# Lecture 5

Multivariate OLS

- Let's return to our bivariate regression world.
- We run $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
- We obtain unbiased estimates if our error term is uncorrelated with x, i.e. if there is not some third variable that is correlated with both

Multivariate OLS

- Consider the case of attendance on student test scores.
- Grades are associated with higher test scores, but we also know that better students tend to show up to class more often
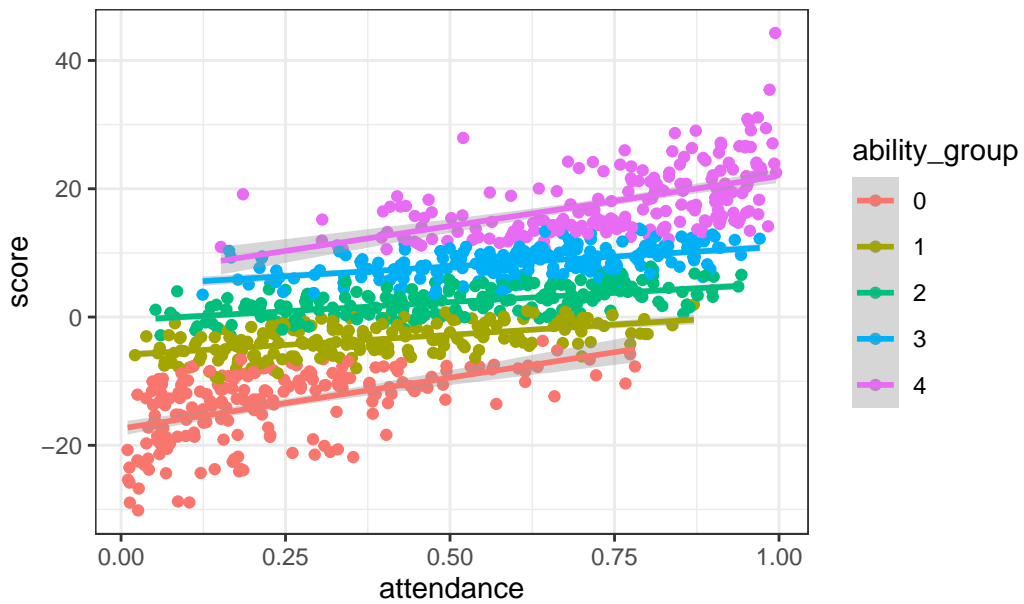- Can we correct for this?

Multivariate OLS - Visual

Attendance on Score – Simulated, Aggregate
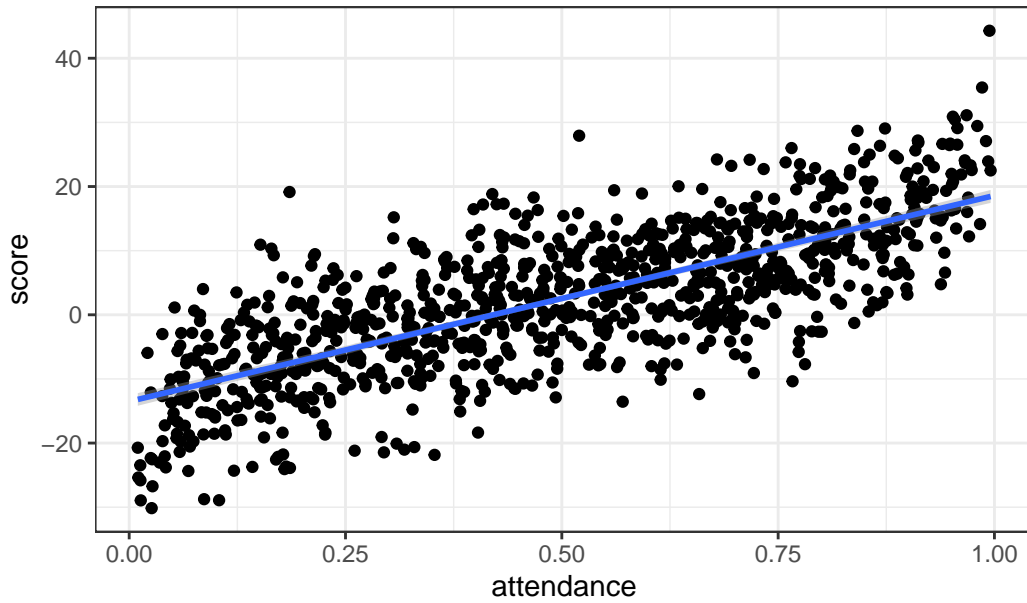
Multivariate OLS - Visual

Attendance on Score – Simulated, Aggregate

Multivariate OLS - formula plus motivation

- For the previous question we can write $grade_i = \beta_0 + \beta_1 attendance_i + \beta_2 baseline_i + \varepsilon_i$
- What is the interpretation of β_1?
  - The change in grade resulting from a 1 unit increase in attendance **holding baseline ability constant**
  - e.g. for two individuals with the same baseline ability, what is the effect of increasing attendance by 1
- This now controls for the previous omitted variable

iClicker

You have two classes of students: gifted students (ability=1) and non-gifted students (ability=0), and students who attend class (attend=1) and those who don't (attend=0). The average scores for these 4 classes of students are displayed in the table below. You want to estimate the following models. Calculate $\hat{\beta}_1$

(1) $score_i = \beta_0 + \beta_1 attend_i + \varepsilon_i$

(2) $score_i = \gamma_0 + \gamma_1 attend_i + \gamma_2 ability_i + \eta_i$

	attend	ability	score	N Students
1:	0	0	40	15
2:	0	1	80	5
3:	1	0	60	5
4:	1	1	100	15

iClicker

You estimate two models. Under what condition will $E[\hat{\beta}_1] = E[\hat{\gamma}_1]$

(1) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

(2) $y_i = \gamma_0 + \gamma_1 x_i + \gamma_2 z_i + \eta_i$

- A When x and z are correlated
- B When z and y are correlated
- C When z is correlated with both x and y
- D When z is uncorrelated with either x or y

Administrative Miscellanea

- Homework 6 Due Friday
- Problem set 2 graded - you can resubmit by this Friday midnight to make up for lost points
    - If you haven't submitted, this Friday at midnight is the final deadline
    - It's worth 80 points (end of class total of 1000)
- Contine Multivariate OLS today
- Quiz 4 Wednesday (multivariate OLS)
- I will announce when problem set 3 is available/due
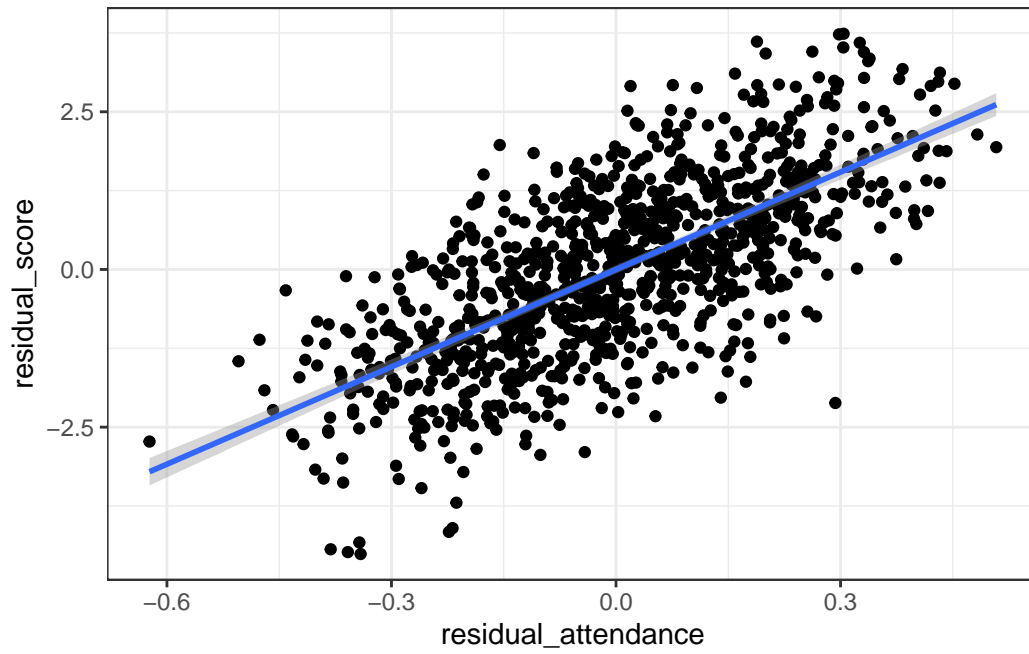
Multivariate OLS - Intuition

- We can make a valid comparison by only comparing students with the same test scores.
    - e.g. for all students who received a score of 90%, we regress grade on attendance
- This approach is really inefficient: instead we can just "net out" the effect of baseline ability. How to do this?

