STA 235H - Multiple Regression: Polynomials

Fall 2023

McCombs School of Business, UT Austin

Some Announcements

- Homework answer key will be posted on Tuesday/Wednesday.
  - Make sure you check it out!
  - Exercises: Multiple regression (e.g. Bechdel Test example), differences in associations between groups (e.g. luxury vs non-luxury cars depreciation).
- Check personalized feedback for JITT 3, if included.
  - Additional videos on material (and some R code) in Resources > Videos

Side note: Difference between percent change and change in percentage points

- Imagine that if you study 4hrs your probability of getting an A is, on average, 70% and if you study for 5hrs that probability increases to 75%.
- Then, we can say that your probability increased by 5 percentage points.
- Why is this not the same as saying that your probability increased by 5%?
- Remember percent change?

$$\frac{y_1 - y_0}{y_0} = \frac{75 - 70}{70} = 0.0714$$

- This means that, in this case, a 5 percentage point increase is equivalent to a 7% increase in probability.

Be aware of the difference in percentage points and percent!

Today

- Roadmap of where we've been and where we're going.
- Nonlinear models:
  - Polynomial terms
- Introduction to Causal Inference
  - Potential Outcomes Framework

Roadmap so far

- Started the class with a review on **simple linear regressions**:
  - Association between a variable X and outcome Y
  - e.g. $Revenue = \beta_0 + \beta_1 Bechdel + \varepsilon$

- Followed by **multiple regression**:
  - *Partial* association between X and Y, when holding other variables constant.
  - e.g. $Revenue = \beta_0 + \beta_1 Bechdel + \beta_2 Revenue + \beta_3 IMDB + \varepsilon$

- What if we want to compare **differences in associations between groups**?:
  - *Compare* the association between X and Y for group $D = 1$ and $D = 0$.
  - e.g. $Price = \beta_0 + \beta_1 Year + \beta_2 Luxury + \beta_3 Year \times Luxury + \varepsilon$

Roadmap so far

- What if our outcome Y is *weird* (e.g. not normally distributed)?
    - *If Y is skewed to the right (log-normal)*: Transform to $log(Y)$ to improve linearity assumption!
    - e.g. $log(Price) = \beta_0 + \beta_1 Year + \beta_2 Luxury + \beta_3 Mileage + \varepsilon$
    - Interpret coefficients as **percent change** (%)

- What if our outcome Y is *weird* (e.g. binary)?
    - e.g. $Employed = \beta_0 + \beta_1 Age + \beta_2 Afam + \beta_3 NKids + \varepsilon$
    - Interpret coefficients as **change in probability** (e.g. percentage points)

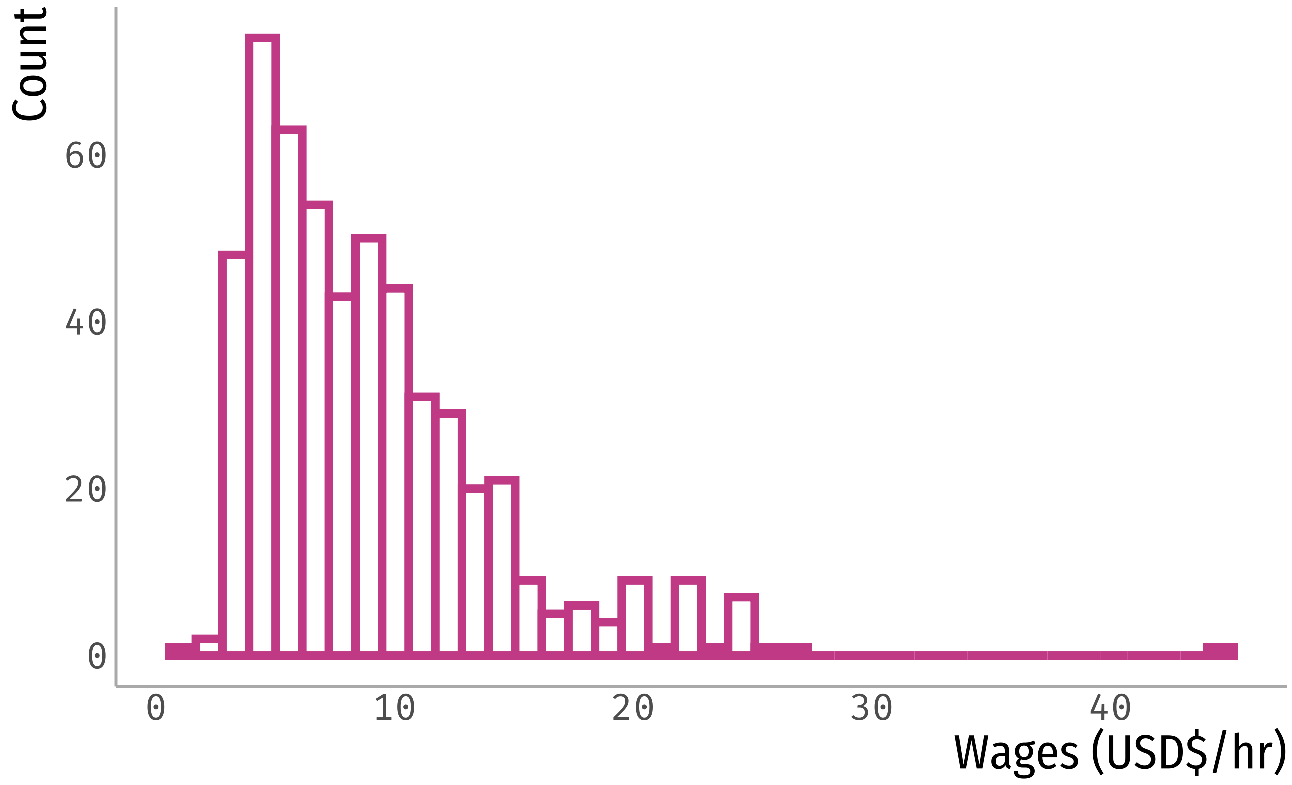
- What if there **isn't a linear relation** between X and Y?
    - Include **polynomial terms** for X

