Lecture 1 - Potential Outcomes, Directed Acylic Graphs, and Structural Models

Paul Goldsmith-Pinkham

January 13, 2026

Not every economics research paper is estimating a causal quantity. But, the implication or takeaway of papers is (almost) always a causal one. Causality lies at the heart of every exercise. [1]

The goals in this lecture are:

- Enumerate tools used to discuss causal questions
- Emphasize a *multimodal* approach
- Set terminology/definitions for future discussions

Concretely, this involves covering three ways, notationally, of considering causal questions:

1. the potential outcomes (PO) framework,
2. the directed acyclic graph (DAG) framework,
3. structural models.

Over the course of describing these, we will also refresh our memories on the difference between the estimator, the estimand and the estimate, and learn the identification condition for the average treatment effect (ATE).

Notation

We will begin by outlining some notation for potential outcomes. When defining treatment effects, this notation is extremely convenient and clear, particularly when considering settings with significant unobserved heterogeneity. However, since so much of the extant literature in economics (and econometrics) is written using more standard structural equations (e.g. $Y = X\beta + \epsilon$), it is important to be able to translate between the two. For the sake of completeness, I also want to expose you to the directed acyclic graph (DAG) framework, (**??**) which is more commonly used in other fields such as epidimiology and computer science.[2] It is much less common in economics, but without getting into broader epistemic debates, it's extremely useful in some settings for clarifying the identifying assumptions (especially in settings relying on a "conditional on observables" assumption).

[1] "We do not have knowledge of a thing until we have grasped its why, that is to say, its cause." – Aristotle

[2] There was a period of time when the debate about DAGs was quite ornery (especially online). I think this has subsided.

Potential Outcomes

We will follow **?** in our notation. I will be slightly looser in my definitions for the purpose of space, but I encourage you to read Chapter 1 of **?** or Chapter 7 of **?** for a more precise treatment.

Consider a sample of N units, indexed by i. Each unit has a treatment status D_i and an outcome Y_i.[3] Sometimes, I will refer to the collection of observations or treatments as $\mathbf{D}$ and $\mathbf{Y}$ to denote a vector of length N with each element corresponding to the treatment or outcome for a given unit. Both $\mathbf{D}$ and $\mathbf{Y}$ are *observed* in our data: we see who is treated ($\mathbf{D}$), and the subsequent outcome (the $\mathbf{Y}$ given the $\mathbf{D}$)

[3] For now, we will assume that there is just a binary treatment, but this can be generalized to multiple or continuous treatments. It will make life more complicated.

Example 1

Many medical examples naturally lend themselves to thinking about potential outcomes. For example, consider the outcome of whether you have a headache in three hours:

$$Y = \begin{cases} 1 & \textit{Have a headache in three hours} \\ 0 & \textit{Do not have a headache in three hours} \end{cases}$$

and the treatment of taking an aspirin:

$$D = \begin{cases} 1 & \textit{Take an aspirin} \\ 0 & \textit{Do not take an aspirin.} \end{cases}$$

We now consider the *potential* outcome for unit i. We can denote this as $Y_i(\mathbf{D})$, which is the outcome for unit i if the set of treatments for the N units is $\mathbf{D}$. Note that this is a complicated function! It depends on the treatment status of all units, not just the treatment status of unit i. This leads us to a first important assumption:

Assumption 1 (Stable Unit Treatment Value Assumption)
If $D_i = D_i'$, then $Y_i(\mathbf{D}) = Y_i(\mathbf{D}')$.

Put in words, it means that your potential outcome is only affected by your own treatment status, and not the treatment status of others.[4] This assumption lets us write our potential outcome as $Y_i(D_i)$, and focus just on how our own treatment affects our outcome. This is a strong assumption; we will discuss how one might consider relaxing it in a few lectures. And of course any macroeconomist will tell you that this is a terrible assumption. But, it is a useful starting point.

[4] Sometimes this is called a "no interference" condition. As we'll see later on, this could also be labeled a spillover in the economics literature.

Example ?? (continued)
We can now consider the potential *outcome in the state of the world where a person takes an aspirin or not: $Y_i(1)$ vs. $Y_i(0)$. Note that it is not fundamentally possible to observe both states of the world: even if a person were observed in different time periods, and in one case they took the aspirin and in another they did not, this would reflect fundamentally different observations. This type of repeated observation could be used to help identify the average potential outcomes, but would require additional assumptions.*
SUTVA is a very natural assumption in our medical example, since others' aspirin treatment decision should have no impact on our headache. However, this is likely not true with vaccines or other interventions.

