Introduction to Causal Inference

Thomas Wiemann
University of Chicago

Econometrics
Econ 21020

Updated: April 18, 2022

The Purpose of Causal Inference

Recall the distinction we drew in Lecture 1:

Descriptive questions are concerned with the *realized* state of the world:

- ▷ *What is* the hourly wage of college graduates?

Causal questions are concerned with the *potential* states of the world:

- ▷ *What* would the hourly wage of college graduates be *if* had they not pursued higher education?

Causal inference is the development of logically consistent answers to causal (“what if”) questions using real data.

The Purpose of Causal Inference (Contd.)

Answers to causal questions are necessary for economists who want to:

P-1. Evaluate the impact of a historical policy.

- ▷ “Did it pay off for Americans born in the 60s to go to college?”

P-2. Forecast the impact of a historical policy in a new environment.

- ▷ “Should 2022 high school graduates be encouraged to pursue higher education?’

P-3. Forecast the impact of a new policy in a new environment.

- ▷ In Feb/Mar 2020: “What will the number of COVID-19 cases in Chicago be with social distancing and indoor mask mandates?”

***Note**: The term “policy” here is broadly defined. It may be a public policy (e.g., tax increases) but also encompasses actions of individuals or firms (e.g., market entry).*

The Purpose of Causal Inference (Contd.)

P-2 and P-3 are strictly more challenging than P-1 because they require transporting insights from one environment (or policy) to other environments (or policies). This process is known as "extrapolation".

This poses a fundamental problem in knowledge:

- "The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past." (Knight, 1921)

As we will see, P-1 is already plenty difficult and will keep us busy for the remainder of the course.

- But extrapolation for P-2 and P-3 is hugely important!
- Economists are not *just* historians.

Three distinct tasks arising in the analysis of causal questions

A careful causal analysis requires three distinct tasks:

Table 1: Three distinct tasks in the analysis of causal questions

Task	Description	Requirements
1	Definition of hypotheticals (or counterfactuals)	A scientific theory
2	Parameter identification from a hypothetical population	Mathematical analysis
3	Parameter estimation and inference from real data	Statistics

Notes. Paraphrased Table 1 from Heckman and Vytlacil (2007).

- ▷ Definition of counterfactuals develops quantified versions of the "what if" questions (i.e., expressed using mathematics).
- ▷ Identification concerns the task of linking hypotheticals to known functions of observables in a logically consistent manner.
- ▷ Estimation and inference is constructs and characterizes useful "guesses" of the hypotheticals from a finite number of data points.

Today: Revisit Task 1 and 2 w/ new tools from probability theory.

Outline

1. Task 1: Definition
   - ▷ The All Causes Model
   - ▷ Potential Outcomes
   - ▷ Common Causal Parameters

2. Task 2: Identification
   - ▷ The Fundamental Problem of Causal Inference
   - ▷ Types of Identification

These notes benefit greatly from Heckman and Vytlacil (2007).

Outline

1. **Task 1: Definition**
   - **The All Causes Model**
   - Potential Outcomes
   - Common Causal Parameters
2. Task 2: Identification
   - The Fundamental Problem of Causal Inference
   - Types of Identification

These notes benefit greatly from Heckman and Vytlacil (2007).

The All Causes Model

Careful analysis of causal questions requires a formal language.

Thankfully, econometricians don't need to reinvent (every) wheel: Mathematics is an extremely useful formal language well suited for our purposes. We thus turn to mathematical modeling of causal relationships.

To begin their analysis, econometricians often consider

$$Y = g(W, U), \tag{1}$$

where

- ▷ $Y \equiv$ an outcome;
- ▷ $W \equiv$ a policy;
- ▷ $U \equiv$ all determinants of Y other than W ;
- ▷ and an economic model $g : \operatorname{supp} W \times \operatorname{supp} U \to \operatorname{supp} Y$.

To capture that the econometrician does not jointly observe (Y, W, U), we consider them as random variables.

The All Causes Model (Contd.)

Note that g is a deterministic map:

- ▷ Given a policy w and other determinants u, the outcome is $g(w, u)$.
- ▷ No uncertainty given (W, U): they contain all causes of Y.
- ▷ The model in (1) is often referred to as the *all causes* model.

The all causes model is a very general framework that places little initial restrictions on (Y, W, U) or g. For example,

- ▷ W may be binary, discrete, continuous or mixed.
- ▷ U can be a random vector of arbitrary dimension.
- ▷ g can be any function.

