UNITED STATES MILITARY ACADEMY
WEST POINT®

Causal Inference in Introductory Statistics Courses: Why, What, and How?

Kevin Cummiskey and Bryan Adams

CAUSEWeb Webinar on October 8th, 2019

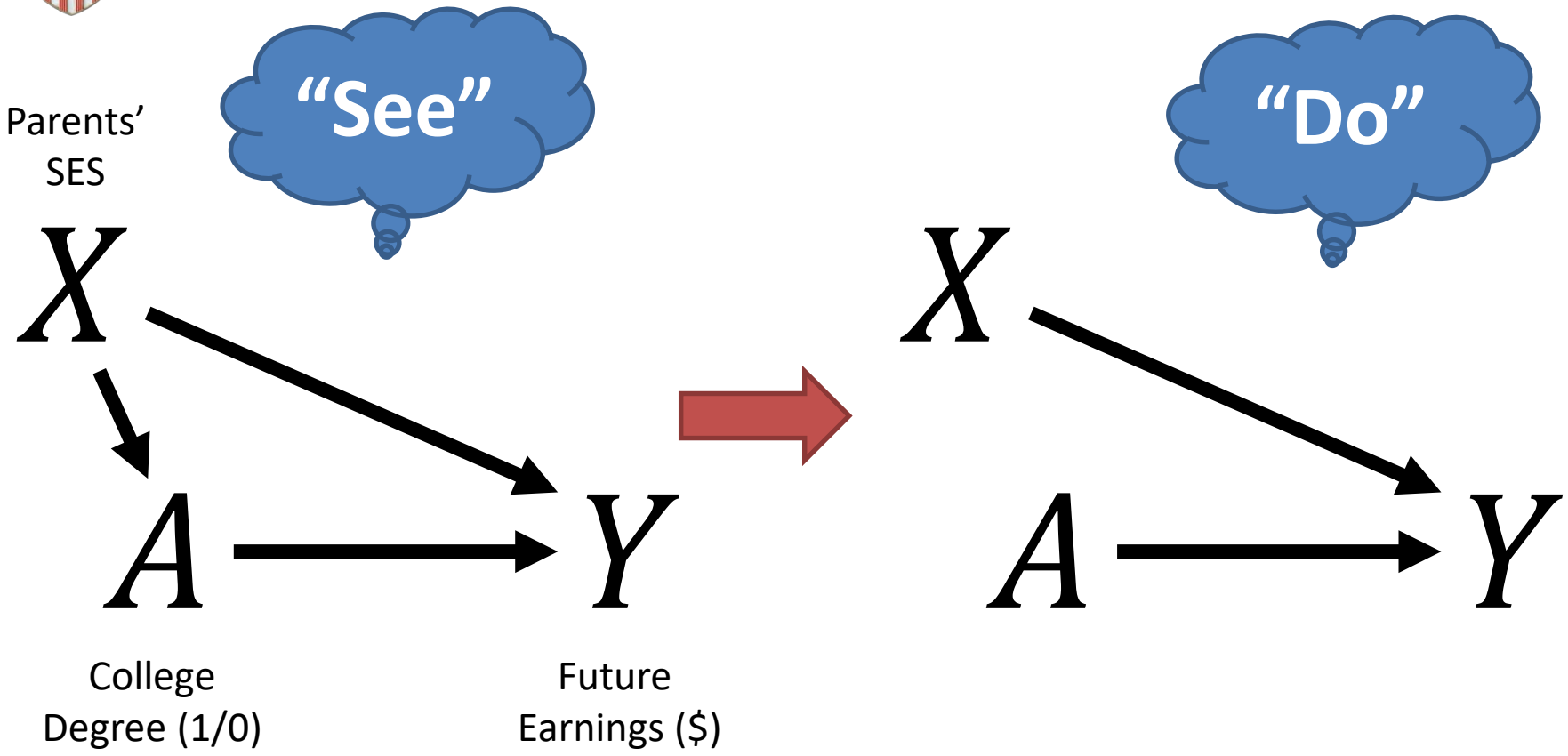
Acknowledgements: James Pleuss, Dusty Turner, Nicholas Clark, and Krista Watts
Acknowledgements #2: Introduction to Statistical Investigations by Tintle et al.
Acknowledgements #3: Michael Kahn, *An exhalent problem for teaching statistics*

Views expressed in this presentation are those of Kevin Cummiskey and Bryan Adams and do not represent the official position of the U.S. Department of Defense, the Army, or West Point.

Agenda

- Brief Intro to Causal Inference
  - Defining causal effects
  - Identifying causal effects in data
  - Using causal diagrams to depict causal assumptions
- Causal inference in introductory statistics courses
  - Why?
  - What?
  - How?

Measures of Association and Causal Effects



Measure of Association:

$$E(Y|A = 1) - E(Y|A = 0)$$

Causal Effect:

?

Notation for Causal Effects

- Potential Outcomes Framework (Rubin 1974)

$A \in \{0,1\}$ (Binary Treatment)