- What if I want to know what is **the effect of X on Y**?
    - Causal Inference!

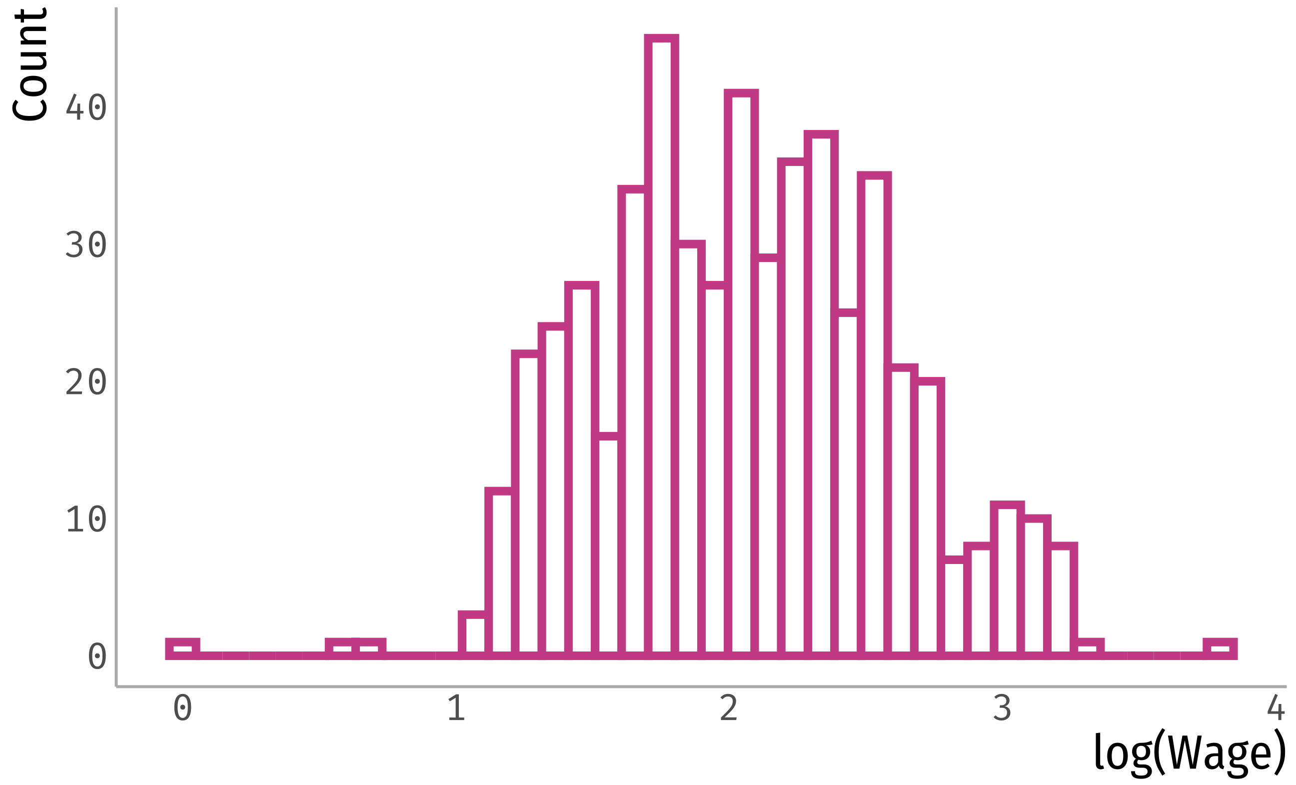
Adding polynomial terms

- 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!

