Causality and Multiple Regression Supplement

Krista Watts and Kevin Cummiskey

2021-03-12

Contents

Preface **5**

1 Introduction **7**

1.1 Goals of quantitative research (describe, predict, cause-and-effect) 7

1.2 Validity 8

2 Causality **11**

2.1 What does it mean for one thing to cause another 12

2.2 Randomized controlled experiments 12

2.3 Observational Studies and Confounding 13

2.4 Causal Diagrams 14

3 Matching **17**

4 (COL Watts) Quantitative explanatory variable with categorical confounding variable **19**

4.1 Unadjusted effect of X on Y 19

4.2 Effect of X and Y adjusting for C. 19

4.3 Assessing model adequacy 19

5 (Kevin) Interactions **21**

6 Activity **23**

7 (COL Watts) Quantitative explanatory variable with quantitative confounding variable 31

7.1 Effect of X and Y adjusting for C. 31

7.2 Assessing model adequacy 31

Preface

This textbook supplement is an introduction to causality in statistics and multivariable methods for students in their first college course in statistics. Traditionally, except for a terse warning from instructors that "correlation does not imply causation", students only encounter causality in statistics if they take graduate-level courses in certain disciplines (economics, epidemiology, etc.). However, given the ubiquity of data-driven arguments in modern society, a deeper understanding of drawing cause-and-effect conclusions from data is a core competency of college graduates and deserves attention in the liberal arts curriculum. For more on causality in the undergraduate curriculum, see our paper (Cummiskey et al., 2020) and others on this subject (Horton, 2015; Kaplan, 2018; Lübke et al., 2020).

After introducing concepts in causality, we use an example-based approach to multiple regression emphasizing how the scientific goal of the study impacts modeling decisions and interpretation of results. Prediction and cause-and-effect are the two main scientific goals of studies using multiple regression. These goals determine how researchers use and interpret models. However, this distinction is rarely made in introductory courses. Too often, multiple regression in the introductory course focuses on prediction and much of the "thinking" outsourced to algorithmic searches. This narrow focus on prediction deprives students of the opportunity to exercise critical judgment and creativity. On the other hand, cause-and-effect studies require tying subject matter knowledge to developing causal models. During this process, students have to carefully consider relationships between variables (which is not particularly important in prediction), thus developing multivariable thinking, an important goal of the revised GAISE report (Carver et al., 2016). Our approach to multiple regression uses the scientific goal to motivate how we think about our data and models. Hernán et al. (2019) argue for this approach in data science education; we believe it is appropriate for introductory statistics courses also.

This etext is appropriate for a wide variety of introductory statistics courses including both algrebra-based and calculus-based courses. It assumes students understand basic concepts in data analysis, inference (confidence intervals and statistical tests for means and proportions), and simple linear regression.

Chapter 1

Introduction

1.1 Goals of quantitative research (describe, predict, cause-and-effect)

Researchers and practitioners in nearly every discipline employ quantitative methods to answer important questions and make decisions. Regardless if it is finance, medicine, psychology, etc., all researchers typically have one of three goals (Cozby and Bates, 2020) when they conduct quantitative research:

- **describe**
- **predict**
- **cause-and-effect**

For example, consider medical researchers investigating a newly discovered variant of the coronavirus. Initially, they focus on describe studies. *How prevalent is the variant in the population? How does the prevalence vary over space and time? What groups of individuals are most vulnerable to the variant? What else is associated with the variant?* Together, these studies help the researchers better understand the variant and identify important variables for refined predict studies.

Predict studies identify individuals most at risk for the variant or for transmitting it to others. Researchers select potential predictors as inputs to algorithms and statistical methods that transform them into estimates of each individual's risk for the variant. These methods rely upon associations between the inputs and the variant. In addition, these methods estimate the error associated with their predictions. Researchers try to ensure error estimates obtained within their study are good estimates of the error when their models are applied to other

populations. While predict studies help us identify who is at risk, they do not tell us the effect if we change something about them (quit smoking, lose weight, etc.) To answer these "what if?" questions, researchers use cause-and-effect studies.