Y_1: potential outcome when individual is treated

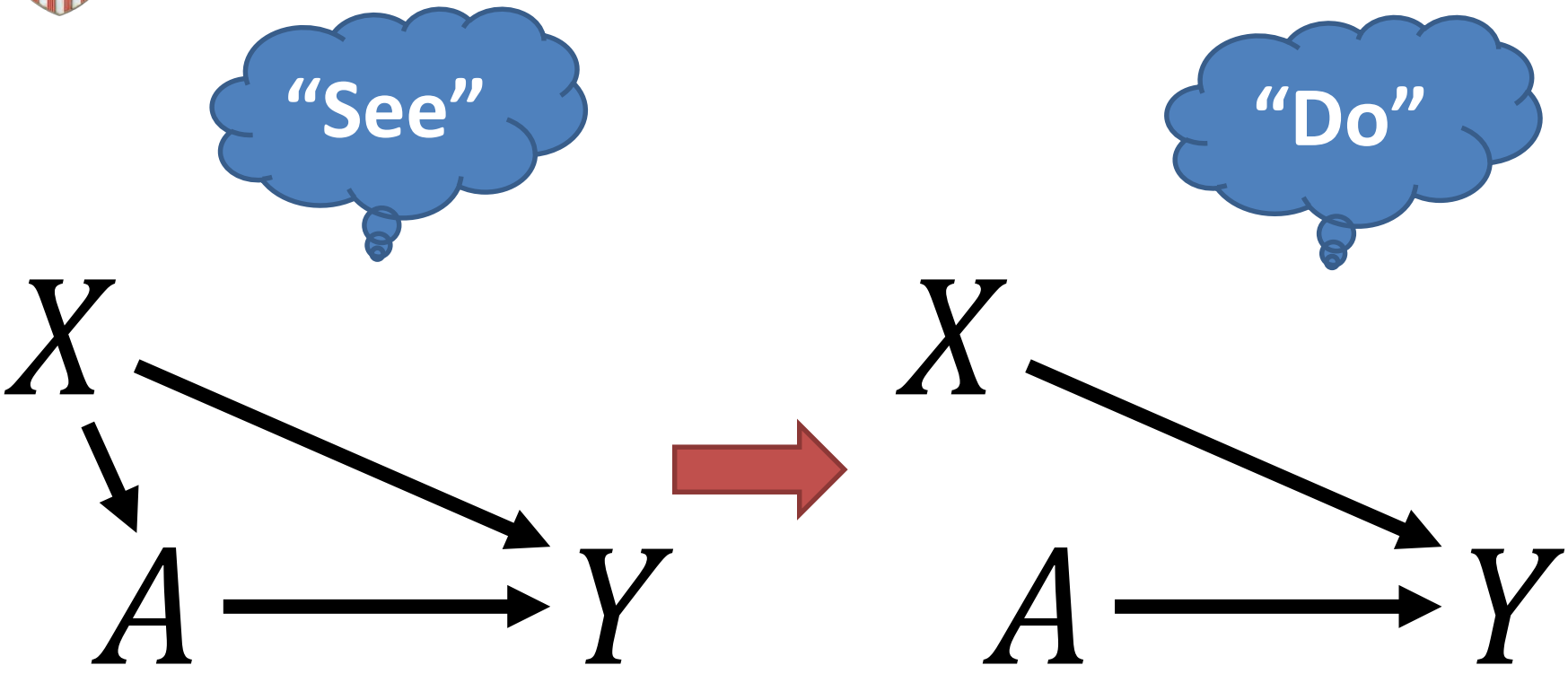
Y_0: potential outcome when individual is treated

Average Causal Effect (ACE): $E(Y_1) - E(Y_0)$

- "Do" Operator (Pearl 1995)

$$E(Y|\mathrm{do}(A = 1)) - E(Y|\mathrm{do}(A = 0))$$

Measures of Association and Causal Effects



Measure of Association:

$$E(Y|A = 1) - E(Y|A = 0)$$

Causal Effect:

$$E(Y_1) - E(Y_0)$$

Identifying Causal Effects

The fundamental problem in causal inference:

- We cannot observe both potential outcomes for each individual.
- Therefore, we cannot directly observe the average causal effect.

$$E(Y_1) - E(Y_0)$$

- We can only observe:

$$E(Y_1|A = 1) - E(Y_0|A = 0)$$

However, sometimes we can identify causal effects from things we can observe. ***Causal diagrams are very helpful.***

Causal Diagrams

Key Idea: Causal diagrams contain all common causes of the treatment and outcome.

Figure 1. A is a cause of Y.

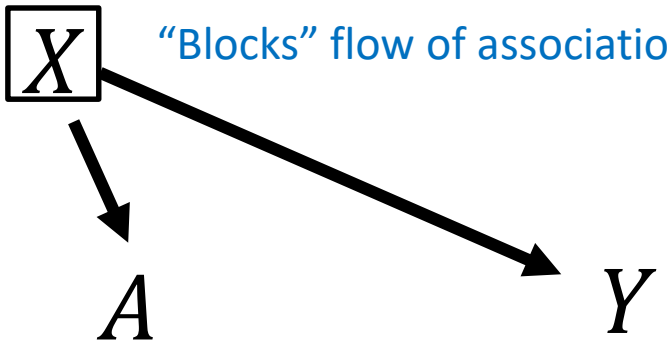


Figure 2. Confounding: X is a confounder of the effect of A on Y.

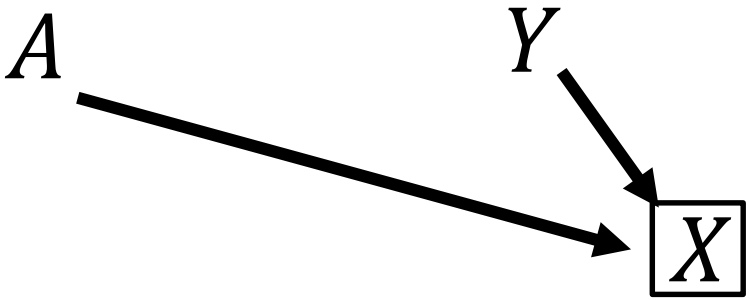


Figure 3. Collider - A and Y are common causes of collider X.

"opens" flow of association

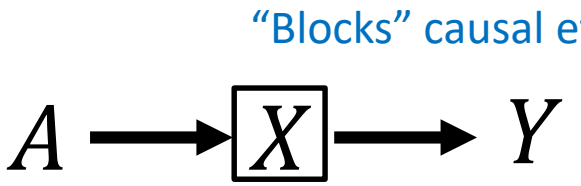
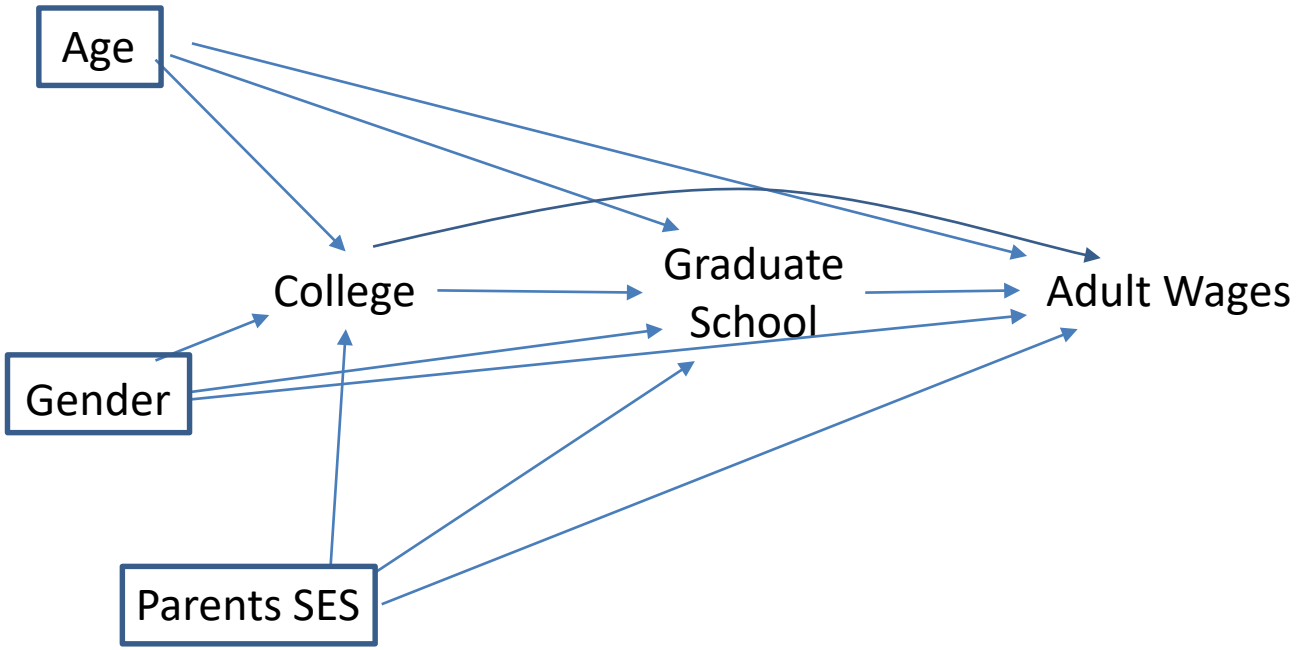