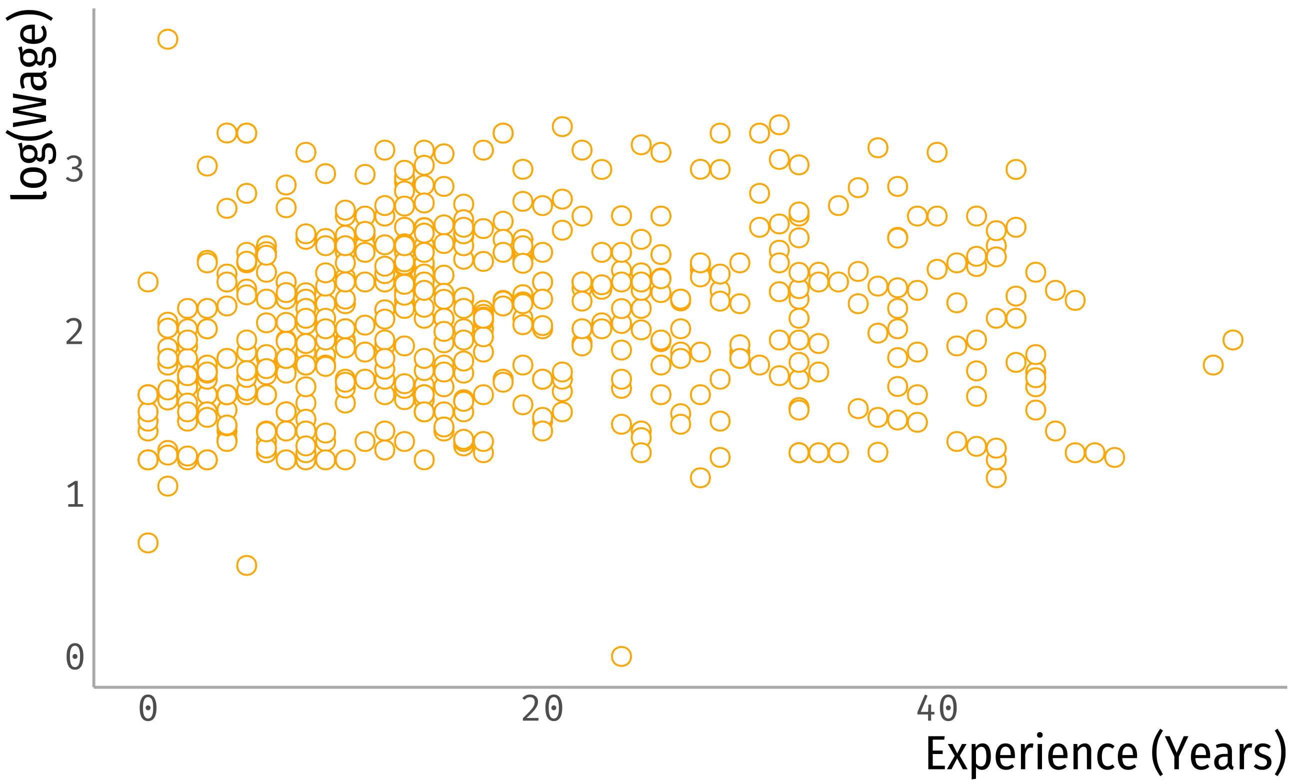
Determinants of wages: CPS 1985



Determinants of wages: CPS 1985

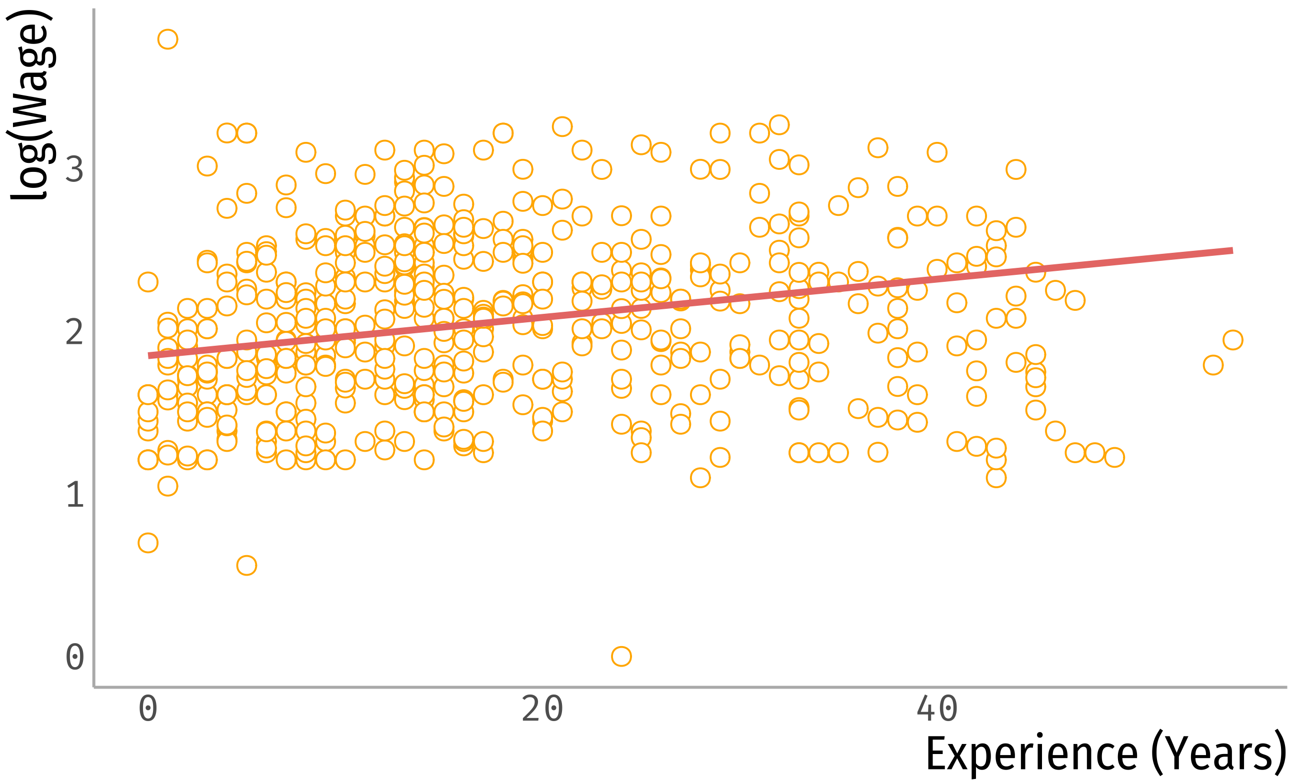


Experience vs wages: CPS 1985