Cause-and-effect studies try to estimate the effect of intervening or changing some aspect of an individual. These studies are tremendously important in public policy (where they are aptly called *intervention studies*) to justify legislation and regulations to reduce risk. Unlike prediction studies, which are highly algorithmic and only require researchers to identify appropriate inputs, cause-and-effect studies use subject matter knowledge of the causal structure of the inputs to select appropriate statistical models (Hernán et al., 2019). For example, to assess the affect of quitting smoking on coronavirus risk, we would have to understand why some people quit smoking and others do not, as these reasons themselves may be responsible for some of the reduction in risk we observe between the two groups. Also, unlike prediction studies which rely upon what we *see*, cause-and-effect studies require us to ask what would happen if we *do* something (Pearl and Mackenzie, 2018). Only human beings are capable of such counterfactual thinking ("what would have happened if") - even our most sophisticated computer algorithms require human supervision to address cause-and-effect questions.

Exercises

1. For the following studies, identify the goal (describe, predict, cause-and-effect) of the study. Briefly explain your choice:

- *A researcher obtains school records to collect information on the extracurricular activities of students.*
- *A researcher obtains school records to determine the impact of extracurricular activities on grades.*
- *A researcher obtains school records on extracurricular activities to identify students who are at risk to graduate.*

2. Select an area of interest to you. Write a sentence describing a study in your area of interest for each of the three goals (describe, predict, cause-and-effect)

1.2 Validity

- internal validity
- external validity

In an area of research, there is often a natural tradeoff between internal and external validity. Our knowledge comes from both observational studies and experiments (trials) – rarely does a single study answer every question and the two types of studies used together increase our overall knowledge. Consider researchers investigating the effectiveness of Covid-19 vaccines. Initial vaccine trials focus on internal validity to understand its effectiveness against the virus and obtain approval for mass distribution. For example, Polack et al. (2020) randomly assigned 43,548 persons 16 years of age or older to receive two doses of either placebo or the BNT162b2 vaccine (Pfizer). They observed eight cases of Covid-19 among participants who received BNT162b2 and 162 cases among those who received the placebo. This was a huge moment in the pandemic! However, the subjects in the study were relatively healthy and had volunteered for the trials. From this study alone, the external validity of these results is not clear and researchers want to know if the vaccine results would hold in the general public. Given the success of the trials, it would be unethical to withhold the vaccine from large segments of the population, so observational studies are required.

Dagan et al. (2021) conducted an observational study after the mass vaccination campaign in Israel using the health records of 4.7 million patients enrolled in its largest integrated health care organization. The researchers matched vaccine recipients to controls on important variables: age, sex, sector, neighborhood of residence, etc.. In this larger sample, they found vaccine effectiveness results consistent with the randomized trial, thus providing evidence of the vaccine's effectiveness in the general public. The authors specifically address the importance of observational studies:

> "Although randomized clinical trials [experiments] are considered the "gold standard" for evaluating intervention effects, they have notable limitations of sample size and subgroup analysis, restrictive inclusion criteria, and a highly controlled setting that may not be replicated in a mass vaccine rollout."

Exercises

1. Read the article "Dozens to be deliberately infected with coronavirus in UK 'human challenge' trials" in Nature News. Write a paragraph discussing issues of internal and external validity.

Chapter 2

Causality

The last few decades have seen a revolution in how statisticians view causality. A hundred years ago, when the statistical methods students learn in introductory courses were developed, causality was considered outside the realm of statistics, except for the case of randomized controlled experiments. In other words, statistics could only answer questions of association, but not of causality. However, this limited view of statistics was at odds with its usage in every day research. For example, the greatest public health triumph of the 20th century was the reduction in cigarette smoking, which exploded after World War II with their mass production and marketing. The statistical evidence of the health effects of smoking comes entirely from observational studies – there has never been a randomized controlled trial for smoking. The American Cancer Society's observational studies beginning in the 1950's were huge undertakings and provided compelling evidence of the harmful effects of cigarette smoking (Hammond and Horn, 1954; Hammond et al., 1966). Clearly, statistical theory and practice had diverged in their understandings of causality.

