ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Introduction to Causality and Research Design: "No Causation Without Manipulation" [1]

Causal effects are of interest to economists (and other social scientists) because we would often like to know what the effects of manipulating a particular program or policy are. Take, for example, the return to schooling — possibly the most heavily analyzed quantity in labor economics (maybe even in all applied microeconomics!). Using CPS data, it is easy to estimate the relationship between schooling and earnings — we can, for example, use linear regression to approximate the expected value of earnings conditional upon years of schooling (see previous lecture). However, this only reveals to us how these two variables covary in the US population. It does not, in general, reveal what effects a policy manipulation that increased schooling by one year for each student in the US might have on earnings. Using the terms that you learned in ARE 212, years of schooling is "endogenously" determined by individual students and their parents. If you are interested in predicting how earnings change when you draw a different individual with a higher level of education from the US population, then it is perfectly reasonable to apply the regression coefficient. The policy manipulation, however, refers to an "exogenous" change in years of schooling. There is therefore no reason that the regression coefficient — estimated using data in which schooling is endogenously determined — should correspond to the effect of an exogenous change in years of schooling. Note that it is not the case that one quantity is "right" and the other is "wrong" — which one is "correct" depends on what question you are trying to answer. Rather, it's simply the case that the two quantities are different, and one cannot be substituted for the other.

Despite the central role that causality plays in answering policy-relevant questions (since

[1] This phrase comes from Holland (1986). For those unfamiliar with recent American history, it refers to the American Colonial protest, "No Taxation Without Representation." That phrase now appears on license plates in the District of Columbia as "Taxation Without Representation," because D.C. residents pay the same Federal income taxes that you and I do but do not have the "luxury" of being able to vote for voting representation in the House of Representatives or the Senate.

a policy intervention implies, almost by definition, some sort of external manipulation), many econometrics courses do not formally present or discuss a model of causality. Instead, they often begin by presenting a structural model of some economic phenomenon — which is implied to have an underlying causal interpretation — and then proceed to discuss the cases in which linear regression (or some other estimator) will estimate this model. This presentation, however, sometimes leaves students thinking that a regression is inherently "wrong" or useless if it doesn't provide unbiased or consistent estimates of an underlying structural model (which, as we observed in the previous lecture, is certainly not the case). It is also true that in some (many?) cases the structural parameters themselves do not correspond to any meaningful causal effect without further transformations or assumptions.

The causal model we discuss today has come to be known as the *Rubin Causal Model* (RCM), in reference to Rubin (1974) and subsequent publications. The RCM relies heavily upon the notion of *potential outcomes* — that is to say, possible outcomes under different values of a variable we shall refer to as the *treatment* — and it is useful for two reasons. First, it is useful when understanding many common estimation techniques, such as instrumental variables, regression discontinuity design, propensity score matching, etc. More importantly, however, it can be useful in framing or understanding what question you are trying to answer or what effect you are trying to estimate. If the quantity cannot be conceptualized as arising from an experimental manipulation of some type of treatment, then it cannot be estimated from a randomized trial, and the techniques that we learn which simulate randomized experiments will be inappropriate.[2]

1 The Rubin Causal Model

Suppose that we have N units, $i = 1, ..., N$, drawn randomly from a large population. We are interested in the effect of some binary treatment variable, D_i, on an outcome, Y_i. We

[2]Of course, the question may still be of interest, but you will have to find a different (possibly easier!) way to answer it, and you should understand that the answer will not correspond to the effect of a policy intervention.

refer to $D_i = 1$ as the *treatment condition* and $D_i = 0$ as the *control condition.* Given these two possiblities — treatment and control — we postulate the existence of two potential outcomes for each unit: $Y_i(0)$ under the control condition and $Y_i(1)$ under the treatment condition.[3] The key here is that, although we will never observe both $Y_i(0)$ and $Y_i(1)$ (we will observe at most one or the other, but never both), it is theoretically possible that we could observe either. In Holland's terminology, every unit must be *potentially exposable* to every value of the treatment variable. If you cannot conceptualize both $Y_i(0)$ and $Y_i(1)$ for the same unit, then D does not correspond to a treatment that is potentially manipulable and we cannot talk about the causal effect of manipulating D without further defining the problem. Holland, for example, argues that race is not something to which each unit is potentially exposable — we do not in general think of race as being something that we can experimentally manipulate, and it is unclear what it would mean to ask what my potential outcomes would be if I changed my race to be, for example, Black.