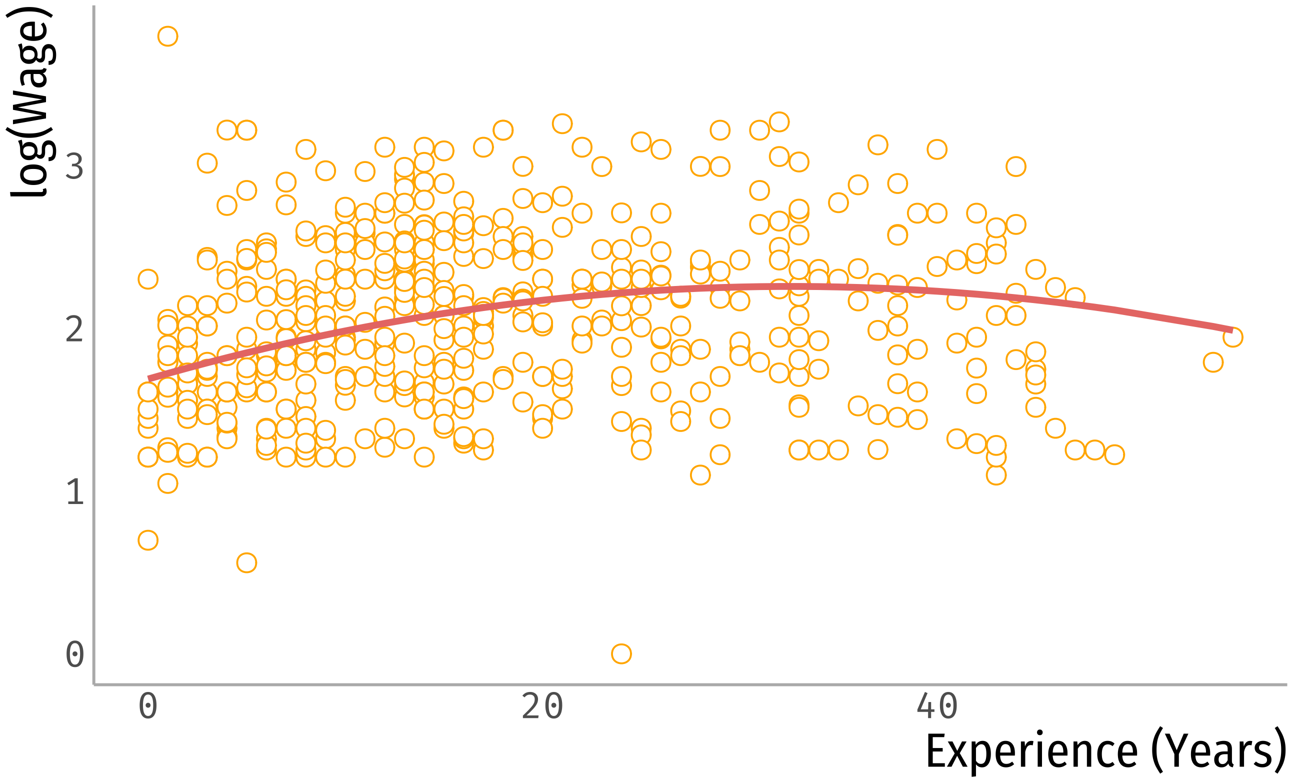
Experience vs wages: CPS 1985

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



Experience vs wages: CPS 1985

