Chapter 1

Causality

Unlike physical laws such as Einstein's mass–energy equivalence $E = mc^2$ and Newton's universal gravitation $F = Gm_1m_2/r^2$, economic phenomena can rarely be summarized in such a minimalistic style. When using experiments to verify physical laws, scientists often manage to come up with smart design in which signal-to-noise ratio is so high that small disturbances are kept at a negligible level. On the contrary, economic laws do not fit a laboratory for experimentation. What is worse, the subjects in economic studies — human beings — are heterogeneous and with many features that are hard to control. People from distinctive cultural and family backgrounds respond to the same issue differently and researchers can do little to homogenize them. The signal-to-noise ratios in economic laws are often significantly lower than those of physical laws, mainly due to the lack of laboratory setting and the heterogeneous nature of the subjects.

Educational return and the demand-supply system are two classical topics in econometrics. A person's incomes is determined by too many random factors in the academic and career path that is impossible to exhaustively observe and control. The observable prices and quantities are outcomes of equilibrium so the demand and supply affect each other.

Generations of thinkers have been debating the definitions of causality. In economics, the long-standing tradition is *structural causality*. Structural causality is a thought experiment. It assumes that there is a DGP that produces the observational data. If we can use data to recover the DGP or some features of the DGP, then we have learned causality or some implications of causality.

In recent years, an alternative approach is the potential outcome framework, mainly coming from biostatistics and spreading into econometrics. The idea is to think about the outcomes as if they are from an experiment where the researcher can "manipulate" the treatment—a factor of interest—and monitor the the difference in the outcomes. In reality when an experiment cannot be literally carried out, we still try to modify or argue for the settings to resemble an experiment.

1.1 Potential Outcome Framework

This a thought experiment. A personal has two potential outcomes $Y^{(1)}$ and $Y^{(0)}$, and his treatment effect is the difference $\Delta = Y^{(1)} - Y^{(0)}$. However, no one can step into the same river twice. One and only one of the potential outcomes will be realized, and therefore Δ is unobservable. Given the treatment status $D \in \{0, 1\}$, in reality we can be observed is

$$Y = DY^{(1)} + (1 - D)Y^{(0)}.$$

We want to use the observable Y to learn the ATE

$$\text{ATE} := \mathbb{E}[\Delta] = \mathbb{E}[Y^{(1)}] - \mathbb{E}[Y^{(0)}].$$

Randomized controlled trials. If the treatment is randomly assigned by flipping a coin (the coin does not need to be even), then

$$\begin{pmatrix} Y^{(1)} \\ Y^{(0)} \end{pmatrix} \perp D. \tag{1.1}$$

It implies

$$\mathbb{E}\left[Y|D=1\right] = \mathbb{E}\left[DY^{(1)} + (1-D)Y^{(0)}|D=1\right] = \mathbb{E}\left[Y^{(1)}|D=1\right] \overset{\text{idp}}{=} \mathbb{E}\left[Y^{(1)}\right].$$

since (Y, D) are observable, the LHS is operational. Independence between $Y^{(1)}$ and D ensures that the conditional expectation $\mathbb{E}\left[Y^{(1)}|D=1\right]$ equals the unconditional expectation $\mathbb{E}[Y^{(1)}]$. Similarly,

$$\mathbb{E}\left[Y|D=0\right] = \mathbb{E}\left[Y^{(0)}|D=0\right] \overset{\text{idp}}{=} \mathbb{E}[Y^{(0)}].$$

Under RCT, we have an operational formula for ATE:

$$\text{ATE} = \mathbb{E}\left[Y|D=1\right] - \mathbb{E}\left[Y|D=0\right].$$

Given the data, we mimic the population average to compute

$$\widehat{\text{ATE}} = \frac{1}{n_1} \sum_{\{i:d_i=1\}} y_i - \frac{1}{n_0} \sum_{\{i:d_i=0\}} y_i$$

where $n_1 = \sum_{i=1}^n \mathbb{I}\{d_i = 1\}$ and $n_0 = \sum_{i=1}^n \mathbb{I}\{d_i = 0\}$.

