

Causality and Multiple Regression Supplement

Krista Watts and Kevin Cummiskey

2021-05-25

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 Confounding adjustment with regression	19
3.1 Categorical confounding variable	19
3.2 Quantitative explanatory variable with quantitative confounding variable	30
4 Interactions	37
4.1 Main effect of house size on sales price	39
4.2 Interaction between house size and location	40
5 Activity	45

Preface

This etext 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 algebra-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 employ quantitative methods to answer important questions and make decisions. Regardless of the discipline or application, they typically have one of three goals (Cozby and Bates, 2020):

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

Consider medical researchers investigating a newly discovered variant of the coronavirus. Initially, they might 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 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 individuals (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 risk for the variant, we need to understand why some people quit smoking and others do not, as these reasons themselves may be responsible for some changes in risk between the two groups. 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.

1.2 Validity

When designing a study, researchers carefully consider its internal and external validity.

- internal validity - How well does the study establish a cause-and-effect relationship between a treatment and outcome?
- external validity - How broadly can the study’s results be applied?

Often, there is a trade off between internal and external validity when designing a study. For example, to increase internal validity, researchers sometimes restrict the study to a subset of similar subjects to eliminate the effects of other variables. However, in doing so, they reduce the external validity of the study if the results do not apply to the other subjects. In an area of research, our knowledge comes from many studies with different levels of internal and external validity – rarely does a single study answer every question. Consider the recent development of Covid-19 vaccines. For example, initial Covid-19 vaccine trials focused on internal validity to understand its effectiveness against the virus and obtain approval for mass distribution. 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! Randomly assigning subjects to receive the vaccine increased the internal validity of the

study. However, for ethical reasons, the subjects had to be relatively healthy and volunteer for the trials. These limitations reduced the external validity of the study because it is not clear whether the vaccine has the same effect in the general population. However, conducting more randomized trials on other populations has ethical implications of its own, as they would require withholding the vaccine from large segments of the population. In this case, an observational study is more appropriate.

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. 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).
3. 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

There was a revolution over the last three decades in how scientists view causality. A hundred years ago, when the statistical methods we 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; Pearl, 1993; Angrist and Imbens, 1995). They developed mathematical language for expressing causation, which cannot be uniquely expressed using the traditional language of association. In addition, they showed 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 can 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 should be richer and more nuanced than this simple rule of thumb.

2.1 What does it mean for one thing to cause another

We say one variable (called the treatment, exposure, or intervention) *causes* another variable (the outcome or response) 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”).

The central challenge in assessing causality is that we cannot observe the outcomes for subjects under both levels of the treatment variable. In other words, we only get to observe each subject in either the treatment or the control group. 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 non-smoker. 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.

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 compare the outcome in one group of subjects with the treatment to 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 *exchangeable*. 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. Would you enroll 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. The distinction between randomized controlled experiments and observational studies is about how subjects are *assigned* to the treatment groups.) Instead, other factors determine subjects' treatment group assignment. For example, adolescents with a history of abuse, family violence, and stressful life events are more likely to smoke (Ellickson et al., 2001; Simantov et al., 2000).

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 systematically off. Increasing the sample size does not 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, we 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.

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. Using simple heuristics for causal diagrams, they help 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. 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 types of variables that are the 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 say there is a *backdoor path* between X and Y through C . 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 . When we control for C , we say the backdoor path from X to Y is “blocked”.

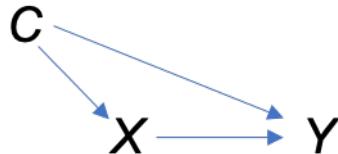


Figure 2.1: A confounding variable C on the effect of X on Y .

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.

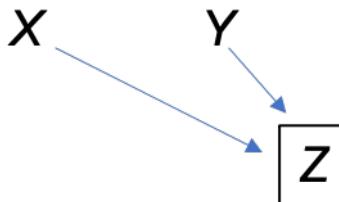


Figure 2.2: Collider Z with no treatment effect of X on Y .

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 a negative association between talent and looks in this group because an individual needs either talent or looks (or both) to become a famous Hollywood actor. Being less talented makes it more likely the actor is good looking because there must be some reason they became a famous Hollywood actor. More seriously, *selection bias* is a huge issue in medical and public health studies resulting in easily misinterpreted associations (Hernán et al., 2004; Cole et al., 2010; Elwert and Winship, 2014). For example, several studies have found a protective effect of smoking on Covid-19 infections (see Exercises).

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 we adjust for M , we will not observe an association between X and Y when there is an effect of X on Y .



Figure 2.3: Mediator M of the effect of treatment 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 preceded by increased blood pressure. Even if there is an effect of the medication on the adverse reaction, 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 they measure, often resulting in poor estimates of effects (Hernán et al., 2002).

Exercises

1. You investigate the effect of a college education on adult earnings by surveying 1000 adults in your area and ask them if they have a bachelor's degree (Y/N) and their yearly income in dollars. You find the college graduates earn about \$30,000 more per year on average than the high school graduates.
 - a) Name the treatment and outcome variables in your study.
 - b) Is this an observational study or randomized controlled experiment? Explain.
 - c) Explain what it means for college education to cause higher adult earnings. In addition, explain how this differs from college education being associated with higher adult earnings.
 - d) Identify and explain a potential confounding variable in your study. (Be sure to explain how your confounding variable is a common cause of both the treatment and outcome).
 - e) Draw a causal diagram depicting the relationship between the treatment, outcome, and your confounding variable.

- f) You also ask whether the individual has a Master's degree (Y/N). What type of variable is having a Master's degree? Redraw your diagram with this variable included in it.
2. Several studies have found a negative association between smoking and Covid-19 infection. Read "The spectre of Berkson's paradox: Collider bias in Covid-19 research" and answer the following questions.
- a) Explain what it means for smoking to cause a reduction in Covid-19 infections.
 - b) What type of variable is hospitalization in Figure 2?
 - c) Explain why many studies observe a negative association between smoking and Covid-19 infection.

Chapter 3

Confounding adjustment with regression

Recall from Chapter 2 that confounding occurs when the treatment variable depends upon another variable that is itself a cause of the outcome. When researchers have not controlled for a confounding variable through study design, they employ statistical methods during analysis to adjust for confounding. One of the most common techniques is multiple regression.

3.1 Categorical confounding variable

Let's say we have treatment variable (X) and response variable (Y) but the relationship between them is confounded by a third variable (C). For instance, suppose we want to investigate the relationship between ice cream sales and hospital admissions for heart attacks in the summer in NYC. In this example, our "treatment" is volume of ice cream sales on any given day in NYC and our response is number of hospital admissions that day for heart attacks. If we only look at those two variables, we might see a strong correlation; as ice cream sales increase, so do heart attacks. But it would be a mistake to infer that eating ice cream causes heart attacks from this analysis alone.

What does it mean for ice cream sales to *cause* heart attacks? How is that different from an association between ice cream sales and heart attacks? Ice cream sales cause heart attacks if high and low ice cream sales on the same days resulted in different heart attack rates. However, we only get to observe each day with one value of ice cream sales. Therefore, a third variable may be a common cause of both ice cream sales and heart attacks. In this case, we say there is an association between ice cream sales and heart attacks.

