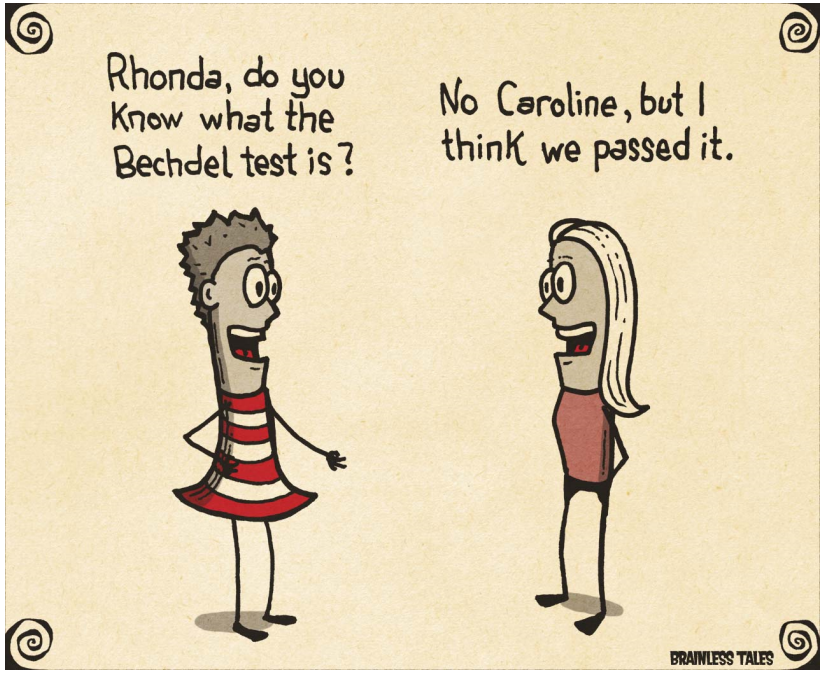




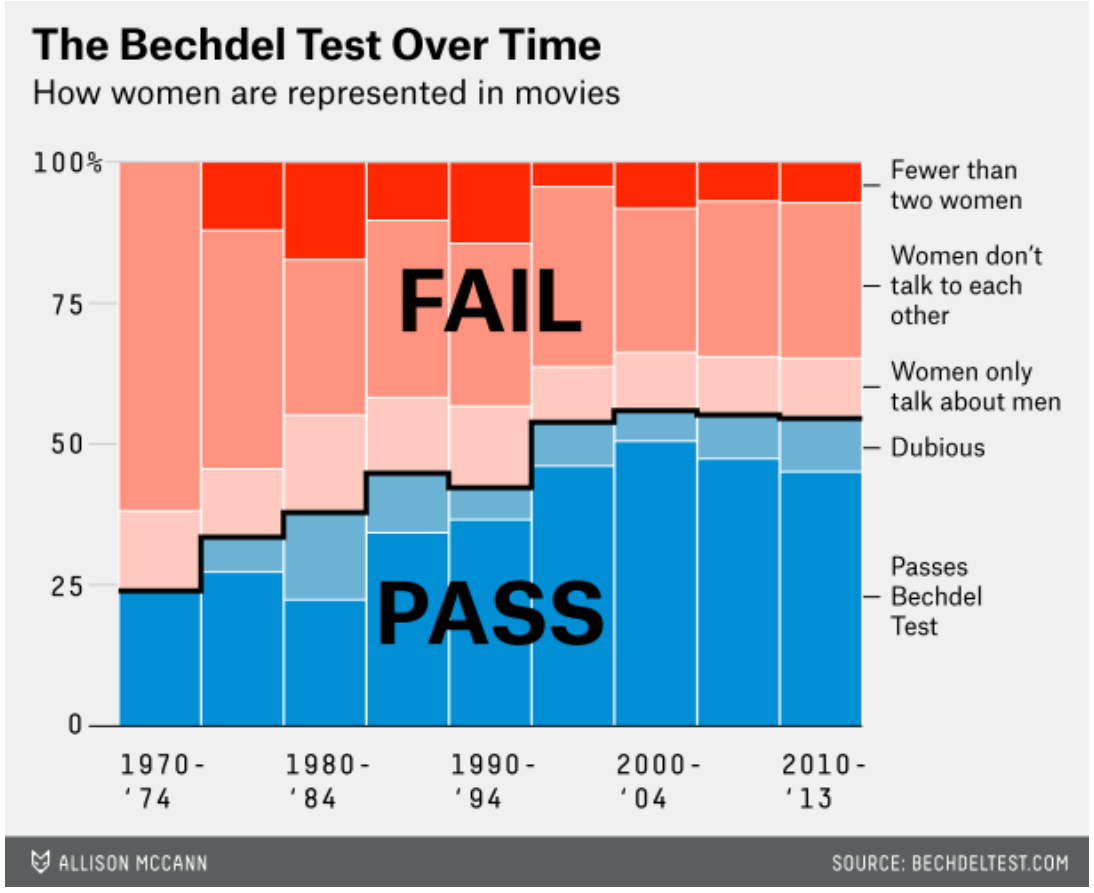
Let's introduce an example: The Bechdel Test