The above approach uses conditioning, which is intuitive. There is an alternative, yet equivalent way to ATE, using only unconditional quantities. Notice that

$$DY = D^2Y^{(1)} + D(1-D)Y^{(0)} = DY^{(1)}$$

since for either $D \in \{0, 1\}$, we have $D^2 = D$ and $D(1-D) = 0$. Again, Y depends on D but $Y^{(1)}$ is independent of D, and therefore

$$\mathbb{E}\left[DY\right] = \mathbb{E}\left[DY^{(1)}\right] \overset{\text{idp}}{=} \mathbb{E}\left[D\right]\mathbb{E}[Y^{(1)}] = \Pr\left[D=1\right]\mathbb{E}[Y^{(1)}]$$

and therefore

$$\mathbb{E}\left[Y^{(1)}\right] = \frac{\mathbb{E}\left[DY\right]}{\Pr\left[D=1\right]} = \frac{\mathbb{E}\left[\mathbb{I}(D=1)Y\right]}{\Pr\left[D=1\right]}$$

if $\Pr\left[D=1\right] \neq 0$. The denominator $\Pr\left[D=1\right]$ is called the **propensity score**—the probability that a person is assigned into the treatment group.

Similarly, if $\Pr\left[D=0\right] \neq 0$ we have

$$\mathbb{E}\left[Y^{(0)}\right] = \frac{\mathbb{E}\left[(1-D)Y\right]}{\Pr\left[D=0\right]} = \frac{\mathbb{E}\left[\mathbb{I}(D=0)Y\right]}{\Pr\left[D=0\right]}.$$

Therefore, if $\Pr\left[D=1\right] \in (0,1)$ the ATE is

$$\text{ATE} = \frac{\mathbb{E}\left[DY\right]}{\Pr\left[D=1\right]} - \frac{\mathbb{E}\left[(1-D)Y\right]}{\Pr\left[D=0\right]}.$$

ATE is the difference of the two ratios.

Given data, we can compute $\mathbb{E}\left[DY\right] / \Pr\left[D=1\right]$ as

$$\frac{(\sum_{i=1}^{n} d_i y_i)/n}{n_1/n} = \frac{\sum_{i=1}^{n} d_i y_i}{n_1} = \frac{1}{n_1} \sum_{\{i:d_i=1\}} y_i.$$

It is easy to see that, the sample version is the same either we compute via the conditioning or the propensity score. Their equivalence is analogous to the fact that the conditional density can be written as the ratio of the joint density and the marginal density.

1.1.1 CATE

At a granular level, the researcher also observe some confounding factor (causal inference term) (alternatively, in plain stat term it is called covariate) W.

The conditional ATE (CATE)

$$\text{ATE}(w) = \mathbb{E}[\Delta|W=w] = \mathbb{E}[Y^{(1)}|w] - \mathbb{E}[Y^{(0)}|w]$$

can vary across different realizations of W. For example, an vaccine is more effective to children than adults. In RCT, if the random treatment assignment is regardless of the age group, we have $(Y^{(1)}(w), Y^{(0)}(w))$ depend W, while $D \perp W$, then we maintains (1.1).

However, for different age groups, we may want to use different treatment assignment probability. For example, we put 70% of the children into the treatment, while we put 40% of the adults into the treatment. In this case, $D(w)$ also depends on W. RCT can be implemented in a stratified approach: inside each age group, the researcher flips a coin to assign treatment. If so, the **unconfoundedness condition** (or **conditional independence assumption**)

$$\begin{pmatrix} Y^{(1)}(w) \\ Y^{(0)}(w) \end{pmatrix} \perp D(w) \Big| W = w$$

is satisfied. The CATE is made operational via

$$\begin{aligned}\text{ATE}(w) &= \mathbb{E}\left[Y^{(1)}|W=w\right] - \mathbb{E}\left[Y^{(0)}|W=w\right] \\ &\overset{\text{cia}}{=} \mathbb{E}\left[Y^{(1)}|D(w)=1,w\right] - \mathbb{E}\left[Y^{(0)}|D(w)=0,W_i=w\right] \\ &= \mathbb{E}\left[Y|D=1,w\right] - \mathbb{E}\left[Y|D=0,w\right]. \qquad (1.2)\end{aligned}$$

Inside each age group, we just need to compute the difference between the averages of those treated and those untreated, respectively.