It's worth remarking on a few things. First, this potential outcome is an function of the individuals' treatment status, and allowed to vary by individual. Second, this outcome itself is not necessarily observed. Indeed, what we observe is

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

Hence, for the untreated units, we observe their $Y_i(0)$, and vice versa for the treated units. This model is often referred to as the Neyman-Rubin Causal model.[5]

The fact that we only observe either $Y_i(1)$ or $Y_i(0)$ is sometimes called the "fundamental problem of causal inference.[6]" Since we can only observe one outcome for a given unit, we cannot trace out the counterfactual outcomes for a single unit. This makes it quite challenging to know what the effect of changing D_i is on a single unit i.[7]

One way to view the fundamental problem of causal inference is as a missing data problem.[8] We will use many different techniques throughout this course to impute a counterfactual outcome such that we can know the causal effect of an intervention.

[5] These were not coauthors - Jerzy Neyman was a Polish statistician who initially proposed the potential outcomes framework to study completely randomized experiments (**?**). This model was adopted and expanded by Donald Rubin in a number of influential papers. This model was coined the Rubin Causal model by Paul Holland, in an influential paper (**?**) about statistics and causality that we will revisit shortly.

[6] This term, again, comes from **?** (which you should read!).

[7] If we assume the treatment effects everyone exactly the same, then it is straightforward. While we might make a homogeneity assumption like this, we don't always believe it in practice.

[8] The treatment by **?** covers this in very nice detail.

Comment 1
It is worth thinking a bit about what causal effect you are interested in estimating. Often this is referred to as the ***estimand****. This could be many things:*

- *A structural parameter (dInvestment/dTaxRate?)*
- *The effect of zoning restrictions on housing supply*
- *A policy evaluation of a renter's assistance program*
- *The existence of underreaction in stock prices to earnings news*

Comment 2
It is important to get these terms straight.

- *Estimand: the quantity to be estimated*
- *Estimate: the approximation of the estimand using a finite data sample*
- *Estimator: the method or formula for arriving at the estimate for an estimand*

For a particularly goofy way to remember this: https://twitter.com/paulgp/status/1275135175966494721?s=20

Identification of the Average Treatment Effect Estimand

We will conclude this lecture by describing sufficient conditions under which we can identify the **Average Treatment Effect** or ATE, a common target estimand for researchers.

Before we do that, we need to define the individual level causal estimand (that is, recall, inherently unknowable). Call this the **Individual Treatment Effect** or ITE. This is the difference between the potential outcomes for a given unit:

$$\tau_i \equiv Y_i(1) - Y_i(0). \tag{2}$$

This can be easily generalized to multiple treatments as well: we will discuss this in a few lectures.

Average Treatment Effect

We now consider the *average* treatment effect over the population. This is, quite simply, the average of the individual treatment effects over all individuals in the overall population.

Definition 1
We define the average treatment effect in our population as

$$\tau_{ATE} \equiv \mathbb{E}(\tau_i) = \mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)).$$

Why do we find the ATE interesting?[9] For one thing, it describes the effect of giving the treatment to everyone in the population. This is often of interest to policymakers, who want to know the effect of a policy on the entire population.

We now consider some additional average treatment estimands. The first is the Average Treatment Effect on the Treated (ATT):

Definition 2

$$\tau_{ATT} \equiv \mathbb{E}(\tau_i|D_i = 1) = \mathbb{E}(Y_i(1)|D_i = 1) - \mathbb{E}(Y_i(0)|D_i = 1).$$

This estimates the effect for individuals who received the treatment.[10] Note that one piece of the ATT is observed: $E(Y_i(1)|D_i = 1)$. This is just the observed outcome for the treated units.

We can also define the conditional average treament effect (CATE). Let X_i be a pre-determined set of covariates. Then, we can define the CATE as:

Definition 3

$$\tau_{CATE}(x) \equiv \mathbb{E}(Y_i(1)|X_i = x) - \mathbb{E}(Y_i(0)|X_i = x).$$

Example ?? (continued)
The ATE is what the effect would be on headaches if every person in the population took aspirin relative to not taking aspirin.
The ATT is what the impact of aspirin has been for those who took aspirin, relative to if they had not taken aspirin.
The CATE is what the impact of aspirin for a particular group, such as older men, would be relative to not taking aspirin.

