Causal Inference

The Mixtape

Scott Cunningham

Yale
UNIVERSITY PRESS
NEW HAVEN & LONDON

Potential Outcomes Causal Model

It's like the more money we come across, the more problems we see.
Notorious B.I.G.

Practical questions about causation have been a preoccupation of economists for several centuries. Adam Smith wrote about the causes of the wealth of nations [Smith, 2003]. Karl Marx was interested in the transition of society from capitalism to socialism [Needleman and Needleman, 1969]. In the twentieth century the Cowles Commission sought to better understand identifying causal parameters [Heckman and Vytlacil, 2007].[1] Economists have been wrestling with both the big ideas around causality and the development of useful empirical tools from day one.

We can see the development of the modern concepts of causality in the writings of several philosophers. Hume [1993] described causation as a sequence of temporal events in which, had the first event not occurred, subsequent ones would not either. For example, he said:

> We may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed.

Mill [2010] devised five methods for inferring causation. Those methods were (1) the method of agreement, (2) the method of difference, (3) the joint method, (4) the method of concomitant variation, and (5) the method of residues. The second method, the method of difference, is most similar to the idea of causation as a comparison among counterfactuals. For instance, he wrote:

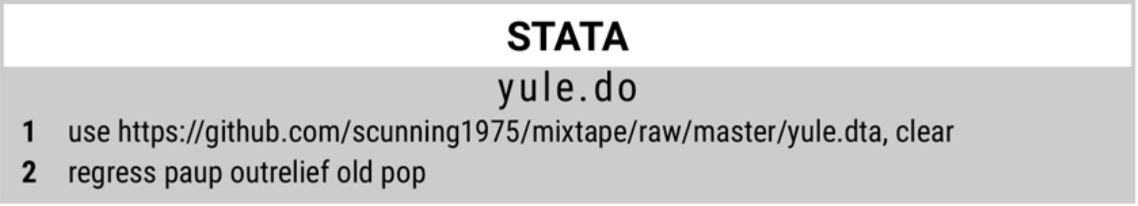
> If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death. [399]

Statistical inference. A major jump in our understanding of causation occurs coincident with the development of modern statistics. Probability theory and statistics revolutionized science in the nineteenth century, beginning with the field of astronomy. Giuseppe Piazzi, an early nineteenth-century astronomer, discovered the dwarf planet Ceres, located between Jupiter and Mars, in 1801. Piazzi observed it 24 times before it was lost again. Carl Friedrich Gauss proposed a method that could successfully predict Ceres's next location using data on its prior location. His method minimized the sum of the squared errors; in other words, the ordinary least squares method we discussed earlier. He discovered OLS at age 18 and published his derivation of OLS in 1809 at age 24 [Gauss, 1809].[2] Other scientists who contributed to our understanding of OLS include Pierre-Simon LaPlace and Adrien-Marie Legendre.

The statistician G. Udny Yule made early use of regression analysis in the social sciences. Yule [1899] was interested in the causes of poverty in England. Poor people depended on either poorhouses or the local authorities for financial support, and Yule wanted to know if public assistance increased the number of paupers, which is a causal question. Yule used least squares regression to estimate the partial correlation between public assistance and poverty. His data was drawn from the English censuses of 1871 and 1881, and I have made his data available at my website for Stata or the Mixtape library for R users. Here's an example of the regression one might run using these data:

$$\text{Pauper} = \alpha + \delta\text{Outrelief} + \beta_1\text{Old} + \beta_2\text{Pop} + u$$

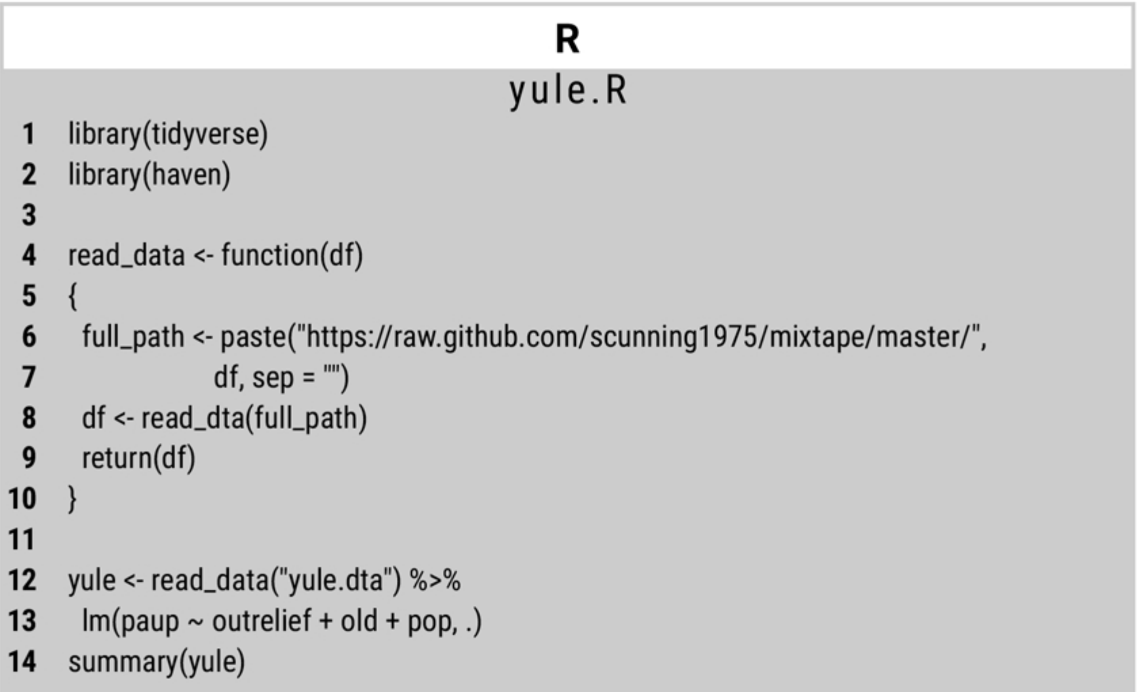
Let's run this regression using the data.