Using the notation above, we define the *causal effect* of treatment $D = 1$ on outcome Y for unit i as:

$$Y_i(1) - Y_i(0) = \tau_i$$

Alternatively, we often refer to τ_i as the treatment effect for unit i. Several things are important to note here. First, the effect of a treatment is always defined in a relative sense — in this case it is the effect of the treatment condition $D = 1$ relative to the potential outcome that would have occurred under the control condition $D = 0$. In medicine, $D = 1$ might correspond to giving a drug (e.g., Lipitor) to a patient, while $D = 0$ corresponds to giving a placebo to the patient. In our field, $D = 1$ might correspond to implementing a specific

[3]Note that the notation here is slightly different than in the excellent Holland (1986) article. In Holland's article, the subscript of $Y_t(i)$ corresponds to treatment/control while the argument inside the parentheses corresponds to the unit number ($1, ..., N$ in our case). In our notation, the subscript corresponds to the unit number while the argument inside the parentheses corresponds to treatment/control. We also deviate slightly from Cameron and Trivedi's notation in that they use subscripts for both treatment/control and unit number. We do this because our notation corresponds to the notation used in seminal articles such as Angrist, Imbens, and Rubin (1996).

carbon tax in California, while $D = 0$ corresponds to not doing so.[4] Second, the effect of the treatment need not be constant across different units, as indicated by the fact that τ is indexed by i — many (probably most) treatments have heterogeneous effects. Finally, we will never observe both $Y_i(1)$ and $Y_i(0)$ for any given unit. This is because, although it is not evident in the notation, treatments also involve a time dimension. When we write $D = 1$ and $D = 0$, we implicitly mean that we are applying the treatment or control condition at a specific point in time. In the medical example, if we administer Lipitor to a patient on his 55th birthday, we cannot simultaneously not administer Lipitor to him at the exact same moment. In the environmental policy example, if we implement a carbon tax in California for the 2021 fiscal year, we cannot simultaneously not implement that carbon tax in California during the same fiscal year. We might choose not to implement the tax in 2020 or 2022 — just as we might choose not to administer Lipitor to the patient on his 54th or 56th birthdays — but since other factors affecting the unit can change during the interim period, we are not guaranteed of observing the outcome that would have occurred had we implemented the control condition in 2021 (or on the 55th birthday).

This inability to observe both $Y_i(0)$ and $Y_i(1)$ for any given unit leads to the following theorem:

Fundamental Problem of Causal Inference: It is impossible to observe the value of $Y_i(0)$ and $Y_i(1)$ on the same unit i and, therefore, it is impossible to observe τ_i, the effect for unit i of the treatment on Y_i. (Holland 1986)

The Fundamental Problem of Causal Inference would appear to rule out any precise estimation of τ_i, and, at the unit level, it is true that we can never observe the exact treatment effect. However, all is not lost. As we mentioned in Lecture 1, we are often interested in relationships that hold "on average," or in expectation. In this context, it is possible to estimate quantities of interest. We define the *average causal effect* or *average treatment effect* (ATE) of the treatment relative to the control as the expected value of the

[4] If the treatment variable can take on more than two values (e.g., 0, 1, or 2), then multiple treatment effects exist for each unit (e.g., $Y_i(1) - Y_i(0)$ and $Y_i(2) - Y_i(1)$), and these effects need not be equal, just as the relationship between a dependent variable and an explanatory variable need not be linear.

difference $Y_i(1) - Y_i(0)$, or

ATE:

$$\bar{\tau} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

With the appropriate research design, it is possible to estimate ATE.

2 Estimation of Treatment Effects: The Randomized Controlled Trial

For each unit i, there *exist* the quantities ($Y_i(0)$, $Y_i(1)$, D_i). However, we only *observe* (Y_i, D_i), where

$$Y_i = (1 - D_i)Y_i(0) + D_i Y_i(1)$$

The distinction between what exists conceptually and what we can actually observe is subtle but tremendously important. Although we can only observe $Y_i(0)$ for untreated units and $Y_i(1)$ for treated units, we can conceive of the counterfactual quantities $Y_i(1)$ for untreated units (i.e., the outcome that control unit i would have realized under the treatment condition) and $Y_i(0)$ for treated units (i.e., the outcome that treated unit i would have realized under the control condition). Understanding the distinction between the observed Y_i and the unobserved-but-still-existent counterfactual quantities ($Y_i(0)$ or $Y_i(1)$) will be crucial in subsequent derivations in this course.