In our case, the third variable is temperature. People tend to buy more ice cream when it is hot; if there also tends to be more heart attacks when it is hot, temperature confounds the relationship between ice cream sales and hospital admissions for heart attack. Recall the confounding diagram from Chapter 2. If we extend it to this example, we have

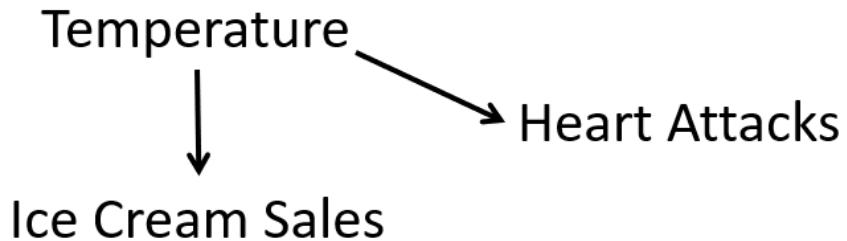


Figure 3.1: Causal diagram for ice cream sales and heart attacks.

Notice that we believe that a change in temperature *causes* a change in both ice cream sales and heart attacks.

If we control for temperature, then we will no longer see an association between ice cream sales and heart attacks. One way to control for temperature is to restrict the analysis to only days in a very narrow temperature range. However, restricting your analysis in this way would result in a lot of lost information. Instead, we turn to statistical methods like regression.

3.1.1 Unadjusted effect of X on Y

There exists a belief that height is protective against high blood pressure; that is, taller people tend to have lower blood pressure. In this section, we investigate this belief using data gathered from 1999 to 2006 by the CDC's National Health and Nutrition Examination Survey (NHANES), a (relatively) representative sample of the adult U.S. population. The data contains variables on 12,671 Americans age 18 and older who were not pregnant or taking high blood pressure medication. We are interested in estimating the effect of height (HT) on mean systolic blood pressure (SBP). We could model this with the following regression equation.

$$\widehat{SBP} = \beta_0 + \beta_1 * HT$$

If we fit a regression model with height as the treatment variable and systolic blood pressure as the response, we get the following results:

```

bp_dat<-read_csv("./data/blood_pressure.csv")
ht_mod=lm(SBP~HT,
           data = bp_dat)
summary(ht_mod)

##
## Call:
## lm(formula = SBP ~ HT, data = bp_dat)
##
## Residuals:
##    Min     1Q   Median     3Q    Max
## -52.876 -13.804  -3.703   9.922 141.605
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 141.22705   2.52887 55.846 < 2e-16 ***
## HT          -0.10276   0.01504 -6.833 8.58e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.09 on 17486 degrees of freedom
## Multiple R-squared:  0.002663, Adjusted R-squared:  0.002606
## F-statistic: 46.69 on 1 and 17486 DF, p-value: 8.575e-12

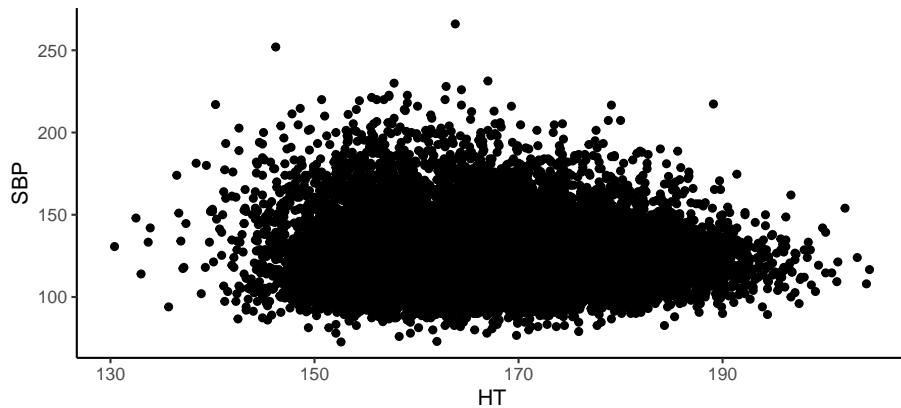
```

Note the coefficient on height is actually positive, meaning that taller people actually tend to have higher blood pressure! In fact, we expect SBP to go up by about 0.05 mmHG for every cm in additional height. This makes sense when we look at the data. There is A LOT of data here, but we see a general increase in SBP as height increases.

```

bp_dat%>%
  ggplot(aes(x=`HT`,
             y=`SBP`))+
  geom_point()+
  theme_classic()

```

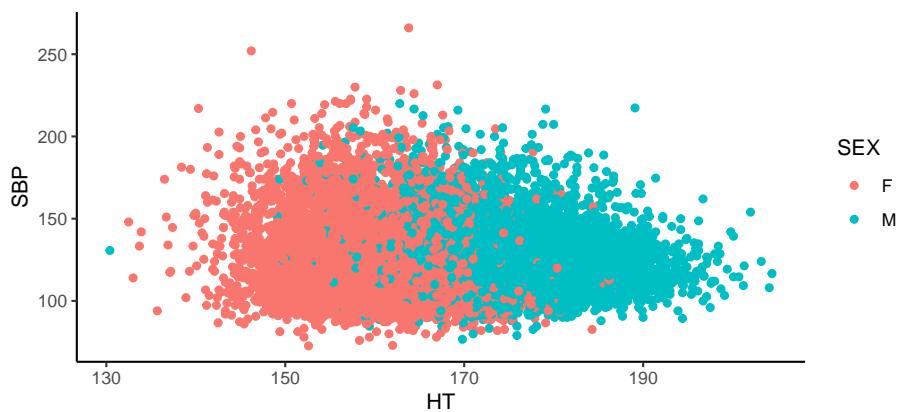


This is contrary to what we expected. So I guess previous research was wrong. Wait – not so fast. Biological sex is a potential confounder.

3.1.2 Effect of X on Y adjusting for C

Men tend to be taller than women (175 cm vs 162cm). Men also tend to have higher mean systolic blood pressure (122 vs 118 mmHG). Therefore, Sex is a possible confounder. Below is our data, colored by Sex.

```
bp_dat%>%
  ggplot(aes(x='HT',
             y='SBP'))+
  geom_point(aes(colour = SEX))+
  theme_classic()
```



Here, we see SBP tends to decrease as height increases *within men*. Likewise, *within women*, we see the same. If we want to estimate the effect of height

on blood pressure, we must first adjust for Sex. We do this using multiple regression. “Multiple” because we have multiple variables on the right hand side of the equation. (This could be multiple explanatory variables or, as we have here a single explanatory variable with one or more confounding variables.) Often we do this simply by adding the other variable to the regression equation. But how do we add a categorical variable to a mathematical equation. Suppose my regression model is

$$\widehat{SBP} = \beta_0 + \beta_1 * HT + \beta_2 * SEX$$

What does it mean to multiply something by “male”? When we want to include categorical variables as explanatory variables in regression models, we often use indicator (also called ‘dummy’) variables. In this case, biological sex has two levels, male and female, and we need one indicator variable to represent Sex. We will let our indicator variable equal 1 if a subject was male and 0 if the subject was female. β_2 is interpreted as the expected increase in SBP when a subject is male.