Figure 4. X is a mediator of the effect of a A on Y.

Using Causal Diagrams to Identify Causal Effects

We can identify causal effects when (known as *backdoor criterion*):

- There are no backdoor paths between treatment and outcome.
- You have measured sufficient confounders to "block" any backdoor paths.



Causal Inference Review

- Goal is to estimate effect of intervening on one variable on another variable.
- Sometimes, we can estimate this effect from observational data.
- Investigators specify causal assumptions a priori using expert knowledge.
- From causal diagrams, you can tell if causal effects are identifiable. They are identifiable if:
  - There are no backdoor paths between treatment and outcome.
  - You have measured sufficient confounders to "block" any backdoor paths.

Why causal inference in introductory statistics?

GAISE Report

Recommendations for introductory statistics courses:

1. Teach statistical thinking.
   - Teach statistics as an ***investigative process*** of problem-solving and decision making.
   - Give students experience with ***multivariable thinking***.
2. Focus on conceptual understanding.
3. Integrate real data with ***a context and purpose***.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

Carver, R., et al. "Guidelines for assessment and instruction in statistics education (GAISE) college report 2016." *Alexandria, VA: American Statistical Association* (2016).

Why Causal in Intro Stats?

- Reason #1: Causal questions require the investigator (student) to consider the entire investigative process.

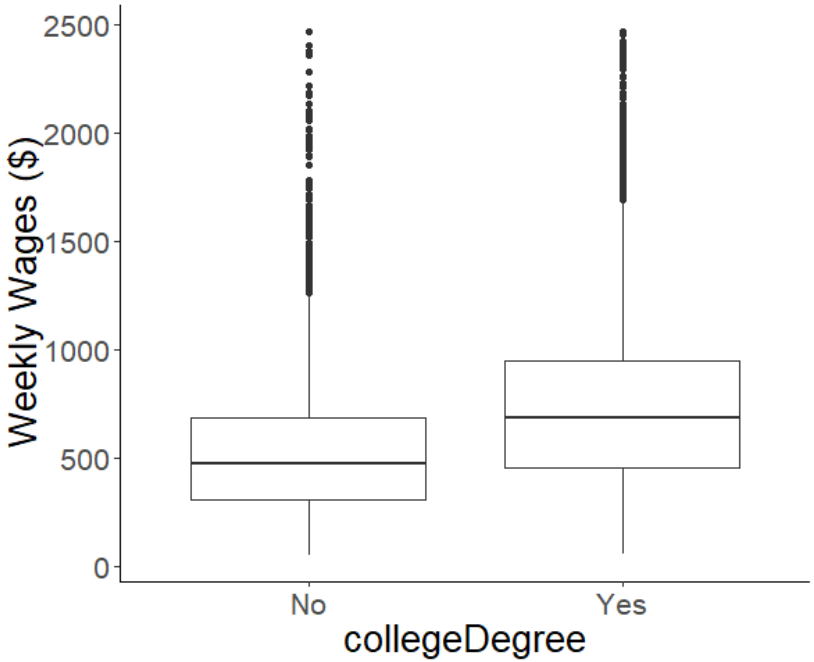


Figure. Weekly wages (1992 $'s) of ~25,000 males in the 1988 U.S. Current Population Survey (source: *Tintle et al* 2016)

Is there an association between collegeDegree and Weekly Wages?
- Size
- Strength

Would obtaining a college degree have resulted in someone earning higher wages?
- What was the study design?
- How was the data collected?
- How was the data analyzed?

Why Causal in Intro Stats?

- Reason #2: Discussing and formally specifying causal assumptions help develop multivariable thinking.

Explanatory Variable: College Degree (Y/N)

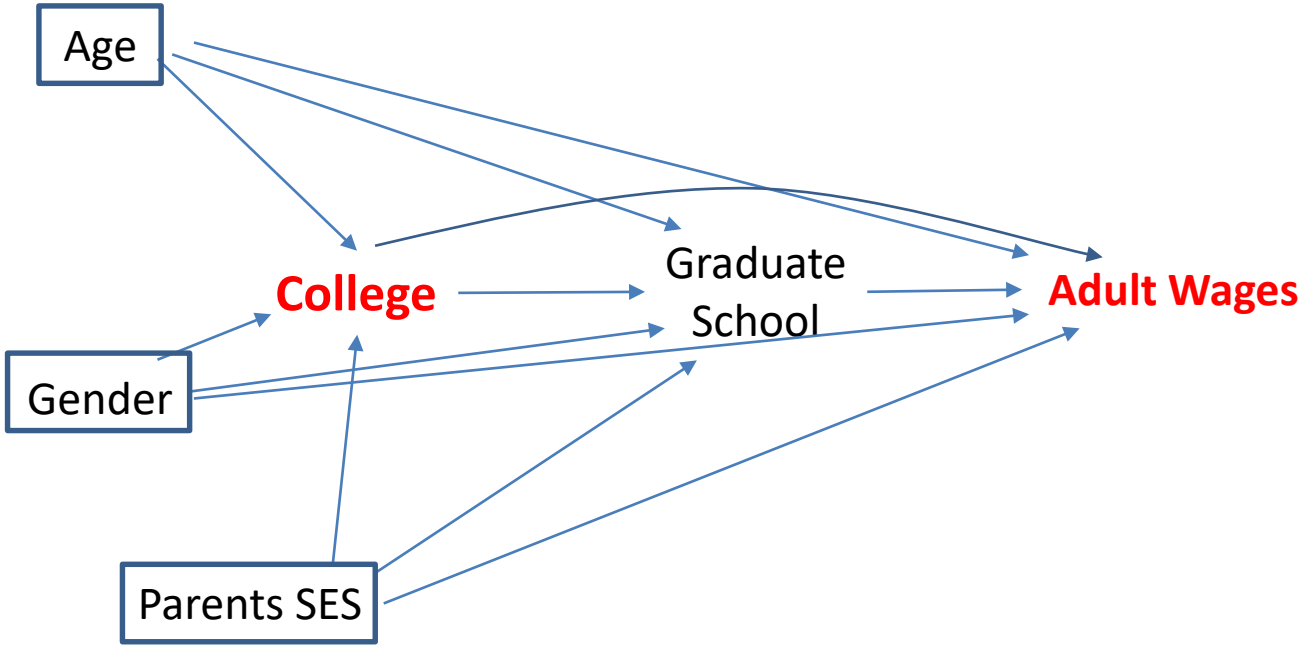
Outcome Variable: Adult Earnings ($)

Other Variables:

- Age
- Gender
- Parents' Education Level
- Parents' Income
- Graduate School

Why Causal in Intro Stats?