STATA

yule.do

```
use https://github.com/scunning1975/mixtape/raw/master/yule.dta, clear
regress paup outrelief old pop
```



R

yule.R

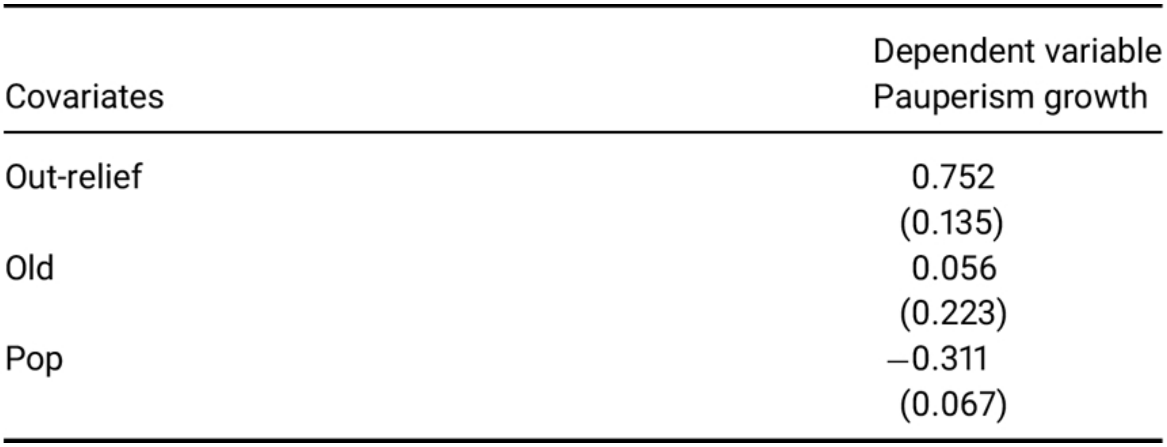
```
library(tidyverse)
library(haven)

read_data <- function(df)
{
  full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

yule <- read_data("yule.dta") %>%
  lm(paup ~ outrelief + old + pop, .)
summary(yule)
```

Each row in this data set is a particular location in England (e.g., Chelsea, Strand). So, since there are 32 rows, that means the data set contains 32 English locations. Each of the variables is expressed as an annual growth rate. As a result, each regression coefficient has elasticity interpretations, with one caveat—technically, as I explained at the beginning of the book, elasticities are actually *causal* objects, not simply correlations between two variables. And it's unlikely that the conditions needed to interpret these as causal relationships are met in Yule's data. Nevertheless, let's run the regression and look at the results, which I report in Table 10.

Table 10. Estimated association between pauperism growth rates and public assistance.



Covariates	Dependent variable Pauperism growth
Out-relief	0.752
	(0.135)
Old	0.056
	(0.223)
Pop	−0.311
	(0.067)

In words, a 10-percentage-point change in the out-relief growth rate is associated with a 7.5-percentage-point increase in the pauperism growth rate, or an elasticity of 0.75. Yule used his regression to crank out the correlation between out-relief and pauperism, from which he concluded that public assistance increased pauper growth rates.

But what might be wrong with this reasoning? How convinced are you that all backdoor paths between pauperism and out-relief are blocked once you control for two covariates in a cross-sectional database for all of England? Could there be unobserved determinants of both poverty and public assistance? After all, he does not control for any economic factors, which surely affect both poverty and the amount of resources allocated to out-relief. Likewise, he may have the causality backwards—perhaps increased poverty causes communities to increase relief, and not merely the other way around. The earliest adopters of some new methodology or technique are often the ones who get the most criticism, despite being pioneers of the methods themselves. It’s trivially easy to beat up on a researcher from one hundred years ago, working at a time when the alternative to regression was ideological make-believe. Plus he isn’t here to reply. I merely want to note that the naïve use of regression to estimate correlations as a way of making causal claims that inform important policy questions has been the norm for a very long time, and it likely isn’t going away any time soon.

Physical Randomization

The notion of physical randomization as the foundation of causal inference was in the air in the nineteenth and twentieth centuries, but it was not until Fisher [1935] that it crystallized. The first historically recognized randomized experiment had occurred fifty years earlier in psychology [Peirce and Jastrow, 1885]. But interestingly, in that experiment, the reason for randomization was *not* as the basis for causal inference. Rather, the researchers proposed randomization as a way of fooling subjects in their experiments. Peirce and Jastrow [1885] used several treatments, and they used physical randomization so that participants couldn't guess what would happen next. Unless I'm mistaken, recommending physical randomization of treatments to units as a basis for causal inference is based on Splawa-Neyman [1923] and Fisher [1925]. More specifically, Splawa-Neyman [1923] developed the powerful potential outcomes notation (which we will discuss soon), and while he proposed randomization, it was not taken to be literally necessary until Fisher [1925]. Fisher [1925] proposed the explicit use of randomization in experimental design for causal inference.[3]

Physical randomization was largely the domain of agricultural experiments until the mid-1950s, when it began to be used in medical trials. Among the first major randomized experiments in medicine—in fact, ever attempted—were the Salk polio vaccine field trials. In 1954, the Public Health Service set out to determine whether the Salk vaccine prevented polio. Children in the study were assigned *at random* to receive the vaccine or a placebo.[4] Also, the doctors making the diagnoses of polio did not know whether the child had received the vaccine or the placebo. The polio vaccine trial was called a *double-blind, randomized controlled trial* because neither the patient nor the administrator of the vaccine knew whether the treatment was a placebo or a vaccine. It was necessary for the field trial to be very large because the rate at which polio occurred in the population was 50 per 100,000. The treatment group, which contained 200,745 individuals, saw 33 polio cases. The control group had 201,229 individuals and saw 115 cases. The probability of seeing such a big difference in rates of polio because of chance alone is about one in a billion. The only plausible explanation, it was argued, was that the polio vaccine caused a reduction in the risk of polio.