We thus require economic theory to give meaning to the all causes model.

Note*: Conventionally, U denotes unobservable determinants of Y and W is the determinant of Y that we are interested in. It is sometimes useful to highlight that there are additional observed determinants of Y that are not of main interest. For this purpose, we can introduce additional varibales X and reformulate the model in* (1) *to*

$$Y = \tilde{g}(W, X, U).$$

The All Causes Model (Contd.)

Example 1

Recall the returns to education example from Lecture 1. We may have

- ▷ $Y \equiv$ hourly wages;
- ▷ $W \equiv$ and indicator for having obtained a college degree;
- ▷ $U \equiv$ determinants of Y other than W, e.g., intellect or connections.
- ▷ $g \equiv$ a labor production function.

Example 2

Suppose you're contemplating how much to study for this course. Then

- ▷ $Y \equiv$ final course score (or: hourly wages?);
- ▷ $W \equiv$ total time spent studying for Econ 21020;
- ▷ $U \equiv$ determinants of Y other than W, e.g., extracurricular stress.
- ▷ $g \equiv$ a study production function.

Outline

1. **Task 1: Definition**
   - ▷ The All Causes Model
   - ▷ **Potential Outcomes**
   - ▷ Common Causal Parameters
2. Task 2: Identification
   - ▷ The Fundamental Problem of Causal Inference
   - ▷ Types of Identification

Potential Outcomes

The model in (1) is useful for defining potential states of the world.

- Often referred to simply as *potential outcomes*.

When $\operatorname{supp} W = \{0, 1\}$, the potential outcomes are

- $g(0, U) \equiv$ the outcome if W would be 0;
- $g(1, U) \equiv$ the outcome if W would be 1.

Note in the binary case, it follows from our model that

$$Y = \tag{2}$$

More generally, the potential outcomes are

$$g(w, U), \quad \forall w \in \operatorname{supp} W. \tag{3}$$

Note*: In the setting with discrete W, i.e., $\operatorname{supp} W = \{w_1, \ldots\}$, the potential outcomes are often denoted by $Y(w) \equiv g(w, U), \forall w \in \operatorname{supp} W$. I avoid this notation for now to emphasize the role of g and U.*

Potential Outcomes (Contd.)

Example 3

Recall the returns to education example 1. The potential outcomes are:

- $g(0, U) \equiv$ the hourly wage if they do not obtain a college degree;
- $g(1, U) \equiv$ the hourly wage if they obtain a college degree.

Example 4

Recall the returns to studying econometrics example 2. The potential outcomes are:

- $g(w, U) \equiv$ the final course score if (one of) you studies w hrs, where w can take any value in $[0, 1416]$ (1416 being the number of hours if you studied 24/7, literally).

Example 4 highlights that arbitrarily "creative" potential outcomes can be defined. Potential outcomes are a thought experiment:

- They need not be plausible (or even possible). Only interesting!

Potential Outcomes (Contd.)

Using the potential outcomes, we can now construct parameters that are informative for the causal question of interest.

A key object is the individual-level *treatment* effect (ITE)

$$\tau(U) \equiv g(1, U) - g(0, U), \tag{4}$$

which is of particular interest in the setting of binary policies.

For non-binary policies, we may instead consider an ITE of the form

$$\tau_{w',w}(U) \equiv g(w', U) - g(w, U), \tag{5}$$

for $w', w \in \operatorname{supp} W$.

The ITE gives the difference in the outcome between two policy actions *holding all other determinants U fixed*.

- ▷ The *ceteris paribus* principle popular in economics since at least Marshal (1890).

***Terminology**: It may seem more natural here to call $\tau(U)$ the "individual-level policy effect." I don't do so given the ubiquity of the "treatment effects" terminology.*

Potential Outcomes (Contd.)

Example 5

Recall the returns to education example 1. The ITE is

$$\tau(U) = g(1, U) - g(0, U), \tag{6}$$

which is the difference in hourly wages from obtaining a college degree for a given individual.

Example 6

Recall the returns to studying econometrics example 2. The ITE for studying 72hrs instead of 18hrs is

$$\tau_{72,18}(U) = g(72, U) - g(18, U), \tag{7}$$

which is the difference in the final course score from the increased study effort for a given individual.