- Reason #3: Causal diagrams are visual tools for structuring multivariable thinking.



Why Causal in Intro Stats?

- Reason #4: Causal diagrams are useful for demonstrating other concepts.

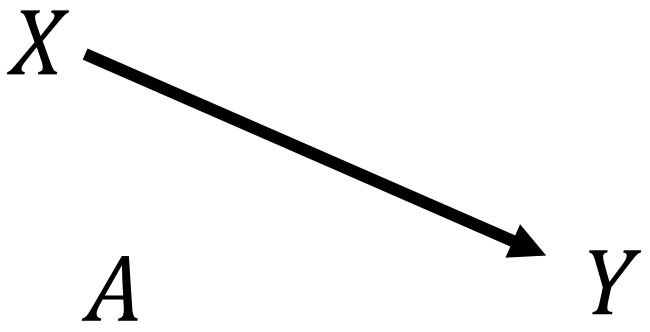


Figure 1. Null Hypothesis.

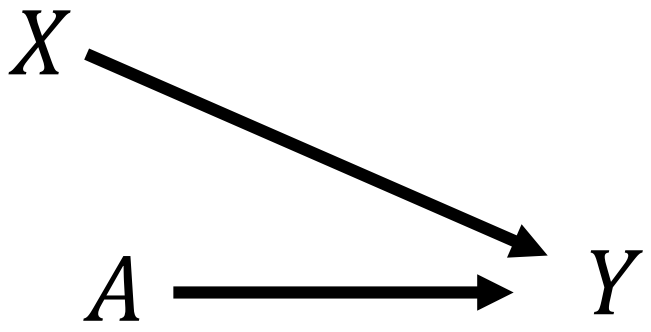


Figure 2. Randomized Controlled Experiment.

Why Causal in Intro Stats?

- Reason #5: “Correlation does not imply causation” is still good advice for students, but there is a lot more our discipline has to say about causality.

What Causal in Intro Stats?

- Topic #1: Difference between association and causal relationships.

Association – "See"

*I **see** a relationship between having a college degree and adult earnings in the data.*

Causal – "Do"

*If I **did** an intervention on college degree, there would have been a change in adult wages.*

What Causal in Intro Stats?

- Topic #2: Confounding.

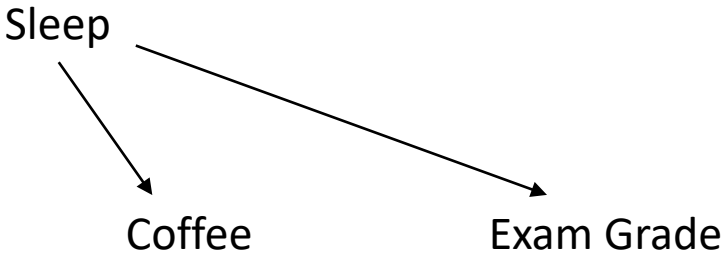


Figure 1. Coffee has no effect on exam performance. However, we observe an association because of confounding (sleep).

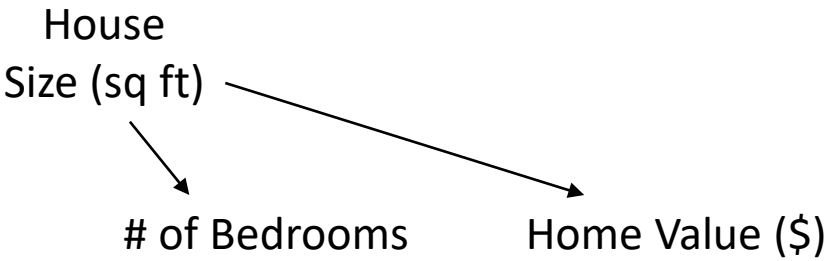
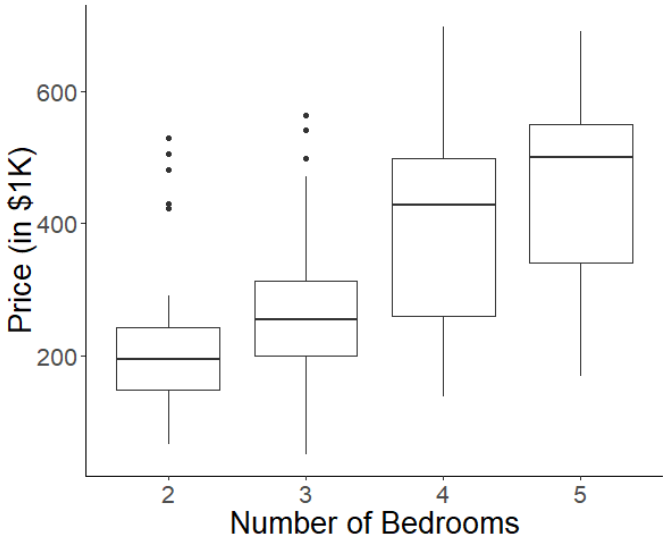
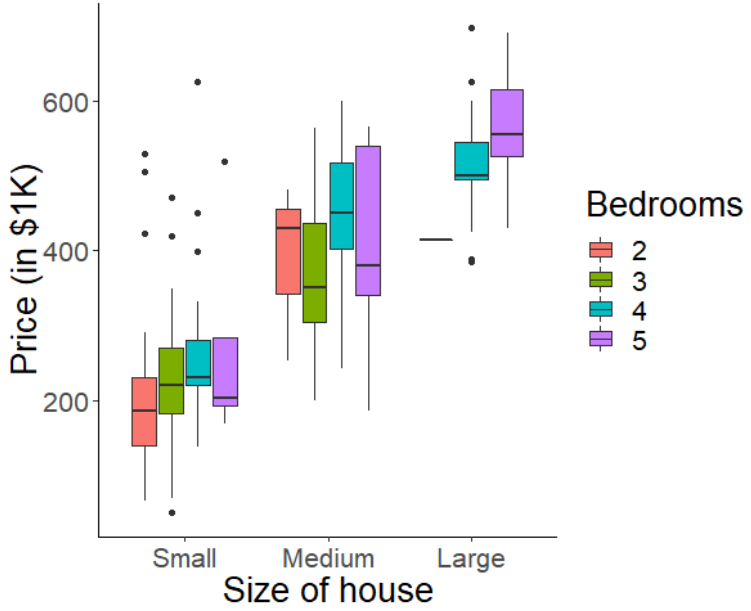


Figure 2. Houses with more bedrooms tend to sell for higher prices. However, adding a bedroom to your house will generally not change its value.



