STA 235H - Multiple Regression: Interactions & Nonlinearity

Fall 2023

McCombs School of Business, UT Austin

Before we start...

- Use the knowledge check portion of the JITT to assess your own understanding:
  - Be sure to answer the question correctly (look at the feedback provided)
  - Feedback are guidelines; Try to use your *own words*.
- If you are struggling with material covered in STA 301H: Check the course website for resources and come to Office Hours.
- Office Hours Prof. Bennett: Wed 10.30-11.30am and Thu 4.00-5.30pm

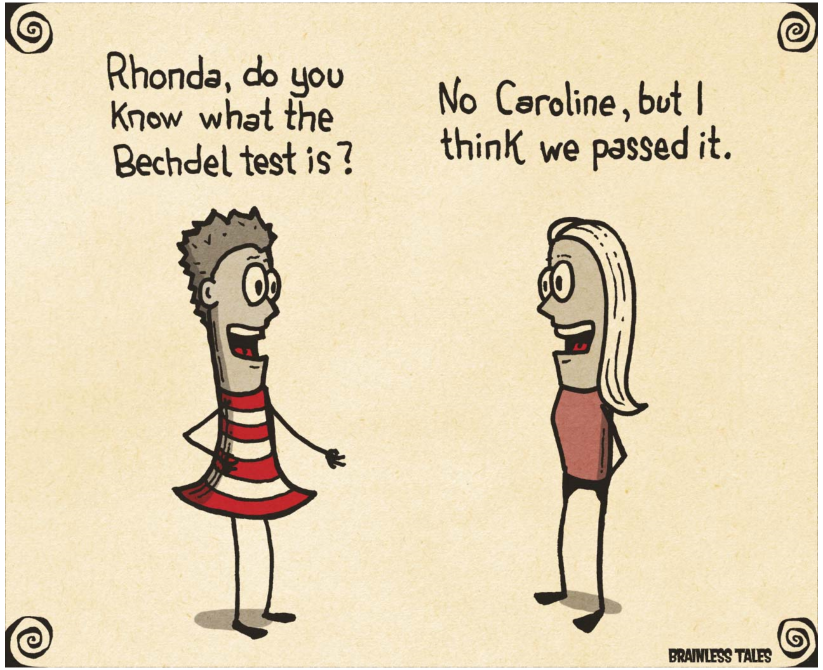
No in-person class next week -- Recorded class

Today

- Quick multiple regression review:
  - Interpreting coefficients
  - Interaction models
- Looking at your data:
  - Distributions
- Nonlinear models:
  - Logarithmic outcomes
  - Polynomial terms

Remember last week's example? The Bechdel Test

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



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

```
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
```

What does each column represent?

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

```
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
```

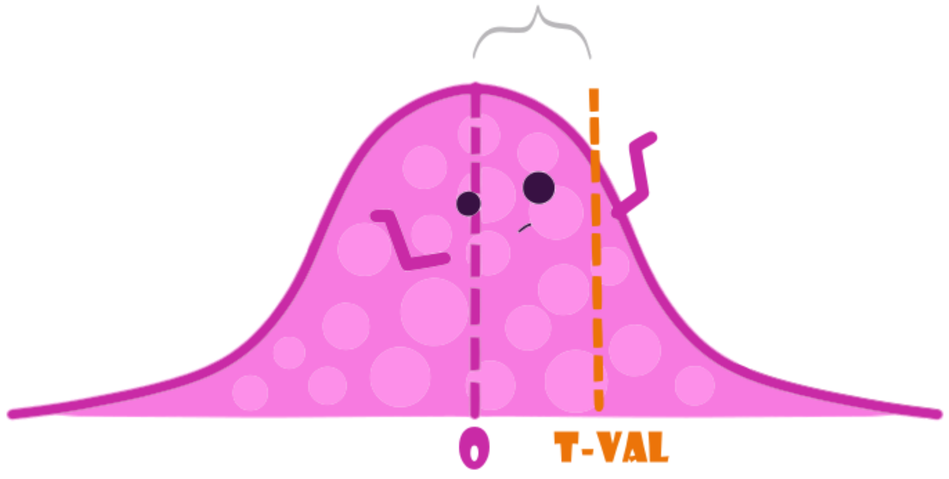
- **"Estimate"**: Point estimates of our paramters β. We call them $\hat{\beta}$.
- **"Standard Error"** (SE): You can think about it as the variability of $\hat{\beta}$. The smaller, the more precise $\hat{\beta}$ is!
- **"t-value"**: A value of the Student distribution that measures how many SE away $\hat{\beta}$ is from 0. You can calculate it as $tval = \frac{\hat{\beta}}{SE}$. It relates to our null-hypothesis $H_0 : \beta = 0$.
- **"p-value"**: Probability of rejecting the null hypothesis and being *wrong* (Type I error). You want this to be a small as possible (statistically significant).

Reminder: Null-Hypothesis

We are testing $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

- "Reject the null hypothesis"

- "Not reject the null hypothesis"



Note: Figures adapted from @AllisonHorst's art

Reminder: Null-Hypothesis

Reject the null if the t-value falls **outside** the dashed lines.

One extra dollar in our budget

- Imagine now that you have an hypothesis that Bechdel movies also get more bang for their buck, e.g. they get more revenue for an additional dollar in their budget.

How would you test that in an equation?

Interactions!

One extra dollar in our budget

Interaction model:

$$Revenue = \beta_0 + \beta_1 Bechdel + \beta_3 Budget + \beta_6 (Budget \times Bechdel) + \beta_4 IMDB + \beta_5 MetaScore + \varepsilon$$

How should we think about this?

- Write the equation for a movie that **does not pass the Bechdel test**. How does it look like?
- Now do the same for a movie that **passes the Bechdel test**. How does it look like?

One extra dollar in our budget

Now, let's interpret some coefficients:

- If $Bechdel = 0$, then:

$$Revenue = \beta_0 + \beta_3 Budget + \beta_4 IMDB + \beta_5 MetaScore + \varepsilon$$

- If $Bechdel = 1$, then:

$$Revenue = (\beta_0 + \beta_1) + (\beta_3 + \beta_6) Budget + \beta_4 IMDB + \beta_5 MetaScore + \varepsilon$$

- What is the **difference** in the association between budget and revenue for movies that pass the Bechdel test vs. those that don't?