Fixing versus Conditioning

In the definition of potential outcomes, we *fixed* the policy W at a particular value w.

- ▷ Because W is a random variable, you may be tempted to think this is equivalent to *conditioning* on $W = w$. This is incorrect!

The distinction was first highlighted Haavelmo (1943). To illustrate, let

$$Y = g(W, U) = W\beta + U, \tag{8}$$

with supp Y = supp $U = \mathbb{R}$, supp $W = \{0, 1\}$, and $\beta \in \mathbb{R}$.

The ITE, with *fixed* W, is given by

(9)

This is different from the difference in, say, *conditional* means as

(10)

The latter does not satisfy the ceteris paribus principle!

The Simpson's Paradox Revisited

The confusion between fixing and conditioning gives rise to the *Simpson's Paradox* that we briefly encountered in Lecture 2B (Example 10).

Consider an outcome Y (e.g., hourly wages), a binary treatment W (e.g., being a college grad), and another binary random variable X (e.g., being from the US). Using the L.I.E., we saw that it's perfectly possible that

$$\begin{aligned} E[Y|W=0,X=1] &< E[Y|W=1,X=1], \\ E[Y|W=0,X=0] &< E[Y|W=1,X=0], \end{aligned} \tag{11}$$

while at the same time

$$E[Y|W=0] > E[Y|W=1]. \tag{12}$$

The paradox appears when these mathematical statements are *incorrectly* translated to English:

- ▷ "Equation (11) shows that going to college increases hourly wages for those from and not from the US, but Equation (12) shows that going to college decreases hourly wages overall. How can that be?"

The Simpson's Paradox Revisited (Contd.)

It turns out: It can't. It is not possible that a treatment is beneficial to all subpopulations (here: being from and not from the US) while at the same time being harmful overall.

- ▷ The error lies in the fact that Equations (11) and (12) don't characterize the effects of the treatment: They are descriptives.

To see that the real Simpsons paradox (i.e., w/o a translation error) cannot happen, consider

$$\begin{aligned} E[g(0,U)|X=1] &< E[g(1,U)|X=1], \\ E[g(0,U)|X=0] &< E[g(1,U)|X=0], \end{aligned} \tag{13}$$

where $g(0,U)$ and $g(1,U)$ denote the potential outcomes as before. Then using the L.I.E., we have

$$E[g(1,U)] = \tag{14}$$

Outline

1. **Task 1: Definition**
    - The All Causes Model
    - Potential Outcomes
    - **Common Causal Parameters**

2. Task 2: Identification
    - The Fundamental Problem of Causal Inference
    - Types of Identification

Common Causal Parameters

Thus far, our discussion focused on individual-level treatment effects.

- ▷ $\tau(U)$ is the most detailed causal parameter...
- ▷ ...but also the most difficult to characterize fully.

Instead of focusing on $\tau(U)$ directly, most causal analyses in econometrics are content with characterizing key properties of its distribution.

Definition 1 (Average Treatment Effect; ATE)

Consider the random vector (Y, W, U) with $\operatorname{supp} W = \{0, 1\}$, whose joint distribution is characterized by

$$Y = g(W, U), \tag{15}$$

with $g : \operatorname{supp} W \times \operatorname{supp} U \rightarrow \operatorname{supp} Y$. The *average treatment effect* is defined as

$$\text{ATE} \equiv E[\tau(U)] = E[g(1, U) - g(0, U)]. \tag{16}$$

Common Causal Parameters (Contd.)

The ATE gives the expected effect of the treatment for *a randomly selected individual*. We may make more detailed statements via conditioning on the treatment itself:

Definition 2 (ATE on the Treated and Untreated)

Consider the random vector (Y, W, U) with $\operatorname{supp} W = \{0, 1\}$, whose joint distribution is characterized by

$$Y = g(W, U), \tag{17}$$

with $g : \operatorname{supp} W \times \operatorname{supp} U \rightarrow \operatorname{supp} Y$. The *average treatment effect on the treated* is defined as

$$\text{ATT} \equiv E[\tau(U)|W = 1] = E[g(1, U) - g(0, U)|W = 1]. \tag{18}$$

The *average treatment effect on the untreated* is defined as

$$\text{ATU} \equiv E[\tau(U)|W = 0] = E[g(1, U) - g(0, U)|W = 0]. \tag{19}$$

Common Causal Parameters (Contd.)