We can also do the computation via the propensity score. Denote the (conditional) propensity score as $p(w) = \Pr\left[D=1|W=w\right]$. Then

$$\text{ATE}(w) = \frac{\mathbb{E}\left[DY|w\right]}{p(w)} - \frac{\mathbb{E}\left[(1-D)Y|w\right]}{1-p(w)}.$$

Compared to 1.2, the above expression makes the role of $p(w)$ explicit.

Aggregating subgroups. Now, suppose the researcher has CATE at hand, and she wants to aggregate the CATE across subgroups into an overall ATE. Then by the law of iterated expectations:

$$\text{ATE} = \mathbb{E}[\text{ATE}(w)] = \mathbb{E}\left\{\mathbb{E}\left[Y|D=1,w\right] - \mathbb{E}\left[Y|D=0,w\right]\right\} \qquad (1.3)$$

Notice that when D depends on w,

$$\mathbb{E}\left\{\mathbb{E}\left[Y|D=d,w\right]\right\} \neq \mathbb{E}\left[Y|D=d\right]$$

and thus

$$\text{ATE} \neq \mathbb{E}\left[Y|D=1\right] - \mathbb{E}\left[Y|D=0\right].$$

That is, we cannot simply get rid of w in the two sides of (1.2); we must take into consideration the different treatment assignment probability when aggregating.

If we use the ratio instead, then

$$\text{ATE} = \mathbb{E}[\text{ATE}(w)] = \mathbb{E}\left[\frac{\mathbb{E}\left[DY|w\right]}{p(w)} - \frac{\mathbb{E}\left[(1-D)Y|w\right]}{1-p(w)}\right], \qquad (1.4)$$

where the different $p(w)$ is explicitly accounted. Although (1.3) and (1.4) are mathematically equivalent, when we use the sample average to mimic the population average, (1.4) is easier to work with as it conditions only on the random variable W, but not D. Suppose the functional form of $p(w)$ is known, and we can use the **inverse probability weighted** (IPW) estimator

$$\text{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{d_i y_i}{p(w_i)} - \frac{(1-d_i)y_i}{1-p(w_i)}\right].$$

Notice

$$\begin{aligned}\mathbb{E}\left[\frac{d_i y_i}{p(w_i)}\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{dy}{p(w)}\middle|W=w\right]\right\} = \mathbb{E}\left\{\frac{1}{p(w)}\mathbb{E}\left[dy|w\right]\right\} = \mathbb{E}\left\{\frac{1}{p(w)}\mathbb{E}\left[dy^{(1)}|w\right]\right\} \\ &\overset{\text{cia}}{=} \mathbb{E}\left\{\frac{\mathbb{E}\left[d|w\right]}{p(w)}\mathbb{E}\left[y^{(1)}|w\right]\right\} = \mathbb{E}\left\{\mathbb{E}\left[y^{(1)}|w\right]\right\} = \mathbb{E}\left[y^{(1)}\right],\end{aligned}$$

and similarly

$$\mathbb{E}\left[\frac{(1-d_i)\,y_i}{1-p(w_i)}\right] = \mathbb{E}\left[y^{(0)}\right].$$

We have

$$\mathbb{E}\,[\text{IPW}] = \mathbb{E}[y^{(1)}] - \mathbb{E}[y^{(0)}] = \text{ATE}.$$

We conclude that IPW is an unbiased estimator of ATE.

