Causality and Multiple Regression Supplement

Authors TBD

2021-03-06

Contents

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).

Prediction and cause-and-effect are the two main scientific goals of studies employing multiple regression. Importantly, researchers use and interpret regression models differently depending upon the goal. 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 judgement 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 emphasizes the scientific goal of the study and how it motivates researchers' thinking about their 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.

(add a mention to the revised GAISE report.)

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**

(discuss external and internal validity here?)

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 methods are applied to other populations. While predict studies help us identify who is at risk, they do not tell us the effect on their risk if we change something about them (quitting smoking, lost 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 expensive 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 Why multiple regression?

(need to edit the below from an email)

You are investigating the relationship between education and adult earnings. (Note this is almost always an observational study – a survey of sampled individuals with variables such as income, education, age, sex, etc.)

Let's assume you've never taken our course (or one like it). You can get pretty far investigating this data with concepts most students learn in high school. For example, you could compare adult earnings in college graduates to those of non-college graduates or find an estimate of earnings per year of education. If you are especially savy, you might recognize that college graduates and noncollege graduates are different in ways (age, sex, etc.) that are also important to adult earnings and repeat your original analyses within the different levels of these variables. To you (and most people outside our discipline), statistics is arithmetic, and it is hard to envision anything beyond arithmetic if that's all you know. However, there are important questions you cannot answer with this limited view of statistics:

(1) What range of population values for measures of association between education and adult earnings are compatible with your data?

(2) If there is no association between education and earnings, how extreme are your observed results?

(3) What is the association between education and earnings holding other important variables constant?

(4) How does the relationship between education and earnings depend upon other variables?

Most people's cynicism towards statistics is rooted in a lack of understanding that our discipline has answers to (1)-(4).

Thus far in our course, we have focused on (1) and (2) for various types of independent and dependent variables. And, if someone were to ask me, "what have your students learned in your course this semester?", my response would start with (1) and (2). These are important concepts…humans are not predisposed to think this way.

However, you still cannot answer (3) and (4) if you're in our course right now. (3) relates to confounding; (4) is interactions. The concepts of confounding and interactions are best understood before approaching multiple regression (a point we discuss at length here: https://www.overleaf.com/read/nkzpvpxwsjzq). However, multiple regression is the most flexible statistical model for estimating quantities in (3) and (4). Being able to answer (3) and (4) is a huge step up in a student's capability. Comparing earnings of college graduates to noncollege graduates is meaningless if these groups are vastly different with regards to age, sex, etc. Unless you have formal education in these methods, you probably don't know (3) and (4) are possible – it's outside what you can do with arithmetic.

Okay, so that's some discussion on why do we teach multiple regression? Now, what do we teach about multiple regression?

Broadly speaking, there are two major aspects to multiple regression: theory, application. Given our students and where this course fits into their education, I am an unabashed proponent of application being the primary focus of our course.

Examples of theory - derive the least squares estimates of the regression coefficients, understand the geometric interpretation of regression, etc.

Examples of application - fit and interpret a model to estimate the association between education and earnings that adjusts for age.

You can have a rich understanding of one of these two aspects and be a complete novice about the other. From my experience, mathematicians are rock stars in theory but haven't seen as much application (don't worry, you have more than enough for this course). For example, in application, a statistician would employ multiple regression in very different ways depending on whether the primary goal of the study is (a) prediction or (b) estimating the effect of an intervention. In (a), the primary concern is making smart use of the data to obtain "good" estimates of the out-of-sample prediction error with little care to what's in the model. In (b), the primary concern is confounding, which is a type of statistical bias, so variables are selected carefully by the researcher to reflect the intervention of interest. The distinction between (a) and (b) is beyond the scope of our course, but I'm just trying to illustrate the richness there is to both the theory and application of multiple regression.

So, back to our course, I think the Michigan house prices example in Intermediate Statistical Analyses by Tintle et. al. and Dan Baller's worksheet sums up well what I think students should get from our course. Specifically, they should get an appreciation that there are statistical models that (1) estimate measures of association while holding other variables constant and (2) see how these associations change based on other variables. For (2), an interaction between a quantitative and categorical variable is about as far as we can reasonably expect most students to get.