Whether we are interested in the ATE, ATT, or ATU (or a different parameter all together) depends on the causal question.

Example 7

Recall the returns to education example 1. Here,

- ▷ the ATE gives the expected returns to education for a randomly selected individual;
- ▷ the ATT gives the expected returns to education for college graduates;
- ▷ the ATU gives the expected returns to education for non-graduates.

There are lots of prospective students touring campus these days. Suppose we are considering to encourage them from pursuing college. Which parameter would be most relevant?

Instead, say we are considering a policy that encourages students from high schools with traditionally low college-enrollment rates to pursue higher education. Which parameter would be most relevant?

Common Causal Parameters (Contd.)

More recently, practitioners are increasingly interested in associating expected treatment effects with other observed characteristics.

Definition 3 (Conditional ATE)

Consider the random vector (Y, W, U) with $\operatorname{supp} W = \{0, 1\}$, whose joint distribution is characterized by

$$Y = g(W, U), \tag{20}$$

with $g : \operatorname{supp} W \times \operatorname{supp} U \to \operatorname{supp} Y$. Let X be another random variable. The *conditional average treatment effect* is defined as

$$\text{CATE}(x) = E[\tau(U)|X = x] = E[g(1, U) - g(0, U)|X = x], \tag{21}$$

$\forall x \in \operatorname{supp} X$. The CATT and CATU can be defined analogously.

The CATE gives the expected effect of the treatment for a randomly selected individual with $X = x$.

- ▷ Particularly useful for targeted policies (e.g., ads!).

Common Causal Parameters (Contd.)

Example 8

Recall the returns to studying econometrics example 2. Let $X = 1$ denote majoring in math and $X = 0$ otherwise. Consider the CATEs given by

$$\begin{aligned} CATE_{72,18}(1) &= E[g(72, U) - g(18, U) | X = 1], \\ CATE_{72,18}(0) &= E[g(72, U) - g(18, U) | X = 0]. \end{aligned}$$

- ▷ $CATE_{72,18}(1)$ gives the expected returns to studying more for a randomly selected math-major;
- ▷ $CATE_{72,18}(0)$ gives the expected returns to studying more for a randomly selected non-math-major.

Which one do you find more interesting?

Outline

1. Task 1: Definition
    - ▷ The All Causes Model
    - ▷ Potential Outcomes
    - ▷ Common Causal Parameters
2. **Task 2: Identification**
    - ▷ **The Fundamental Problem of Causal Inference**
    - ▷ Types of Identification

Observables and Unobservables

Causal parameters are functions of (the distribution of) U.

- ▷ the ATE is an expectation w.r.t. the marginal distribution of U;
- ▷ the ATT and ATU are expectations w.r.t. the conditional distribution of U given the policy W;
- ▷ the CATE is an expectation w.r.t. the conditional distribution of U given the observable X.

In our analysis, (Y, W) are considered *observable* but U is *unobservable*:

- ▷ We *can* collect a sample from the joint of (Y, W);
- ▷ We *cannot* collect a sample from U (or the joint of (Y, W, U)).

The *best* we can hope for is a sample $(Y_1, W_1), \ldots, (Y_n, W_n) \overset{iid}{\sim} (Y, W)$.

- ▷ (Sample analogue) estimates of expectations w.r.t. to distributions of U are impossible to construct!

Then how can we learn anything about causal parameters from data?

Observables and Unobservables (Contd.)

Identification is the task of expressing a causal parameters as known functions of (the distribution of) the observables.

- ▷ If we can do that, (sample analogue) estimates can be constructed!
- ▷ Identification is the sense of Hurwicz (1950).

The question on how to learn about causal parameters form data thus turns into the question of identification.

As we will see, identification *requires* additional assumptions:

- ▷ Using data alone, it is impossible to learn about causal parameters.
- ▷ It is a common misconception to think otherwise (recall the NYT).

The Fundamental Problem of Causal Inference

A characteristic of potential outcomes is that not all of them are realized:

- Potential outcomes that are never realized are "counterfactuals".

Indeed, for any given individual, we can only observe one potential outcome. This gives rise to the fundamental problem of causal inference:

- *It is impossible to observe $g(w', U)$ and $g(w, U)$ for the same unit whenever $w' \neq w$. Hence, it is impossible to observe $\tau_{w',w}(U)$.*

Holland (1986) coined the fundamental problem of causal inference.