Let's put some data into it

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

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -124.1997    17.4932 -7.0999   0.0000
## bechdel_test              7.5138     6.4257  1.1693   0.2425
## Adj_Budget                1.0926     0.0513 21.2865   0.0000
## Metascore                 7.1424     1.9126  3.7344   0.0002
## imdbRating               15.2268     3.4069  4.4694   0.0000
## bechdel_test:Adj_Budget   0.0546     0.0737  0.7416   0.4585
```

- What is the association between budget and revenue for movies that pass the Bechdel test?
- What is the difference in the association between budget and revenue for movies that pass vs movies that don't pass the Bechdel test?
- Is that difference statistically significant (at conventional levels)?

Let's look at another example

Cars, cars, cars

- Used cars in South California (from this week's JITT)

```
cars <- read.csv("https://raw.githubusercontent.com/maibennett/sta235/main/exampleSite/content/Classes/Week2/1_OLS/data/Sc

names(cars)
```

```
##  [1] "type"      "certified" "body"      "make"      "model"     "trim"
##  [7] "mileage"   "price"     "year"      "dealer"    "city"      "rating"
## [13] "reviews"   "badge"
```

Data source: "Modern Business Analytics" (Taddy, Hendrix, & Harding, 2018)

Luxury vs. non-luxury cars?

Do you think there's a difference between how price changes over time for luxury vs non-luxury cars?

How would you test this?

Let's go to R

Models with interactions

- You include the interaction between two (or more) covariates:

$$\widehat{Price} = \beta_0 + \hat{\beta}_1 Rating + \hat{\beta}_2 Miles + \hat{\beta}_3 Luxury + \hat{\beta}_4 Year + \hat{\beta}_5 Luxury \times Year$$

- $\hat{\beta}_3$ and $\hat{\beta}_4$ are considered the **main effects** (no interaction)
- The coefficient you are interested in is $\hat{\beta}_5$:
  - Difference in the **price change** for one additional year between **luxury vs non-luxury cars**, holding other variables constant.

Now it's your turn

- Looking at this equation:

$$\widehat{Price} = \beta_0 + \hat{\beta}_1 Rating + \hat{\beta}_2 Miles + \hat{\beta}_3 Luxury + \hat{\beta}_4 Year + \hat{\beta}_5 Luxury \times Year$$

1) What is the association between price and year for non-luxury cars?

2) What is the association between price and year for luxury cars?

Looking at our data

- We have dived into running models head on. Is that a good idea?

What should we do before we ran any model?

Inspect your data!

Some ideas:

- Use `vtable`:

```
library(vtable)

vtable(cars)
```

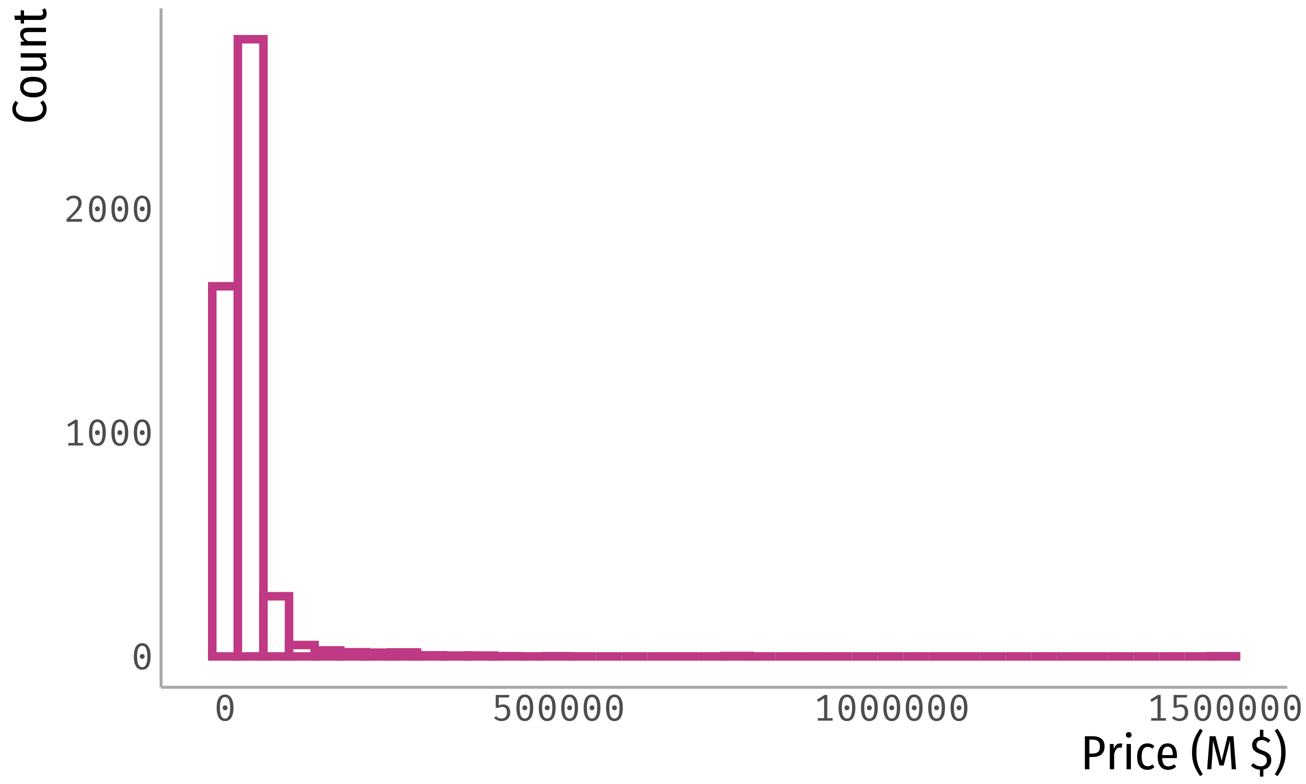
- Use `summary` to see the min, max, mean, and quartile:

```
cars %>% select(price, mileage, year) %>% summary(.)
```

```
##      price             mileage             year     
##  Min.   :   1790   Min.   :     0   Min.   :1966  
##  1st Qu.:  16234   1st Qu.:     5   1st Qu.:2017  
##  Median :  23981   Median :    56   Median :2019  
##  Mean   :  32959   Mean   : 21873   Mean   :2018  
##  3rd Qu.:  36745   3rd Qu.: 36445   3rd Qu.:2020  
##  Max.   :1499000   Max.   :292952   Max.   :2021  
```

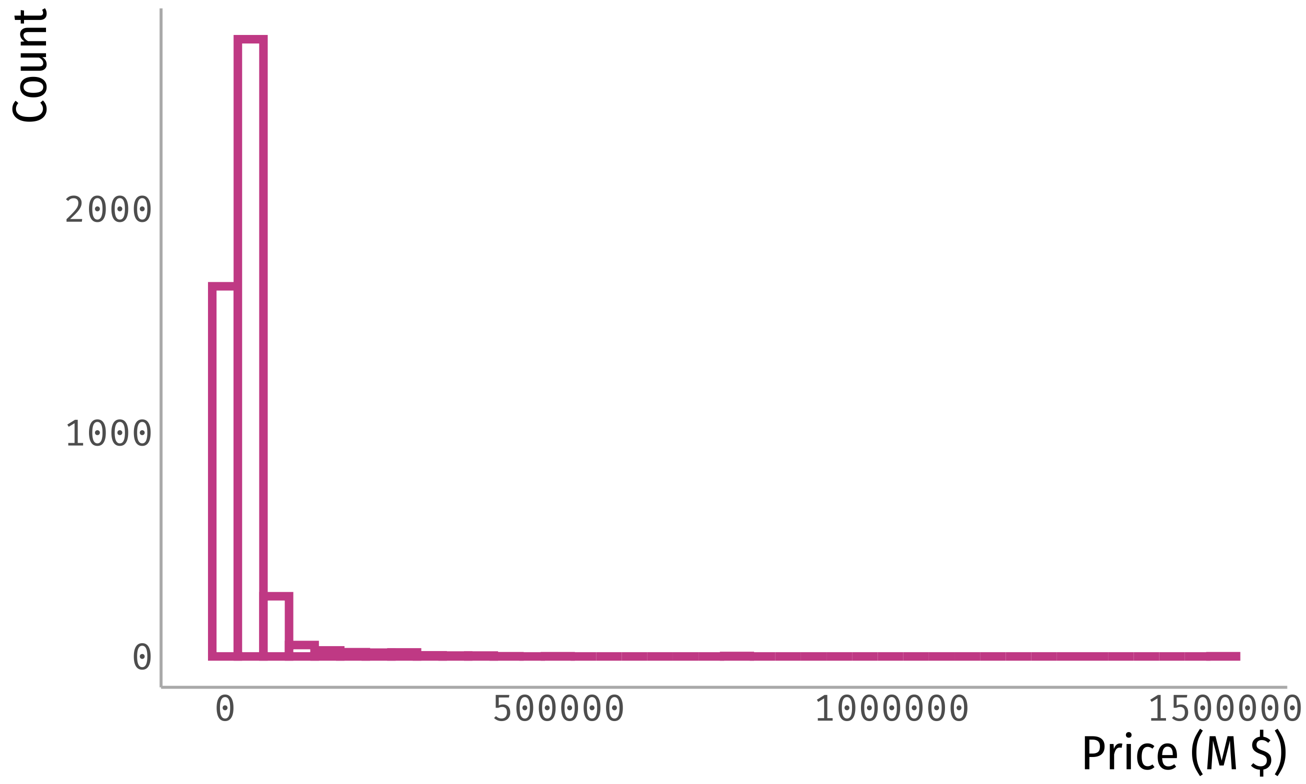
- Plot your data!

Look at the data

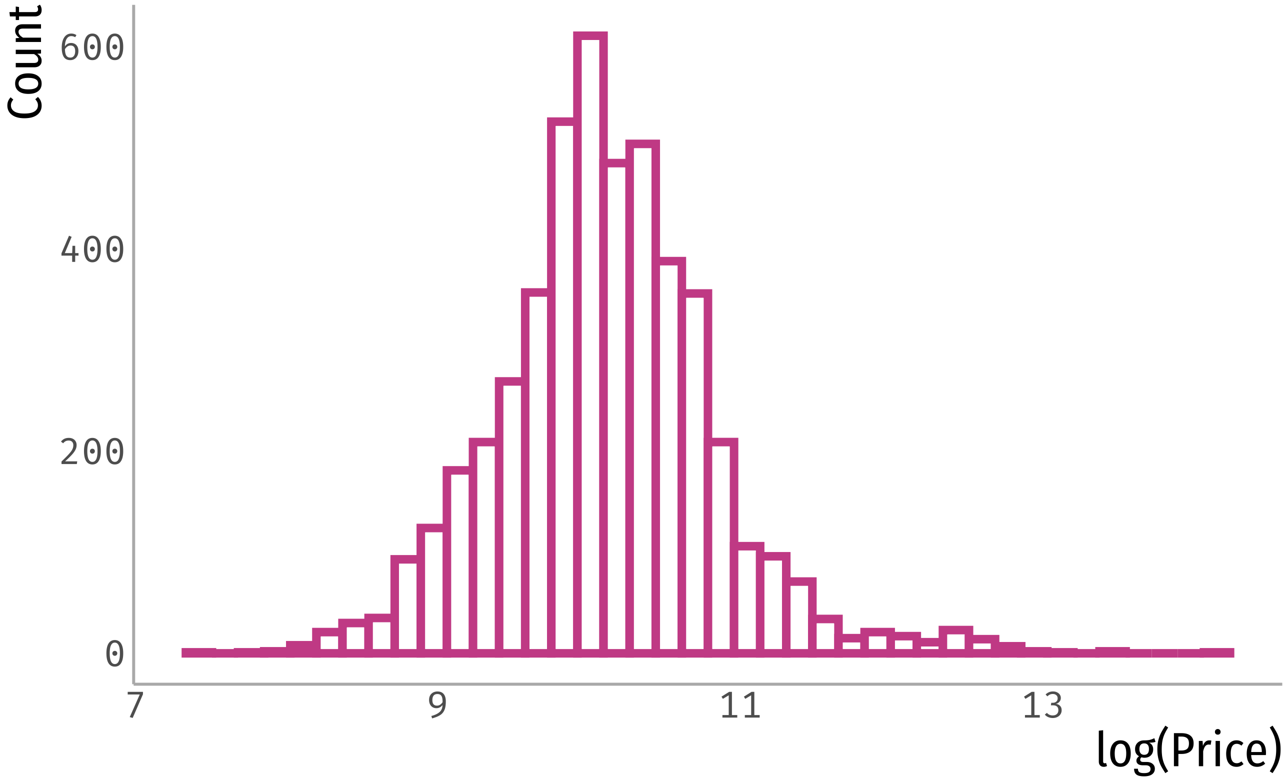


Look at the data

What can you say about this variable?



Logarithms to the rescue?



How would we interpret coefficients now?

- Let's interpret the coefficient for $Miles$ in the following equation:

$$\log(Price) = \beta_0 + \beta_1 Rating + \beta_2 Miles + \beta_3 Luxury + \beta_4 Year + \varepsilon$$

- Remember: β_2 represents the average change in the outcome variable, $\log(Price)$, for a one-unit increase in the independent variable $Miles$.
    - *Think about the units of the dependent and independent variables!*

A side note on log-transformed variables...

$$\log(Y) = \hat{\beta}_0 + \hat{\beta}_1 X$$

We want to compare the outcome for a regression with $X = x$ and $X = x + 1$

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

and

$$\log(y_1) = \hat{\beta}_0 + \hat{\beta}_1 (x + 1) \qquad (2)$$

- Let's subtract (2) - (1)!

A side note on log-transformed variables...

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

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

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

A side note on log-transformed variables...

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

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

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

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

An important approximation

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

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

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

$$\approx \frac{y_1 - y_0}{y_0} = \hat{\beta}_1$$

$$\boxed{\rightarrow \%\Delta y = 100 \times \hat{\beta}_1}$$

How would we interpret coefficients now?

- Let's interpret the coefficient for $Miles$ in the following equation:

$$\log(Price) = \beta_0 + \beta_1 Rating + \beta_2 Miles + \beta_3 Luxury + \beta_4 Year + \varepsilon$$

- For an additional 1,000 miles (*Note: Remember Miles is measured in thousands of miles*), the logarithm of the price increases/decreases, on average, by $\hat{\beta}_2$, holding other variables constant.

- For an additional 1,000 miles, the price increases/decreases, on average, by $100 \times \hat{\beta}_2\%$, holding other variables constant.

How would we interpret coefficients now?

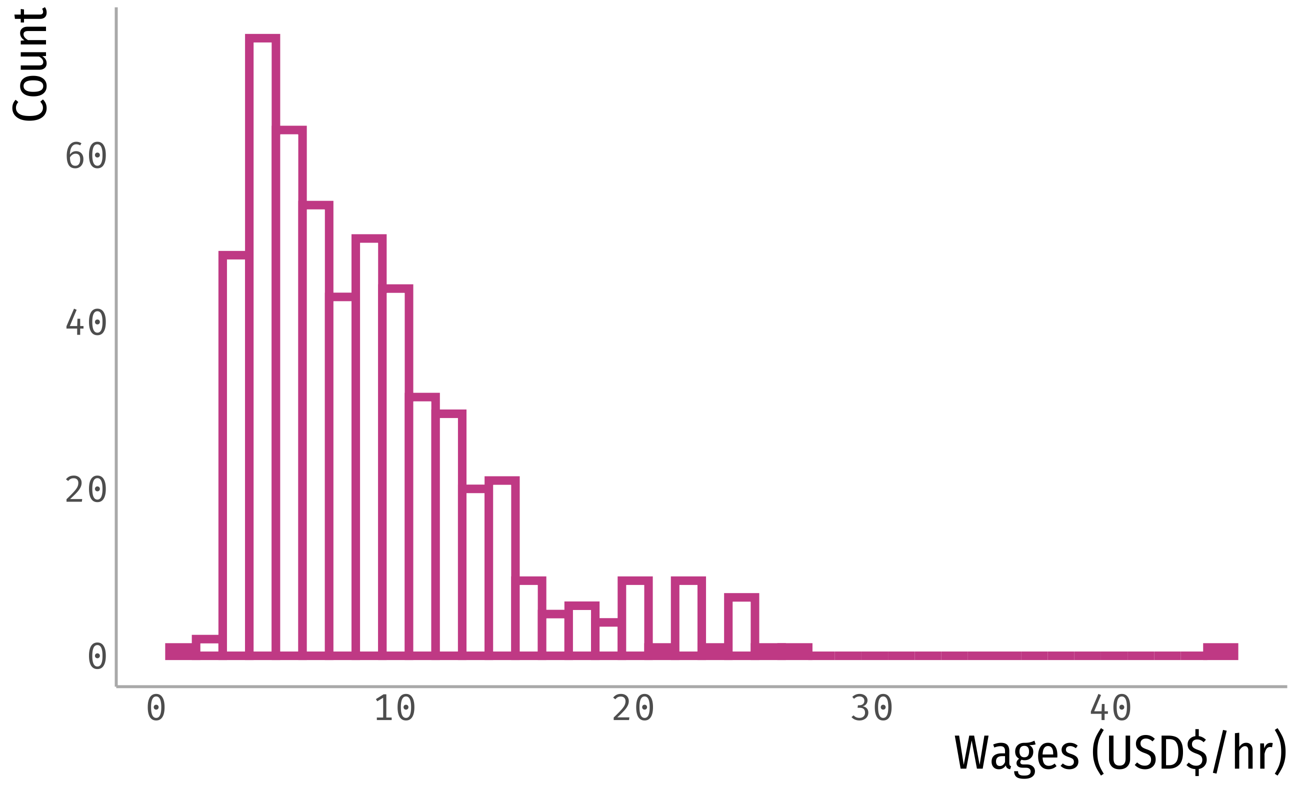
```
summary(lm(log(price) ~ rating + mileage + luxury + year, data = cars))
```