- **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?

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
```

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)

$$X$$

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

- Will this change our estimates? How?

Transform the data

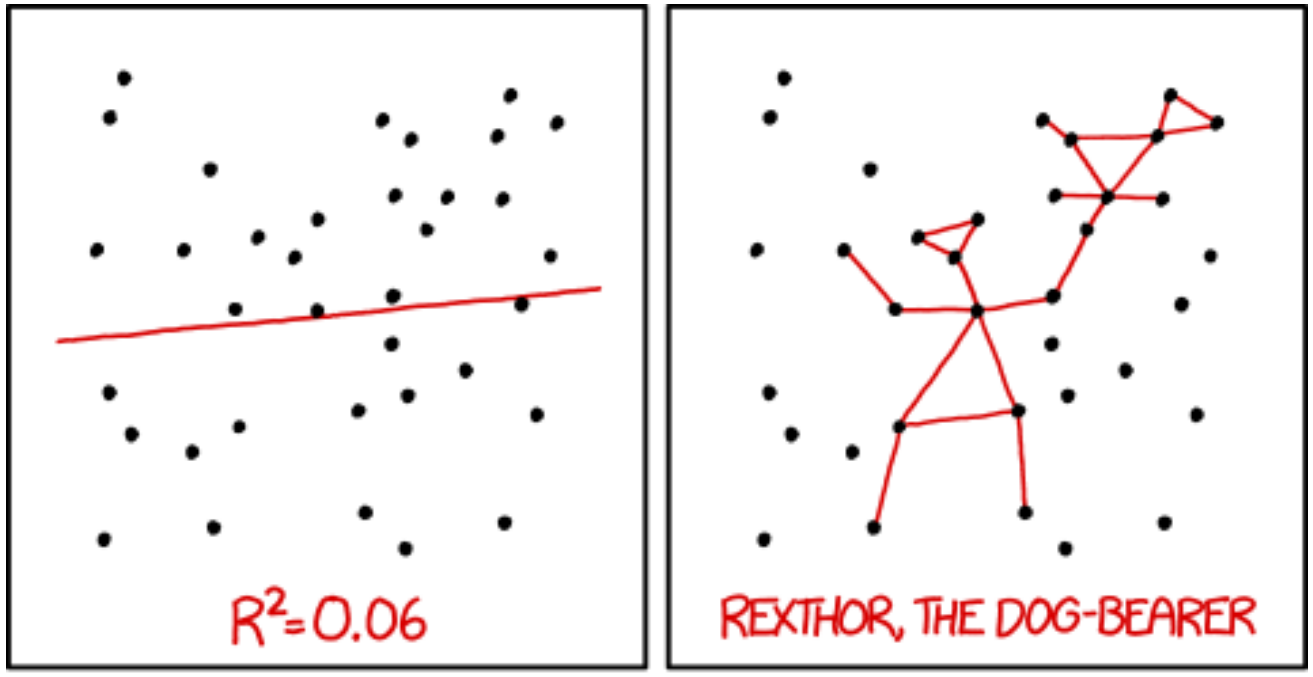
```
scale2 <- function(x, na.rm = FALSE) (x - mean(x, na.rm = na.rm)) / sd(x, na.rm)