A possible multiple regression equation with SBP as our response, height as the treatment adjusting for Sex as a confounder is

$$\widehat{SBP} = \beta_0 + \beta_1 * HT + \beta_2 * SEX$$

When we add Sex to the model, we get the following results.

```
ht_gen_mod=lm(SBP~HT+SEX,
               data = bp_dat)
summary(ht_gen_mod)

##
## Call:
## lm(formula = SBP ~ HT + SEX, data = bp_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.085 -13.547  -3.851   9.538 144.487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 183.54338    3.24339   56.59 <2e-16 ***
## HT          -0.37869    0.02006  -18.87 <2e-16 ***
## SEXM        8.30226    0.40559   20.47 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 19.85 on 17485 degrees of freedom
## Multiple R-squared:  0.026, Adjusted R-squared:  0.02589
## F-statistic: 233.4 on 2 and 17485 DF, p-value: < 2.2e-16
```

We see that the relationship between height and SBP is now negative. Specifically, for every extra centimeter of height, we expect SBP to go down by 0.24 mmHG. Sex confounds the relationship between height and blood pressure. We can also note that the coefficient for Sex is a little over 9, meaning that on average, men have a SBP 9 mmHG higher than women.

Hmm, what are we forgetting...? Model diagnostics!

3.1.3 Assessing model adequacy

We have just done a theory-based test but we have not assessed whether the validity conditions are met. The validity conditions for multiple regression are analogous to those for simple regression. 1. Independence; once I have accounted for everything in the model, the responses can be considered independent of each other. 2. The model is linear; when plotting the residuals vs predicted values, there does not appear to be a pattern. 3. The residuals have constant variance; when plotting the residuals vs predicted values, there is a constant width. 4. Normality; a histogram of the residuals is approximately normal.

In practice, it is difficult to check for independence in the data. Unless there is a specific pattern to the way the data was collected, violations of this assumption may not be evident in our residual plots. The best way to verify this assumption is to know how the data was collected. Is it reasonable to consider it a random sample? If, for instance, we had repeated measurements on the same people, this assumption may not be reasonable. One way to check is to plot the residuals in the order they appear in your data. But if the dependence is not related to the order the data was stored, issues may not be visible in this plot.

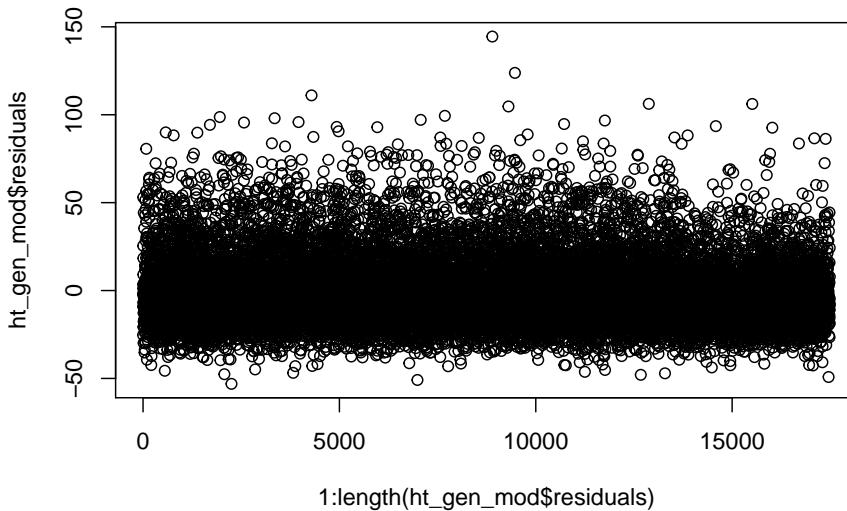
To check for linearity, we look at the residuals vs fitted values plot. Here we are looking for any pattern in the residuals. They should be scattered roughly evenly above and below the $y = 0$ line across the entire graph. (To check more rigorously in multiple regression we would actually check each explanatory variable individually using something called a partial residual plot or component plus residual plot. That is beyond the scope of this course.)

To check for constant variance, we again look at the residuals vs predicted values. We are checking to see if the variability in the residuals is approximately the same for different values of the response. A widening of our residuals as y_{hat} increases, for instance, would indicate this assumption is not met.

Finally, to check for normality, we look at a histogram of the residuals to see if it looks at least approximately normal.

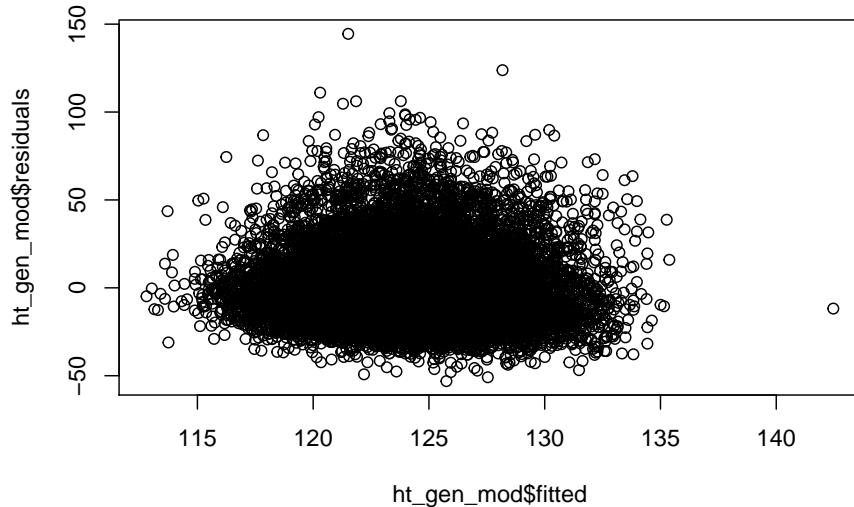
For our blood pressure example, we only have one observation on each person and we have a pseudo-random sample of the U.S. (Details on the exact sampling plan can be found on the CDC's NHANES website.) Additionally, if we plot the residuals vs their index (i.e. in the order they are in the data) we see no clear pattern to the residuals. It seems reasonable that our independence validity condition is met.

```
plot(1:length(ht_gen_mod$residuals),ht_gen_mod$residuals)
```



The following figure depicts the residuals vs fitted values from our model with both height and Sex. The residuals do not show a pattern and are scattered roughly equally above and below the $y = 0$ line. It seems reasonable to believe that the linearity validity condition is met.