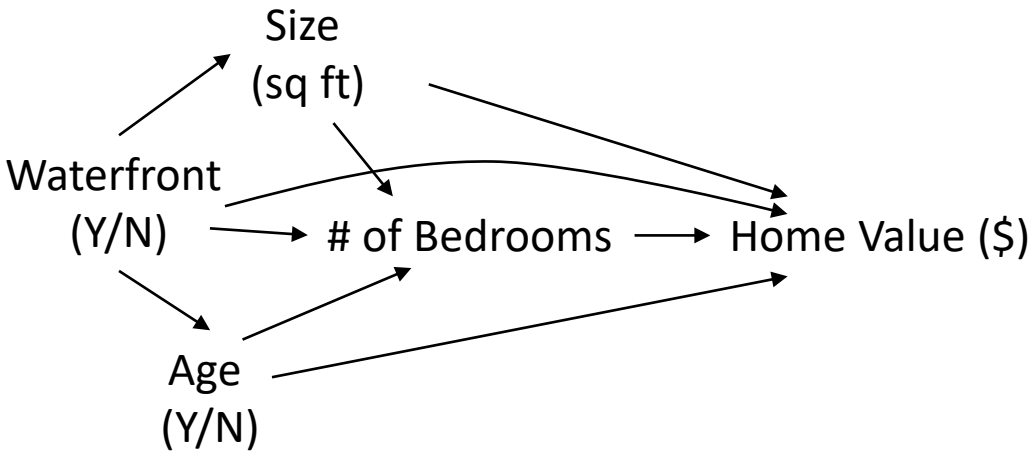


Homes in Cornwall, New York (source: redfin.com)

What Causal in Intro Stats?

- Topic #3: Causal Diagrams

What is the effect of adding a bedroom to a house?

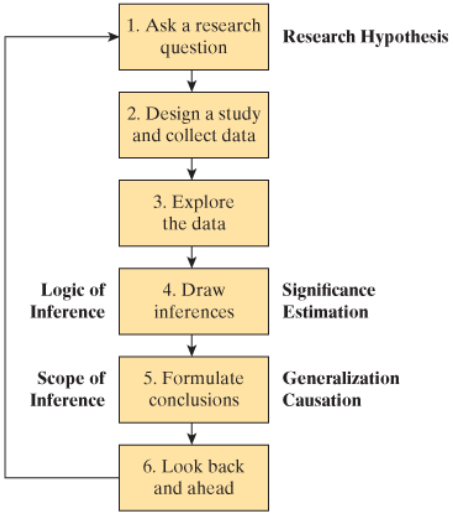


What Causal in Intro Stats?

- Topic #4. Methods for confounding adjustment
  - Stratifying on a confounding variable
  - Regression
  - Matching

How?

Guided Lab



- *Introduction to Statistical Analysis* by Tintle et al.

Background for the guided lab

- How does childhood smoking effect lung function? A 1970's study in Boston collected data related to this topic. The researchers followed a cohort of children in East Boston, MA for seven years to determine, among other things, the effect of childhood smoking on lung function.

1. Ask a research questions?

   What is the effect of smoking on lung function in young people?

2. Design a study and collect data

a. Is this an observational study or experiment? Explain.

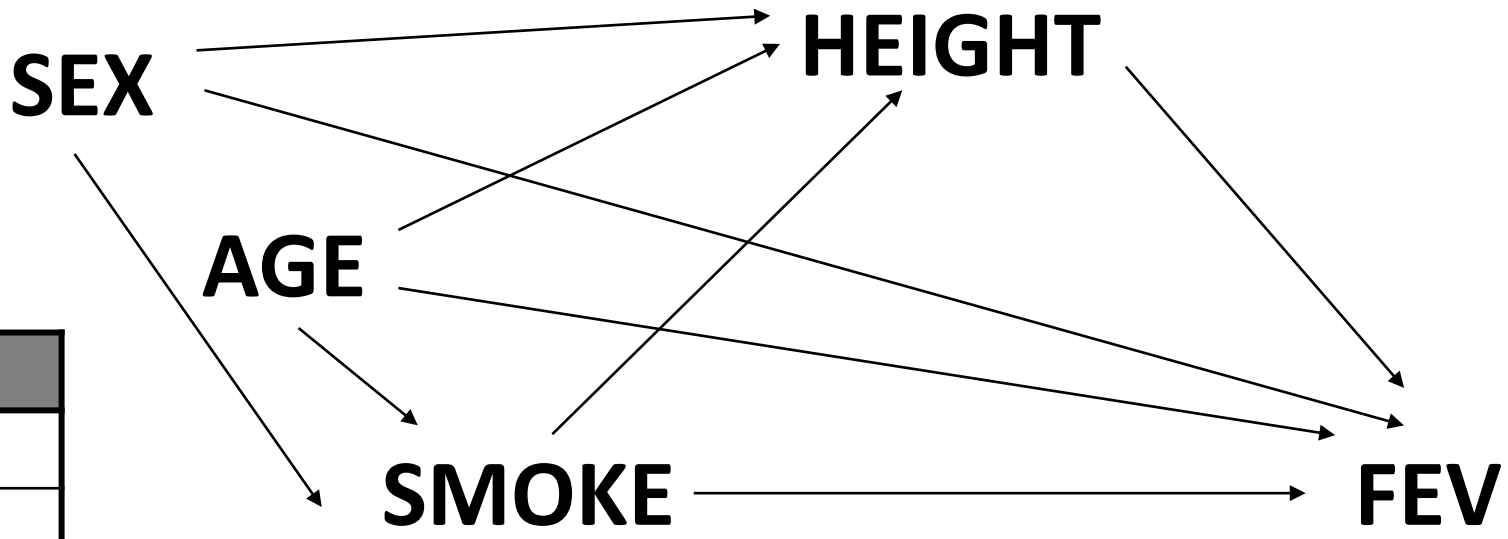
This is an observational study because we are not able to randomize whether a subject smoked or not

b. If you were to conduct this study what variables would you record?