To see its implications for the identification of causal parameters, consider the ATT. We have

$$\text{ATT} = \tag{22}$$

where $E[Y|W = 1]$ is a known function of the observables but $E[g(0, U)|W = 1]$ is a function of U for which the sampling process provides no information:

- $E[g(0, U)|W = 1]$ may take any value on $\mathbb{R}$,
- thus the ATT may take any value on $\mathbb{R}$. That's rather useless!

The Fundamental Problem of Causal Inference (Contd.)

Example 9

Recall the returns to education example 1. Consider

$$\text{ATT} = E[g(1,U)|W=1] - E[g(0,U)|W=1]. \tag{23}$$

Here,

- $E[g(1,U)|W=1] \equiv$ expected hourly wages of college graduates had they obtained a college degree;
- $E[g(0,U)|W=1] \equiv$ expected hourly wages of college graduates had they not obtained a college degree.

The sampling process reveals $E[g(1,U)|W=1] = E[Y|W=1]$, but it is impossible to observe $E[g(0,U)|W=1]$:

- There are no college graduates w/o a college degree.

Even if we "knew" $E[Y|W=1]$, we couldn't learn about the ATT.

- $E[g(0,U)|W=1]$ may take any value on $\mathbb{R}$.

Outline

1. Task 1: Definition
   - ▷ The All Causes Model
   - ▷ Potential Outcomes
   - ▷ Common Causal Parameters

2. **Task 2: Identification**
   - ▷ The Fundamental Problem of Causal Inference
   - ▷ **Types of Identification**

Types of Identification

To make progress, we consider so-called *identifying assumptions*:

- ▷ These are restrictions on the joint distribution of (Y, W, U).

The extent to which we may learn about the causal parameters depends on the particular identifying assumption considered. There are three broad settings:

1. *Not identified*: The assumption did not shrink the feasible set of values for the causal parameter of interest.

2. *Partially-identified*: The assumption shrunk the feasible set of values for the causal parameter of interest to a non-singleton set.

3. *Point-identified*: The assumption shrunk the feasible set of values for the causal parameter of interest to a singleton set.

Types of Identification (Contd.)

Example 10

Recall the returns to education example 1. Consider

$$\text{ATT} = E[g(1,U)|W=1] - E[g(0,U)|W=1]. \tag{24}$$

1. Consider the assumption supp $U = [0,1]$.
   - ▷ Nothing is implied for $E[g(0,U)|W=1]$ unless we know g. The ATT is *not identified*: it can take any value in $\mathbb{R}$.
2. Consider the assumption that g is weakly increasing in W.
   - ▷ Then

     The ATT is *partially-identified*: it can take any value in
3. Consider the assumption that $E[g(0,U)|W=1] = E[g(0,U)|W=0]$.
   - ▷ Then

     The ATT is *point-identified*: it can only take a single value.

Types of Identification (Contd.)

Notice that a unique value for the causal parameter of interest is only implied in the point-identified case.

- This is the strongest identification result, and one that practitioners seem to have preferences for.
- But there are no free lunches: stronger identification results require stronger identifying assumptions.
- Researchers trade-off strength of the identification results (desirable) with strength of the identifying assumptions (undesirable).

This no-free-lunch theorem is known as the *law of decreasing credibility*:

- *The credibility of inference decreases with the strength of the assumptions maintained.* (Manski, 2003)

Common Identifying Assumptions

Which identifying assumption to consider crucially depends on the economic context:

- Identification results are useful only to the extend that the identifying assumptions appear plausible.
- An assumption plausible in one context may be implausible in another.

The search for identifying assumptions for different economic contexts motivates an immense literature in econometrics.

A few identifying assumptions that practitioners most often rely on are:

- Random Assignment (see Lecture 5);
- Selection on Observables (see Lecture 7);
- Instrumental Variables (see Lecture 9).

There are many, many others...

- ... but you'll need to take a more advanced econometrics class to learn about them. (Hopefully this one motivates you to do so!)

References

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, pages 1–12.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 6, chapter 70, pages 4779–4874. Elsevier, Amsterdam.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hurwicz, L. (1950). Generalization of the concept of identification. In Koopmans, T. C., editor, *Statistical inference in dynamic economic models*, volume 6, chapter 4, pages 245–257. Wiley, New York.

Knight, F. (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin Company, New York.

Manski, C. F. (2003). *Partial identification of probability distributions*, volume 5. Springer.