A similar large-scale randomized experiment occurred in economics in the 1970s. Between 1971 and 1982, the RAND Corporation conducted a large-scale randomized experiment studying the causal effect of health-care insurance on health-care utilization. For the study, Rand recruited 7,700 individuals younger than age 65. The experiment was somewhat complicated, with multiple treatment arms. Participants were randomly assigned to one of five health insurance plans: free care, three plans with varying levels of cost sharing, and an HMO plan. Participants with cost sharing made fewer physician visits and had fewer hospitalizations than those with free care. Other declines in health-care utilization, such as fewer dental visits, were also found among the cost-sharing treatment groups. Overall, participants in the cost-sharing plans tended to spend less on health because they used fewer services. The reduced use of services occurred mainly because participants in the cost-sharing treatment groups were opting not to initiate care.[5]

But the use of randomized experiments has exploded since that health-care experiment. There have been multiple Nobel Prizes given to those who use them: Vernon Smith for his pioneering of the laboratory experiments in 2002, and more recently, Abhijit Bannerjee, Esther Duflo, and Michael Kremer in 2019 for their leveraging of field experiments at the service of alleviating global poverty.[6] The experimental design has become a hallmark in applied microeconomics, political science, sociology, psychology, and more. But why is it viewed as important? Why is randomization such a key element of this design for isolating causal effects? To understand this, we need to learn more about the powerful notation that Splawa-Neyman [1923] developed, called “potential outcomes.”

Potential outcomes. While the potential outcomes notation goes back to Splawa-Neyman [1923], it got a big lift in the broader social sciences with Rubin [1974].[7] As of this book's writing, potential outcomes is more or less the lingua franca for thinking about and expressing causal statements, and we probably owe Rubin [1974] for that as much as anyone.

In the potential outcomes tradition [Rubin, 1974; Splawa-Neyman, 1923], a causal effect is defined as a comparison between two states of the world. Let me illustrate with a simple example. In the first state of the world (sometimes called the "actual" state of the world), a man takes aspirin for his headache and one hour later reports the severity of his headache. In the second state of the world (sometimes called the "counterfactual" state of the world), that same man takes nothing for his headache and one hour later reports the severity of his headache. What was the causal effect of the aspirin? According to the potential outcomes tradition, the causal effect of the aspirin is the difference in the severity of his headache between two states of the world: one where he took the aspirin (the actual state of the world) and one where he never took the aspirin (the counterfactual state of the world). The difference in headache severity between these two states of the world, measured at what is otherwise the same point in time, is the causal effect of aspirin on his headache. Sounds easy!

To even ask questions like this (let alone attempt to answer them) is to engage in storytelling. Humans have always been interested in stories exploring counterfactuals. What if Bruce Wayne's parents had never been murdered? What if that waitress had won the lottery? What if your friend from high school had never taken that first drink? What if in *The Matrix* Neo had taken the blue pill? These are fun hypotheticals to entertain, but they are still ultimately storytelling. We need Doctor Strange to give us the Time Stone to answer questions like these.

You can probably see where this is going. The potential outcomes notation expresses causality in terms of counterfactuals, and since counterfactuals do not exist, confidence about causal effects must to some degree be unanswerable. To wonder how life would be different had one single event been different is to indulge in counterfactual reasoning, and counterfactuals are not realized in history because they are hypothetical states of the world. Therefore, if the answer requires data on those counterfactuals, then the question cannot be answered. History is a sequence of observable, *factual* events, one after another. We don't know what would have happened had one event changed because we are missing data on the *counterfactual outcome*.[8] Potential outcomes exist ex ante as a set of possibilities, but once a decision is made, all but one outcome disappears.[9]

To make this concrete, let's introduce some notation and more specific concepts. For simplicity, we will assume a *binary* variable that takes on a value of 1 if a particular unit *i* receives the *treatment* and a 0 if it does not.[10] Each unit will have two *potential outcomes*, but only one observed outcome. Potential outcomes are defined as Y_i^1 if unit *i* received the treatment and as Y_i^0 if the unit did not. Notice that both potential outcomes have the same *i* subscript—this indicates two separate states of the world for the exact same person in our example at the exact same moment in time. We'll call the state of the world where no treatment occurred the *control* state. Each unit *i* has exactly two potential outcomes: a potential outcome under a state of the world where the treatment occurred (Y^1) and a potential outcome where the treatment did not occur (Y^0).

Observable or "actual" outcomes, Y_i, are distinct from potential outcomes. First, notice that actual outcomes do not have a superscript. That is because they are not potential outcomes—they are the realized, actual, historical, empirical—however you want to say it—outcomes that unit *i* experienced. Whereas potential outcomes are hypothetical random variables that differ across the population, observable outcomes are factual random variables. How we get from potential outcomes to actual outcomes is a major philosophical move, but like any good economist, I'm going to make it seem simpler than it is with an equation. A unit's observable outcome is a function of its potential outcomes determined according to the *switching equation*:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0 \tag{4.1}$$

where D_i equals 1 if the unit received the treatment and 0 if it did not. Notice the logic of the equation. When $D_i = 1$, then $Y_i = Y_i^1$ because the second term zeroes out. And when $D_i = 0$, the first term zeroes out and therefore $Y_i = Y_i^0$. Using this notation, we define the unit-specific treatment effect, or causal effect, as the difference between the two states of the world:

$$\delta_i = Y_i^1 - Y_i^0$$

Immediately we are confronted with a problem. If a treatment Y_i^0, effect requires knowing two states of the world, Y_i^1 and but by the switching equation we observe only one, then we cannot calculate the treatment effect. Herein lies the fundamental problem of causal inference—*certainty* around causal effects requires access to data that is and always will be missing.