Chapter 2

(Kevin) Causality

2.1 What does it mean for one thing to cause another

2.2 Randomized controlled experiments

2.3 Observational Studies

2.4 Confounding

2.5 Causal Diagrams

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.

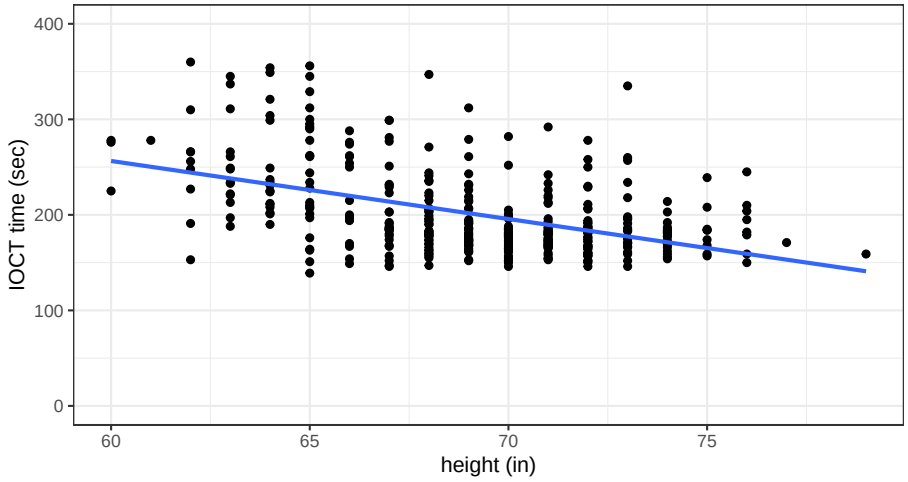
3. What is the response variable in this study? Classify the variable as quantitative or categorical.

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.

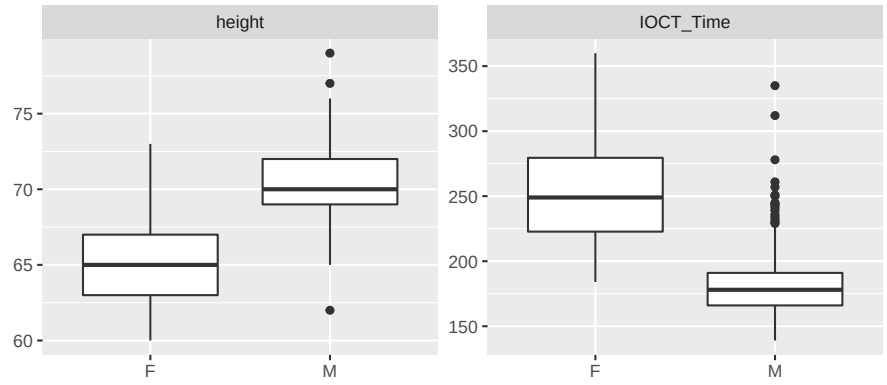


Figure 6.1: Height (inches) and IOCT time (seconds) by sex.

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.

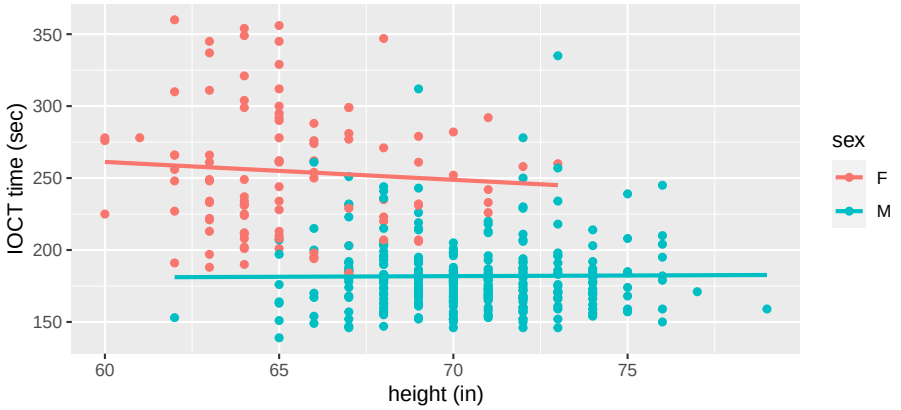


Figure 6.2: Indoor Obstacle Course Test (IOCT) times versus height by sex (n = 384).

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.

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.

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.