Multivariate OLS - FWL

- Run a regression of grade on baseline performance to get a predicted grade from baseline
- Then, regress attendance on baseline ability to get a predicted attendance from baseline
- Finally, take the residuals from these two and run that regression.
    - The residuals are the components of grade and attendance that are **NOT** correlated with ability. We've just controlled for this
- This is equivalent to what is obtained for β_1 in the multivariate regression $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Multivariate OLS: net effect


```
Call:
lm(formula = score ~ attendance + ability, data = dt)

Coefficients:
(Intercept)    attendance        ability
   -0.05624       5.14514       10.02367
```
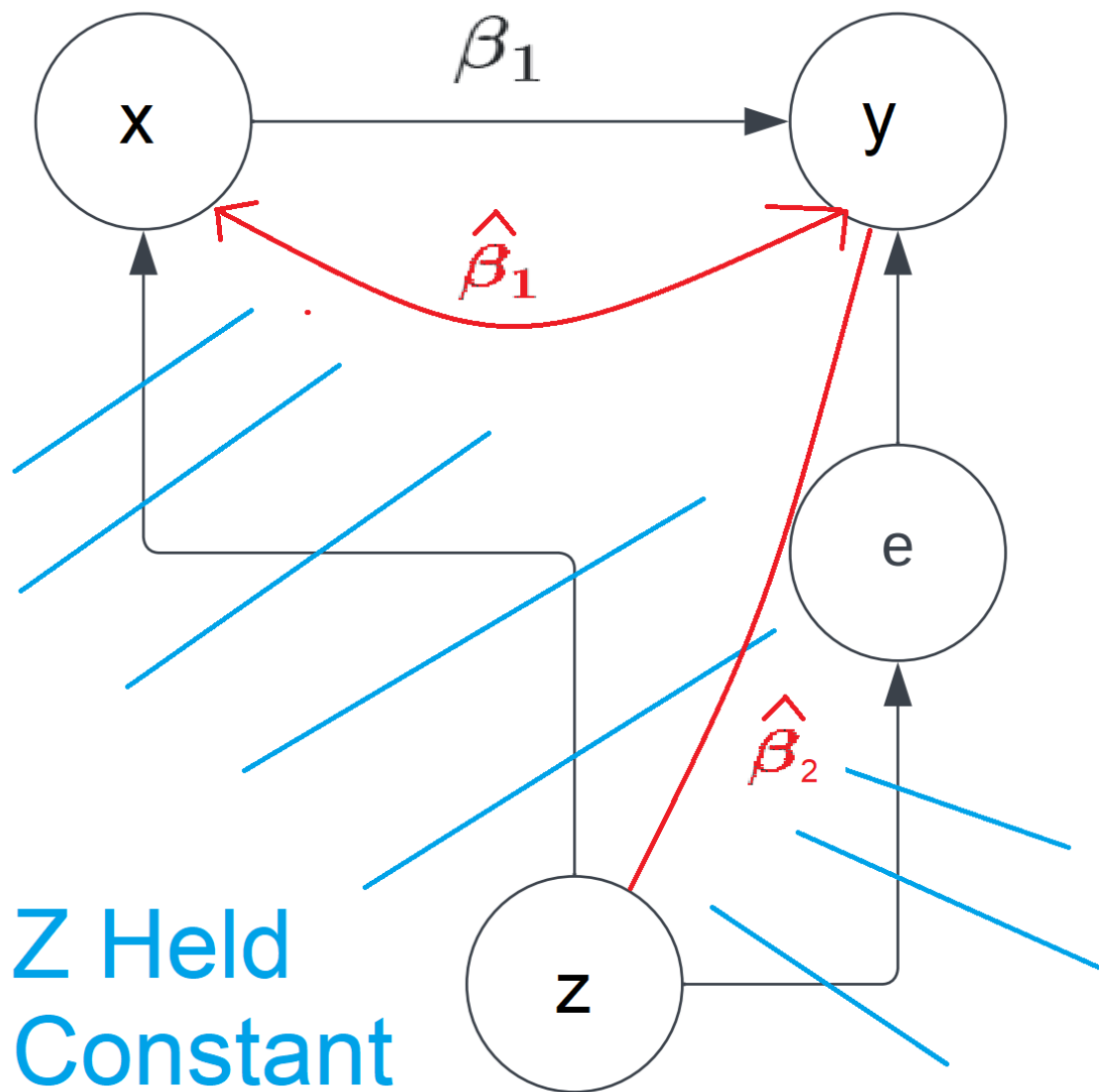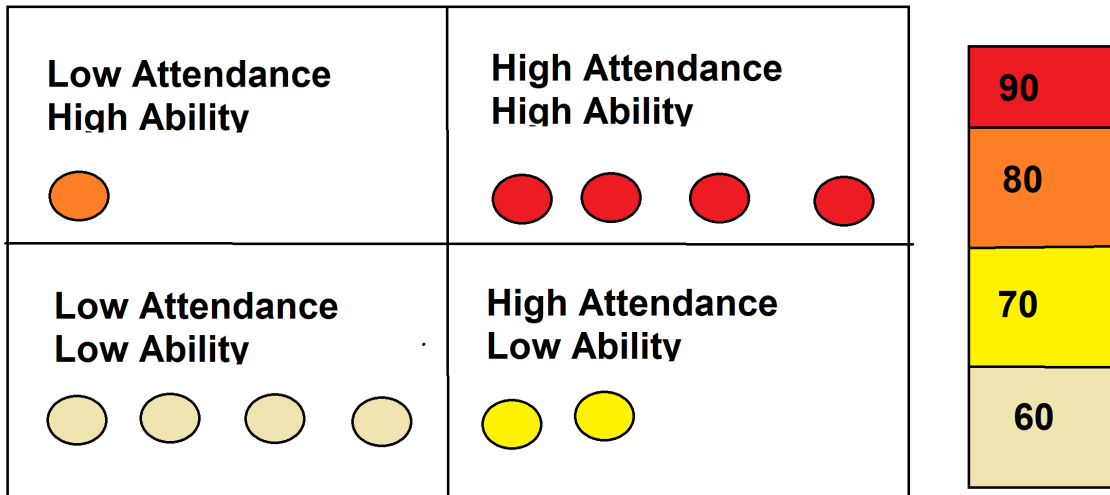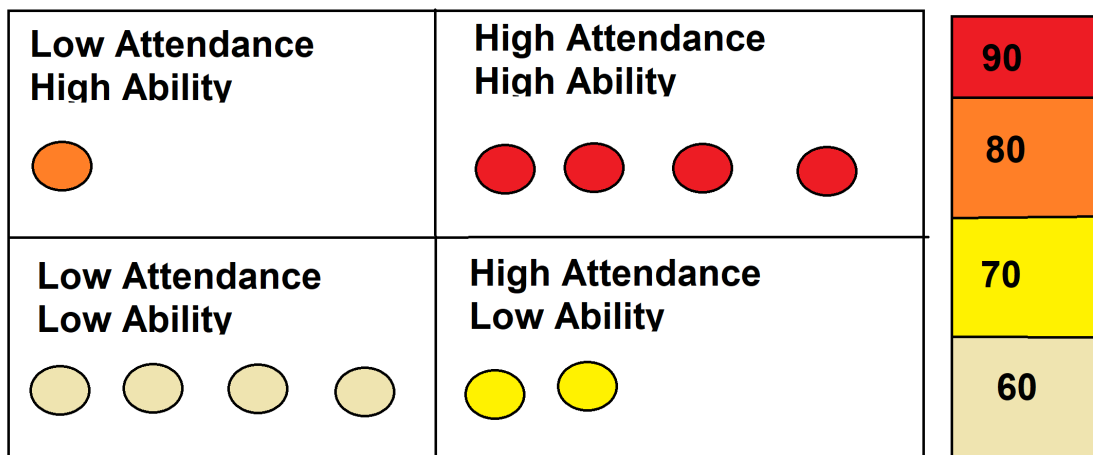
$$\beta_1$$

x → y

$$\widehat{\beta}_1$$

e

$$\widehat{\beta}_2$$

Z Held Constant

z

Multivariate OLS - Diagram

Low Attendance High Ability	High Attendance High Ability		90
🟠	🔴 🔴 🔴 🔴		80
Low Attendance Low Ability	High Attendance Low Ability		70
⬜ ⬜ ⬜ ⬜	🟡 🟡		60