Variable	Description
AGE	The age of the subject in years
FEV	Forced expiratory volume (L), a common measure of lung function
HEIGHT	The height of the subject in inches
SEX	Biological sex of the subject: Female (0) Male (1)
SMOKE	Whether the subject had ever smoked or not: No (0), Yes (1)

2. Design a study and collect data continued

c. Draw a causal diagram describing the relationship between variables

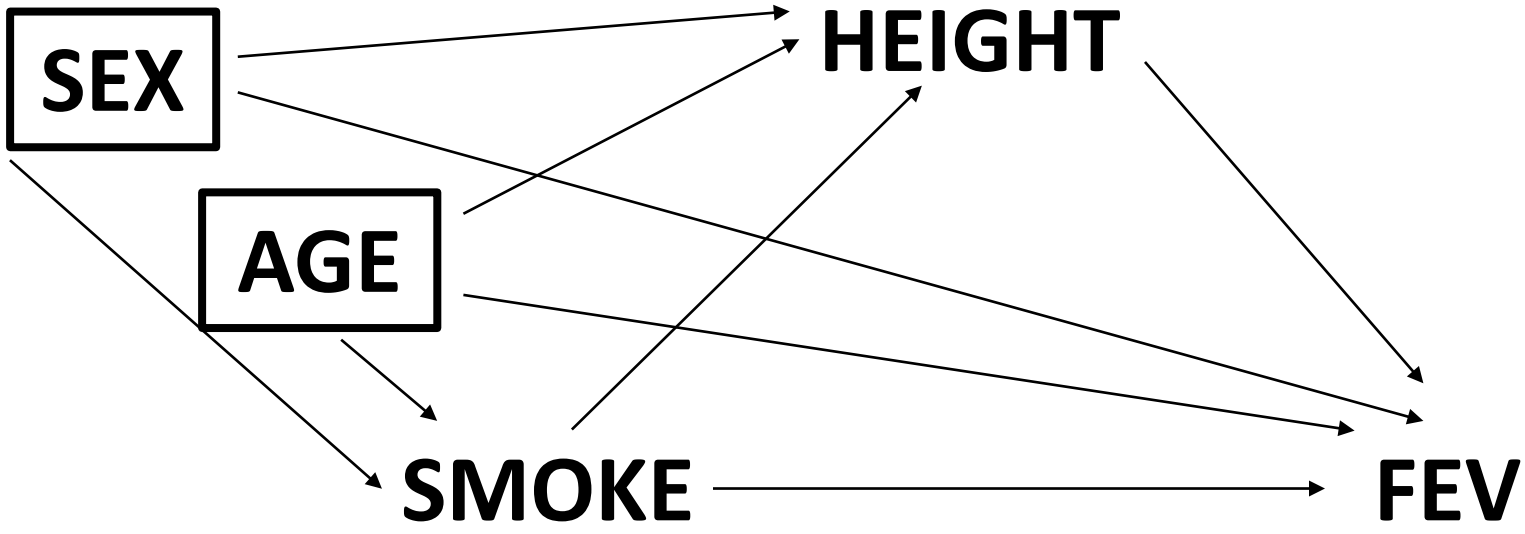


Variable
AGE
FEV
HEIGHT
SEX
SMOKE

2. Design a study and collect data continued

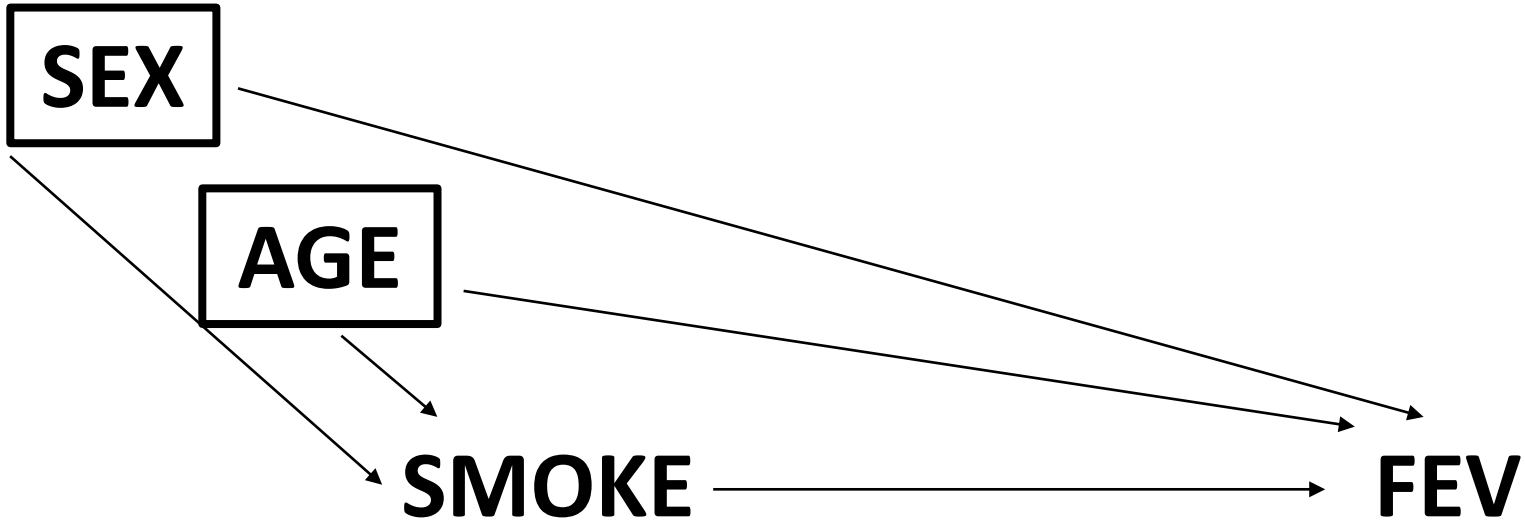
d. Based on your causal diagram, which variables are potential confounders of the effect of smoking on lung function? Explain why you selected these variables.

d. Redraw your causal diagram with a box around each confounder. You will adjust for these variables in your analysis.