By definition,

$$E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1]$$

$$E[Y_i|D_i = 0] = E[Y_i(0)|D_i = 0]$$

Note that in general $E[Y_i(0)|D_i = 0] \neq E[Y_i(0)|D_i = 1]$ (and $E[Y_i(1)|D_i = 1] \neq E[Y_i(1)|D_i = 0]$). That is to say, people who select into the control condition generally have

different outcomes under the control condition ($Y_i(0)$) than people who do not select into the control condition. Thus, the average control outcome for the control unit $E[Y_i(0)|D_i = 0]$ need not equal the average control outcome for all units $E[Y_i(0)]$, which is a combination of both control and treated units. The fact that we do not observe control outcomes ($Y_i(0)$) for any of the treated units, however, does not prevent us from imagining the existence of these counterfactual outcomes. In the context of our medical example, Y is cholesterol level and D represents treatment with Lipitor. Patients who choose to take Lipitor ($D_i = 1$) are likely to have high cholesterol levels in the absence of Lipitor (i.e., $Y_i(0)$ is high, though we do not observe $Y_i(0)$ for them). Patients who choose not to take Lipitor ($D_i = 0$) are likely to have low cholesterol levels in the absence of Lipitor (i.e., $Y_i(0)$ is low, and for these patients we observe $Y_i(0)$ since $Y_i = (1 - D_i)Y_i(0) + D_iY_i(1) = Y_i(0)$). The average untreated cholesterol level for patients not taking Lipitor, $E[Y_i(0)|D_i = 0]$, is therefore less than both the average untreated cholesterol level for treated patients, $E[Y_i(0)|D_i = 1]$, and the average untreated cholesterol level for all patients, $E[Y_i(0)]$.

There is, however, an important case in which $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1] = E[Y_i(0)]$ (and $E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0] = E[Y_i(1)]$). Suppose that the treatment assignment, D, is *randomly assigned*. In that case, D is *independent* of both $Y(0)$ and $Y(1)$. The conditional distribution of $Y_i(0)$ (and $Y_i(1)$) given D_i is therefore equal to the unconditional distribution, and it must be the case that

$$E[Y_i(0)|D_i = 0] = E[Y_i(0)]$$

$$E[Y_i(1)|D_i = 1] = E[Y_i(1)]$$

The average causal effect, $\bar{\tau}$, is thus

$$\bar{\tau} = E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

We can easily estimate $\bar{\tau}$ by taking the difference between the average value of Y_i in the treatment group and the average value of Y_i in the control group. Because it allows

estimation of ATE, the randomized controlled trial is considered the "gold standard" of evidence in medicine, and in many areas of social science as well.

In some instances we may be willing to assume that $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1] = E[Y_i(0)]$ but not that $E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0] = E[Y_i(1)]$. In other words, we may be willing to assume that the untreated potential outcomes are mean-independent of the treatment assignment, but not that the treated potential outcomes are mean-independent of the treatment assignment. This is equivalent to saying that there is no selection into treatment based on the level of untreated outcomes, but there is selection into treatment based on the potential gains of being treated. You could probably write down an economic model that would give this result, but to be honest I doubt it would be a palatable assumption in most empirical settings. Regardless, under this slightly weaker assumption, you can still identify

TOT:

$$\bar{\tau}_{TOT} = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]$$

This quantity is commonly referred to as "the effect of the treatment on the treated," or TOT (treatment-on-treated) or ATOT (average treatment-on-treated) or some other strange permutation of those letters. It is the causal effect of the treatment on those who select into treatment.