Average treatment effects. From this simple definition of a treatment effect come three different parameters that are often of interest to researchers. They are all population means. The first is called the *average treatment effect*:

$$\begin{aligned} ATE &= E[\delta_i] \\ &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned} \tag{4.2}$$

Notice, as with our definition of individual-level treatment effects, that the average treatment effect requires both potential outcomes for each *i* unit. Since we only know one of these by the switching equation, the average treatment effect, or the *ATE*, is inherently unknowable. Thus, the ATE, like the individual treatment effect, is not a quantity that can be calculated. But it can be *estimated*.

The second parameter of interest is the *average treatment effect for the treatment group*. That's a mouthful, but let me explain. There exist two groups of people in this discussion we've been having: a treatment group and a control group. The average treatment effect for the treatment group, or *ATT* for short, is simply that population mean treatment effect for the group of units that had been assigned the treatment in the first place according to the switching equation. Insofar as δ_i differs across the population, the ATT will likely differ from the ATE. In observational data involving human beings, it almost always will be different from the ATE, and that's because individuals will be endogenously sorting into some treatment based on the gains they expect from it. Like the ATE, the ATT is unknowable, because like the ATE, it also requires two observations per treatment unit *i*. Formally we write the ATT as:

$$\begin{aligned} ATT &= E[\delta_i \mid D_i = 1] \\ &= E[Y_i^1 - E_i^0 \mid D_i = 1] \\ &= E[Y_i^1 \mid D_i = 1] - E[Y_i^0 \mid D_i = 1] \end{aligned} \tag{4.3}$$

The final parameter of interest is called the average treatment effect for the control group, or *untreated* group. It's shorthand is *ATU*, which stands for average treatment effect for the untreated. And like ATT, the ATU is simply the population mean treatment effect for those units who sorted into the control group.[11] Given heterogeneous treatment effects, it's probably the case that the *ATT* = *ATU*, especially in an observational setting. The formula for the ATU is as follows:

$$\begin{aligned} ATU &= E[\delta_i \mid D_i = 0] \\ &= E[Y_i^1 - Y_i^0 \mid D_i = 0] \\ &= E[Y_i^1 \mid D_i = 0] - E[Y_i^0 \mid D_i = 0] \end{aligned} \tag{4.4}$$

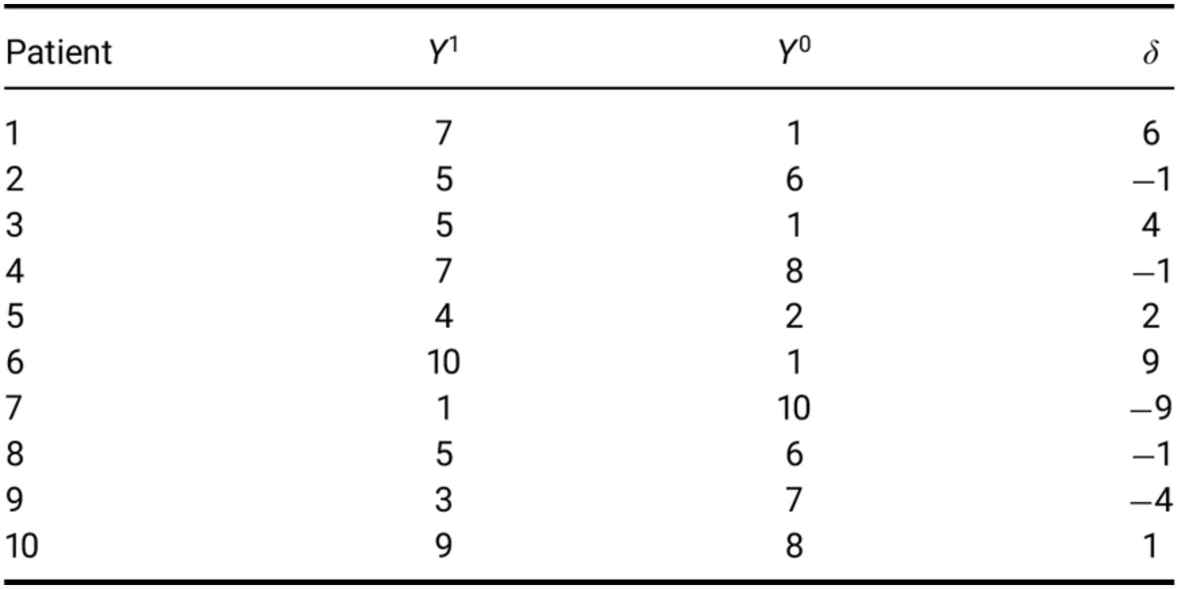
Depending on the research question, one, or all three, of these parameters is interesting. But the two most common ones of interest are the ATE and the ATT.

Simple difference in means decomposition. This discussion has been somewhat abstract, so let's be concrete. Let's assume there are ten patients *i* who have cancer, and two medical procedures or treatments. There is a surgery intervention, $D_i = 1$, and there is a chemotherapy intervention, $D_i = 0$. Each patient has the following two potential outcomes where a potential outcome is defined as post-treatment life span in years: a potential outcome in a world where they received surgery and a potential outcome where they had instead received chemo. We use the notation Y^1 and Y^0, respectively, for these two states of the world.