```
##
## Call:
## lm(formula = log(price) ~ rating + mileage + luxury + year, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14398 -0.29213 -0.02541  0.26465  2.28644
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.5110336  0.1518738  16.534  < 2e-16 ***
## rating       0.0302667  0.0057670   5.248 1.69e-07 ***
## mileage     -0.0098415  0.0004327 -22.745  < 2e-16 ***
## luxury       0.5527371  0.0228132  24.229  < 2e-16 ***
## year         0.0118467  0.0030083   3.938 8.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4361 on 2085 degrees of freedom
## Multiple R-squared:  0.4692,    Adjusted R-squared:  0.4682
## F-statistic: 460.7 on 4 and 2085 DF,  p-value: < 2.2e-16
```

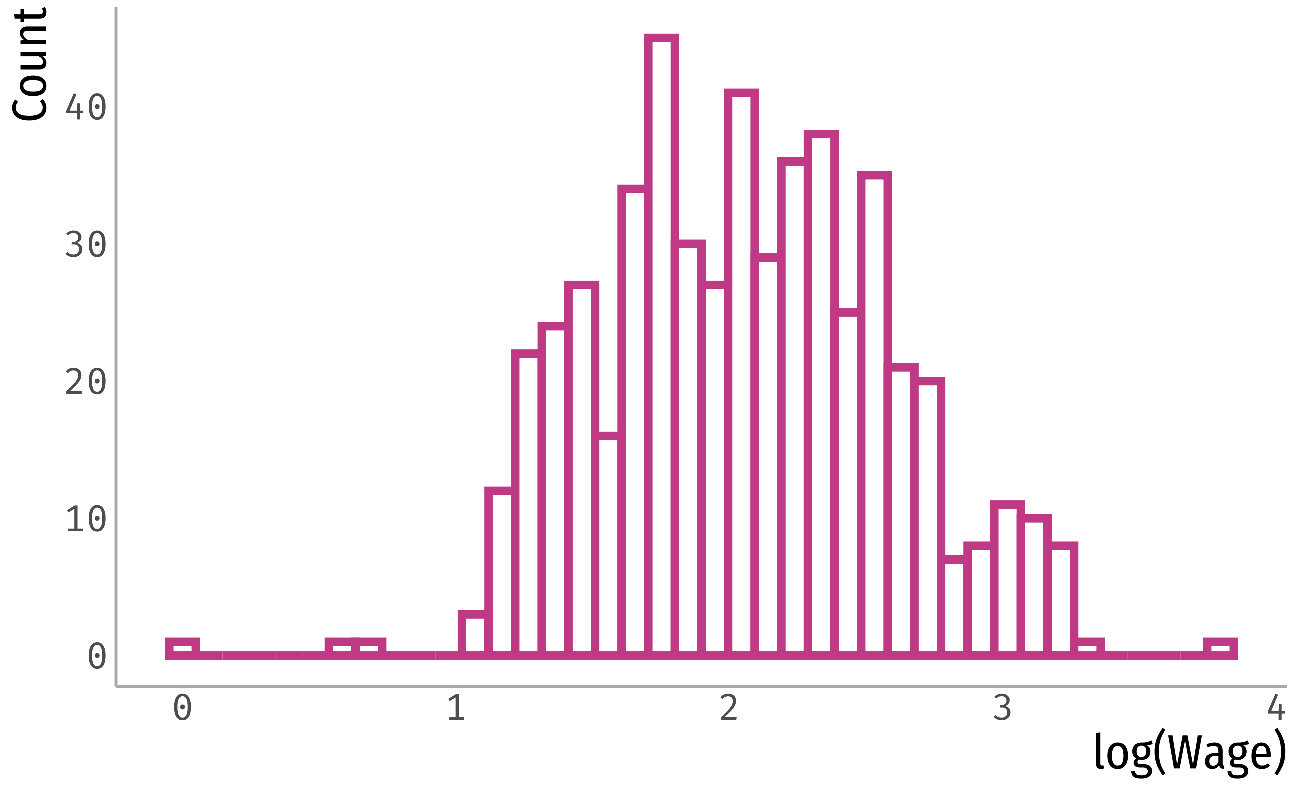
Adding polynomial terms

- Another way to capture **nonlinear associations** between the outcome (Y) and covariates (X) is to include **polynomial terms**:
    - e.g. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
- Let's look at an example!

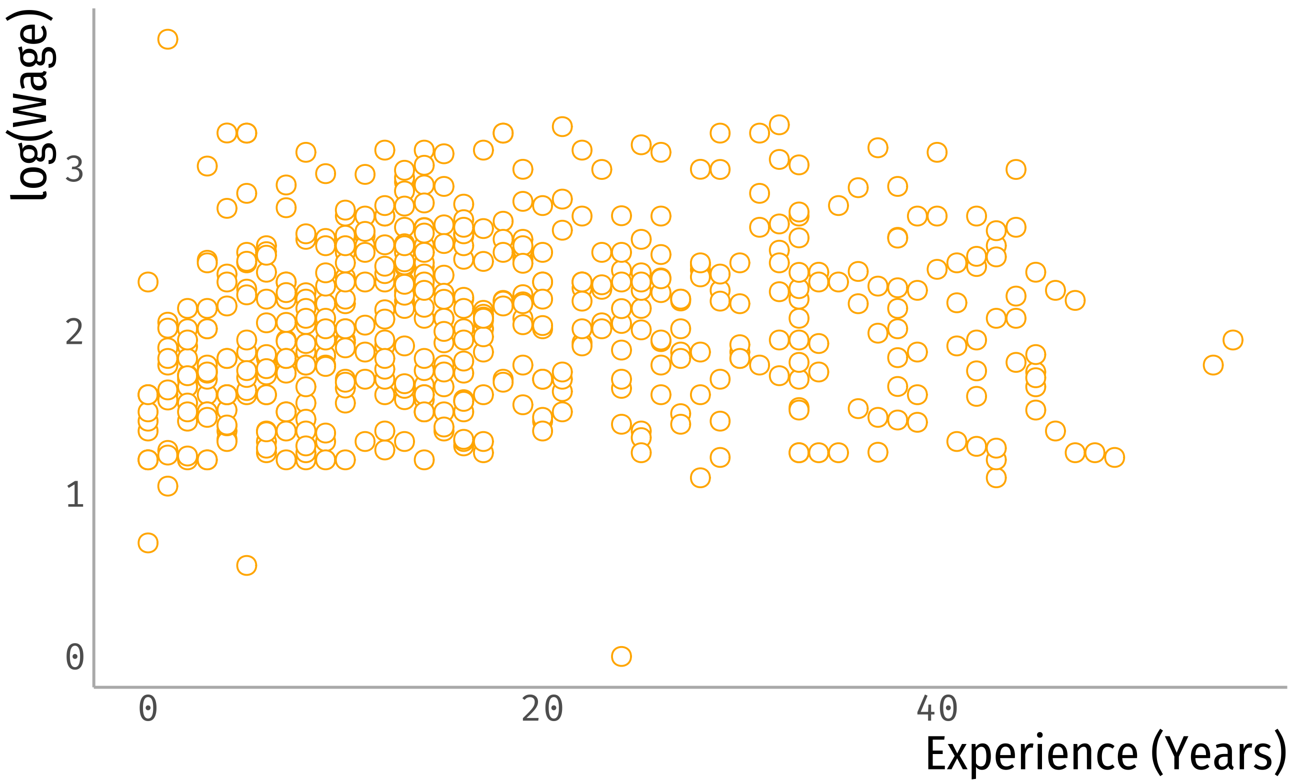
Determinants of wages: CPS 1985



Determinants of wages: CPS 1985

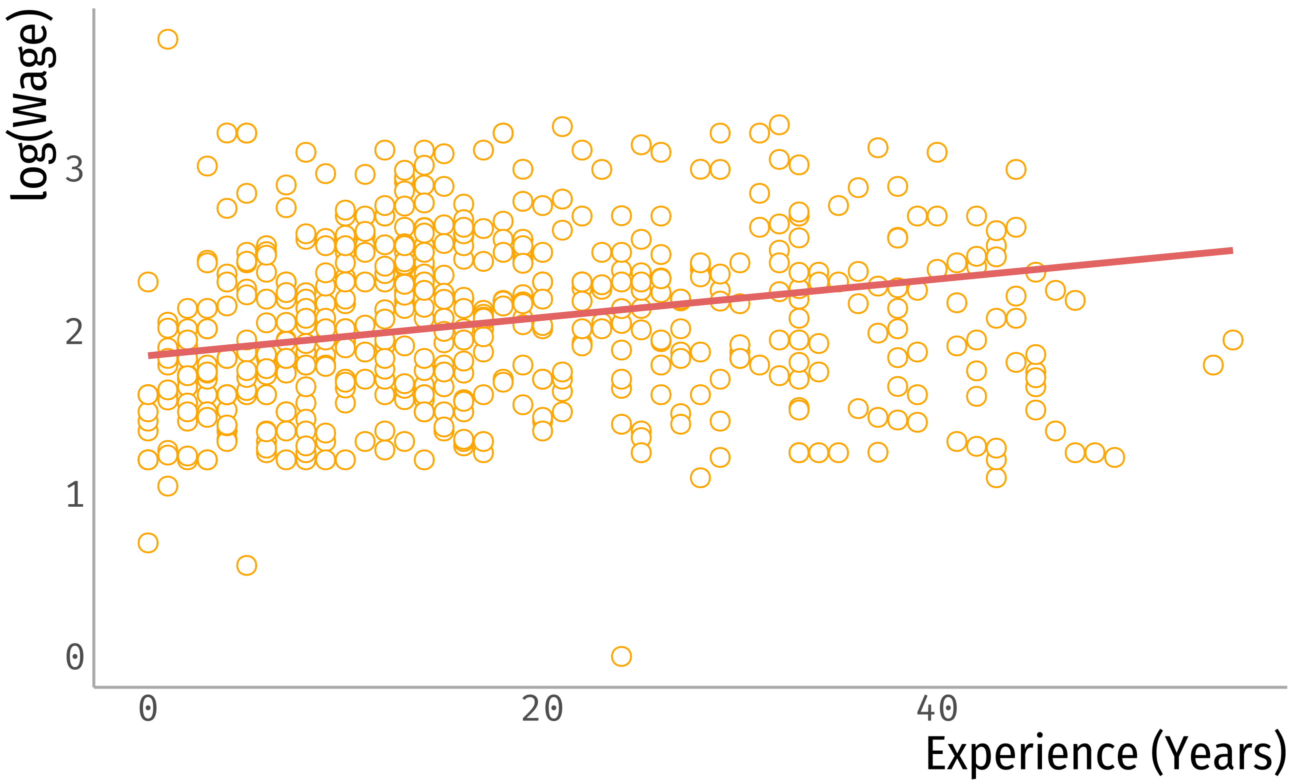


Experience vs wages: CPS 1985



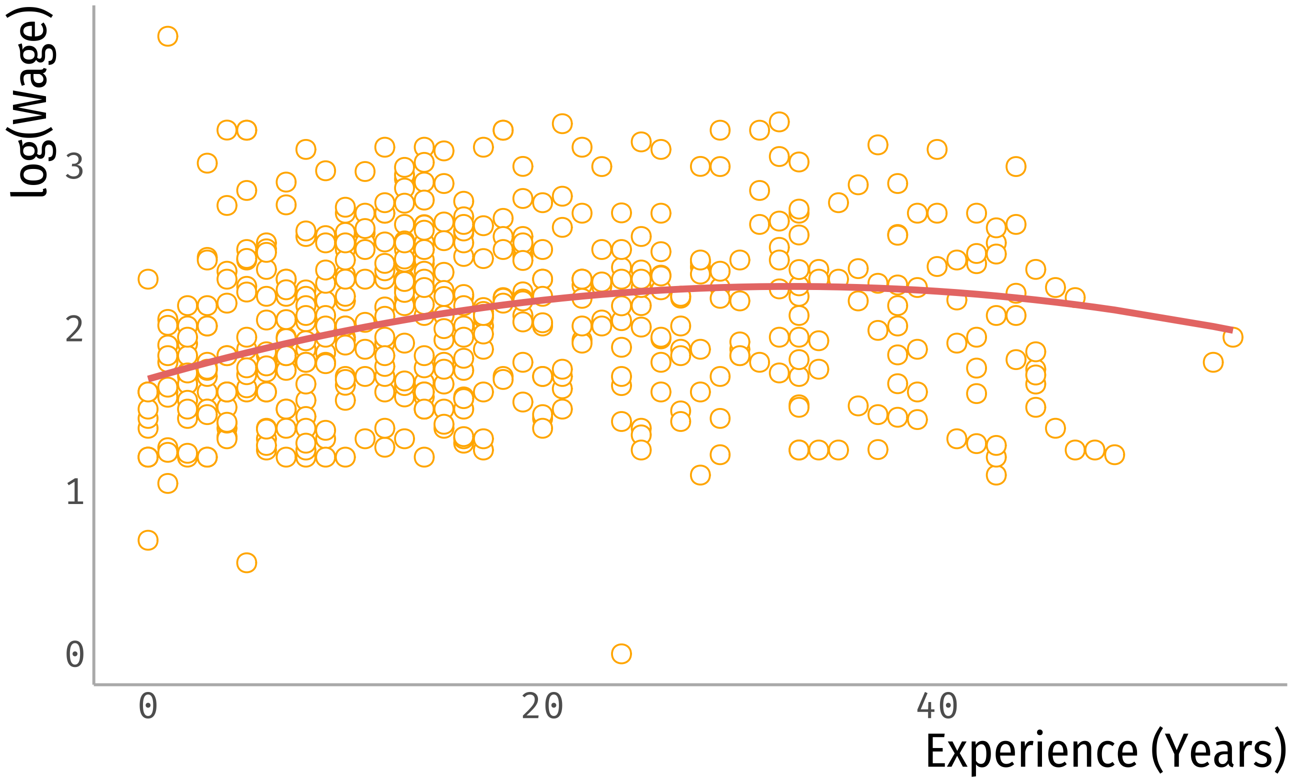
Experience vs wages: CPS 1985

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exp + \varepsilon$$



Experience vs wages: CPS 1985