3 The Stable Unit Value Treatment Assumption: SUTVA

Beyond the assumption of random assignment of D, there is an implicit assumption embedded in the previous section that is known rather awkwardly as the *stable unit treatment value assumption*, or *SUTVA*. Let $\mathbf{D}$ be a $N \times 1$ column vector that contains the treatment values for all N units. Formally, SUTVA states that

$$\text{If } D_i = D_i', \text{ then } Y_i(\mathbf{D}) = Y_i(\mathbf{D}').$$

We have not yet defined what $Y_i(\mathbf{D})$ is, but it is exactly analogous to our definition of $Y_i(D_i)$ (i.e., $Y_i(0)$ and $Y_i(1)$). That is to say, $Y_i(\mathbf{D})$ is the potential outcome for unit i under treatment regime $\mathbf{D}$. Now, instead of just specifying whether unit i is receiving the treatment or the control, we are specifying values of D_i for all units in the sample. For this reason, SUTVA is often referred to as the "no interference" assumption, since it states that unit i's potential outcomes are unaffected by whether unit j ($j \neq i$) is treated or untreated. A classic example of SUTVA *not* holding is the case of vaccines. If D_i represents inoculation of unit i with the measles vaccine, and Y_i represents whether unit i gets measles, clearly $Y_i(D_i)$ depends on the values of the entire vector $\mathbf{D}$. In particular, if $D_j = 1$ for all $j \neq i$, then $Y_i(0)$ will likely be 0 despite the fact that unit i is unprotected, because there are no other unprotected units to spread the disease to unit i. If $D_j = 0$ for all $j \neq i$, however, then $Y_i(0)$ might change to 1. Another example of SUTVA not holding is the carbon tax scenario presented earlier. If $Y_c(0)$ is the average temperature in California (in the year 2100) in the absence of a California carbon tax, it should be clear that $Y_c(0)$ depends on whether other states or countries implement carbon taxes. If SUTVA does not hold, then there is not just one treatment effect, τ_i, per unit but rather a multitude of treatment effects (one for each different permutation of $\mathbf{D}$). More importantly, it may be impossible to estimate the treatment effect relative to the "no intervention" scenario (i.e., the scenario in which $D_i = 0$ for all units i), because as soon as one unit is treated, all are potentially affected (so it is impossible to construct an unbiased estimate of $E[Y_i(0)]$ with the data).

Rubin (1986) discusses SUTVA in the context of poorly defined treatments. That is to say, he focuses on cases in which, even if $\mathbf{D} = \mathbf{D}'$, it is still the case that $Y_i(\mathbf{D}) \neq Y_i(\mathbf{D}')$. This occurs because the treatment, and in particular the assignment mechanism, is not precisely defined, so even though $\mathbf{D} = \mathbf{D}'$, it's really not the same treatment (we will discuss some examples shortly). In subsequent years, however, interest has focused on the "no interference" aspect of SUTVA — in many cases, treating one unit indirectly affects other units, and SUTVA does not hold.

4 Applications and Discussion

4.1 Poorly Defined Treatments

When does it make sense to talk about D as a cause and when does it not? Holland (1986) has a nice discussion on pp. 954-955 that I urge you to read, as does Rubin's comment to that article. In Holland's example, there are three hypothetical scenarios:

(1) She scored highly on the exam because she is female.

(2) She scored highly on the exam because she studied.

(3) She scored highly on the exam because her teacher tutored her.

In scenario (3), it is clear that the treatment is well-defined: the teacher tutors her. We can easily conceive of manipulating whether or not this tutoring occurs. In scenario (1), Holland argues that the student's sex cannot be considered a "cause" because we cannot manipulate it. It is certainly the case that the treatment is not well-defined in this case and cannot fit within the causal framework, although Rubin points out that further refinements could allow the scenario to fit within the RCM. For example, if we said, "She scored highly on the exam because she received gender reassignment surgery," then we would have a clearly defined treatment (though in the case I am using "female" to refer to the female sex, not the female gender). Scenario (2) is the most problematic because it involves a voluntary activity that the student can choose to do. Although we could certainly conceive of an intervention that might *prevent* the student from studying (anesthesia, for example, would be a pretty good bet), it is hard to imagine a manipulation that would *force* the student to study (or at least force her to study as well as she would if she voluntarily studied). Since we cannot manipulate this attribute (studying for the exam), we cannot think of it as cause, at least not within the potential outcomes framework. Hence Holland's phrase, "No Causation Without Manipulation." It should also be clear from this discussion why there is such a close linkage between the potential outcomes framework and policy relevance. If you can't conceive of manipulating a particular attribute, then by definition you cannot design a policy that would

manipulate that attribute!