Guided Lab

- What the final casual diagram will look like



Guided Lab

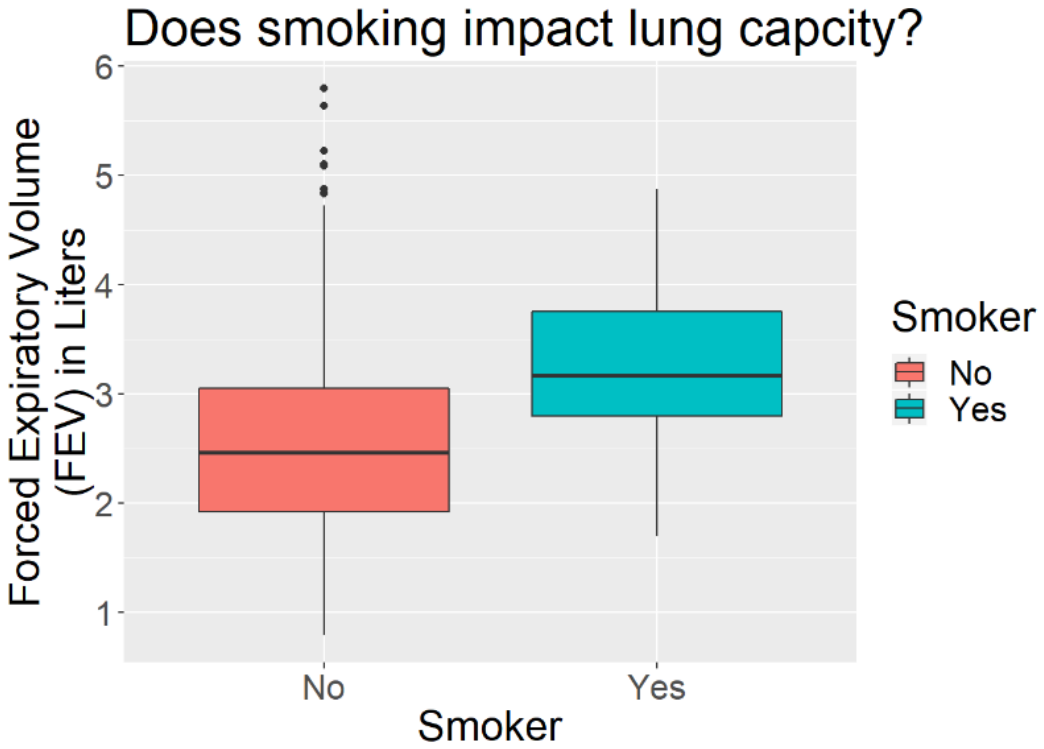
3. Explore the data.

   a. Perform data analysis and comment on at least two interesting features

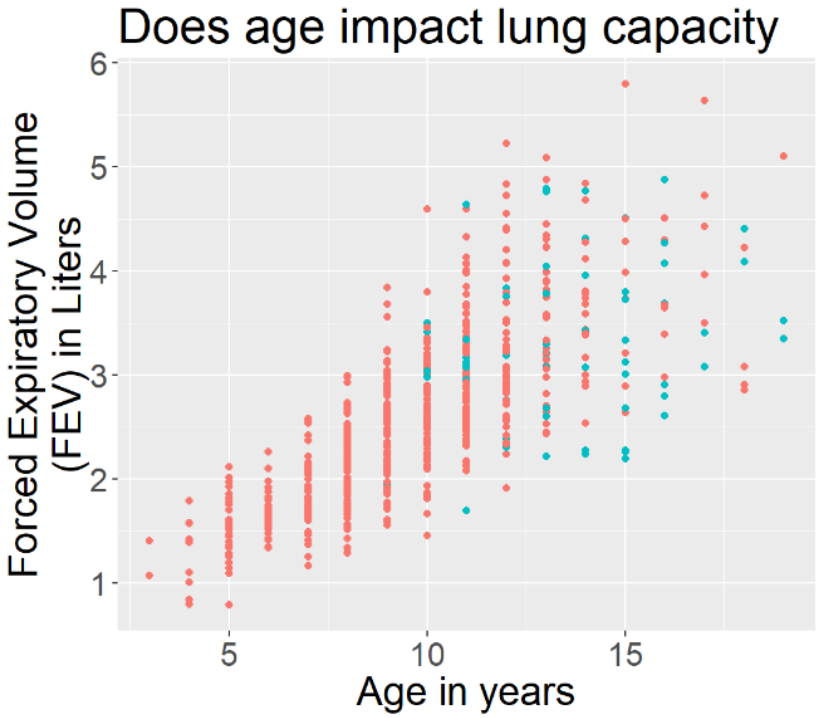
To do this we use the tidyverse package in R

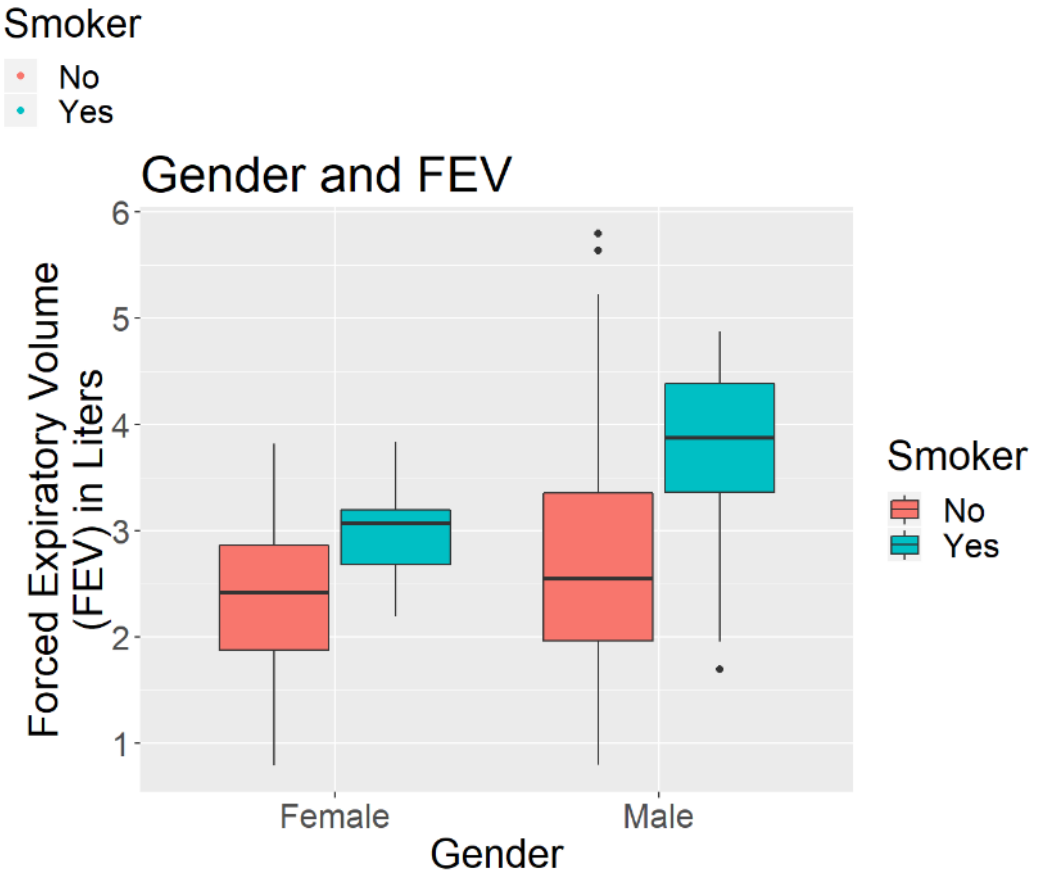
Guided Lab

```
fev_data%>%
  mutate(smoke = case_when(
    smoke == 1 ~ "Yes",
    TRUE ~ "No"
  ))%>%
  rename(Smoker = smoke)%>%
  ggplot(aes(x = Smoker,
             y = fev, fill = Smoker))+
  geom_boxplot()+
  labs(title = "Does smoking impact lung capcity?", y = "Forced Expiratory Volume\n(FEV) in Liters", x = "
Smoker")+
  theme(text = element_text(size = 20))
```