It is useful to note the following relationship between the ATE and the CATE:

$$\tau_{ATE} = \int \tau_{CATE}(x) f(x) dx.$$

If X_i is discrete with values in $\mathcal{X}$, more simply this is

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \tau_{CATE}(x) Pr(X_i = x).$$

We now discuss under what conditions we can identify the ATE.

[9] This is sometimes called the *population average treatment effect* or PATE. This is then contrasted with the *sample average treatment effect* or SATE. The SATE is the ATE defined for the sample of N individuals we observe, while the PATE is the ATE for the population of individuals we can draw the sample from. We will discuss this in more detail later in the class, but for now I will just refer to the ATE to refer to the average treatment effect over the sample, and assume that the SATE and PATE are similar. Typically if samples are randomly draw, this is a reasonable assumption, and the difference is mainly in the inference. See **?** for an example discussion.

[10] It will be a little while until we discuss cases when the ATT and ATE are different. A notable example is difference-in-differences. Many examples where we use *models* to estimate our counterfactual outcome will lead to cases where we can only identify the ATT and not the ATE.

Identification of the ATE

Comment 3
What is identification? Intuitively, for a given estimand to be identified, it means that in a world with no uncertainty about data, can we always identify the value of our estimand from the data we observe? To quote **?***: "Econometric identification really means just one thing: model parameters or features being uniquely determined from the observable population that generates the data."*

Note that without further assumptions, the ATE is not identified from the observed data, $(\mathbf{Y}, \mathbf{D})$. Why? Consider the following estimator of the ATE:

$$\tau = E(Y_i|D_i = 1) - E(Y_i|D_i = 0) \tag{3}$$

which compares the treated units' average outcome to the untreated units' average outcome. Rewriting using our potential outcomes,

$$\tau = E(Y_i(1)|D_i = 1) - E(Y_i(0)|D_i = 0),$$

we see that our key challenge is that the two expectations condition on *different* values of D_i. Hence, if there is correlation between D_i and $(Y_i(1), Y_i(0))$, these two averages are not comparable.

Example 2
Imagine I am a researcher studying the effect of a wage training program (D_i) on wages (Y_i). I have a sample of workers, and I observe their wages and whether they participated in the training program. I want to know the effect of the training program on wages. If I use **??** *to compare the wages of those who take the program to those who do not, I may be comparing individuals who are very different.*

For example, if the training program is voluntary, then it is likely that those who take the program are more motivated or knowledgeable about the labor force, and hence would have higher wages even if they did not take the program. In this case, the naive estimator would overstate the effect of the training program on wages. Let U_i be a binary variable capturing their motivation or knowledge of the labor force. If $E(D_i|U_i = 1) - E(D_i|U_i = 0) > 0$, and $E(\tau_i|U_i = 1) - E(\tau_i|U_i = 0) > 0$, then the naive estimator will overstate the effect of the training program on wages.

Comment 4
As an exercise, prove that the naive estimator is biased in **??**.

We are now ready for our first identification result. We first define **strong ignorability**:

Definition 4
We say that D_i is strongly ignorable conditional on a vector $\mathbf{X}_i$ if

1. $Y_i(0), Y_i(1) \perp D_i | \mathbf{X}_i$
2. $\exists \varepsilon > 0$ *such that* $\varepsilon < Pr(D_i = 1|\mathbf{X}_i) < 1 - \varepsilon$.

The first part of **??** is sometimes referred to unconfoundeness (or in economics, exogeneity): we assume that the choice of treatment is independent (conditional on $\mathbf{X}$) of the units' potential outcome. This means a unit can't select into the treament based on their potential benefits.[11]

The second condition asserts that there is some variation in treatment. This is sometimes called the common support or overlap condition . It is a bit stronger than we need, but it is a convenient way to ensure that we can compare the treated and untreated units.

[11] Strong ignorability is a much more precise term than exogeneous, but tends to be used less in economics. When communicating with an economics audience, you might say that D_i is conditionally randomly assigned, or D_i is exogeneous – but this would omit the second condition (which is that the treatment is not too rare or too common).