However, in the 1980's and 1990's, researchers in different disciplines began revisiting causality (Greenland and Robins, 1986). They developed mathematical language for expressing causation, which cannot be uniquely expressed using the traditional language of association. In addition, they showed that randomized controlled trials are special cases of more general situations when the researcher has full knowledge of the assignment mechanism. The *assignment mechanism* is the process in which subjects are assigned to different levels of the treatment. Lastly, they showed that causal effects could be estimated from observational studies under a wide variety of circumstances when the assignment mechanism is known. Today, while "correlation does not imply causation" is still useful advice when assessing causal claims in observational studies, statistical theory and practice suggest our assessment of causal claims in observational studies should be much richer and nuanced than this simple rule of thumb.

(add citations)

2.1 What does it mean for one thing to cause another

We say one variable (the treatment or intervention) *causes* another variable (the outcome) if there is a change in the average outcome between subjects when they receive the treatment and the same subjects when they do not receive the treatment. This definition differs from *association*, which is a change in the average outcome between subjects who received the treatment and different subjects who did not receive the treatment. Thus, causation is a comparison of observed outcomes and their counterfactuals ("what would have happened if the subject were in the other treatment group").

Unfortunately, in most cases, we cannot observe both outcomes for subjects. For example, when estimating the effect of smoking on long term health outcomes, it is impossible to observe the same subject as a smoker and as a nonsmoker. We only observe one of the outcomes. However, under certain circumstances, we can obtain good estimates of effects without observing both outcomes for each individual. The most famous of these is the randomized controlled experiment.

(add see and do from Pearl here?)

2.2 Randomized controlled experiments

One of the most important scientific discoveries of the early 20th century was the randomized controlled trial (RCT). In its simplest form, researchers randomly assign subjects to receive the treatment or be in the control group. If they observe a difference in average outcomes between the two groups, then we would say the treatment caused the outcome. *Why does assigning subjects to groups by the simple action of flipping a coin result in such a radical difference in how we interpret the results?* The answer lies in the definition of causation above. Causation compares the outcomes between the same subjects. When we randomize the treatment, we end up comparing the outcomes between one group of subjects with the treatment and another group of subjects without the treatment who we expect to be very similar. In fact, when we have large enough sample sizes, it would be very unusual for the two groups to differ much. We refer to the two groups as *exchangeble*. In other words, we would expect the control group to have had similar results as the treatment group if they were the treatment group, and vice versa.

However, an overly restrictive view of causality followed this important discovery. That is, causality can *only* be shown with RCTs. This placed a huge limitation on the types of research questions statistics could address. Frequently, RCTs are not ethical, feasible, or desirable. Imagine enrolling in a study where you could be randomly assigned to be a smoker for the next 20 years. Towards

the end of the 20th century, researchers began taking a more expansive view of causality in observational studies.

2.3 Observational Studies and Confounding

In observational studies, researchers do not intervene on the assignment of subjects to treatment and control groups. (Note: a common misconception is that observational studies do not have treatment and control groups. This is not true. It is about how subjects are *assigned* to the treatment groups.) Instead, other factors determine subjects' group (treatment or control) assignment. Confounding occurs when these other factors determining assignment are themselves causes of the outcome. In other words, the treatment and control groups are different in ways that are important to the outcome. The two groups are not exchangeable. An observed association between the treatment and outcome could mean (1) the treatment caused the outcome, (2) other factors causing group assignment caused the outcome, or (3) both. Furthermore, with only information on the treatment and outcome, it is not possible to identify which of the three is the correct explanation. Confounding is a form of statistical bias – using the observed association as an estimate of the treatment effect will be systemically off. Increasing the sample size does not help fix bias, you just get a more precise, wrong estimate.