We can calculate the average treatment effect if we have this matrix of data, because the average treatment effect is simply the mean difference between columns 2 and 3. That is, $E[Y^1] = 5.6$, and $E[Y^0] = 5$, which means that $ATE = 0.6$. In words, the average treatment effect of surgery across these specific patients is 0.6 additional years (compared to chemo).

But that is just the average. Notice, though: not everyone benefits from surgery. Patient 7, for instance, lives only one additional year post-surgery versus ten additional years post-chemo. But the ATE is simply the average over these heterogeneous treatment effects.

Table 11. Potential outcomes for ten patients receiving surgery Y^1 or chemo Y^0.



Patient	Y^1	Y^0	δ
1	7	1	6
2	5	6	–1
3	5	1	4
4	7	8	–1
5	4	2	2
6	10	1	9
7	1	10	–9
8	5	6	–1
9	3	7	–4
10	9	8	1

To maintain this fiction, let's assume that there exists the perfect doctor who knows each person's potential outcomes and chooses whichever treatment that maximizes a person's post-treatment life span.[12] In other words, the doctor chooses to put a patient in surgery or chemotherapy depending on whichever treatment has the longer post-treatment life span. Once he makes that treatment assignment, the doctor observes their post-treatment actual outcome according to the switching equation mentioned earlier.

Table 12 shows only the observed outcome for treatment and control group. Table 12 differs from Table 11, which shows each unit's potential outcomes. Once treatment has been assigned, we can calculate the average treatment effect for the surgery group (ATT) versus the chemo group (ATU). The ATT equals 4.4, and the ATU equals −3.2. That means that the average post-surgery life span for the surgery group is 4.4 additional years, whereas the average post-surgery life span for the chemotherapy group is 3.2 fewer years.[13]

Table 12. Post-treatment observed life spans in years for surgery D = 1 versus chemotherapy D = 0.

Patients	Y	D
1	7	1
2	6	0
3	5	1
4	8	0
5	4	1
6	10	1
7	10	0
8	6	0
9	7	0
10	9	1

Now the ATE is 0.6, which is just a weighted average between the ATT and the ATU.[14] So we know that the overall effect of surgery is positive, although the effect for some is negative. There exist heterogeneous treatment effects, in other words, but the net effect is positive. What if we were to simply compare the average post-surgery life span for the two groups? This simplistic estimator is called the simple difference in means, and it is an *estimate* of the ATE equal to

$$E[Y^1 \mid D=1] - E[Y^0 \mid D=0]$$

which can be estimated using samples of data:

$$\begin{aligned} SDO &= E[Y^1 \mid D=1] - E[Y^0 \mid D=0] \\ &= \frac{1}{N_T}\sum_{i=1}^{n}(y_i \mid d_i=1) - \frac{1}{N_C}\sum_{i=1}^{n}(y_i \mid d_i=0) \end{aligned} \tag{4.5}$$

which in this situation is equal to 7−7.4=−0.4. That means that the treatment group lives 0.4 fewer years post-surgery than the chemo group when the perfect doctor assigned each unit to its best treatment. While the statistic is true, notice how misleading it is. This statistic without proper qualification could easily be used to claim that, on average, surgery is harmful, when we know that's not true.

14 $ATE = p \times ATT + (1-p) \times ATU = 0.5 \times 4.4 + 0.5 \times -3.2 = 0.6.$

It's biased because the individuals units were optimally sorting into their best treatment option, creating fundamental differences between treatment and control group that are a direct function of the potential outcomes themselves. To make this as clear as I can make it, we will decompose the simple difference in means into three parts. Those three parts are listed below:

$$\begin{aligned} E[Y^1 \mid D=1] - E[Y^0 \mid D=0] = ATE & \\ & + E[Y^0 \mid D=1] - E[Y^0 \mid D=0] \\ & + (1-\pi)(ATT - ATU) \end{aligned} \tag{4.6}$$

To understand where these parts on the right-hand side originate, we need to start over and decompose the parameter of interest, *ATE*, into its basic building blocks. ATE is equal to the weighted sum of conditional average expectations, *ATT* and *ATU*.

$$\begin{aligned} ATE &= \pi ATT + (1-\pi)ATU \\ &= \pi E[Y^1 \mid D=1] - \pi E[Y^0 \mid D=1] \\ &+ (1-\pi)E[Y^1 \mid D=0] - (1-\pi)E[Y^0 \mid D=0] \\ &= \Big\{\pi E[Y^1 \mid D=1] + (1-\pi)E[Y^1 \mid D=0]\Big\} \\ &- \Big\{\pi E[Y^0 \mid D=1] + (1-\pi)E[Y^0 \mid D=0]\Big\} \end{aligned}$$

where π is the share of patients who received surgery and $1 - \pi$ is the share of patients who received chemotherapy. Because the conditional expectation notation is a little cumbersome, let's exchange each term on the left side, *ATE*, and right side with some letters. This will make the proof a little less cumbersome:

$$\begin{aligned}
E[Y^1 \mid D=1] &= a \\
E[Y^1 \mid D=0] &= b \\
E[Y^0 \mid D=1] &= c \\
E[Y^0 \mid D=0] &= d \\
ATE &= e
\end{aligned}$$

Now that we have made these substitutions, let's rearrange the letters by redefining ATE as a weighted average of all conditional expectations

$$\begin{aligned}
e &= \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\} \\
e &= \pi a + b - \pi b - \pi c - d + \pi d \\
e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d}) \\
0 &= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)a - (1-\pi)b + (1-\pi)d - (1-\pi)c \\
\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1-\pi)(a-c) - (1-\pi)(b-d)
\end{aligned}$$