Multivariate OLS - Diagram

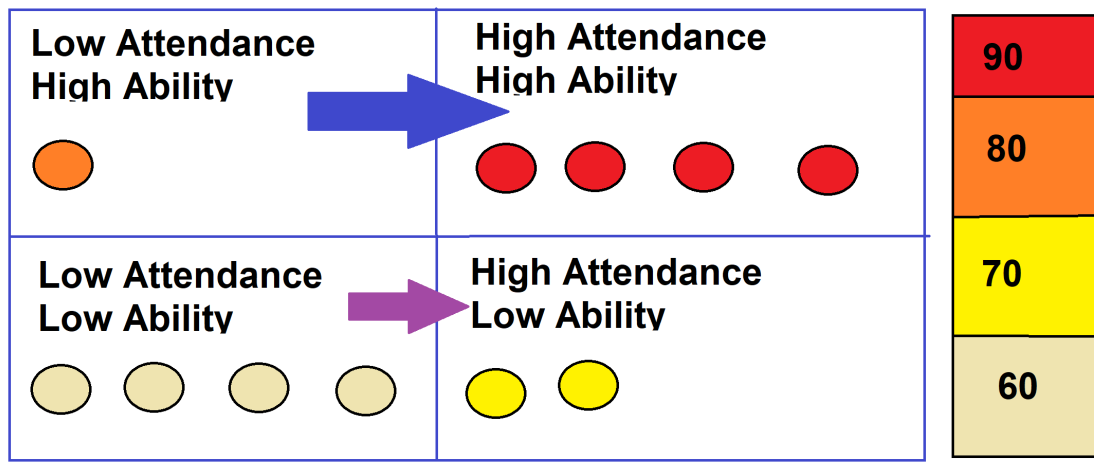| Low Attendance High Ability	High Attendance High Ability		90
🟠	🔴 🔴 🔴 🔴		80
Low Attendance Low Ability	High Attendance Low Ability		70
⬜ ⬜ ⬜ ⬜	🟡 🟡		60

Low Attendance average: 64
High Attendance average: 83
OLS Estimate: ~19

Multivariate OLS - Diagram

For high ability: high attendance has +10 avg score
For low ability: high attendance has +10 avg score
Multivariate OLS Estimate: 10

Multivariate OLS Diagram: confirmation Bivariate

```
Call:
lm(formula = grade ~ attendance, data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-13.333  -4.000  -4.000   6.667  16.000

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)   64.000      4.355  14.697 0.000000135 ***
attendance    19.333      5.896   3.279     0.00955 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.737 on 9 degrees of freedom
Multiple R-squared:  0.5443,    Adjusted R-squared:  0.4937
```

```
F-statistic: 10.75 on 1 and 9 DF,  p-value: 0.009545
```

Multivariate OLS Diagram: confirmation Multivariate

```
Call:
lm(formula = grade ~ attendance + ability, data = dt)

Residuals:
               Min                       1Q                  Median
-0.000000000000022812 -0.000000000000001200  0.000000000000002400
                 3Q                      Max
 0.000000000000003902  0.000000000000007513

Coefficients:
                       Estimate           Std. Error           t value
(Intercept) 59.999999999999985789  0.000000000000004320  13890374101471608
attendance  10.000000000000001776  0.000000000000006323   1581488661539333
ability     20.000000000000003553  0.000000000000006323   3162977323078666
                    Pr(>|t|)
(Intercept) <0.0000000000000002 ***
attendance  <0.0000000000000002 ***
ability     <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.000000000000009236 on 8 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 1.098e+31 on 2 and 8 DF,  p-value: < 0.00000000000000022
```

Multivariate OLS Diagram: confirmation FWL

```
   attendance ability grade residgrade residattendance
1:          0       0    60  -3.333333      -0.3333333
2:          0       1    80  -8.000000      -0.8000000
3:          1       0    70   6.666667       0.6666667
4:          1       1    90   2.000000       0.2000000
```

Multivariate OLS Diagram: Confirmation FWL

```
Call:
lm(formula = residgrade ~ residattendance, data = dt)

Residuals:
                Min                      1Q                 Median
-0.000000000000022662 -0.000000000000001297   0.000000000000002301
                 3Q                     Max
 0.000000000000004978   0.000000000000006645

Coefficients:
                          Estimate              Std. Error
(Intercept)     -0.0000000000000003013   0.0000000000000026033
residattendance 10.0000000000000035527   0.0000000000000059114
                          t value              Pr(>|t|)
(Intercept)                -0.116                  0.91
residattendance 1691660616295972.000  <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.000000000000008634 on 9 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 2.862e+30 on 1 and 9 DF,  p-value: < 0.00000000000000022
```

What if we have multiple omitted variables

- If we run $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ we could still have other variables that are correlated with both x_1 and y, conditional on x_2
- We can just add every omitted variable in our regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$
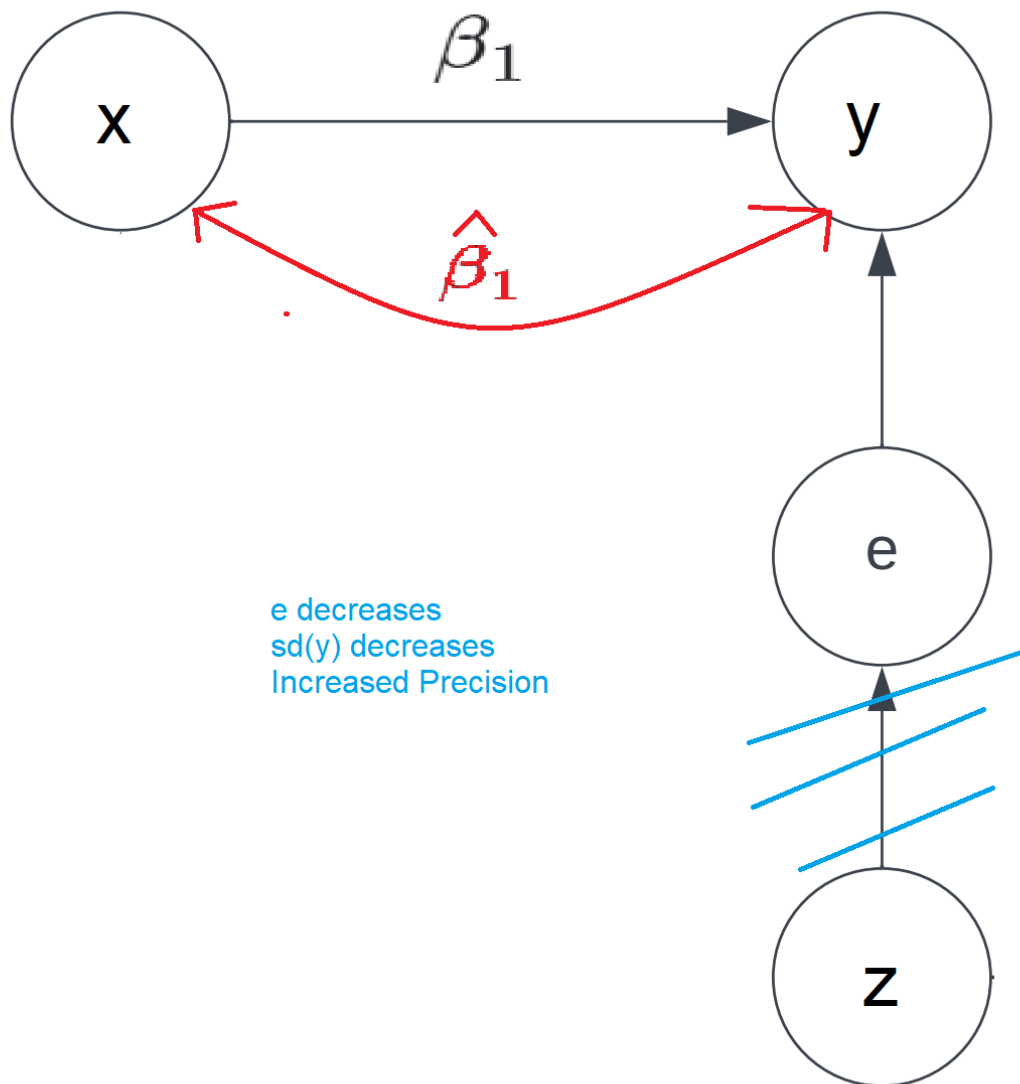- Are there issues with this?

