Lecture 04 - Theories of Causality & Potential Outcomes

Samuel DeWitt, Ph.D.

Theories of Causality & The Potential Outcomes Model

Before we dive further into advanced causal methods, it is helpful to review the dominant theories of causality in modern criminology.

We will also discuss the potential outcomes model, a popular framework for thinking about causality in situations where random assignment is not possible.

What Does Causality Mean?

A very general definition has been taken from Hume:

"[w]e 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 first.* Or, in other words, *where, if the first object had not been, the second never had existed.*

Though the two statements may appear the same, they are not. They both imply different ways to approach causal analysis.

One Object Following Another...

This approach to causality is predicated upon observing both the cause and the effect (which is a good first start) but aspects we have found to be necessary for causal inference are missing.

Specifically, while an analysis operating with this definition may satisfy the need for temporal order (the cause precedes the effect in time) it does not:

1) Establish an empirical association more broadly
2) Account for other influences on both the cause and effect
3) Explain the mechanism which accounts for the relationship
4) Describe the contextual features necessary to observe the relationship

What's Missing? - Empirical Association

What does it mean to establish an empirical association?

Well, that we have put what we think is a causal relationship to extended empirical scrutiny.

Namely, that we have observed variation in both the purported **cause** and **effect** and have observed a regularity that variation in the former precedes **consistent** variation in the latter.

Consistency would be judged by both *magnitude* and *direction*.

What's Missing? - Nonspuriousness

Do we account for other influences on the cause and/or effect? If we do, we satisfy this *nonspuriousness* requirement.

Often, though, when we do not have experimental data, we do not satisfy this requirement. This means our causal estimates are **biased**.

Example: we can observe that police per capita and the crime rate co-vary, but which causes which?

What else is important to observe other than just those variables?

What's Missing? - Mechanism

Though not required, an understanding of **why** a relationship exists between two variables is important toward revealing valuable information about causation.

Although we may observe an empirical association which is not spurious and satisfies temporal order, this **does not** mean that we have a complete understanding about the dynamics of that relationship.

A prominent example is the relationship between youth employment and crime. We tended to think it would be protective (reduce crime), but it has been shown to increase it.

We need to think carefully about the causal relationship, collect additional data, and fill these gaps in to obtain a complete understanding of the causal process.

What's Missing? - Context

Even when we have all of the above, we cannot guarantee that the causal relationship is true across time, geography, or other characteristics that allow for the causal relationship to exist.

This has to do with the **contextual** characteristics which could be necessary to observe a causal relationship.

As an example, one may surmise that a causal relationship exists between the unemployment rate and crime.

What if, however, that relationship is only apparent when certain segments of the labor market are affected? What if there's a healthy under-the-table labor market?

The Second Statement...

As Lewis (1973) notes, these statements are thought to have been two different ways to say the same thing, but a more critical analysis of these statements reveals an important difference.

The first statement refers to observing empirical regularities whereas the second statement refers to what has become to be known as a **counterfactual**

A **counterfactual** is an alternative, unobserved state of the world where an individual unit (individual, city, etc...) that was **treated** was instead **untreated**.

The question then becomes, what would that unit's value have been on the outcome of interest if they weren't exposed to treatment?

Big Problem: Or, It's Difficult to See Alternate Realities

This should jump out as a significant issue to you - how can we come to know a value that, by its definition, we cannot actually observe?

Well, since we don't have a window to alternate universes (yet) we need to start with this perfect **counterfactual** case and work down from there to something that is the next best substitute.

Ideally, this *stand-in* **counterfactual** shares many features with the observed case with the primary difference between the units being that one is **treated** with some intervention of interest and the other remains **untreated**.

Potential Outcomes Framework

This leads me to a popular framework which has developed around causal inference - the potential outcomes model.

Think of any type of treatment we may find interesting in criminal justice (e.g., policing strategies, rehabilitation services), but keep your mind open to things we cannot also control as researchers (e.g., sentencing, gang-membership).

In the simplest case, **treatment** has two values - 0 if the unit is untreated and 1 if the unit is treated.

There are then two **potential outcomes** we can observe for each unit - their outcome if they are **untreated** (0) or their outcome if they are **treated** (1).

Potential Outcomes Framework - Notation

Below is the mathematical notation for the **potential outcomes model** of causal inference.

Outcome for unit i: Y_i

Treatment status for unit i: D_i where $D_i = 1$ if they receive treatment and 0 otherwise.

Potential outcomes: Y_{1i} when unit i is **treated** and Y_{0i} when unit i is **untreated**

Potential Outcomes Framework

As above, though - we cannot observe **both** potential outcomes - only one!

So, the problem of causal inference is then interpreted as a **missing data** issue.

We have outcome data for one observation, but are lacking it for the other. We need to find a suitable value to put there in order to calculate a causal effect.

Potential Outcomes Framework and Experiments

This is our **best-case scenario** in the social sciences for causal inference.

Why do you think this is? That is, what feature of an experiment satisfies the need for the best possible **counterfactual** available?

Potential Outcomes Framework and Experiments

If you said **random assignment** you are right.

Random assignment allows us to assume, across the sample of **treated** and **untreated** units that the only systematic difference between them is treatment.

Therefore, any change in the outcome of interest, assuming the experiment is controlled well enough, is due to treatment, and not some other variable or variables.

In essence, a true experiment fulfills the **counterfactual** requirement by making assignment to treatment random - everything about untreated units is expected to be no different than treated units as it concerns characteristics other than their **treated** status.

Practical Example

As an example you will come to know well (soon), let's discuss the impact of joining a gang on criminal behavior among youth in the NLSY97 sample.

Using the terminology above, we can conceptualize gang membership as a form of **treatment** where, in the strict sense of the term, one unit is **treated** with being a gang member and the other is **untreated** - they stay out of the gang.

Therefore, we have two potential expectations for the dichotomous delinquency variable: the delinquent status of a youth if they **are** a gang member ($D = 1; Y_{1i}$) or their delinquent status if they **are not** a gang member ($D = 0; Y_{0i}$).

Issues with the Practical Example

In order to be a good counterfactual, the non-gang youth need to be closely similar to gang youth on any relevant characteristic we think may co-vary with delinquency.

How plausible do you think it is that we can satisfy this assumption?

Issues with the Practical Example

If you said *not very plausible* you are right.

For reasons we will begin exploring over the next several weeks, there are numerous pre-existing differences between these groups that influence their criminal behavior **and** whether or not they join a gang.

In order to say with confidence that *gangs lead to an increase in delinquency* we have to be certain that our comparison group of non-gang youth we use as a substitute counterfactual were equally as likely to join a gang at that point in time.

Issues with the Practical Example

We may have **extensive** reasons not to believe that assumption.

Therein lies the problem of causal inference in the social sciences - ethically, many **treatments** we are interested in understanding the impact of may not be subjected to random assignment (i.e., joining a gang, getting sent to prison).

But, without random assignment, causal inference is **significantly** more difficult and necessitates techniques like the ones we will discuss for the remainder of the semester.