bechdel_std <- bechdel %>% select(log_Adj_Revenue,log_Adj_Budget,
                                  bechdel_test, Metascore, imdb) %>%
  mutate_at(c("log_Adj_Revenue","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)      0.3384     0.0138 24.5385   0.0000
## bechdel_test     0.0476     0.0139  3.4229   0.0006
## log_Adj_Budget   0.4683     0.0139 33.6160   0.0000
## Metascore        0.0706     0.0205  3.4512   0.0006
## imdb             0.0342     0.0205  1.6716   0.0948
```

Does this model explain well the observed data?



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Some measures to quantify uncertainty

- $\mathbf{R^2}$: Correlation between the outcome $\mathbf{Y}$ and the predicted variable $\mathbf{\hat{Y}}$.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where:

- $RSS = \sum (y - \hat{y})^2$ is the residual sum of the squares
- $TSS = \sum (y - \bar{y})^2$ is the total sum of squares (proportional to the variance of the data)

- $\mathbf{R^2}$ is non-decreasing when adding new variables.

Some measures to quantify uncertainty

- $\mathbf{R^2}$: Correlation between the outcome $\mathbf{Y}$ and the predicted variable $\mathbf{\hat{Y}}$.

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where:

- $RSS = \sum (y - \hat{y})^2$ is the residual sum of the squares
- $TSS = \sum (y - \bar{y})^2$ is the total sum of squares (proportional to the variance of the data)

- $\mathbf{R^2}$ is non-decreasing when adding new variables.

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{(n - p - 1)}$$

Some measures to quantify uncertainty (cont.)

- **Residual Squared Error (RSE)**: Measure of the SD of error term ε

$$RSE = \sqrt{\frac{1}{n-p+1}RSS}$$

- What happens if p increases?

Some measures to quantify uncertainty (cont.)

```
summary(lm(log(Adj_Revenue) ~ bechdel_test +log(Adj_Budget) + Metascore + imdb,
           data=bechdel), digits = 4)
```