What if we have multiple omitted variables

- If we don't observe a variable, we can't control for it.
- Example: ability. This is called selection on observables vs selection on un-observables
- If we control for an irrelevant variable, does it bias $\hat{\beta}_1$? e.g. if we add control for the day of the month the student was born?
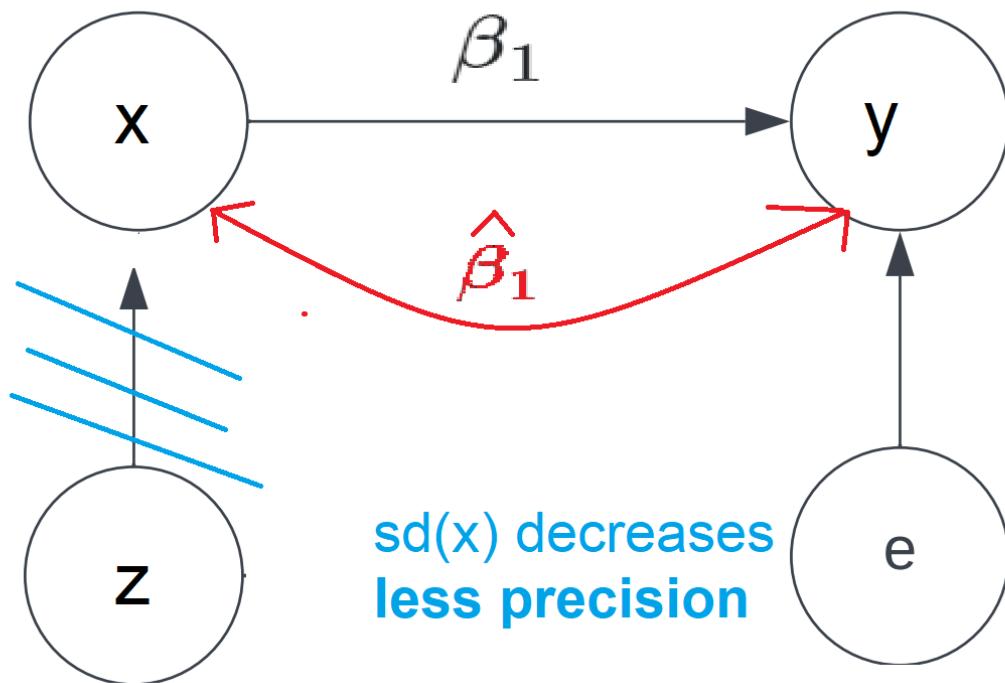- No, it just changes our standard errors

What if we have multiple omitted variables

- Are there controls that can bias our estimate
- Yes. It's complicated.
- The short answer is that we don't want to control for any intermediate pathways (e.g. if x causes z, and z causes y, we don't control for y)
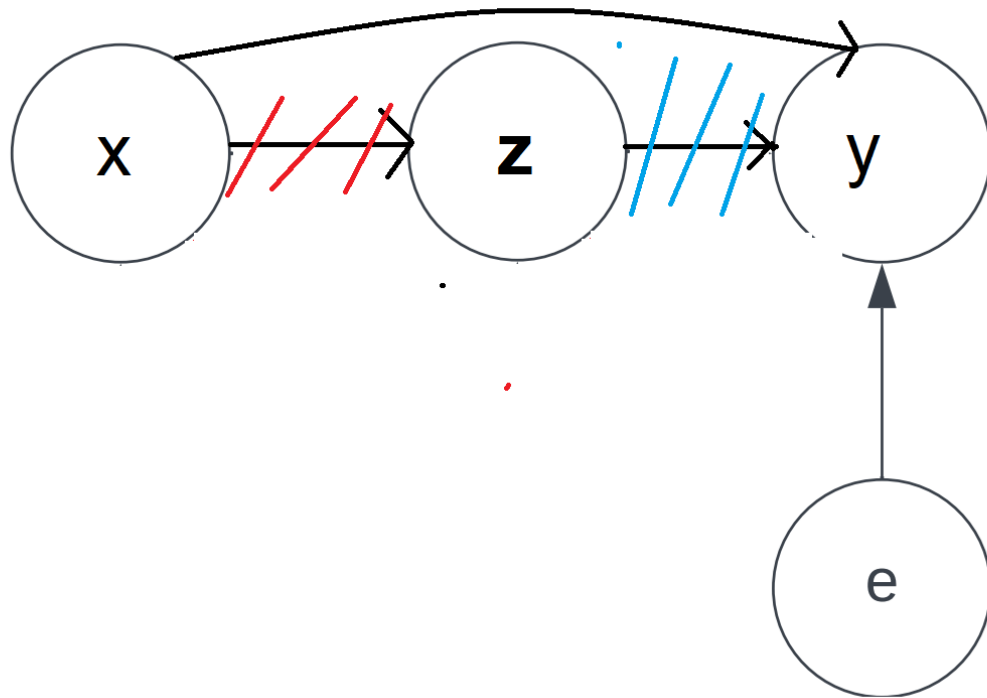
Different Controls

β_1

x   y

$\widehat{\beta_1}$

e

e decreases
sd(y) decreases
Increased Precision

z

Different Controls

$$\beta_1$$

x → y

$$\widehat{\beta_1}$$

z

sd(x) decreases
less precision

e

13

Different Controls

iClicker

You are interested in knowing whether teachers with a master's degree are more effective than teachers with a bachelor's degree. You run the regression $score_{it} = \beta_0 + \beta_1 degree_{jt} + \beta_2 score_{i,t-1} + \varepsilon_{ijt}$ where $score_{it}$ is student i's standardized test score, $degree_{jt}$ is 1 if teacher

j has a master's degree, and $score_{i,t-1}$ is student i's standardized test score last year. What additional controls should you add to this model?

iClicker

$score_{it} = \beta_0 + \beta_1 degree_{jt} + \beta_2 score_{i,t-1} + \varepsilon_{ijt}$

What additional controls should you add to this model?

- A Student gender
- B Student race
- C Student attendance
- D Teacher experience
- E Class Size

An Example: Twitter Opinion Predicts Corporate Earnings?

TABLE 4

Earnings Surprises and Twitter Opinion

Panel A: All Tweets

$$ESURP = \alpha + \beta_1*OPI + \beta_2*PRIOR\_ESURP + \beta_3*EXRET_{[-10;-2]} + \beta_4*RP\_OPI + \beta_5*SIZE + \beta_6*MB + \beta_7*ANL$$
$$+ \beta_8*INST + \beta_9*Q4 + \beta_{10}*LOSS + \varepsilon$$

	Coefficient (t-statistic)			
	ESURP = SUE		**ESURP = FE**	
	BAYES	*VOCAB*	*BAYES*	*VOCAB*
Variable	**Model I**	**Model II**	**Model III**	**Model IV**
Intercept	−1.0572***	−1.3467***	−0.3084***	−0.3256***
	(−11.03)	(−13.23)	(−5.57)	(−5.66)
OPI	**0.0275***	**0.3138***	**0.0075**	**0.0193**
	(2.84)	**(13.98)**	**(2.53)**	**(2.40)**
PRIOR_ESURP	0.2831***	0.2801***	0.1871***	0.1871***
	(40.47)	(42.91)	(10.98)	(11.14)
EXRET$_{[-10;-2]}$	0.0219***	0.0213***	0.0084***	0.0085***
	(6.40)	(6.66)	(4.60)	(4.42)
RP_OPI	0.3545***	0.3465***	0.0582***	0.0585***
	(5.27)	(5.27)	(2.68)	(2.72)
SIZE	0.1114***	0.1448***	0.0242***	0.0260***
	(7.59)	(9.39)	(3.17)	(3.32)
MB	−0.0112**	−0.0045	0.0018	0.0023
	(−2.10)	(−0.87)	(0.85)	(1.04)
ANL	−0.1111***	−0.0709**	0.0099	0.0123
	(−3.69)	(−2.41)	(0.53)	(0.65)
INST	0.1928**	0.0992	0.1975***	0.1918***
	(2.51)	(1.31)	(4.79)	(4.73)
Q4	0.2968***	0.3437***	−0.0924***	−0.0892***
	(6.23)	(6.99)	(−5.02)	(−4.85)
LOSS	0.8547***	0.8801***	−0.0577**	−0.0570**
	(18.21)	(18.96)	(−2.27)	(−2.18)
n	30,815	30,815	28,784	28,784
Adj. R^2 (%)	9.02	9.60	5.18	5.18

Worked Example: Alcohol and Mortality

- You are measuring the effects of alcohol consumption on life expectancy. Your naive regression is $life_i = \beta_0 + \beta_1 drinks_i + \varepsilon_i$. When you run this regression you obtain $\hat{\beta}_1 = 1$, but you know there are omitted variables like social networks that can bias this results. To fix this, you now use a multivariate regression controlling for social status (measured as number of close friends), marital status, age, BMI, and self reported health status

- 1) Write the estimating equation for this multivariate OLS model

Worked Example: Alcohol and Mortality

- $life_i = \beta_0 + \beta_1 drinks_i + \varepsilon_i$, $\hat{\beta}_1 = 1$. you now use a multivariate regression controlling for social status (measured as number of close friends), marital status, age, BMI, and self reported health status

- 2) You obtain a value of $\hat{\beta}_1 = -0.5$ in the multivariate model. Interpret $\hat{\beta}_1$ in the context of the research question

Worked Example: Alcohol and Mortality

- $life_i = \beta_0 + \beta_1 drinks_i + \varepsilon_i$, $\hat{\beta}_1 = 1$. you now use a multivariate regression controlling for social status (measured as number of close friends), marital status, age, BMI, and self reported health status

- 3) Which estimate of $\hat{\beta}_1$ is a better estimate of the causal effect of alcohol consumption on life expectancy? Is $\hat{\beta}_1 = -0.5$ unbiased? Why or why not?