For example, consider an observational study investigating long term health effects of smoking. In many populations, males are more likely to be smokers. In addition, males have different risks for long term health outcomes than females, regardless of whether they smoke. If we observe an association between smoking and an outcome without information on sex, we cannot distinguish the effect of smoking from the effect of sex.

However, researchers in the late 20th century had a key insight. If we know the assignment mechanism and measure a sufficient set of confounding variables, you can obtain good estimates of treatment effects from observational studies. This was huge! Observational studies and RCTs are not fundamentally different. Estimating effects requires understanding the assignment mechanism, and RCTs are just a special case with a simple, known assignment mechanism. Given this insight, researchers became more comfortable making causal claims from observational studies when they have knowledge of the assignment mechanism. *How do we identify a sufficient set of confounding variables?* For that question, we turn to causal diagrams.

(discuss well-defined interventions?)

2.4 Causal Diagrams

Causal diagrams are useful tools for depicting the assignment mechanism (also called the causal model). Experts use subject matter knowledge in their field to specify the causal model. They typically specify the causal model prior to collecting data to identify confounding variables to measure. Importantly, there is no way to determine the presence of unmeasured confounding using the data. In addition, using simple heuristics for causal diagrams, they identify a sufficient set of confounding variables to control for during design and analysis.

Causal diagrams are directed, acyclic graphs (DAGs) where the nodes are variables and a directed edge (arrow) connecting two nodes indicates the node at the arrow's tail is a cause of the node at the arrow's head. The graphs are directed because the arrows point in one direction. They are acyclic because you can never get back to where you started by following arrows. The convention in many disciplines is to order the variables temporally from left to right – we adopt that convention here.

There are three building blocks of causal diagrams.

2.4.1 Confounding variable

Figure 2.1 depicts the confounding variable C of the effect of treatment X on outcome Y. We will observe an association between X and Y even if there is no treatment effect. The levels of X differ in terms of C and C itself is a cause of Y. Without measuring and controlling for C, we cannot distinguish the effect of X on Y from the association through confounding variable C. However, if C is the only confounding variable, controlling for it will result in good estimates of the effect of X on Y.

(backdoor path)

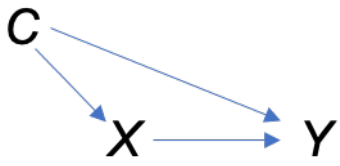


Figure 2.1: A confounding variable C on the effect of X on Y.

For example, let's say researchers investigate the effect of a master's degree on adult earnings. They survey a large number of individuals and record whether they have a master's degree and their earnings. Socioeconomic status is a confounding variable. Individuals with higher socioeconomic status are more likely

to earn master's degrees. In addition, they are more likely to have higher adult earnings, regardless of whether they have a master's degree. For this example, let's assume socioeconomic status is the only confounding variable. If the researchers observe an association between master's degrees and earnings without information on socioeconomic status, it is not possible to determine if the association is due to an effect of master's degrees or the effect of socioeconomic status.

2.4.2 Collider

Figure 2.2 depicts treatment X with no effect on outcome Y. X and Y are both causes of collider Z (the two incoming arrows *collide*). The box around Z indicates conditioning upon Z in the analysis. In this case, there will be an association between X and Y even though there is no treatment effect. Figure 2.2 is a common way to depict *selection bias* where an association in subjects selected for the study is not present in the general population. In the selection bias diagram, Z is an indicator of selection into the study with a box around it because researchers only observe subjects selected into the study.

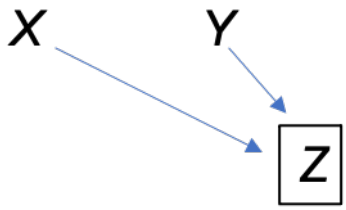


Figure 2.2: Collider Z with no treatment effect of X on Y.