Guided Lab


Does age impact lung capacity
Forced Expiratory Volume (FEV) in Liters
Age in years
1
2
3
4
5
6
5
10
15


Smoker
No
Yes
Gender and FEV
Forced Expiratory Volume (FEV) in Liters
Gender
Female
Male
1
2
3
4
5
6
Smoker
No
Yes

Guided Lab

4. Draw inferences
   a. Simple Linear Regression (Unadjusted for Age and Sex)
   b. Multiple Linear Regression (Age and Sex Adjusted)
   c. Matched (Age and Sex Adjusted)

Simple Linear Regression (Unadjusted for Age and Sex)

```
not_a_good_model = lm(fev~smoke, data = fev_data)

summary(not_a_good_model)
```

```
##
## Call:
## lm(formula = fev ~ smoke, data = fev_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7751 -0.6339 -0.1021  0.4804  3.2269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
## smoke        0.71072    0.10994   6.464 1.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8412 on 652 degrees of freedom
## Multiple R-squared:  0.06023,    Adjusted R-squared:  0.05879
## F-statistic: 41.79 on 1 and 652 DF,  p-value: 1.993e-10
```

Multiple Linear Regression (Age and Sex Adjusted)

```
fev_model = lm(fev ~ smoke + age + sex, data = fev_data%>%
                 mutate(sex = case_when(
                   sex == 1 ~ "Male",
                   TRUE ~ "Female"
                 )))

summary(fev_model)
```

```
##
## Call:
## lm(formula = fev ~ smoke + age + sex, data = fev_data %>% mutate(sex = case_when(sex ==
##     1 ~ "Male", TRUE ~ "Female")))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46707 -0.35426 -0.03811  0.32199  1.94943
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.237771   0.080228   2.964  0.00315 **
## smoke       -0.153974   0.077977  -1.975  0.04873 *
## age          0.226794   0.007884  28.765  < 2e-16 ***
## sexMale      0.315273   0.042710   7.382  4.8e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5432 on 650 degrees of freedom
## Multiple R-squared:  0.6093, Adjusted R-squared:  0.6075
## F-statistic: 337.9 on 3 and 650 DF,  p-value: < 2.2e-16
```

Matched (Age and Sex Adjusted)

```
library(MatchIt)

smoke_match = matchit(smoke ~ age + sex, data = fev_data,
                      method = "exact")

matched_data = match.data(smoke_match)

matched_model = lm(fev ~ smoke, data = matched_data, weights = weights)


summary(matched_model)
```

```
##
## Call:
## lm(formula = fev ~ smoke, data = matched_data, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4566 -0.4972 -0.2292  0.1417  3.9347
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.34307    0.04234  78.959   <2e-16 ***
## smoke       -0.07114    0.10507  -0.677    0.499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7633 on 386 degrees of freedom
## Multiple R-squared:  0.001186,   Adjusted R-squared:  -0.001402
## F-statistic: 0.4584 on 1 and 386 DF,  p-value: 0.4988
```

Guided Lab

4. Draw inferences

Model	Estimate	95% CI
Age and sex Unadjusted	0.71	(0.49, 0.93)
Age and sex adjusted (regression)	-0.15	(-0.31, 0.00)
Age and sex adjusted (matching)	-0.08	(-0.22,0.08)

5. Formulate conclusions

   a. Is your result statistically significant? Is your result practically significant?

6. Look back and ahead

   a. Describe two other confounding variables we should have considered in this analysis

      Diet, Physical activity

Future Work

- Materials are available at: https://github.com/kfcaby/causalLab

- Paper forthcoming:

  Cummiskey, Adams, Pleuss, Turner, Clark, and Watts (2019). Causal Inference in Introductory Statistics Courses.

- Joint work with Shonda Kuiper for a game-based lab using Defenders Game (causal lab not available yet) https://www.stat2games.sites.grinnell.edu/

Questions

References and Resources

1. Carver, Robert, et al. "Guidelines for assessment and instruction in statistics education (GAISE) college report 2016." (2016).
2. Chance, Beth L. "Components of statistical thinking and implications for instruction and assessment." *Journal of Statistics Education* 10.3 (2002).
3. Greenland, Sander, Judea Pearl, and James M. Robins. "Causal diagrams for epidemiologic research." *Epidemiology* 10 (1999): 37-48.
4. Hernan, Miguel, and J. M. Robins. "Causal inference book. 2016." (2017) https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
5. Hernan, Miguel. "Causal Diagrams: Draw your assumptions before your conclusions" Online course available at https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your-conclusions.
6. Hernán, Miguel A., et al. "Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology." *American journal of epidemiology* 155.2 (2002): 176-184.
7. Kahn, Michael. "An exhalent problem for teaching statistics." *Journal of Statistics Education* 13.2 (2005).
8. Pearl, Judea. "Causal diagrams for empirical research." *Biometrika* 82.4 (1995)
9. Pearl, Judea, and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.
10. Pearl, Judea. "An introduction to causal inference." *The international journal of biostatistics* 6.2 (2010).
11. Rubin, Donald B. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66.5 (1974): 688.
12. Tintle, Nathan, et al. *Introduction to statistical investigations*. John Wiley & Sons, 2015.

Additional Slides
(Not used in presentation)

Further Assumptions (Not Discussed)

- Stable Unit Treatment Value Assumption (SUTVA)
  - One individual's treatment level does not affect other individuals' outcomes
  - Well-defined treatment levels

Treatment Assignment

- Treatment Assignment Mechanism
  - Assumptions about why some individuals were more likely to receive treatment
  - Specified a priori using subject-area expertise
  - Also called “causal assumptions”
  - Frequently depicted with causal diagrams

Identifying Causal Effects

So, when can we identify causal effects?

- Conditional Ignorability

$$(Y_1, Y_0) \perp A|\boldsymbol{X}$$

- Check for “backdoor paths” in causal diagrams