Worked Example: Alcohol and Mortality

- $life_i = \beta_0 + \beta_1 drinks_i + \varepsilon_i$, $\hat{\beta}_1 = 1$. you now use a multivariate regression controlling for social status (measured as number of close friends), marital status, age, BMI, and self reported health status

- 4) Which regression has a higher value of R^2? Why?

Multiple hypothesis testing

- In a multivariate setting nothing changes when testing the hypothesis of $\beta_1 = 0$
- We can also do a joint test to see if every variable is equal to 0: $H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0$

Multiple hypothesis testing

- The test statistic is called an F-statistic. The actual calculations and tables are burdensome, but R will calculate this for us
- The p-value is also obtained from R, once we have the p-value the hypothesis test is exactly identical
- We obtain both estimates for individual $\hat{\beta}$ estimates and for our full model

- In econometrics we usually care about the former, while predictive analytics cares about the latter

iClicker

We are interested in studying the effect of getting a felony conviction on that individual's unemployment rate. We control for the individual's age, sex, race, parental salary, state, educational attainment, and IQ score. Is our estimate of $\hat{\beta}_1$ likely to be unbiased?

- A Yes
- B No

iClicker

We are interested in studying the effect of being put on academic probation in college on completing a bachelor's degree on time (within 6 years). You are placed on academic probation if your GPA falls below 2.0. For this reason, we control for GPA in our regression. Is our estimate of $\hat{\beta}_1$ likely to be unbiased?

- A Yes
- B No

Dummy variables

- We now have almost everything we need for before getting into experimental designs, but so far we have only dealt with numeric data
- Recall that we can also have ordinal (categorical, non-numeric) data, e.g. gender.
- To study these we can convert them to a numeric value using a dummy variable.
- These are the building blocks of how we handle any ordinal data

Gender as a dummy variable

- Suppose we want to know the effect of gender on earnings. For now set aside causality and run the bivariate OLS regression $wage_i = \beta_0 + \beta_1 gender_i + \varepsilon_i$
- What values should we put for gender?
- We can encode (arbitrarily) male=0, female=1
- What is a a "1 unit increase" in gender?

Gender as a dummy variable

- $wage_i = \beta_0 + \beta_1 gender_i + \varepsilon_i$
- Suppose we obtain $\hat{\beta}_0 = 18$, $\hat{\beta}_1 = -3$. Interpret $\hat{\beta}_0, \hat{\beta}_1$
    - $\hat{\beta}_0$ is the average value of wage when $gender = 0$. But $gender = 0$ means for a male, ie the average male makes \$18/hour
    - $\hat{\beta}_1 = -3$ means that when $gender = 1$, $wage = 18 - 3 * 1 = 15$, ie the average wage for a female is \$15/hour. We can also directly interpret this as women earn \$3/hour less, on average
    - Note that we don't use causal terms

Gender as a dummy variable

- What if instead of encoding male as 0, we did male=1,female=0?

Fitting Dummy Variables Graphically
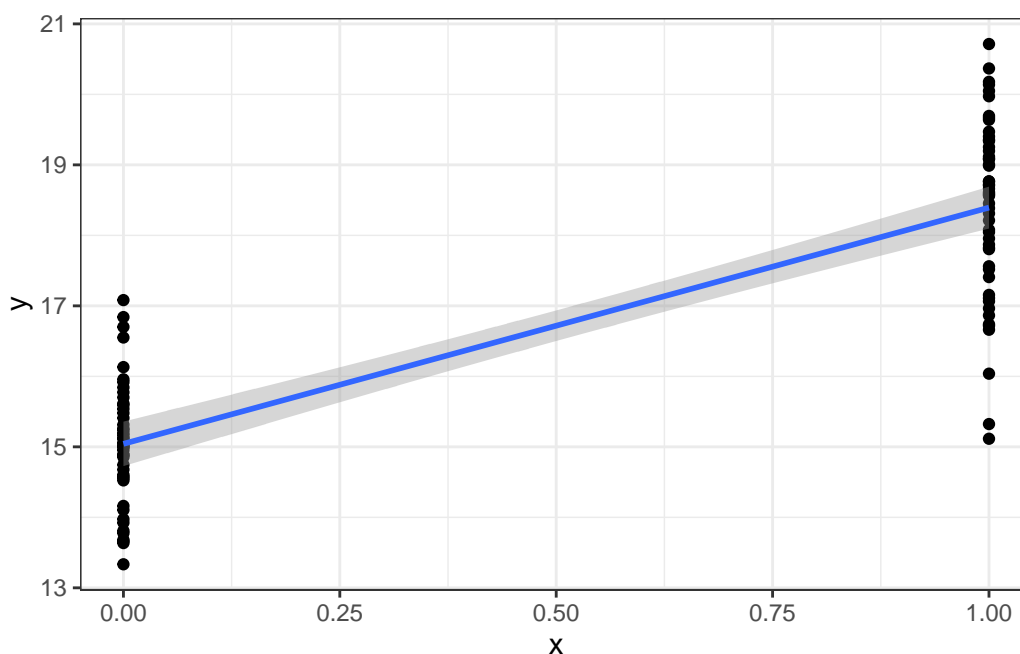
Dummy variable calculation

- $wage_i = \beta_0 + \beta_1 gender_i + \varepsilon_i$
- Calculate $\hat{\beta}_1$

```
    gender college_degree wage   N
1:      0                0   10 100
2:      0                1   30 100
3:      1                0   10 100
4:      1                1   20 100
```

Dummy variable calculation

- $wage_i = \beta_0 + \beta_1 gender_i + \beta_2 degree_i + \varepsilon_i$
- Calculate $\hat{\beta}_1$

```
    gender college_degree wage   N
1:      0                0   10 100
2:      0                1   30 100
3:      1                0   10 100
4:      1                1   20 100
```

A note on dummy variables

- Note that when we fit a dummy variable, we are just comparing the average value of y for our two groups.
    - In some sense this makes them much easier than numeric variables

Dummies with controls

- Suppose we run the same regression of wage on gender, but now we control for age, ie $wage_i = \beta_0 + \beta_1 gender_i + \beta_2 age_i$
- What does $\hat{\beta}_1$ represent?
    - Still the mean wage of females minus the mean wage of males, but now conditional on age
- Under what conditions do we expect our age control to change $\hat{\beta}_1$?
    - The average age of men and women in the workforce needs to differ

Controls as dummies

- Suppose instead we were interested in the effect of wage on age, but we're now controlling for gender: $wage_i = \beta_0 + \beta_1 wage_i + \beta_2 gender_i$
- Nothing has changed from our multivariate OLS, but now we're literally just subtracting out the average wage by gender before doing our regression
  - This is called a fixed effect: we're removing all variation from gender. Before we were only taking out the linear component

Ordinal variables as dummies

- Suppose instead of gender we have education: we have high school dropouts, high school graduates, and college graduates. How can we model this?
- Use multiple dummies: 0/1 for HS dropout, 0/1 for HS graduate, and 0/1 for college graduate. Are there any issues with this?