Now, substituting our definitions, we get the following:

$$\begin{aligned}
E[Y^1 \mid D=1] - E[Y^0 \mid D=0] = ATE& \\
&+ \Big(E[Y^0 \mid D=1] - E[Y^0 \mid D=0]\Big) \\
&+ (1-\pi)(ATT - ATU) \qquad (4.7)
\end{aligned}$$

And the decomposition ends. Now the fun part—let’s think about what we just made! The left side can be estimated with a sample of data, as both of those potential outcomes become actual outcomes under the switching equation. That’s just the simple difference in mean outcomes. It’s the right side that is more interesting because it tells us what the simple difference in mean outcomes is by definition. Let’s put some labels to it.

$$\begin{aligned}\underbrace{\frac{1}{N_T}\sum_{i=1}^{n}(y_i \mid d_i=1) - \frac{1}{N_C}\sum_{i=1}^{n}(y_i \mid d_i=0)}_{\text{Simple Difference in Outcomes}} &= \underbrace{E[Y^1]-E[Y^0]}_{\text{Average Treatment Effect}} \\ &+ \underbrace{E[Y^0 \mid D=1]-E[Y^0 \mid D=0]}_{\text{Selection bias}} \\ &+ \underbrace{(1-\pi)(ATT-ATU)}_{\text{Heterogeneous treatment effect bias}}\end{aligned}$$

Let’s discuss each of these in turn. The left side is the simple difference in mean outcomes, and we already know it is equal to −0.4. Since this is a decomposition, it must be the case that the right side also equals −0.4.

The first term is the average treatment effect, which is the parameter of interest, and we know that it is equal to 0.6. Thus, the remaining two terms must be the source of the bias that is causing the simple difference in means to be negative.

The second term is called the *selection bias*, which merits some unpacking. In this case, the selection bias is the inherent difference between the two groups if both received chemo. Usually, though, it’s just a description of the differences between the two groups if there had never been a treatment in the first place. There are in other words two groups: a surgery group and a chemo group. How do their potential outcomes under control differ? Notice that the first is a counterfactual, whereas the second is an observed outcome according to the switching equation. We can calculate this difference here because we have the complete potential outcomes in Table 11. That difference is equal to −4.8.

The third term is a lesser-known form of bias, but it's interesting. Plus, if the focus is the ATE, then it is always present.[15] The *heterogeneous treatment effect bias* is simply the different returns to surgery for the two groups multiplied by the share of the population that is in the chemotherapy group at all. This final term is 0.5×*(*4.4−*(*−3.2*))* or 3.8. Note in case it's not obvious that the reason π = 0.5 is because 5 out of 10 units are in the chemotherapy group.

Now that we have all three parameters on the right side, we can see why the simple difference in mean outcomes is equal to −0.4.

$$-0.4 = 0.6 - 4.8 + 3.8$$

What I find interesting—hopeful even—in this decomposition is that it shows that a contrast between treatment and control group technically "contains" the parameter of interest. I placed "contains" in quotes

because while it is clearly visible in the decomposition, the simple difference in outcomes is ultimately not laid out as the sum of three parts. Rather, the simple difference in outcomes is nothing more than a number. The number is the sum of the three parts, but we cannot calculate each individual part because we do not have data on the underlying counterfactual outcomes needed to make the calculations. The problem is that that parameter of interest has been masked by two forms of bias, the selection bias and the heterogeneous treatment effect bias. If we knew those, we could just subtract them out, but ordinarily we don't know them. We develop strategies to negate these biases, but we cannot directly calculate them any more than we can directly calculate the ATE, as these biases depend on unobservable counterfactuals.

The problem isn't caused by assuming heterogeneity either. We can make the strong assumption that treatment effects are constant, $\delta_i = \delta \; \forall i$, which will cause *ATU* = *ATT* and make *SDO* = *ATE* + selection bias. But we'd still have that nasty selection bias screwing things up. One could argue that the entire enterprise of causal inference is about developing a reasonable strategy for negating the role that selection bias is playing in estimated causal effects.

Independence assumption. Let's start with the most credible situation for using *SDO* to estimate *ATE*: when the treatment itself (e.g., surgery) has been assigned to patients *independent* of their potential outcomes. But what does this word "independence" mean anyway? Well, notationally, it means:

$$(Y^1, Y^0) \perp\!\!\!\perp D \tag{4.8}$$

What this means is that surgery was assigned to an individual for reasons that had *nothing* to do with the gains to surgery.[16] Now in our example, we already know that this is violated because the perfect doctor specifically chose surgery or chemo based on potential outcomes. Specifically, a patient received surgery if $Y^1 > Y^0$ and chemo if $Y^1 < Y^0$. Thus, in our case, the perfect doctor ensured that *D depended* on Y^1 *and* Y^0. All forms of human-based sorting—probably as a rule to be honest—violate independence, which is the main reason naïve observational comparisons are almost always incapable of recovering causal effects.[17]