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



Mincer equation

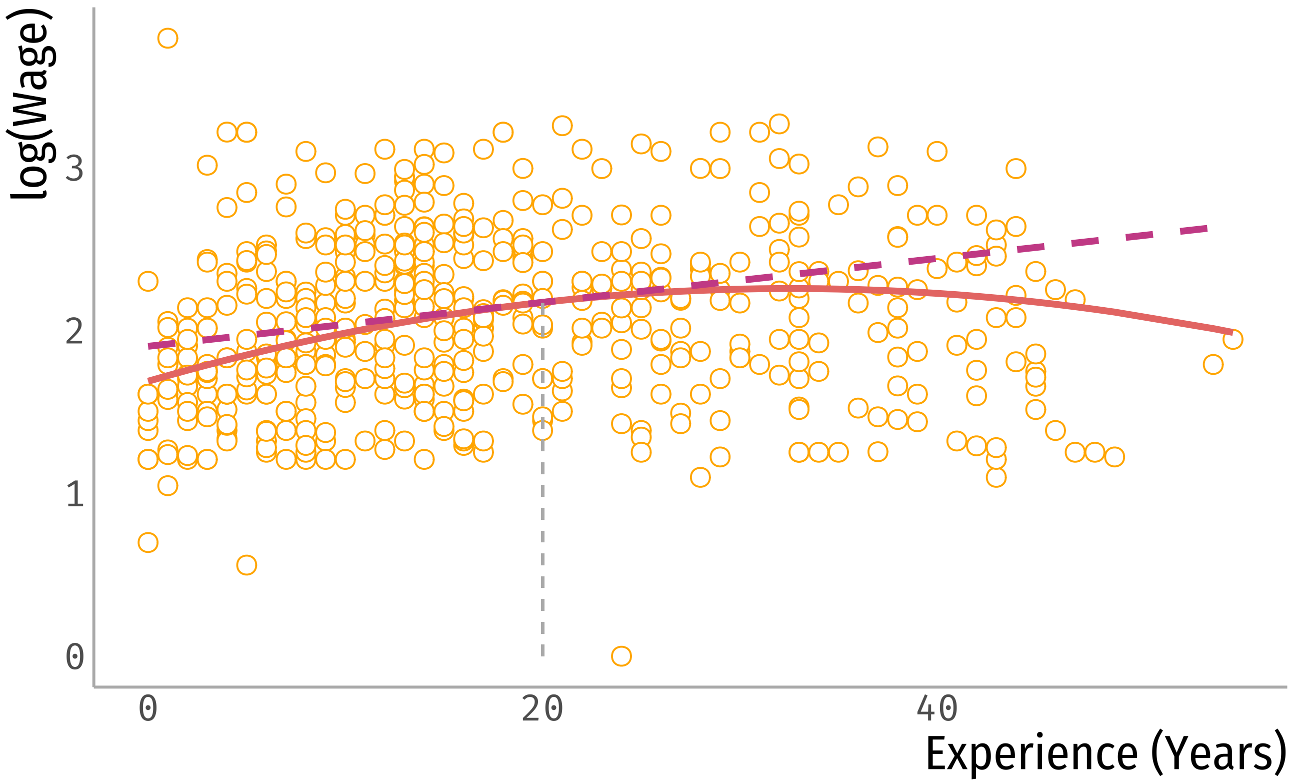
$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exp + \beta_3 Exp^2 + \varepsilon$$