Ordinal variables as dummies

- Once we know the value of the first two, we know the value of the third - these are multicollinear
- We always drop 1 dummy variable. We can do this arbitrarily
- Note that For gender we can have a dummy for male or a dummy for female. The only difference is interpretation.
- The level left out is the reference level

Ordinal Variable Example

- We wish to regress wage on education
- Education = 0 if HS dropout, 1 if HS graduate, 2 if some college, 3 if college degree
- How do we write this regression equation

Oridinal Variable Example

```
Call:
lm(formula = wage ~ educ, data = dt)

Coefficients:
(Intercept)          educ
```
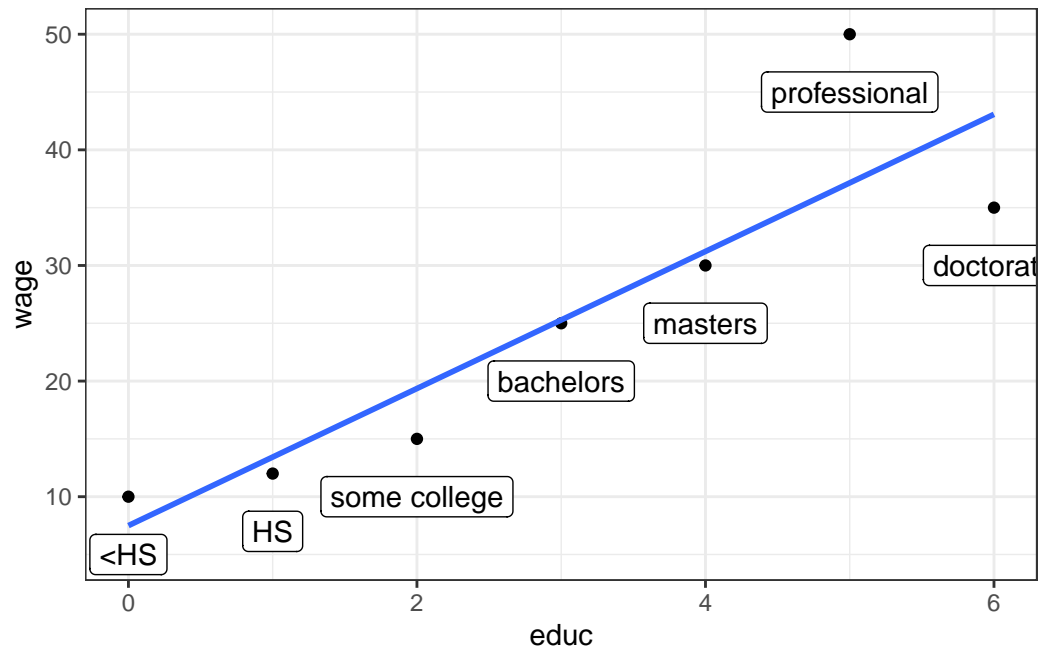
```
       7.500           5.929
```

Oridinal Variable Example

Oridinal Variable Example

```
Call:
lm(formula = wage ~ label, data = dt)

Coefficients:
      (Intercept)        labelbachelors        labeldoctorate              labelHS
               10                    15                    25                    2
    labelmasters  labelprofessional  labelsome college
               20                    40                    5
```

Oridinal Variable Example

```
Call:
lm(formula = wage ~ label, data = dt)
```

```
Coefficients:
    (Intercept)            label<HS      labelbachelors      labeldoctorate
             30                 -20                  -5                   5
        labelHS  labelprofessional  labelsome college
            -18                 20                 -15
```

Dummies as the dependent variable

- We can have a dummy variable as our y variable instead.
- We are interested in knowing the effect of education on employment and run $employment_i = \beta_0 + \beta_1 educ_i + \varepsilon_i$, where $educ_i$ is the number of years of education obtained
- We obtain $\hat{\beta}_1 = .05$ How do we interpret this?

Dummies as the dependent variable

- Linear probability model: every year of education increases our probability of being employed by 5%

Interaction variables

- Suppose we have both gender and education (for simplicity: an indicator for having a bachelor's degree) and are interested in wages.

- We can run $wage_i = \beta_0 + \beta_1 gender_i + \beta_2 education_i$
- What if we think the effect of education is different for males and females?

Interaction variables

- We can interact these variables by multiplying them: $wage_i = \beta_0 + \beta_1 gender_i + \beta_2 education_i + \beta_3 gender_i * education_i + \varepsilon_i$

- We obtain $\hat{\beta}_1 = 1, \hat{\beta}_2 = 2, \hat{\beta}_3 = -1$. How do we interpret these?

- You need to think through all combinations (drop the error term for now and interpret these as averages for notation simplicity):

Interaction variables

- $gender = 0, education = 0 \implies wage = \hat{\beta}_0$
  - $\hat{\beta}_0$ is the average wage for uneducated women
- $gender = 1, education = 0 \implies wage = \hat{\beta}_0 + \hat{\beta}_1$
  - $\hat{\beta}_0 + \hat{\beta}_1$ is the average wage for uneducated men
  - $\hat{\beta}_1$ is the wage differential for uneducated men vs uneducated women

Interaction variables

- $gender = 0, education = 1 \implies wage = \hat{\beta}_0 + \hat{\beta}_2$
  - $\hat{\beta}_0 + \hat{\beta}_2$ is the average wage for educated women
  - $\hat{\beta}_2$ is the wage differential for educated women vs uneducated women
- $gender = 1, education = 1 \implies wage = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$
  - $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$ is the average wage for educated men
  - $\hat{\beta}_3$ is the wage differential for educated men vs uneducated men

Interaction variables

- In other words, we obtain two different effects: the effect of education on earnings for both men and women separately

Interaction variables: Some algebra

- Before we had $wage_i = \beta_0 + \beta_1 gender_i + \beta_2 education_i + \beta_3 gender_i education_i$
- If we want the overall effect of gender we can factor this out
- $wage_i = \beta_0 + (\beta_1 + \beta_3 education_i) gender_i + \beta_2 education_i$
- ie the males earn $\beta_1 + \beta_3 education_i$ more per hour.
- This means uneducated males earn β_1 more, while educated males make $\beta_1 + \beta_3$ more, on average

Interaction variables: Some algebra

- If we want the overall effect of education we can factor it differently
- $wage_i = \beta_0 + \beta_1 gender_i + (\beta_2 + \beta_3 gender_i) education_i$
- ie educated individuals earn $\beta_2 + \beta_3 gender_i$ more, on average
- This means educated women earn β_2 more while educated men earn $\beta_2 + \beta_3$ more

Another interaction: difference-in-differences

- Recreational cannabis was legalized in Illinois in 2020. We can create a dummy variable for whether it was 2020 or later. This variable is often called *post* in difference-in-differences model
- Indiana borders Illinois, but did not have recreational cannabis legalized. We can have another variable called *treat* that equals 1 if the state is Illinois, and 0 if the state is Indiana (assuming we only use these two states)

Another interaction: difference-in-differences

- Suppose we wish to analyze the effect of cannabis legalization on emergency room visits, measured as emergency room visits per 1000 population per year. We filter our data to Illinois and use $ER_i = \beta_0 + \beta_1 post_i + \varepsilon_i$. We obtain $\hat{\beta}_1 = 1$
- Interpret $\hat{\beta}_1$
    - After legalization of cannabis, ER visits increased by 1 visit per 1000 population per year

Another interaction: difference-in-differences

- Is this likely to capture the causal effect of cannabis legalization on emergency room visits?
    - No, ER visits could have been trending over time naturally for unrelated reasons

Another interaction: difference-in-differences

- Suppose instead you use both Illinois and Indiana and run $ER_i = \beta_0 + \beta_1 treat_i + \beta_2 post_i + \beta_3 treat_i * post_i$
- What does $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ represent?

Another interaction: difference-in-differences

- β_0 is the average for Indiana in the pre-period
- β_1 is the difference between Illinois and Indiana in the pre-period
- β_2 is the difference between Indiana in the pre-period and post-period

Another interaction: difference-in-differences

- β_3 is the difference between Illinois in the pre-period and post-period, minus the difference in Indiana in the pre and post period
- Illinois in the post period is $\beta_0 + \beta_1 + \beta_2 + \beta_3$. Illinois pre-period is $\beta_0 + \beta_1$
- Indiana in the post period is $\beta_0 + \beta_2$, and in the pre period is β_0
- So (Illinois_post-Illinois_pre) - (indiana_post-indiana_pre) $= (\beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_1)) - (\beta_0 + \beta_2 - \beta_0) = \beta_3$

Another interaction: difference-in-differences

- Is this a reasonable causal estimate? Under what conditions?
- Yes, so long as they're trending at the same rate.

Example: School fixed effects

- Suppose we are interested in the effect of teachers effectiveness on students. We use value-added as the increase in test scores of a student over their prior year in student-standard-deviation units
- We can run $grade_{ij} = \beta_0 + \beta_1 teacherVA_{ij} + \beta_2 baseline_{ij} + \varepsilon_{ij}$ for student i in school j
- Concern: Students are sorted into teachers, e.g. because parents move to locations with better teachers.

Example: School fixed effects

- Solution: School fixed effects. $grade_{ij} = \beta_0 + \beta_1 teacherVA_{ij} + \beta_2 baseline_{ij} + \lambda_j + \varepsilon_{ij}$
- Now we only use variation within a school
- As long as students aren't sorted to teachers within a school (e.g. from honors classes or special education) then we are likely to capture a causal effect
- This is more likely to occur in elementary school, so long as we filter out special education classes

Example: School fixed effects

- What if you have multiple years of data and there is grade inflation?
- School and year fixed effects
- $grade_{ijt} = \beta_0 + \beta_1 teacherVA_{ijt} + \beta_2 baseline_{ijt} + \lambda_j + \mu_t + \varepsilon_{ij}$
- What if grade inflation differs by school?

Example: School fixed effects

- School-by-year fixed effects
- $grade_{ijt} = \beta_0 + \beta_1 teacherVA_{ijt} + \beta_2 baseline_{ijt} + \theta_{jt} + \varepsilon_{ij}$
- What variation is left?