4.2 Poorly Defined Treatments II: Assignment Matters

While labor, public, and development economics have clearly moved more towards reduced form empirical work that often attempts to uncover causal effects of interventions (as defined by the RCM), industrial organization has become highly structural and rarely uses the potential outcomes framework at all. Although there is something to be said for hysteresis (as well as available data sets), I believe to a significant degree this divergence is due to the difficulty in defining treatments in much of the IO context. In short, it can be important to think about how the assignment mechanism affects not just selection into the treatment, but the actual effect of the treatment itself.

In many strategic situations with incomplete information, we are interested in how a unit reacts to the actions of other units, not just because those actions directly affect the unit in question, but because the action may reveal new information to the unit. For example, consider natural resource extraction scenarios — oil drilling, fisheries, etc. — in which the location of the resources is unknown. We may want to test whether extractor i take cues from the location of other extractors in order to "learn" about the distribution of resources. In this case, the assignment mechanism for the location of other extractors (the treatment) becomes very important. Random assignment of the other extractors will not allow us to estimate the effect in question because extractor i will react differently if (s)he realizes that the other extractors are being randomly assigned. This can be true even if extractor i does not know which extractors are being randomly assigned — the only way his or her behavior will be unchanged is if he does not realize that there is any (additional) randomization being applied.[5] It can therefore be exceedingly difficult to find valid natural experiments in these types of situations. In general, any time that an action sends a signal, and our interest is in estimating the response to that signal, then randomization of that action will not yield the

[5]The same thing applies to randomly revealing valid location information to some of the other extractors. Extractor i's behavior can change if (s)he knows that additional valid information is being inserted into the game.

estimate that we are looking for (unless the signal's target has no idea that the randomization is being introduced). This may be one reason why the potential outcomes framework, with its emphasis on RCTs, is relatively uncommon in IO. Assignment matters.

4.3 Effects of Causes vs. Causes of Effects

Holland writes that "an emphasis on the effects of causes rather on the causes of effects is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation and directly opposes more traditional analyses of causation." This distinction between effects of causes and causes of effects may seem somewhat pedantic. It is not.

A concrete example should help clarify the distinction. Consider the obesity "epidemic," an issue of great importance to both agricultural economists (on the input side) and health economists (on the output side). Researchers in different fields cite a myriad of "causes" of this epidemic: increased consumption of snack foods, larger portions at restaurants, more frequent consumption at restaurants, more sedentary jobs, the introduction of high fructose corn syrup, etc. Under these explanations, however, it is rarely clear what the counterfactual is. Take, for example, the increased consumption of snack foods over the last 30 years. One possible counterfactual is what would have happened if people had not increased their consumption of snack foods while everything else remained unchanged. It is unclear, however, what policy manipulation could enforce that counterfactual scenario. Although it is easy to imagine limiting snack food consumption (through a quota or a tax, for example), it is impossible to imagine doing so while simultaneously preventing individuals from compensating in any other manner. Another possible counterfactual is to imagine a world in which the complex set of technologies and changes in consumer preferences that led to the increase in snacking had been inhibited from developing. Even if it were possible to imagine this sort of manipulation, however, there is no guarantee that total caloric consumption would fall by exactly the amount that snack food consumption has increased (in fact, it almost surely would not). It becomes clear that the answer to the question of what has "caused" the obesity epidemic is every single thing about the world that affects weight and has changed

since 1970. But even this turns out to be an incomplete answer because the distribution of weight in 1970 is itself a cause of the distribution of weight today, so the obesity epidemic is in fact "caused" by everything in the history of the world that has ever had an effect on weight. As Holland (1986) notes, there is really no definable answer to this type of question.

Another example may make the conundrum even clearer. What "caused" the loss of life in New Orleans during the flooding from Katrina? Was it the hurricane? The fact that the levies were not constructed well enough? The fact that not all residents chose (or had the means to) evacuate? The fact that Bush appointed Michael Brown, a former commissioner of the International Arabian Horses Association, as head of FEMA (where, according to Bush, he proceeded to do a "heck of a job" during the rescue effort)? The fact that a pumping system was built in the early 20th century that allowed development of below-sea level areas? The fact that the city even existed at all following the Louisiana Purchase? The list of possibilities is endless.

In contrast to the causes of effects — which is effectively an unlimited exercise in accounting and description — the effects of causes are clearly defined under the RCM. Even if we cannot measure them with existing data, we can at least conceive of what they are.

5 Additional References

Rubin, Donald. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies." *Journal of Educational Psychology*, 1974, 66, 688-701.