- Interpret the coefficient for **education**

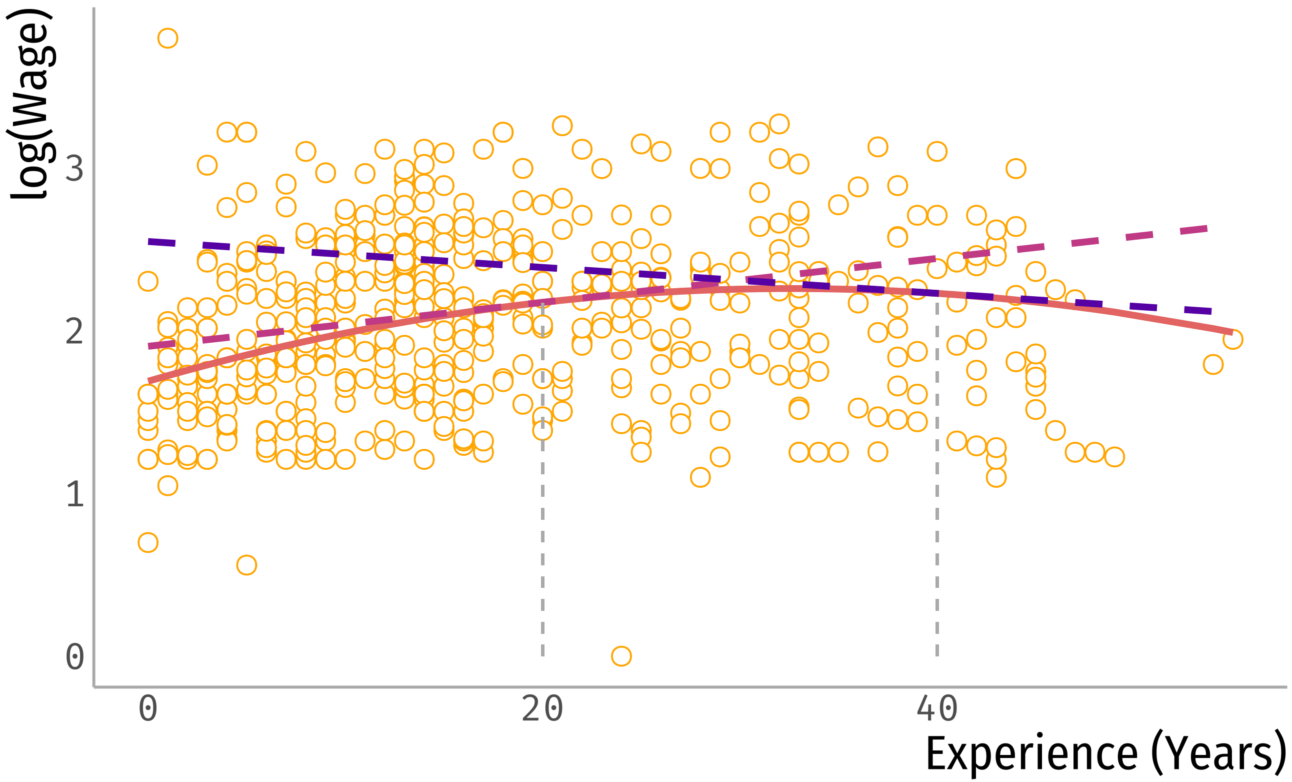
$$\log(Wage) = 0.52 + 0.09 \cdot Educ + 0.034 \cdot Exp - 0.0005 \cdot Exp^2$$

- What is the association between **experience and wages**?