But what if he hadn't done that? What if he had chosen surgery in such a way that did not depend on Y^1 or Y^0? How does one choose surgery independent of the expected gains of the surgery? For instance, maybe he alphabetized them by last name, and the first five received surgery and the last five received chemotherapy. Or maybe he used the second hand on his watch to assign surgery to them: if it was between 1 and 30 seconds, he gave them surgery, and if it was between 31 and 60 seconds, he gave them chemotherapy.[18] In other words, let's say that he chose some method for assigning treatment that did not depend on the values of potential outcomes under either state of the world. What would that mean in this context? Well, it would mean:

$$E[Y^1 \mid D=1] - E[Y^1 \mid D=0] = 0 \tag{4.9}$$

$$E[Y^0 \mid D=1] - E[Y^0 \mid D=0] = 0 \tag{4.10}$$

In other words, it would mean that the mean potential outcome for Y^1 or Y^0 is the same (in the population) for either the surgery group or the chemotherapy group. This kind of *randomization* of the treatment assignment would eliminate both the selection bias and the heterogeneous treatment effect bias. Let's take it in order. The selection bias zeroes out as follows:

$$E[Y^0 \mid D=1] - E[Y^0 \mid D=0] = 0$$

And thus the *SDO* no longer suffers from selection bias. How does randomization affect heterogeneity treatment bias from the third line? Rewrite definitions for ATT and ATU:

$$ATT = E[Y^1 \mid D=1] - E[Y^0 \mid D=1]$$

$$ATU = E[Y^1 \mid D=0] - E[Y^0 \mid D=0]$$

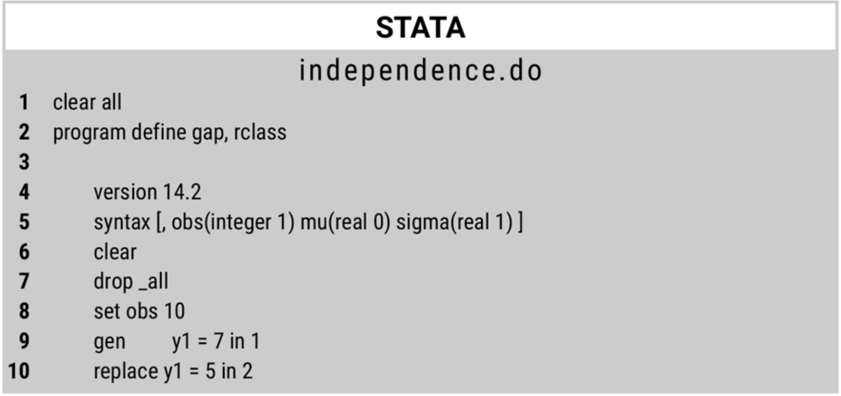
Rewrite the third row bias after 1−*π*:

$$\begin{aligned} ATT - ATU &= \mathbf{E[Y^1 \mid D=1]} - E[Y^0 \mid D=1] \\ &\quad - \mathbf{E[Y^1 \mid D=0]} + E[Y^0 \mid D=0] \\ &= 0 \end{aligned}$$

If treatment is independent of potential outcomes, then:

$$\begin{aligned} \frac{1}{N_T}\sum_{i=1}^{n}(y_i \mid d_i=1) - \frac{1}{N_C}\sum_{i=1}^{n}(y_i \mid d_i=0) &= E[Y^1] - E[Y^0] \\ SDO &= ATE \end{aligned}$$

What's necessary in this situation is simply (a) data on observable outcomes, (b) data on treatment assignment, and (c) $(Y^1,Y^0) \perp D$. We call (c) the independence assumption. To illustrate that this would lead to the SDO, we use the following Monte Carlo simulation. Note that *ATE* in this example is equal to 0.6.



STATA

independence.do

```
clear all
program define gap, rclass

        version 14.2
        syntax [, obs(integer 1) mu(real 0) sigma(real 1) ]
        clear
        drop _all
        set obs 10
        gen        y1 = 7 in 1
        replace y1 = 5 in 2
```

(continued)

STATA *(continued)*

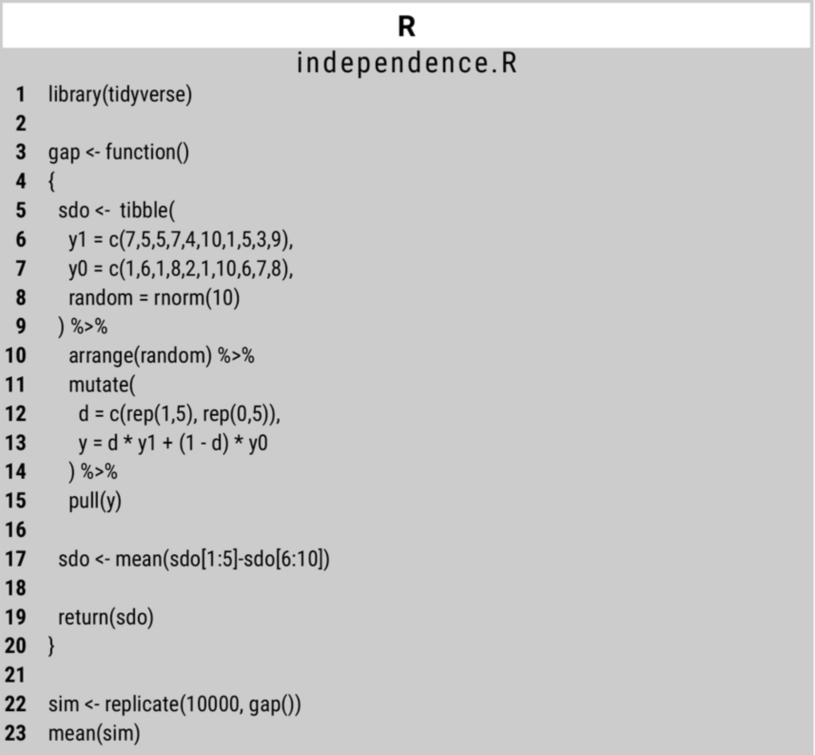
```
    replace y1 = 5 in 3
    replace y1 = 7 in 4
    replace y1 = 4 in 5
    replace y1 = 10 in 6
    replace y1 = 1 in 7
    replace y1 = 5 in 8
    replace y1 = 3 in 9
    replace y1 = 9 in 10

    gen     y0 = 1 in 1
    replace y0 = 6 in 2
    replace y0 = 1 in 3
    replace y0 = 8 in 4
    replace y0 = 2 in 5
    replace y0 = 1 in 6
    replace y0 = 10 in 7
    replace y0 = 6 in 8
    replace y0 = 7 in 9
    replace y0 = 8 in 10
    drawnorm random
    sort random

    gen     d=1 in 1/5
    replace d=0 in 6/10
    gen     y=d*y1 + (1-d)*y0
    egen sy1 = mean(y) if d==1
    egen sy0 = mean(y) if d==0
    collapse (mean) sy1 sy0
    gen sdo = sy1 - sy0
    keep sdo
    summarize sdo
    gen mean = r(mean)
    end

simulate mean, reps(10000): gap
su _sim_1