A very famous example appears in Pearl and Mackenzie (2018). In the general population, we would not expect to see an association between an individual's talent and looks. More talented people are not better looking and vice versa. However, if we look only at famous Hollywood actors (Z), we would see an association between talent and looks in this group because an individual needs either talent or looks (or both) to become a famous Hollywood actor. More seriously, *selection bias* is a huge issue in medical and public health studies resulting in easily misinterpretated associations (Hernán et al., 2004; Cole et al., 2010; Elwert and Winship, 2014).

2.4.3 Mediator

Figure 2.3 depicts mediator M of the effect of treatment X on outcome Y. In this case, X is a cause of M, M is a cause of Y, and there is no effect of X on

Y that cannot be explained by the effect of M on Y. A common mistake is to adjust for M. If M is categorical, we will not observe an association between X and Y within levels of M. However, there is still an effect of X on Y.

Figure 2.3: Mediator M of the effect of treament X on outcome Y.

For example, let's say researchers are investigating an adverse reaction to a medication using a treatment group and a control group where the adverse reactions are always preceeded by increased blood pressure. Even if there is an effect of the medication on the adverse reason, when we condition upon experiencing increased blood pressure, we will not observe an association between the medication and adverse reactions.

Unfortunately, some researchers condition upon everything that they measure, often resulting in poor estimates of effects (Hernán et al., 2002).

Chapter 3

Matching

Chapter on matching

Chapter 4

(COL Watts) Quantitative explanatory variable with categorical confounding variable

Introduce multiple regression by adding a categorical confounding variable

Explanatory (X) - quantitative

Response (Y) - quantitative

Confounding (C) - categorical

4.1 Unadjusted effect of X on Y

short review of simple regression

4.2 Effect of X and Y adjusting for C.

4.3 Assessing model adequacy

Chapter 5

(Kevin) Interactions

Extend the example in the last chapter...introduce in terms of effect modification. We have a new research question, "does the affect of X on Y change based on another variable of interest?"

Chapter 6

Activity

Title: The Indoor Obstacle Course Test (IOCT)

Topics}*: Confounding, Causal Diagrams, Simple Linear Regression, Confidence Intervals

Background: Cadets at West Point must pass the Indoor Obstacle Course Test (IOCT) to graduate. The IOCT begins with a series of floor and climbing obstacles and ends with several laps around an indoor track. It is an exhausting test of endurance and strength. In addition to being a graduation requirement, cadets receive a letter grade that is factored into their class rank.

Shorter cadets often argue they are at a disadvantage on the obstacle course. Many obstacles appear to favor taller cadets because they are easier to reach. In this study, we will investigate the effect of height on IOCT times.

1. Watch the video of Cadet Madaline Kenyon running the IOCT. In your opinion, do some obstacles favor taller cadets? Explain.

The file `obstacle_course.csv` contains height (inches), IOCT times (seconds), biological sex (M/F), and whether the cadet played an intercollegiate sport for a sample of 384 cadets who ran the IOCT course in the last five years.