```
##
## Call:
## lm(formula = log(Adj_Revenue) ~ bechdel_test + log(Adj_Budget) +
##     Metascore + imdb, data = bechdel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1728 -0.4743  0.1103  0.6396  4.4767
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.37982    0.51255   2.692 0.007194 **
## bechdel_test     0.22750    0.06646   3.423 0.000639 ***
## log(Adj_Budget)  0.85936    0.02556  33.616  < 2e-16 ***
## Metascore        0.10115    0.02931   3.451 0.000576 ***
## imdb             0.08643    0.05170   1.672 0.094844 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.159 on 1276 degrees of freedom
##   (671 observations deleted due to missingness)
## Multiple R-squared:  0.477,    Adjusted R-squared:  0.4753
## F-statistic: 290.9 on 4 and 1276 DF,  p-value: < 2.2e-16
```

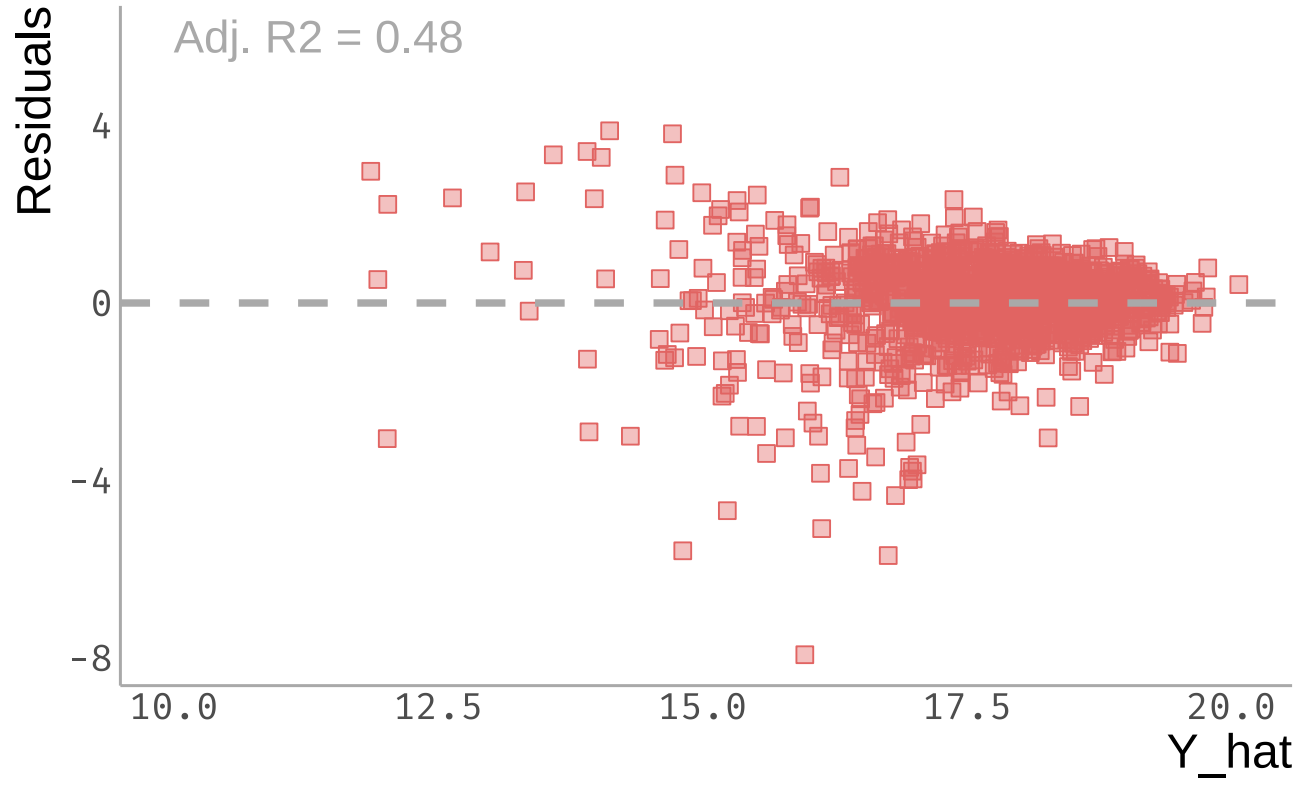
Let's look at the residuals

```
bechdel_fitted <- augment(lm(log(Adj_Revenue) ~ bechdel_test+log(Adj_Budget) + Metascore + imdb
                             , data = bechdel))