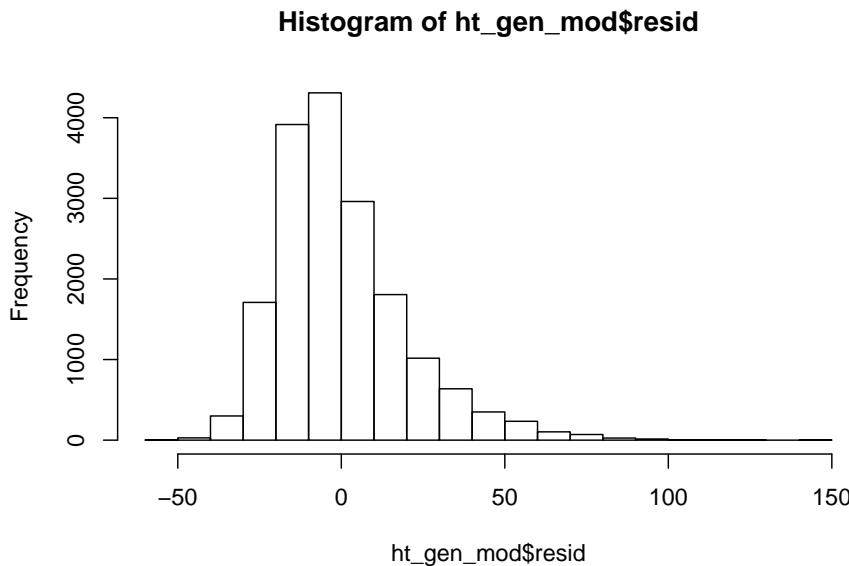
```
plot(ht_gen_mod$fitted,ht_gen_mod$residuals)
```



From this same plot, we see the variance of the residuals is relatively constant as our predicted SBP increases. The validity condition requiring equal variance seems reasonable.

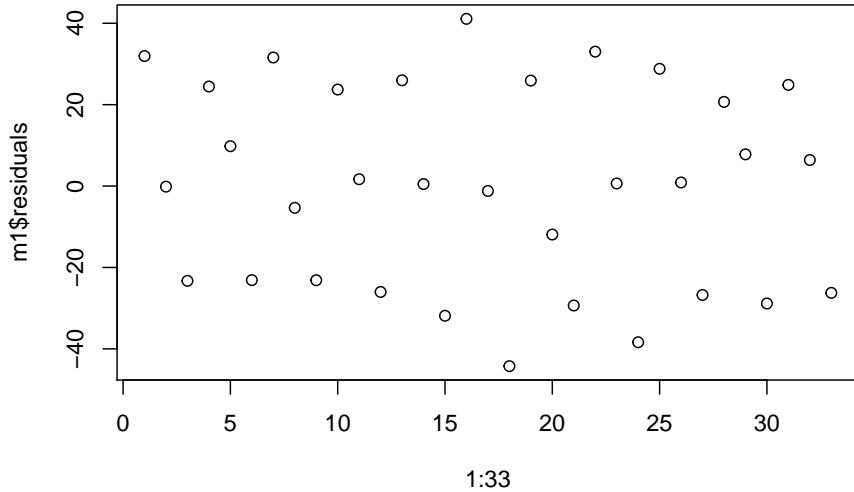
The histogram of our residuals looks relatively normal although it does seem to have a heavy right tail. Our residuals may not be strictly normal. However, linear regression is “robust” to departures from normality. That means that slight, or even moderate, departures from normality generally do not pose a problem. Our residuals look “normal enough” that I feel comfortable concluding that our validity conditions are met.

```
hist(ht_gen_mod$resid)
```



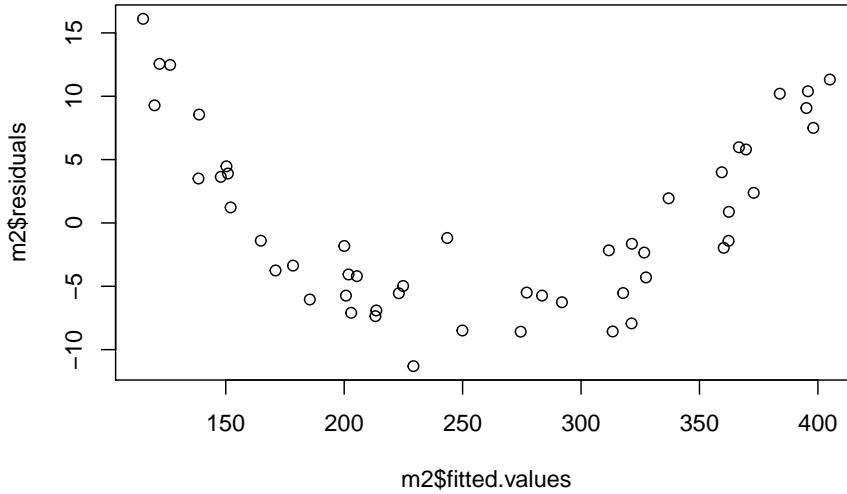
What would we expect to see if the validity conditions were violated? Below are some examples of model diagnostic plots that indicate a departure from our validity conditions. We simulated the data used to create these plots so we know which conditions are not valid. These are also relatively extreme examples for illustrative purposes; in real data you often will not see patterns this obvious. Violations of our model's validity conditions can be especially hard to detect when we have a small sample size.

First, we look at independence. The following plot of residuals in order show a clear pattern. The residuals seem to be in triplets; first we have a positive residual, then a residual around zero, then a negative residual. This pattern indicates that the model is missing an important factor, like time. Perhaps the data was a reading collected three times a day - morning, noon and night - and we expect the reading will differ based on time of day but we did not include that in the model. Departures from independence are rarely this easy to spot.



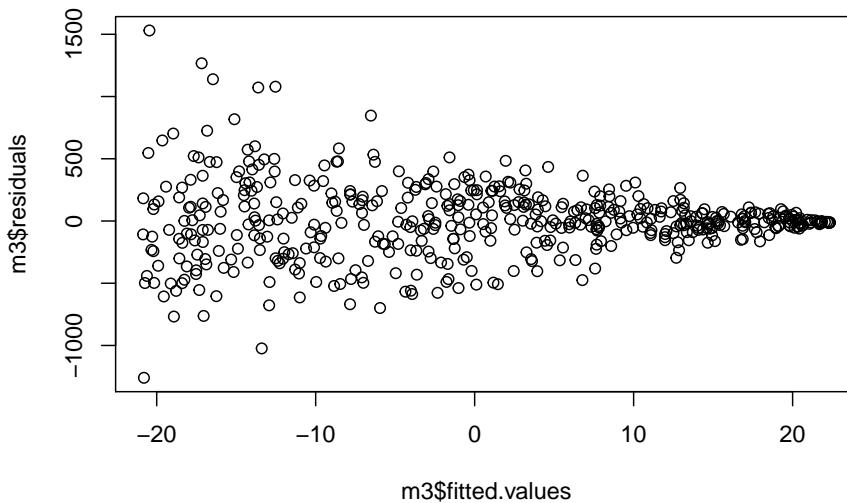
->

Next we look at the linearity assumption. Below we see a residuals vs fitted values plot that shows a clear pattern. Fitted values below about 150 or above about 350 tend to have positive residuals (i.e. be underestimated) whereas fitted values between 150 and 350 tend to have negative residuals (i.e. be overestimated). This is an indication that our linearity assumption is violated; the right hand side of the model is not correctly specified.