```



R

independence.R

```
library(tidyverse)

gap <- function()
{
  sdo <-  tibble(
    y1 = c(7,5,5,7,4,10,1,5,3,9),
    y0 = c(1,6,1,8,2,1,10,6,7,8),
    random = rnorm(10)
  ) %>%
    arrange(random) %>%
    mutate(
      d = c(rep(1,5), rep(0,5)),
      y = d * y1 + (1 - d) * y0
    ) %>%
    pull(y)

  sdo <- mean(sdo[1:5]-sdo[6:10])

  return(sdo)
}

sim <- replicate(10000, gap())
mean(sim)
```

This Monte Carlo runs 10,000 times, each time calculating the average SDO under independence—which is ensured by the random number sorting that occurs. In my running of this program, the ATE is 0.6, and the SDO is on average equal to 0.59088.[19]

Before we move on from the SDO, let's just emphasize something that is often lost on students first learning the independence concept and notation. Independence does not imply that $E[Y^1 \mid D = 1]-E[Y^0 \mid D= 0]=0$. Nor does it imply that $E[Y^1 \mid D= 1]-E[Y^0 \mid D= 1]=0$. Rather, it implies

$$E[Y^1 \mid D=1]-E[Y^1 \mid D=0]=0$$

in a large population.[20] That is, independence implies that the two groups of units, surgery and chemo, have the same potential outcome on average in the population.

How realistic is independence in observational data? Economics—maybe more than any other science—tells us that independence is unlikely to hold observationally. Economic actors are always attempting to achieve some optima. For instance, parents are putting kids in what they perceive to be the best school for them, and that is based on potential outcomes. In other words, people are *choosing* their interventions, and most likely their decisions are related to the potential outcomes, which makes simple comparisons improper. Rational choice is always pushing against the independence assumption, and therefore simple comparison in means will not approximate the true causal effect. We need unit randomization for simple comparisons to help us understand the causal effects at play.

SUTVA. Rubin argues that there are a bundle of assumptions behind this kind of calculation, and he calls these assumptions the *stable unit treatment value assumption*, or SUTVA for short. That's a mouthful, but here's what it means: our potential outcomes framework places limits on us for calculating treatment effects. When those limits do not credibly hold in the data, we have to come up with a new solution. And those limitations are that each unit receives the same sized dose, no spillovers ("externalities") to other units' potential outcomes when a unit is exposed to some treatment, and no general equilibrium effects.

First, this implies that the treatment is received in homogeneous doses to all units. It's easy to imagine violations of this, though—for instance, if some doctors are better surgeons than others. In which case, we just need to be careful what we are and are not defining as the treatment.

Second, this implies that there are no externalities, because by definition, an externality spills over to other untreated units. In other words, if unit 1 receives the treatment, and there is some externality, then unit 2 will have a different Y^0 value than if unit 1 had not received the treatment. We are assuming away this kind of spillover. When there are such spillovers, though, such as when we are working with social network data, we will need to use models that can explicitly account for such SUTVA violations, such as that of Goldsmith-Pinkham and Imbens [2013].

Related to this problem of spillovers is the issue of general equilibrium. Let's say we are estimating the causal effect of returns to schooling. The increase in college education would in general equilibrium cause a change in relative wages that is different from what happens under partial equilibrium. This kind of scaling-up issue is of common concern when one considers extrapolating from the experimental design to the large-scale implementation of an intervention in some population.

Replicating "demand for learning HIV status." Rebecca Thornton is a prolific, creative development economist. Her research has spanned a number of topics in development and has evaluated critically important questions regarding optimal HIV policy, demand for learning, circumcision, education, and more. Some of these papers have become major accomplishments. Meticulous and careful, she has become a leading expert on HIV in sub-Saharan Africa. I'd like to discuss an ambitious project she undertook as a grad student in rural Malawi concerning whether cash incentives caused people to learn their HIV status and the cascading effect of that learning on subsequent risky sexual behavior [Thornton, 2008].

Thornton's study emerges in a policy context where people believed that HIV testing could be used to fight the epidemic. The idea was simple: if people learned their HIV status, then maybe learning they were infected would cause them to take precautions, thus slowing the rate of infection. For instance, they might seek medical treatment, thus prolonging their life and the quality of life they enjoyed. But upon learning their HIV status, maybe finding out they were HIV-positive would cause them to decrease high-risk behavior. If so, then increased testing could create frictions throughout the sexual network itself that would slow an epidemic. So commonsense was this policy that the assumptions on which it rested were not challenged until Thornton [2008] did an ingenious field experiment in rural Malawi. Her results were, like many studies, a mixture of good news and bad.

Attempting to understand the demand for HIV status, or the effect of HIV status on health behaviors, is generally impossible without an experiment. Insofar as individuals are optimally choosing to learn about their type or engaging in health behaviors, then it is unlikely