Theorem 1 (Identification of the ATE)
If D_i is strongly ignorable conditional on $\mathbf{X}_i$, then

$$\mathbb{E}(\tau_i) = \sum_{x \in \text{Supp } X_i} \Big(\mathbb{E}(Y_i | D_i = 1, \mathbf{X}_i = x) - \mathbb{E}(Y_i | D_i = 0, \mathbf{X}_i = x) \Big) Pr(\mathbf{X}_i = x)$$

Proof 1
Note that by strong ignorability,

$$\mathbb{E}(Y_i(0)|\mathbf{X}_i) = \mathbb{E}(Y_i(0)|D_i = 0, \mathbf{X}_i) = \mathbb{E}(Y_i|D_i = 0, \mathbf{X}_i).$$

In essence, independence of D_i and $(Y_i(0), Y_i(1))$ lets us interchange counterfactuals and realized data in conditionals. The rest follows by the law of iterated expectations.

This result is quite powerful, and describes a non-parametric condition for when we can identify (and estimate) the ATE. A corrolary of this theorem is that we can also identify conditional average treatment effects as well (by assumption).

Identification through Directed Acylic Graphs

Above, we encoded random variables' relationships functionally, using potential outcomes. An alternative approach does this graphically. I will not cover this in significant detail, but want to give an

example of how to think about identification using Directed Acylcic Graphs (DAGs).

We can encode the relationship between D and Y using an *arrow* in a graph. The direction emphasizes that D causes Y, and not vice versa.

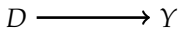


Figure 1: D has a causal effect on Y

We can also allow for the unobservable U, which drove identification concerns above in **??**. In this case, U is termed a *confounder*. We can look at the paths by which D links to Y:

- The standard direct effect $D \rightarrow Y$
- The "back door" path $D \leftarrow U \rightarrow Y$

Note that the back-door is *not* causal. We know from above that the effect of D on Y is not identified under this setup, but this provides a graphical intuition as well – there is a path connecting D and Y but it does not flow in the right direction.

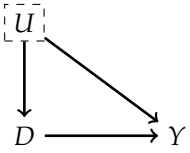


Figure 2: D's effect on Y is confounded by U

Now, we can replace U with an observable X. X is still a confounder, but since it is observable, we can condition on it and identify our effect (as in **??**). As before, examine the paths by which D links to Y:

- The standard direct effect $D \rightarrow Y$
- The "back door" path $D \leftarrow X \rightarrow Y$.

In a DAG, conditioning on a variable along the path "blocks" the path, such that we would block the back door path.

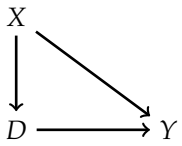
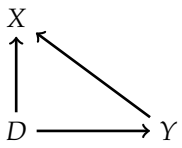


Figure 3: D's effect on Y is confounded by an observable X

Finally, let's consider a more complicated example. X is now a "collider", such that D and Y both affect X.

As before, examine the paths by which D links to Y:



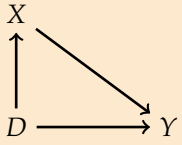
- The standard direct effect $D \to Y$
- The indirect path $D \to X \leftarrow Y$.

This is not called a backdoor path because X does not point into D.

The key difference in this setting is that since X does not cause Y, it is automatically blocked (all effects on X occur through our main effect). However, if you condition on X, you open the path!

Example ?? (continued)
Return to the example of a job training program. We want to study the impact of the program on wages, and we condition on whether a person has a car. If a person's wages affects their likelihood of having a car, we will have created a biased comparison: we will first consider the effect of the job training program among those who have a car (which may be small), and then among those who do not (which may also be small). If much of the effect of the program affects individuals' ability to buy a car, then we will underestimate the effect of the program.

Comment 5
A last example is what's called a mediator. *This is another variable that is affected by the treatment, and affects the outcome. In this case, we can think of the treatment as having two effects: a direct effect, and an indirect effect through the mediator.*



It is possible to control for a mediator in order to estimate only the direct effect – this is sometimes referred to as mediation analysis. However this is very *sensitive to functional form, and not recommended.*

I will not give an exhaustive approach on how to deal with DAGs for identification, but you can hopefully see that there is a great deal of intuitive value in writing down the DAG in some problems. This is

particularly true when dealing with *colliders*.