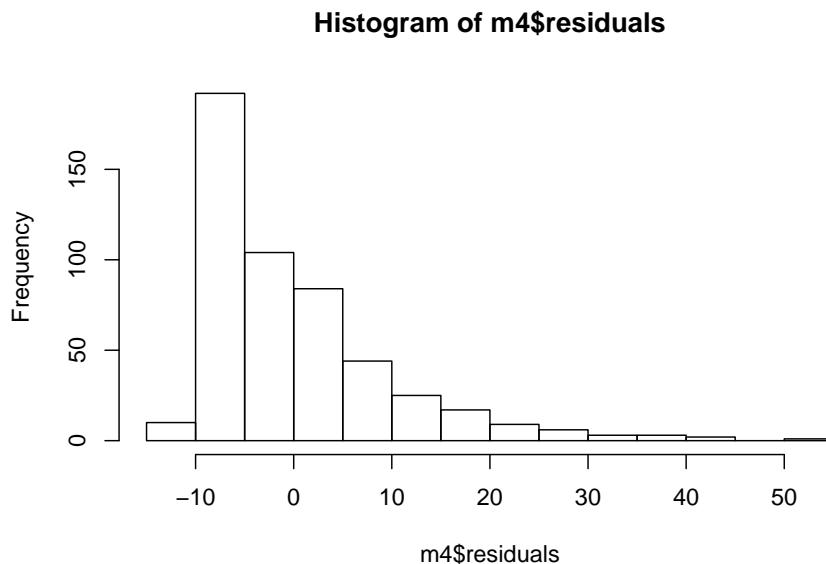


->

Next we investigate our assumption of constant variance. In the plot below, we see a clear fanning out of the residuals; as \hat{y} increases, the variance of the residuals decreases.



And finally, we investigate our normality assumption. In the plot below, we see that the residuals are heavily skewed to the right. The assumption of normality is violated.



In our example, researchers were specifically interested in the relationship between height and blood pressure so height was our treatment variable. Sex was related to both height and blood pressure so we need to include it in our model, but we are not particularly interested in the relationship between Sex and blood pressure from a clinical or scientific standpoint. So Sex was a confounding variable. Notice that we treat these variables the same from a statistical standpoint. It is possible to have multiple exposures or treatments, multiple confounders, or both. The distinction between the two is not a statistical one; it is a scientific one.

3.2 Quantitative explanatory variable with quantitative confounding variable

Confounding variables can, of course, be quantitative as well. Here we will explore the situation where you have a quantitative treatment variable (X), a quantitative response variable (Y), and a quantitative confounding variable (C). In fact, the ice cream example from the chapter 4 was just such a situation.

3.2.1 Unadjusted effect of X on Y

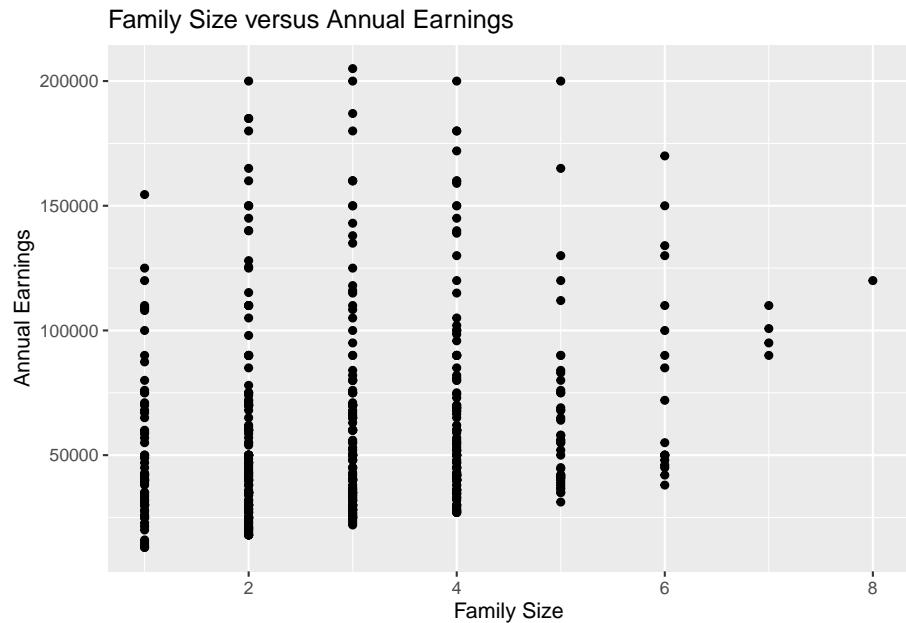
The file wage_data.csv on the course website contains a subset of wage data from the Center for Economic and Policy Research. It includes individual annual income from wages and salary (Earnings) for people who are above the poverty line but below the “top 2%”. [Do you think I should cite info for poverty line and top 2%?] We wish to explore the relationship between family size and earnings. A scatterplot indicates that there does appear to be a positive relationship between the two variables.

```
wage_dat<-read_csv("./data/wage_data.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Education = col_character(),
##   Sex = col_character(),
##   MaritalStatus = col_character(),
##   Race = col_character(),
##   FamilyMakeup = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
wage_dat%>%
  ggplot(aes(x=FamilySize, y=Earnings))+
  geom_point()+
  labs(x="Family Size", y = "Annual Earnings", title = "Family Size versus Annual Earnings")
```



If we were to consider only the treatment variable `FamilySize` and the response variable `Earnings`, we could use the following regression equation:

$$\widehat{Earnings} = \beta_0 + \beta_1 * FamilySize$$

Fitting this in R, we get the following results:

```
size_mod = lm(Earnings ~ FamilySize, data = wage_dat)
summary(size_mod)

##
## Call:
## lm(formula = Earnings ~ FamilySize, data = wage_dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41825 -28222 -13021  13383 141580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49012      4212 11.636 < 2e-16 ***
## FamilySize   4802       1272  3.775 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 39120 on 498 degrees of freedom
## Multiple R-squared:  0.02782,   Adjusted R-squared:  0.02587
## F-statistic: 14.25 on 1 and 498 DF,  p-value: 0.0001792
```

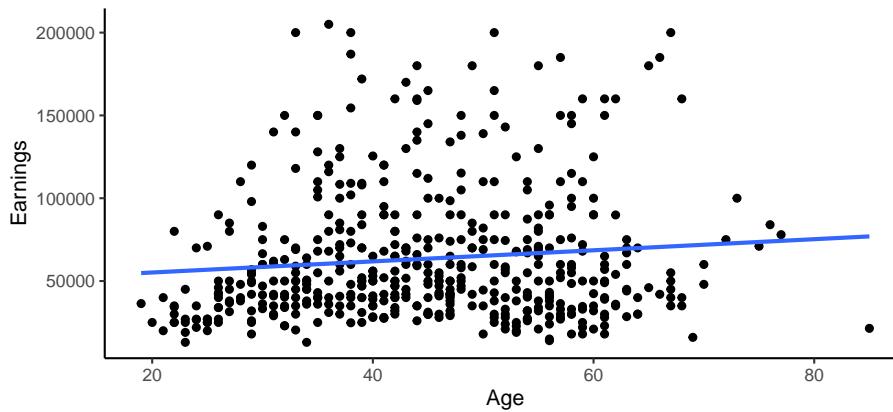
We see that there is a very significant effect of family size on earnings; for every additional member of the family, we expect to earn about \$48702 more.