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exp + \beta_3 Exp^2 + \varepsilon$$



Mincer equation

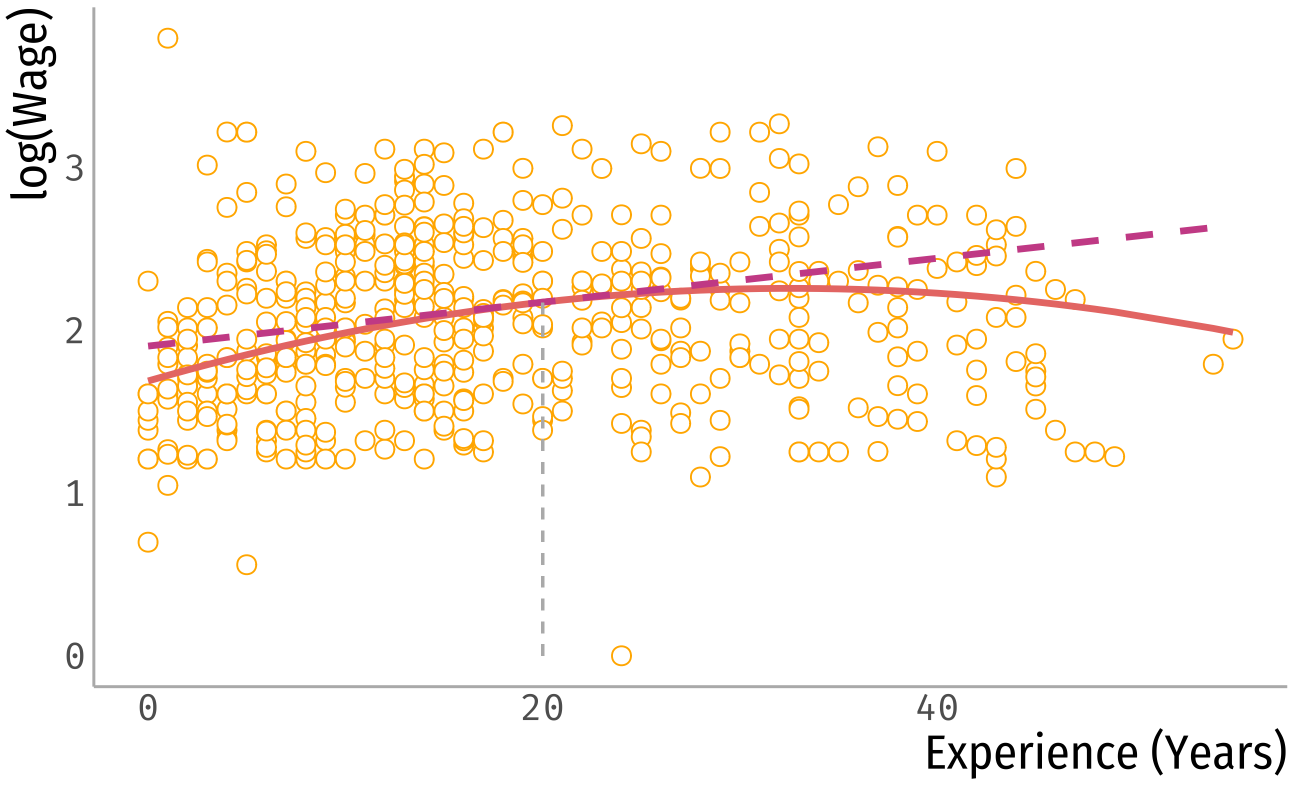
$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exp + \beta_3 Exp^2 + \varepsilon$$

- Interpret the coefficient for **education**

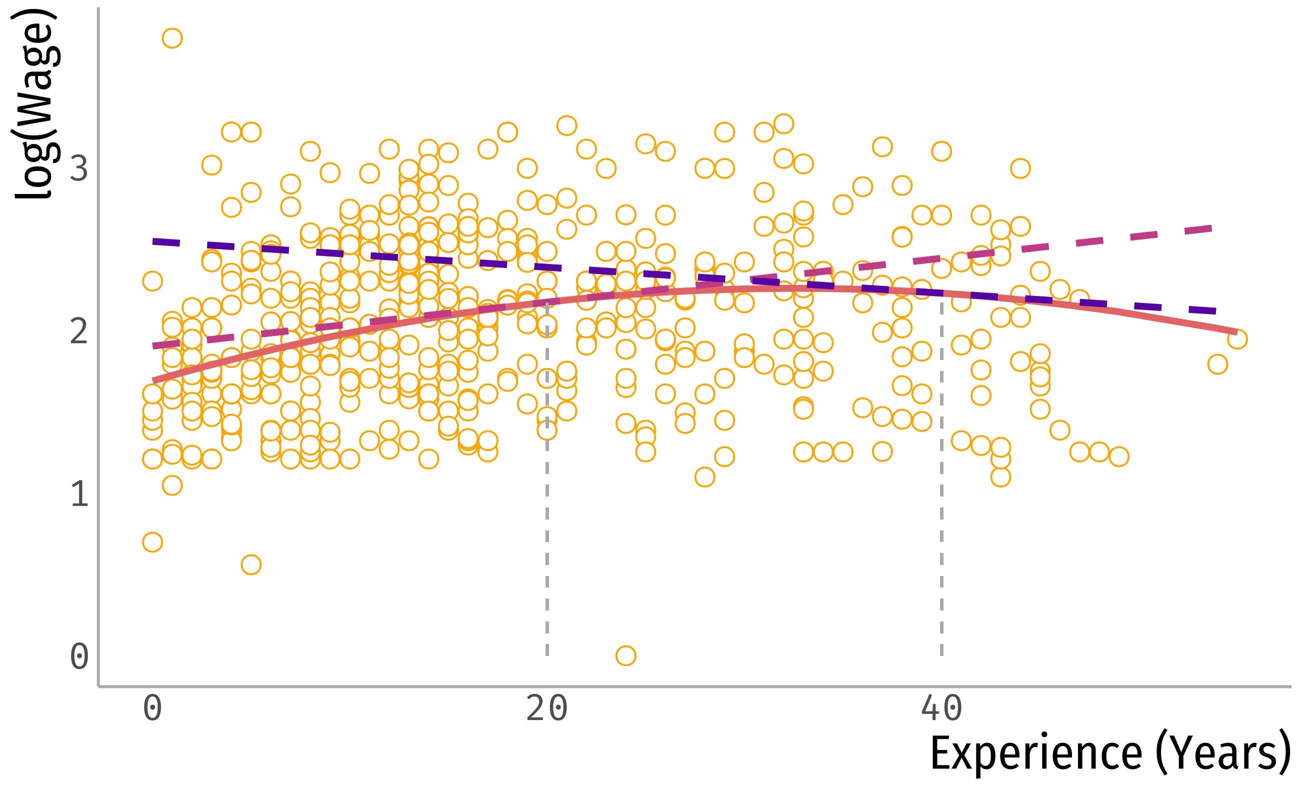
One additional year of education is associated, on average, to $\hat{\beta}_1 \times 100\%$ increase in hourly wages, holding experience constant

- What is the association between experience and wages?

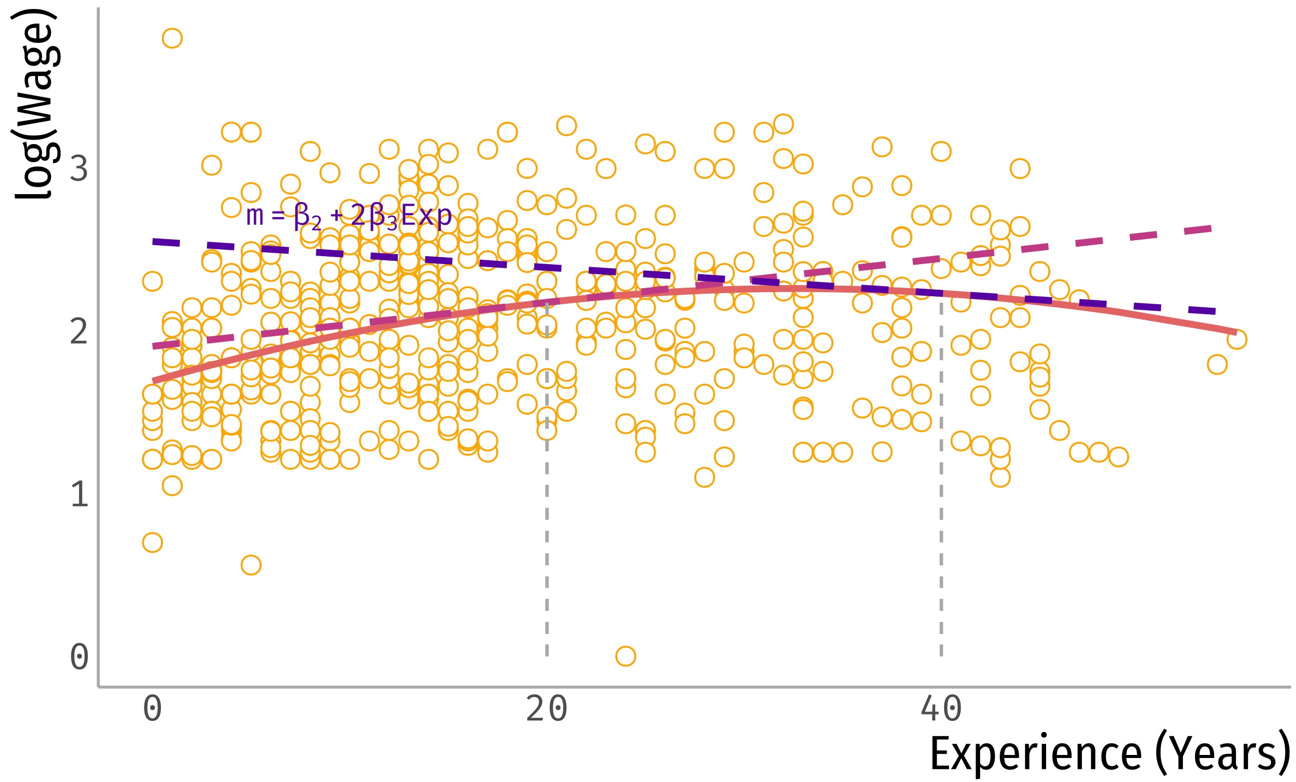
Interpreting coefficients in quadratic equation



Interpreting coefficients in quadratic equation



Interpreting coefficients in quadratic equation



Interpreting coefficients in quadratic equation

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exp + \beta_3 Exp^2 + \varepsilon$$

What is the association between experience and wages?

- Pick a value for Exp_0 (e.g. mean, median, one value of interest)

Increasing work experience from Exp_0 to $Exp_0 + 1$ years is associated, on average, to a $(\hat{\beta}_2 + 2\hat{\beta}_3 \times Exp_0)100\%$ increase on hourly wages, holding education constant

Increasing work experience from 20 to 21 years is associated, on average, to a $(\hat{\beta}_2 + 2\hat{\beta}_3 \times 20)100\%$ increase on hourly wages, holding education constant

Let's put some numbers into it