2. What is the explanatory variable in this study? Classify the variable as quantitative or categorical.

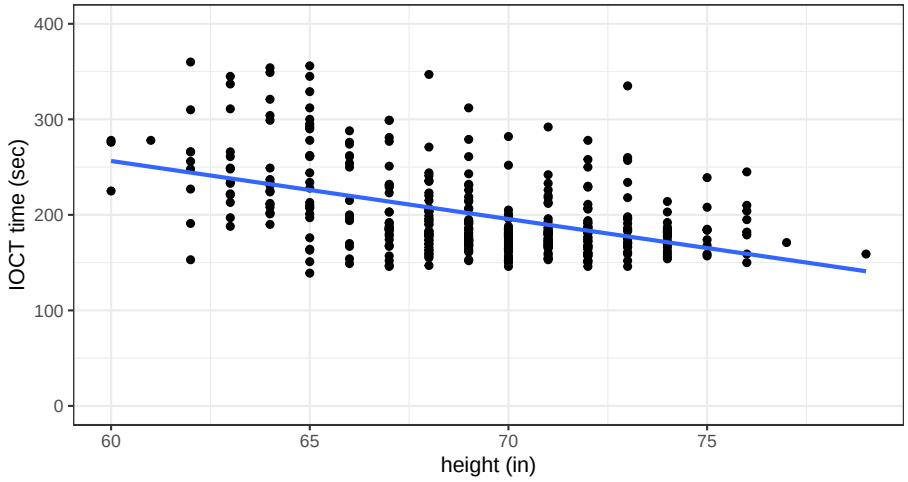
3. What is the response variable in this study? Classify the variable as quantitative or categorical.

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

Table 6.1: Linear regression output for IOCT times and height.

term	estimate	std.error	statistic	p.value
(Intercept)	621.00	39.98	15.53	0
height	-6.08	0.58	-10.52	0

Figure 1 depicts IOCT times in seconds versus height in inches. Table 1 contains information from the linear regression model.



5. Interpret the estimate of the height coefficient in Table 1.

6. Calculate and interpret a 95% confidence interval for the slope coefficient.

7. The p-value for height in Table 1 indicates there is strong evidence of an association between height and IOCT time. Taller cadets tend to do better on the IOCT. Some people would say the result is *statistically significant*. However, statistical significance and practical signifigance are different. Review the grade scale for the IOCT. In your opinion, does the observed association have practical significance? Explain.

8. A shorter cadet argues Figure 1 shows evidence the IOCT is unfair based on height. Do you agree or disagree? Explain.

9. Briefly explain the difference between these two conclusions.

- *Height is associated with faster IOCT times.*
- *Height causes faster IOCT times.*

10. Based on the analysis presented thus far, is it possible to distinguish between these two explanations? Explain.

11. Draw a causal diagram depicting the relationship between height, IOCT time, and sex. Explain your decisions to include/exclude arrows in the diagram.

12. Based on your diagram, identify the confounding variable.

Below are boxplots of height in inches and IOCT times in seconds by sex.

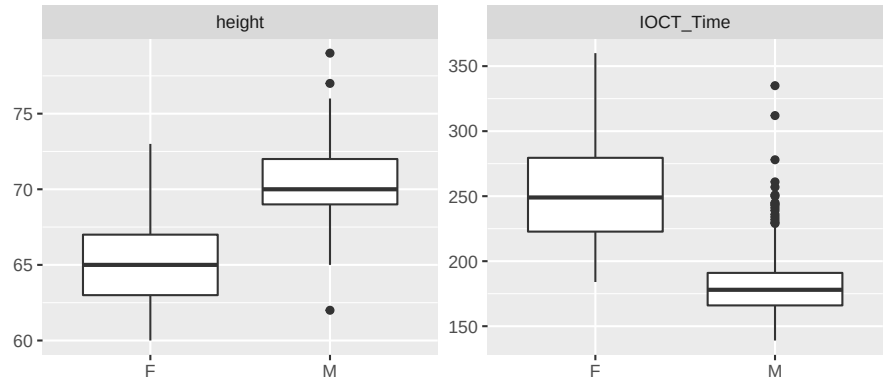


Figure 6.1: Height (inches) and IOCT time (seconds) by sex.

13. Based on Figure 2, is the estimate of the effect of height on IOCT time in Table 1 confounded by sex? If so, is the effect of height smaller or larger than that reported in Table 1? Explain.

Table 6.2: Regression results for female cadets.

term	estimate	std.error	statistic	p.value
(Intercept)	335.76	107.44	3.12	0.00
height	-1.24	1.64	-0.76	0.45

Table 6.3: Regression results for male cadets.

term	estimate	std.error	statistic	p.value
(Intercept)	175.61	40.92	4.29	0.00
height	0.09	0.58	0.15	0.88

Figure 3 depicts the association between IOCT time and height by sex. Tables 2 and 3 depict regression results for female and male cadets, respectively.