3.2.2 Effect of X on Y adjusting for C

However, there are a number of factors that are related to both earnings and family size, perhaps in a causal way. We will look at age. We expect to earn more as we age and get more experience in the workforce and possibly even more education. We also expect that, in general, family size will increase - or at least stay the same - as we get older. We can visually inspect our data to check this. Below is a graph of earnings versus age.

```
wage_dat%>%
  ggplot(aes(x= `Age` ,
             y= `Earnings` ))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  theme_classic()

## `geom_smooth()` using formula 'y ~ x'
```



Age is a potential *confounder*. When we add age to the model with family size, we have the following regression model:

$$\widehat{Earnings} = \beta_0 + \beta_1 * FamilySize + \beta_2 * Age$$

When we estimate the coefficients for this model in R, we get the following results.

```
size_age_mod = lm(Earnings~FamilySize+Age, data = wage_dat)
summary(size_age_mod)

##
## Call:
## lm(formula = Earnings ~ FamilySize + Age, data = wage_dat)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -48975 -26711 -12103  14107 145564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26643.8     8322.0   3.202  0.00145 **
## FamilySize   5534.1     1283.0   4.314 1.94e-05 ***
## Age          449.7      144.7   3.107  0.00199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38790 on 497 degrees of freedom
## Multiple R-squared:  0.04635,   Adjusted R-squared:  0.04251
## F-statistic: 12.08 on 2 and 497 DF,  p-value: 7.551e-06
```

We see that there is still a significant positive effect of family size on earnings, although its magnitude is slightly larger. We expect to earn about \$5534 less per year for every additional member of the family *after adjusting for age*. We also note that the coefficient for age is 449.7, meaning that, for every year older a person is, we expect them to earn \$450 more on average, after adjusting for family size. Of course, the relationship between family size and earnings is a complex one that would require a much more extensive analysis to fully understand.

3.2.3 Assessing model adequacy

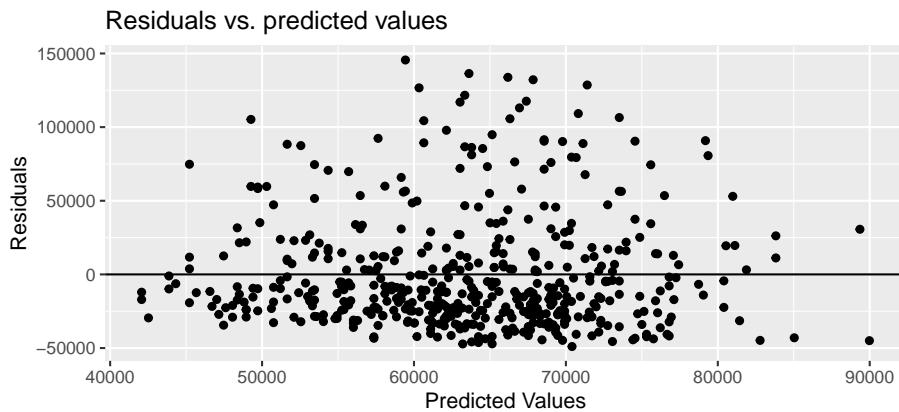
Just like with previous regression models, these results might not mean much if our validity conditions aren't met.

Here, we may be willing to believe that our assumption of independence is met by the way the data was collected. The CEPR data is from the Current Population Survey's Annual Social and Economic Supplement (CPS ASEC or March Supplement). Surveyors use a random sample to survey about 60,000 occupied households by both telephone and in person surveys. Households come

from all 50 states and the District of Columbia and are only in the sample pool for a limited time. Since respondents responses to survey questions are not dependent on another respondents, independence can be assumed.

To check for linearity and equal variance, we can plot the predicted values vs. the residuals.

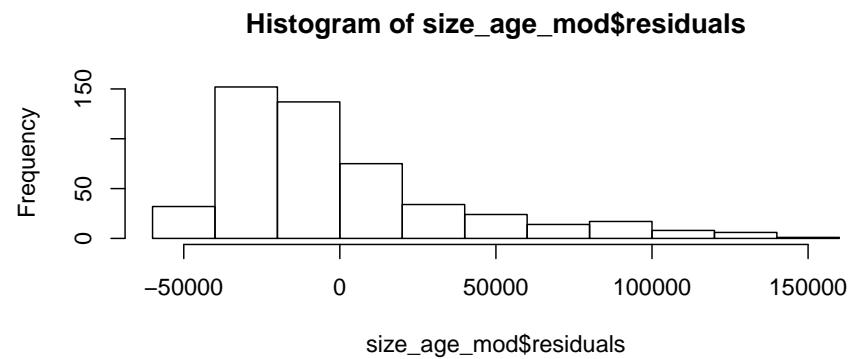
```
size_age_mod%>%
  fortify(wage_dat)%>%
  ggplot(aes(x = .fitted, y = .resid))+
  geom_point()+
  geom_hline(yintercept = 0)+
  labs( x = "Predicted Values", y = "Residuals", title = "Residuals vs. predicted values")
```



The linearity and constant variance conditions appear to be met. Although we have some large positive residuals, there does not appear to be a pattern and the width is relatively constant.

Our histogram, however, shows that the residuals seem to be skewed relatively heavily to the right. Our fourth validity condition, normality of the residuals, is likely not met. Although we can be comfortable in our estimates of the regression coefficients, we should be careful about drawing any conclusions about their statistical significance. (For a much more in depth assumption about assessing model adequacy, to include methods to handle violations of the validity conditions, take MA376!)

```
hist(size_age_mod$residuals)
```



Chapter 4

Interactions

Last chapter, we investigate the effect of a treatment on an outcome. In this chapter, we investigate how the effect of a treatment on an outcome changes with another variable. If the effect changes with this variable, we say there is an *interaction* between this variable and the treatment. The distinction between the two types of research questions is important.

Last chapter, the research question of interest is: *what is the effect of height on blood pressure?* Instead, we might ask: *is the effect of height on blood pressure different for males and females?* In other words, is there an interaction between height and Sex. This research question is fundamentally different because the primary interest is in how Sex *changes* the relationship between height and blood pressure. (Note some disciplines refer to Sex as an *effect modifier* because Sex *changes* or *modifies* the relationship between height and blood pressure.)

Identify which of the following research questions involve interactions.

- a) How does studying affect academic performance?
- b) Does the effect of studying on academic performance depend upon sleep?
- c) How much does a 500 square foot addition to a house increase its sales price?
- d) Is the increase in sales price for a 400 square foot addition different in the Town of Cornwall than the City of Newburgh?