```r
summary(lm(log(wage) ~ education + experience + I(experience^2), data = CPS1985))
```

```
##
## Call:
## lm(formula = log(wage) ~ education + experience + I(experience^2),
##     data = CPS1985)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12709 -0.31543  0.00671  0.31170  1.98418
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.5203218  0.1236163   4.209 3.01e-05 ***
## education        0.0897561  0.0083205  10.787  < 2e-16 ***
## experience       0.0349403  0.0056492   6.185 1.24e-09 ***
## I(experience^2) -0.0005362  0.0001245  -4.307 1.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4619 on 530 degrees of freedom
## Multiple R-squared:  0.2382,	Adjusted R-squared:  0.2339
## F-statistic: 55.23 on 3 and 530 DF,  p-value: < 2.2e-16
```

- Increasing experience from 20 to 21 years is associated with an average increase in wages of 1.35%, holding education constant.

Main takeaway points

- The model you fit **depends on what you want to analyze**.
- **Plot your data!**
- Make sure you capture associations that **make sense**.

Next week

- Issues with regressions and our data:
  - Outliers?
  - Heteroskedasticity
- Regression models with discrete outcomes:
  - Probability linear models

References

- Ismay, C. & A. Kim. (2021). “Statistical Inference via Data Science”. Chapter 6 & 10.
- Keegan, B. (2018). "The Need for Openess in Data Journalism". *Github Repository*