Structural equations and causal effects

It is important to not lose sight of the fact that these should be estimates that inform our *economic* model. Since much of our background is traditionally in structural equations (that often map to economic models) it can often be more familiar to write out outcome equation as:

$$Y_i = \alpha + \beta D_i + \varepsilon_i.$$

It is quite helpful to see how this maps back to the potential outcome framework:

$$\begin{aligned}
Y_i &= Y_i(0)(1 - D_i) + Y_i(1)D_i \\
&= Y_i(0) + \tau_i D_i \\
&= Y_i(0) + \tau D_i + (\tau_i - \tau)D_i \\
&= \underbrace{E(Y_i(0)|D_i = 0)}_{\alpha} + \underbrace{\tau}_{\beta} D_i + \underbrace{(\tau_i - \tau)D_i + (Y_i(0) - E(Y_i(0)|D_i = 0))}_{\varepsilon_i}
\end{aligned}$$

Consider now what $E(Y_i|D_i)$ will recover:

$$\begin{aligned}
E(Y_i|D_i = 1) &= \alpha + \tau + E(\varepsilon_i|D_i = 1) \\
E(\varepsilon_i|D_i = 1) &= (E(\tau_i|D_i = 1) - \tau) + E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0) \\
E(Y_i|D_i = 0) &= \alpha + E(\varepsilon_i|D_i = 0) \\
E(\varepsilon_i|D_i = 0) &= 0.
\end{aligned}$$

So, we can see that we will recover the average treatment effect as β if D_i is randomly assigned (or strongly ignorable). This is a special case where the coefficient β in the linear regression case will give the average treatment effect, the constant will give the average for the untreated, and the error term will capture the rest. We will suffer from omitted variable bias if $E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0) \neq 0$ – e.g. if there is selection into treatment based on your control potential outcome. Notice that if $E(\tau_i|D_i = 1) \neq \tau$, then we will also not estimate the ATE, but we will estimate the ATT.

More generally, however, it's just useful to see that there are one-to-one mappings between the potential outcome framework and structural regressions. In many ways, the potential outcome framework is helpful because it emphasizes the relevant counterfactual state more than many linear models.

Phil Haile has some lovely slides discussing the importance of structure in economics. One of the key issues he pushes back on is

the idea where many applied researchers estimating treatment effects say they are being "model-free." In other words, rather than writing down a structural model and attempting to estimate something complicated with a functional form, they view their treatment effects as model-agnostic. This is sometimes referred to as the "reduced form".

What is the reduced form from a structural estimation perspective? Following Haile, a reduced form relationship is one where the endogeneous variable is a function of *exogeneous* variables and unobserved structural error terms. Exogeneous here means variables that satisfy the necessary independence assumptions with the structural error terms.

Example 3
Consider a supply and demand system:

$$Q_d = D(P, X, U_d)$$
$$Q_s = S(P, Z, U_s).$$

These are simultaneous equations where the observed price we see in the market is the price where $Q_d = Q_s$. Often, supply (Q_s) will be written in terms of price (which is a function of marginal cost):

$$Q = D(P, X, U_d)$$
$$P = S(Q, Z, U_s).$$

Since P and Q are endogeneous, these are structural *equations.*

The reduced form version of these equations would have the form

$$Q = d(X, Z, U_d, U_s)$$
$$P = s(X, Z, U_d, U_s).$$

In economics, we may consider estimating the effect of price on quantity (e.g. a labor demand elasticity), which is a parameter in the structural demand equation. When we use instrumental variables and two-stage least squares (to be discussed further in a later class), the first stage *will be the reduce form, and the second stage is a structural model.*

Overall, it's important to remember that many of our estimation approaches imply a particular structural model. We may be approximating something more complicated, but we're typically making some kind of modeling decision.

Discussion Questions 1

1. *Consider the potential outcome framework in the context of individuals. We are thinking about annual earnings Y_i for an individual i. Often, we study the earnings gap between men and women. Is it reasonable to consider the potential outcome $Y_i(1)$ vs. $Y_i(0)$ for $D_i = 1$ when i is a woman vs. when i is a man?*

2. *Consider the linear model from above:*

$$Y_i = \alpha + \beta D_i + \varepsilon_i.$$

When would we expected homoskedasticity to hold?