Above, b) and d) are research questions involving interactions. In a), the primary question of interest is the effect of a treatment variable (time studying) on an outcome variable (academic performance). In b), we are interested in the interaction between sleep and time studying. In other words, we want to

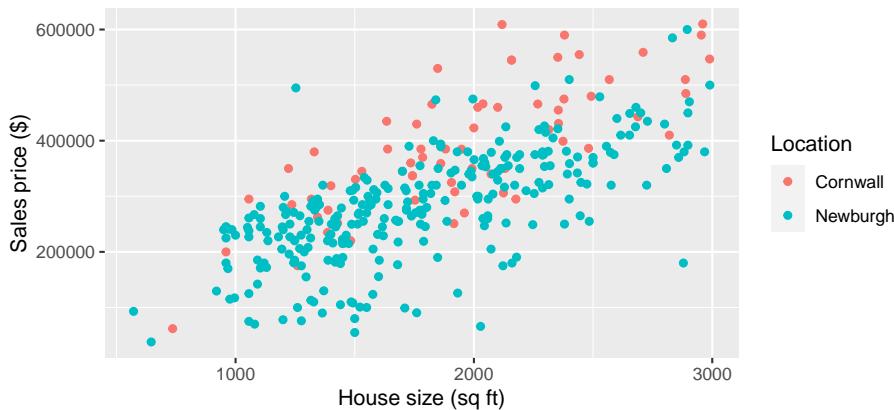
know if the amount of sleep modifies the effect of studying on academic performance. It seems likely that studying is more effective when you get more sleep and evidence of this interaction tells us something about how the brain retains knowledge. c) and d) follow a similar pattern. In c), we are interested in the effect of house size on sales price. In d), we are interested in the interaction between house size and location. In other words, we want to know if the location of the house changes the relationship between size and sales price.

A very common misconception is that an interaction between two variables implies the two variables are associated with each other. To see why this is not the case, consider research question d) above. An interaction between house size and location means the average sales price per one square foot increase in house size is different in Cornwall than Newburgh. An association between house size and location means houses are, on average, bigger in either Cornwall or Newburgh. Clearly, houses being more expensive per square foot increase in size in one location (an interaction) does not imply the houses are bigger in either location (association).

In the next section, we investigate an example of an interaction between house size and location in real estate transactions. Newburgh, New York is a small city (population: 30,000) along the Hudson River about 60 miles north of New York City. Nearby, the town of Cornwall, New York (population: 13,000) is also along the Hudson River with similar commute times to NYC. Despite their similar locations, real estate prices in the two locations are very different. The file houses.csv contains records of 356 single family homes with three or four bedrooms sold in Cornwall or Newburgh from May 2020 to May 2021 (source: redfin.com). The figure below depicts the sales price versus house size by location.

```
houses <- read_csv(file = "./data/houses.csv")

houses %>%
  ggplot(aes(x = `SQUARE FEET`,
             y = PRICE,
             color = CITY)) +
  geom_point() +
  labs(x = "House size (sq ft)",
       y = "Sales price ($)",
       color = "Location")
```



4.1 Main effect of house size on sales price

First, let's estimate the effect of house size on sales price using the model below. We'll refer to β_1 in this model as the *main effect* of house size.

$$\widehat{Price} = \beta_0 + \beta_1 Size$$

where *Price* is the sales price of the house in dollars and *Size* is the size of the house in square feet.

```
model_main_effect <- houses %>%
  lm(PRICE ~ `SQUARE FEET`, data = .)

summary(model_main_effect)

##
## Call:
## lm(formula = PRICE ~ `SQUARE FEET`, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -279057  -38960   2236  46628  274449 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 36757.744  14889.155   2.469   0.014 *  
## `SQUARE FEET` 146.683      7.963  18.421  <2e-16 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  
## Residual standard error: 78710 on 354 degrees of freedom  
## Multiple R-squared:  0.4894, Adjusted R-squared:  0.488  
## F-statistic: 339.3 on 1 and 354 DF,  p-value: < 2.2e-16
```

From the model output, the estimate of the main effect of house size on sales price is \$146.68 per square foot increase. In other words, the average sales price increases by \$146.68 per square foot increase in house size. If a homeowner adds a 20ft x 20ft addition, they can expect to increase the sales price of their home by about \$60000.

4.2 Interaction between house size and location

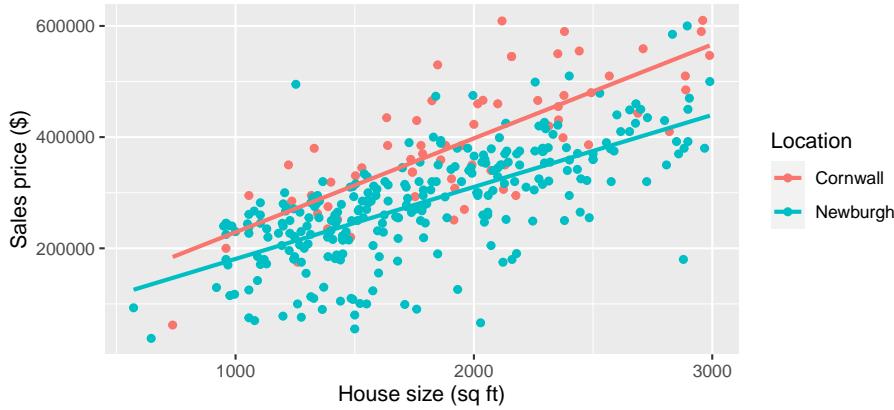
In the last section, we found the average sales price increases by \$146.68 for each square foot increase in house size. In this section, we ask whether there is an interaction between house size and location. In other words, is the increase in average sales price per square foot increase in house size the same in Newburgh as Cornwall?

The figure below depicts the sales price versus house size with a regression line for each location and shows some evidence of the following.

- The average sales price increases with house size because the lines have positive slopes. *We would say there is a main effect of house size.*
- The average sales price is lower in Newburgh than Cornwall because the blue line is lower than the red line. *We would say there is a main effect of location.*
- There is an interaction between house size and location because the two lines are not parallel. In other words, the increase in average sales price per square foot of house size is smaller in Newburgh than Cornwall. The slopes of the Newburgh and Cornwall lines are \$129.5 and \$168.84 per square foot increase, respectively. This is a huge difference! A homeowner in Newburgh building a 20ft x 20ft addition would expect to increase their sales price by about \$16000 less than a homeowner in Cornwall building the same addition.

```
houses %>%  
  ggplot(aes(x = `SQUARE FEET`,  
             y = PRICE,  
             color = CITY)) +  
  geom_point() +  
  labs(x = "House size (sq ft)",
```

```
y = "Sales price ($)" ,
color = "Location") +
geom_smooth(method = "lm", se = FALSE)
```



We can construct confidence intervals and perform hypothesis tests on the interaction (difference in slopes) between house size and location using a regression model. Below is a regression model with an interaction between house size and location. The quantity of interest in our analysis is β_3 , the coefficient on the interaction term.

$$\widehat{Price} = \beta_0 + \beta_1 Size + \beta_2 Location + \beta_3 Location * Size$$

where *Location* equals one if the house is in Newburgh and equals zero if in Cornwall. Interpretations of each term in the model are:

- β_0 : the average sales price of a house in Cornwall with zero square feet.
- β_1 : the change in average sales price for a one square foot increase in size for a house in Cornwall.
- β_2 : the difference in average sales price comparing a zero square foot home in Newburgh to a zero square foot home in Cornwall.
- β_3 : the difference in the change in average sales price for a one square foot increase in size comparing Newburgh to Cornwall.

Generally, we are not interested in β_0 , β_1 , and β_2 when the research question is about an interaction. Too often, even experienced researchers misinterpret these coefficients. For example, β_2 in this model is **not** the effect of location.