ggplot(data = bechdel_fitted, aes(x = .fitted, y = .std.resid)) +
  geom_point()
```

- You can also use the functions `predict(lm(Y ~ X + Z))` and `resid(lm(Y ~ X + Z))` to obtain the fitted values and residuals, respectively.

Let's look at the residuals (cont.)

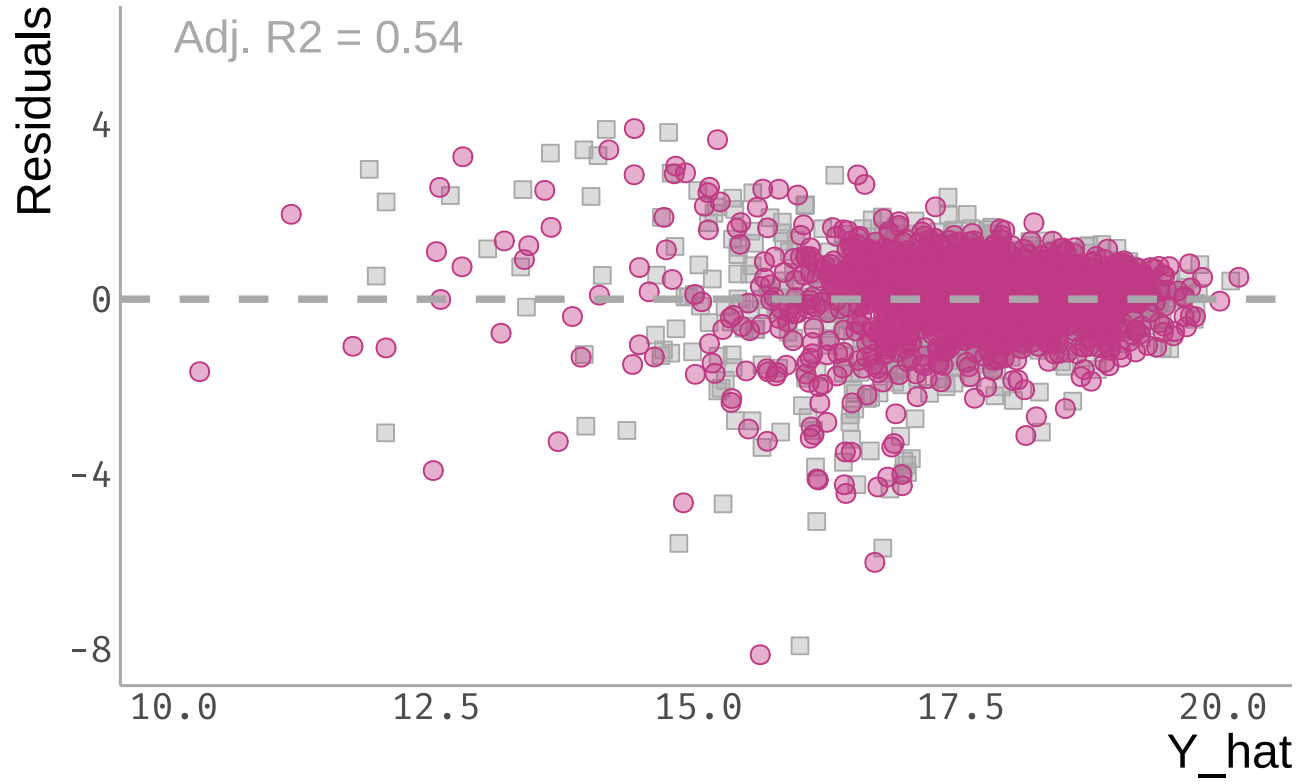


What happens if we add additional variables?

```
bechdel_fitted2 <- augment(lm(log(Adj_Revenue) ~ bechdel_test+log(Adj_Budget) + Metascore + imdb +
                                Year + English + USA + factor(Rated), data = bechdel))


ggplot(data = bechdel_fitted2, aes(x = .fitted, y = .std.resid)) +
  geom_point()
```

What happens with additional variables? (cont.)



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!**

Next class

- Finishing with **multiple regression models**:
  - Statistical adjustment and collinearity
- Model with discrete responses: **Binary outcomes**
- In the following weeks, we will start with **Causal Inference**.

References

- Heiss, A. (2020). "Course: Program Evaluation for Public Service". *Slides for Regression and Inference*.
- James, G. et al. (2017). "Introduction to Statistical Learning with Applications in R". *Chapter 3*
- Keegan, B. (2018). "The Need for Openess in Data Journalism". *Github Repository*