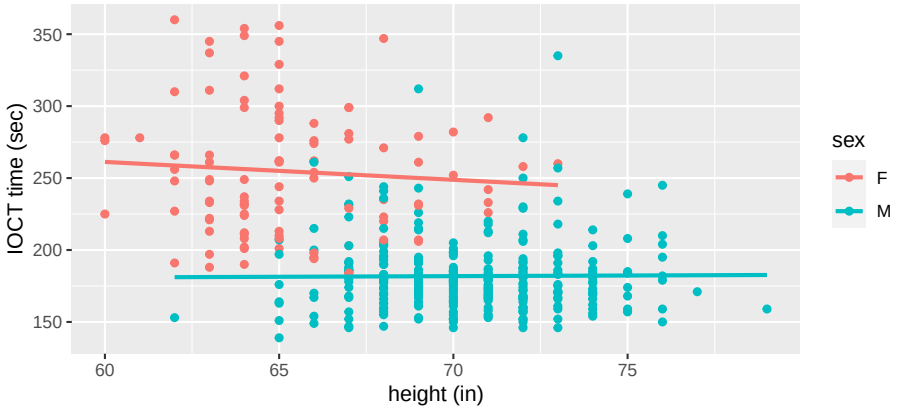


Figure 6.2: Indoor Obstacle Course Test (IOCT) times versus height by sex (n = 384).

14. Based on Figure 3 and Tables 2 and 3, does it appear there is an association between IOCT time and height within levels of sex? Explain.

15. In your opinion, is there much evidence that height is an advantage on the IOCT (in other words, is height the *cause* of better IOCT times)? Explain.

16. Briefly discuss two ways you could improve this study to better assess whether there is a height advantage.

Chapter 7

(COL Watts) Quantitative explanatory variable with quantitative confounding variable

Explanatory (X) - quantitative

Response (Y) - quantitative

Confounding (C) - quantitative

7.1 Effect of X and Y adjusting for C.

7.2 Assessing model adequacy

Bibliography

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G. H., Velleman, P., Witmer, J., et al. (2016). Guidelines for assessment and instruction in statistics education (gaise) college report 2016.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., and Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International journal of epidemiology*, 39(2):417–420.

Cozby, P. C. and Bates, S. (2020). *Methods in behavioral research*. McGraw-Hill Education.

Cummiskey, K., Adams, B., Pleuss, J., Turner, D., Clark, N., and Watts, K. (2020). Causal inference in introductory statistics courses. *Journal of Statistics Education*, 28(1):2–8.

Dagan, N., Barda, N., Kepten, E., Miron, O., Perchik, S., Katz, M. A., Hernán, M. A., Lipsitch, M., Reis, B., and Balicer, R. D. (2021). Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, 40:31–53.

Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, 15(3):413–419.

Hammond, E. C. et al. (1966). Smoking in relation to the death rates of one million men and women. *Natl Cancer Inst Monogr*, 19(166):127–204.

Hammond, E. C. and Horn, D. (1954). The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. *Journal of the American Medical Association*, 155(15):1316–1328.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, pages 615–625.

Hernán, M. A., Hernández-Díaz, S., Werler, M. M., and Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American journal of epidemiology*, 155(2):176–184.

Hernán, M. A., Hsu, J., and Healy, B. (2019). A second chance to get causal inference right: a classification of data science tasks. *Chance*, 32(1):42–49.

Horton, N. J. (2015). Challenges and opportunities for statistics and statistical education: looking back, looking forward. *The American Statistician*, 69(2):138–145.

Kaplan, D. (2018). Teaching stats for data science. *The American Statistician*, 72(1):89–96.

Lübke, K., Gehrke, M., Horst, J., and Szepannek, G. (2020). Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education*, 28(2):133–139.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., et al. (2020). Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615.