Why is β_3 the quantity of interest in this study? Recall we want to determine the difference between the slopes of the Newburgh and Cornwall lines. In the model above, the expected sales price of a home in Newburgh is:

$$\widehat{Price}_{newburgh} = \beta_0 + \beta_1 Size + \beta_2 * (1) + \beta_3 * (1) * Size = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Size$$

The slope of the Newburgh line is $\beta_1 + \beta_3$. The expected sales price of a home in Cornwall is:

$$\widehat{Price}_{cornwall} = \beta_0 + \beta_1 Size + \beta_2 * (0) + \beta_3 * (0) * Size = \beta_0 + \beta_1 Size$$

The slope of the Cornwall line is β_1 . Therefore, the difference in slopes comparing Newburgh to Cornwall is $(\beta_1 + \beta_3) - \beta_1 = \beta_3$.

The R code below fits the interaction model.

```
model_interaction <- houses %>%
  lm(PRICE ~ `SQUARE FEET` * CITY, data = .)

summary(model_interaction)

## 
## Call:
## lm(formula = PRICE ~ `SQUARE FEET` * CITY, data = .)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -248040  -33210    5729   46740   281326 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             60123.88   34377.89   1.749   0.0812 .  
## `SQUARE FEET`            168.84     16.92    9.976  <2e-16 *** 
## CITYNewburgh           -8718.48   37440.43  -0.233   0.8160    
## `SQUARE FEET`:CITYNewburgh -39.34     18.76   -2.096   0.0368 *  
## ---                     
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 71350 on 352 degrees of freedom
## Multiple R-squared:  0.5828, Adjusted R-squared:  0.5792 
## F-statistic: 163.9 on 3 and 352 DF,  p-value: < 2.2e-16
```

We see the change in average sales price per square foot increase in house size is about \$40 smaller in Newburgh than in Cornwall. The p -value for this coefficient indicates there is strong evidence of an interaction between house size and location.

Note the coefficient on $CITY$ has a large p -value. Many researchers misinterpret this to mean there is no effect of location on sales price. From the figure, it is very clear sales prices are higher on average in Cornwall than Newburgh for houses of similar size. The large p -value for the $CITY$ coefficient simply tells us that there is not evidence houses *with zero square feet* are more expensive in Newburgh than Cornwall. It is a meaningless quantity often misinterpreted as something important.

Chapter 5

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.

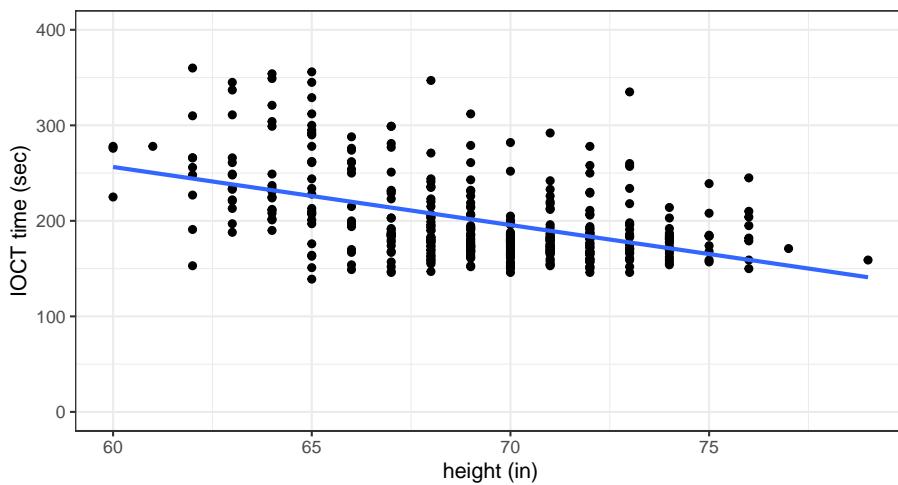
3. What is the response variable in this study? Classify the variable as quantitative or categorical.

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

Table 5.1: 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 significance 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.

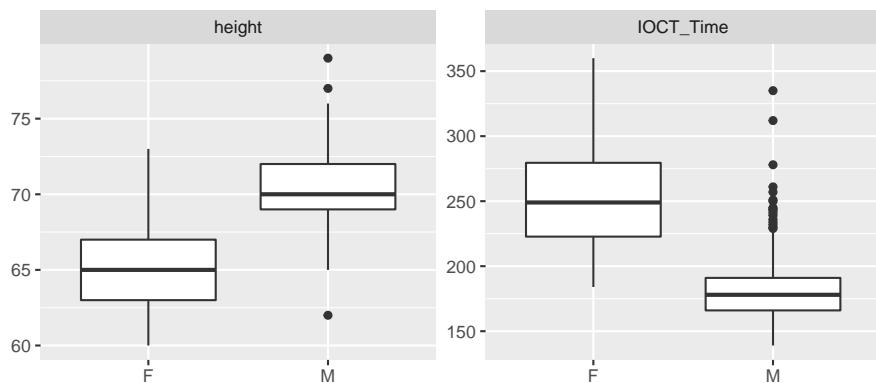


Figure 5.1: Height (inches) and IOCT time (seconds) by sex.

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.

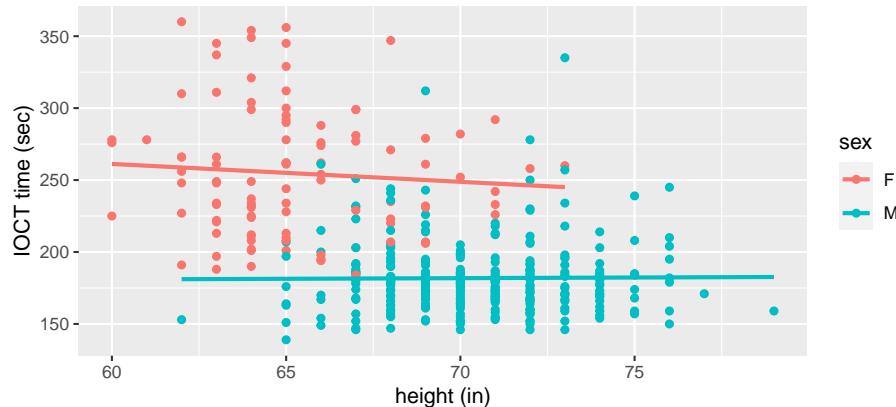
Table 5.2: Regression results for female cadets.

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 5.3: 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 5.2: 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.

Bibliography

- Angrist, J. and Imbens, G. (1995). Identification and estimation of local average treatment effects.
- 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*.
- Ellickson, P. L., Tucker, J. S., and Klein, D. J. (2001). High-risk behaviors associated with early smoking: results from a 5-year follow-up. *Journal of Adolescent health*, 28(6):465–473.
- 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. (1993). [bayesian analysis in expert systems]: Comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.
- 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.
- Simantov, E., Schoen, C., and Klein, J. D. (2000). Health-compromising behaviors: why do adolescents smoke or drink?: identifying underlying risk and protective factors. *Archives of pediatrics & adolescent medicine*, 154(10):1025–1033.