Example: School fixed effects

- Do we lose anything by using school fixed effects?
- What if better schools have better teachers, but students are randomly assigned to schools?
- Tradeoff between internal validity and external validity

Model Transformations

- When we turned education into two levels for whether an individual had a bachelors degree we already did an example of a variable transformation
- More generally, we can transform variables to provide a better fit to the data
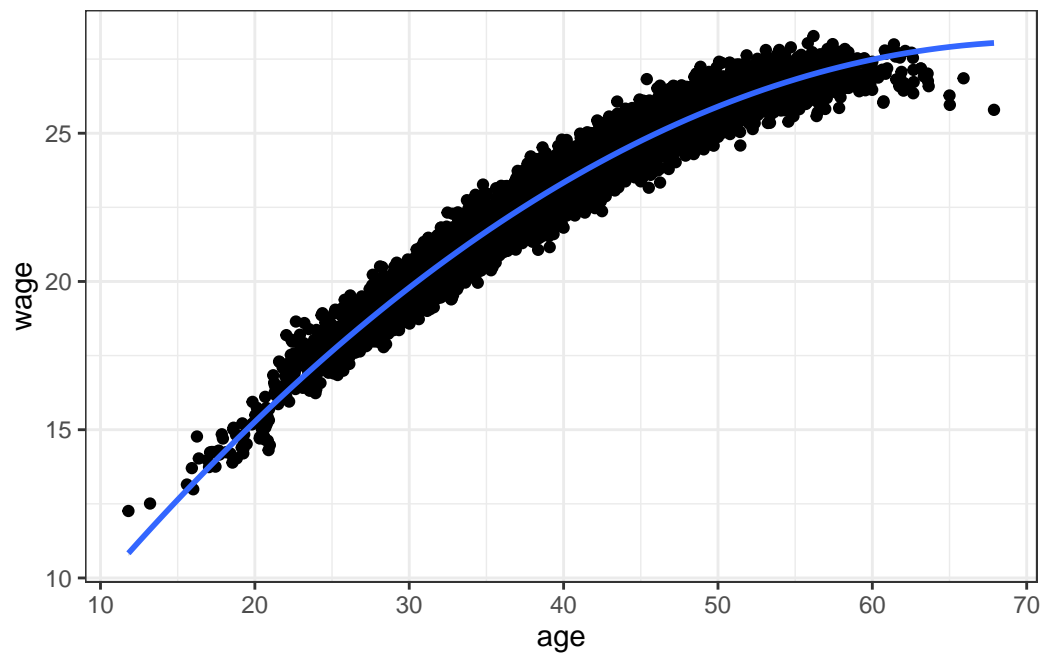- Let's consider how wages evolve with age as our example

Wage vs age
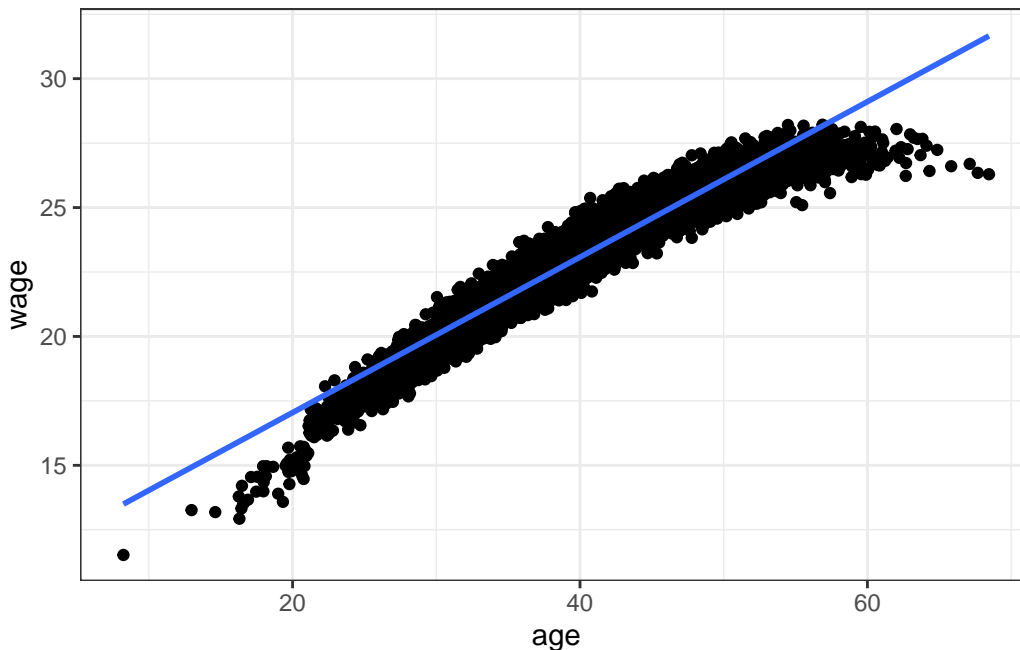
Wage vs age Quadratic Fit
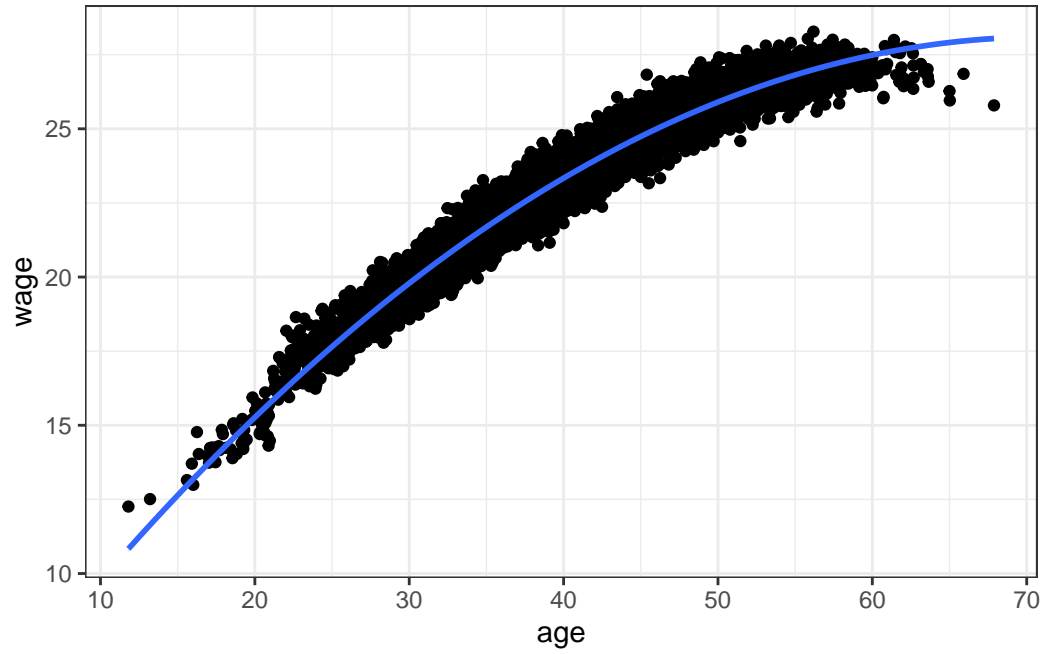
Interpreting Quadratic

```
Call:
lm(formula = wage ~ age, data = dt)

Coefficients:
(Intercept)           age
    11.0291        0.3007


Call:
lm(formula = wage ~ age + I(age^2), data = dt)

Coefficients:
(Intercept)           age      I(age^2)
   3.280827      0.697292     -0.004898
```

- $wage_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2$
- How do we interpret β_1, β_2?

Interpreting Quadratic

- Derivatives: slope is $\beta_1 + 2\beta_2 age$. On average wage increases by $\beta_1 + 2\beta_2 age$ for every additional year of age
- Note that the slope now depends on the age (it's not constant)

Linear-Log Transformations

- Logs are the inverses of exponential functions: $log(k^x) = x$. They have the property that $log(xy) = log(x) + log(y)$
- Suppose we have $y = \beta_0 + \beta_1 log(x) + \varepsilon$. How do we interpret a 1 unit change in log(x)?