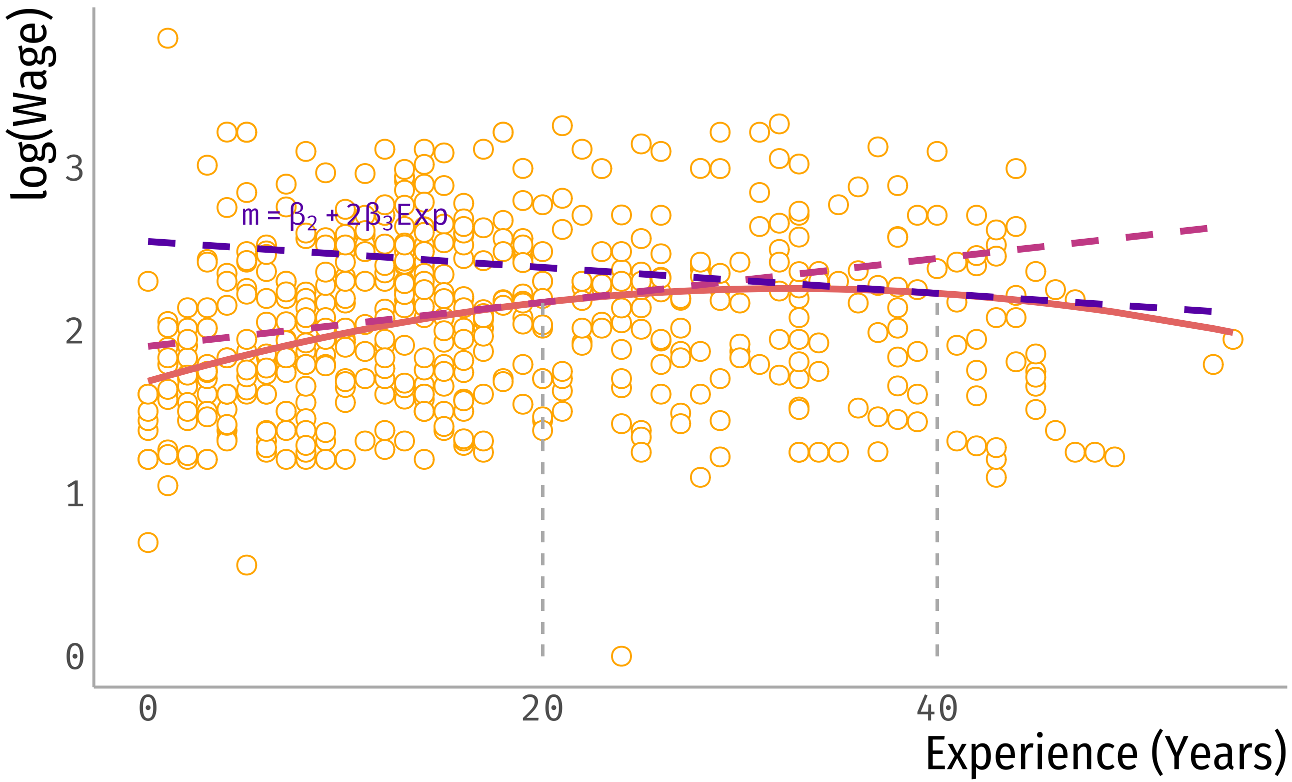
Interpreting coefficients in quadratic equation



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

Let's put some numbers into it:

$$\log(Wage) = 0.52 + 0.09 \cdot Educ + 0.034 \cdot Exp - 0.0005 \cdot Exp^2$$

Increasing work experience from 20 to 21 years is associated, on average, to a $(0.034 - 2 \times 0.0005 \times 20) \times 100 = 1.34\%$ increase on hourly wages, holding education constant

Note that in this case we are interpreting the association between Experience and Wages as a percent change, because Wages is in a logarithm!

Let's go to R

References

- Ismay, C. & A. Kim. (2021). "Statistical Inference via Data Science". Chapter 6 & 10.