Linear-Log Transformations

- If $log(x)$ increases by 1, then that means that $log(k^{x+1}) = log(k * k^x)$
  - ie it means that x increases by a factor of k
- The base of our logarithm is arbitrary - a 1 unit increase in log(x) results in a multiplicative change to x
  - In other words, a 1 unit increase in log(x) causes a $100\beta_1$ percent change in y
  - (approximately: $log(1 + x) \approx x$)

Linear-Log Transformations

- Example: $y = \beta_0 + 10log(x) + \varepsilon$
- a 1% increase in x is associated with a 0.1 unit increase in y

Log-linear Transformations

- We an also have our y variable as a log: $log(y) = \beta_0 + \beta_1 x + \varepsilon$
- This means that a 1 unit increase in 1 leads to a 1 unit increase in log(y), or a 100% increase in y

Log-Log Transformations

- We can also have logs on both sides : $log(y) = \beta_0 + \beta_1 log(x) + \varepsilon$. What does β_1 represent?
- A 1% increase in x results in a 1% increase in y
- This is elasticity if you covered it in microeconomics

Example: Income
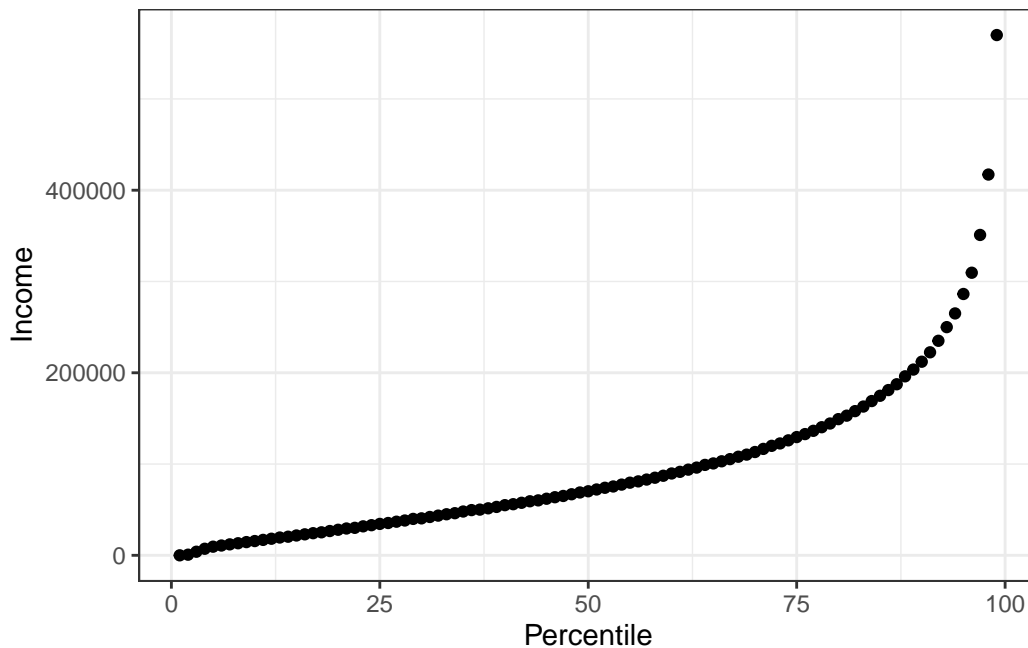
Example: Income

```
Call:
lm(formula = Income ~ Percentile, data = dt)

Coefficients:
(Intercept)    Percentile
    -40367          2751
```

```
Call:
lm(formula = log(Income + 1) ~ Percentile, data = dt)

Coefficients:
(Intercept)    Percentile
     8.9293        0.0404
```

```
Call:
lm(formula = log(Income + 1) ~ log(Percentile), data = dt)

Coefficients:
    (Intercept)   log(Percentile)
          5.588             1.478
```

Different Types of interpretations

- Interpret $\hat{\beta}_1$ in each of the following:
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
    - 1 unit increase in x associated with an average β_1 unit increase in y
- $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$
    - " " holding z constant
- $y_i = \beta_0 + \beta_1 d_i + \varepsilon_i$
    - The average difference between the group characterized by $d_i = 1$ compared to $d_i = 0$

Different Types of interpretations

- $y_i = \beta_0 + \beta_1 d_i + \beta_2 z_i + \varepsilon_i$
  - " " holding z constant
- $d_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - 1 unit increase in x associated with average of $100\beta_1\%$ increase in probability of attaining $d_i = 1$
- $d_i = \beta_0 + \beta_1 d2_i + \varepsilon_i$
  - Average difference in probability of attaining $d_i = 1$ for group $d2 = 1$

Different Types of interpretations

- $log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - 1 unit increase in x associated with $100\beta_1\%$ increase in y
- $y_i = \beta_0 + \beta_1 log(x_i) + \varepsilon_i$
  - 100% increase in x associated with β_1 unit increase in y
- $log(y_i) = \beta_0 + \beta_1 log(x_i) + \varepsilon_i$
  - 1% increase in x associated with β_1 percent increase in y

Different Types of interpretations

- $y_i = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_1 d_1 * d_2 + \varepsilon_i$
  - average difference for group $d_1 = 1$ vs $d_1 = 0$ when $d_2 = 0$
- $y_{it} = \beta_0 + \beta_1 x_{it} + \lambda_t + \varepsilon_{it}$
  - ... holding time fixed ("only using variation within years, not across")

Fixed Effect models: A formalization

- In the past we've run regressions where observations vary at the individual level, but sometimes it is important to think about observations varying at multiple levels.
- For instance individual, time, and regional level variation may exist in some data but not others (e.g. a panel dataset vs a cross sectional one).
- More formally, in a cross sectional dataset we observe $y_i = \beta_0 + beta_1 x_i + \varepsilon_i$, while in a different dataset we may observe individuals at the regional level: $y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$

Fixed Effect models: A formalization

- This matters for fixed effects: We may observe the region but decide to only use variation within a region rather than across regions.
- Notationally we do this by adding a regional fixed effect: $y_{ij} = \beta_0 + \beta_1 x_{ij} + \lambda_j + \varepsilon_{ij}$
- Mechanically we fit this using dummy variables, one for each region (except 1): $y_{ij} = \beta_0 + \beta_1 x_{ij} + D_{j1} + D_{j2} + ...D_{jk} + \varepsilon_{ij}$

Fixed Effect models: A formalization

- This is identical to the demeaned version: $Y_{ij} - \bar{Y}_j = \beta_1(X_{ij} - \bar{X}_j)$
- Identical to running a regression for each region of x on y, then averaging the results (using a weighted average)

Two Way Fixed Effects (TWFE)

- Suppose we're looking at the average income of individuals in different cities over time and are interested in how the crime rate affects that: $log(income_{it}) = \beta_0 + \beta_1 crime_{it} + \varepsilon_{it}$ where i indexes the city and t the year.
- Both crime rates and incomes vary over cities and over time, so these are obvious sources of endogeneity. We can include both fixed effects in our model:
- $log(income_{it}) + \beta_0 + \beta_1 crime_{it} + \lambda_t + \mu_i + \varepsilon_j$

Two Way Fixed Effects (TWFE)

- $log(income_{it}) + \beta_0 + \beta_1 crime_{it} + \lambda_t + \mu_i + \varepsilon_j$
- This means that we have removed variation across cities and across time. What variation is left?
- We still have the interaction between city and time, i.e. the differential trend in cities
- This is the exact same variation we use in a difference-in-difference model: the difference in trend between our treatment and control group

Two Way Fixed Effects (TWFE)

- $log(income_{it}) + \beta_0 + \beta_1 crime_{it} + \lambda_t + \mu_i + \varepsilon_j$

- Suppose we run this and obtain $\hat{\beta}_1 = -1$. How do we interpret this (assume that crime is measured in violent crimes per 1000 people per year)
- each additional crime per 1000 people is associated with a 1% decrease in income

Endogeneity in Two Way Fixed Effects Estimators

- By including fixed effects for city and year we've removed all sources of variation associated with these
- Are there still potential endogeneity issues?
- Yes, if there are any endogenous factors (that impact both income and crime) that trend differentially across city
- i.e. what we haven't removed with our fixed effects.

Endogeneity in Two Way Fixed Effects Estimators

- Example: suppose high income individuals move to areas of lower crime.
- This will lead to a differential trend across cities that is not controlled for with our fixed effects
- Question: Can we solve for this endogeneity by also controlling for the interaction between city and year?

Endogeneity in Two Way Fixed Effects Estimators

- This is called a city-by-year fixed effect. In general we can, but once we've controlled for this we no longer have variation in crime rate or income if we only observe data at the city level.
- If we observe crime and income at a more granular level we can still use these fixed effects

Endogeneity in Two Way Fixed Effects Estimators

- In our two way fixed effects estimation, can we measure the effect of whether a city is near a lake on income levels?
- No, because whether a city is near a lake does not vary over time within a city, and therefore has no variation

Endogeneity in Two Way Fixed Effects Estimators

- We are concerned that city size might be an omitted variable (urban areas have higher crime rates and higher incomes). Should we control for city size?
- No need - this is absorbed into city fixed effects. Within a city there is no variation in population
- What about race, e.g. percent of black population?
- This only matters if racial composition of a city is trending over time. It's a weak control.