1.2 Structure and Identification

A key issue to resolve before looking at the realized sample is *identification*. We say a model or DGP is *identified* if the each possible parameter of the model under consideration generates distinctive features of the observable data. A model is *under-identified* if more than one parameter in the model can generate exact the same features of the observable data. In other words, a model is under-identified if from the observable data we cannot trace back to a unique parameter in the model. A correctly specified model is the prerequisite for discussion of identification. In reality, all models are wrong. Thus when talking about identification, we are indulged in an imaginary world. If in such a thought experiment we still cannot unique distinguish the true parameter of the data generating process, then identification fails. We cannot determine what is the true model no matter how large the sample is.

1.2.1 ATE and CEF

Consider a continuous treatment D. Suppose the DGP, or the structural model, is $Y = h\,(D, W, U)$ where D and W are observable and U is unobservable. It is natural to define ATE with the continuous treatment (Hansen's book Chapter 2.30 calls it *average causal effect*) as

$$\text{ATE}\,(d,w) = \mathbb{E}\left[\lim_{\Delta\to 0}\frac{h\,(d+\Delta, w, U) - h\,(d, w, U)}{\Delta}\right] = \mathbb{E}\left[\frac{\partial}{\partial d} h\,(d, w, U)\right],$$

where the continuous differentiability of $h\,(d, w, u)$ at d is implicitly assumed. Unlike the binary treatment case, here d explicitly shows up in ATE (d, w) because the effect can vary at different values of d. ATE here is the average effect in the population of individuals if we hypothetically move d a tiny bit around d, keeping w constant.

If we do not intend to model the underlying economic mechanism $h\,(d, w, u)$, can we still learn ATE (d, w) which bears the structural causal interpretation from the *conditional mean function* (CEF) $m\,(d, w) = \mathbb{E}\,[Y|d, w] = \int y f\,(y|d, w)\,\mathrm{d}y$, which is a mechanical statistical object? The answer is positive under CIA:

$$U \perp D\,|\,W.$$

Notice

$$\begin{aligned}\frac{\partial}{\partial d}m\left(d,w\right) = \frac{\partial}{\partial d}\mathbb{E}\left[Y|d,w\right] = \frac{\partial}{\partial d}\mathbb{E}\left[h\left(d,w,U\right)|d,w\right] &= \frac{\partial}{\partial d}\int h\left(d,w,u\right)f\left(u|d,w\right)\mathrm{d}u\\ &= \int\frac{\partial}{\partial d}\left[h\left(d,w,u\right)f\left(u|d,w\right)\right]\mathrm{d}u\\ &= \int\left[\frac{\partial}{\partial d}h\left(d,w,u\right)\right]f\left(u|d,w\right)\mathrm{d}u + \int h\left(d,w,u\right)\left[\frac{\partial}{\partial d}f\left(u|d,w\right)\right]\mathrm{d}u,\end{aligned}$$

where the second line implicitly assumes interchangeability between the integral and the partial derivative. Under CIA, $\frac{\partial}{\partial d}f\left(u|d,w\right)=0$ and the second term drops out. Thus

$$\frac{\partial}{\partial d}m\left(d,w\right) = \int\left[\frac{\partial}{\partial d}h\left(d,w,u\right)\right]f\left(u|d,w\right)\mathrm{d}u = \mathbb{E}\left[\frac{\partial}{\partial d}h\left(d,w,u\right)|d,w\right] = \mathrm{ATE}\left(d,w\right).$$

It says that if CIA holds, we can learn the causal effect of D on Y by the partial derivative of conditional expectation function (CEF). In particular, if we further assume a linear CEF $m\left(d,w\right)=\mu+\Delta d+\beta_w' w$, then the causal effect is the coefficient Δ.

CIA is the key condition that links the CEF and the causal effect. CIA is not an innocuous assumption. In applications, our causal results are credible only when we can convincingly defend CIA.

Exercise 1. Sometimes applied researchers assume by brute force that $Y=m\left(D,W\right)+U$ is the DGP and $\mathbb{E}\left[U|D,W\right]=0$, where D is the variable of interest and W is the covariate. Under these assumptions,

$$\mathrm{ATE}\left(d,w\right) = \mathbb{E}\left[\frac{\partial}{\partial d}\left(m\left(d,w\right)+u\right)|d,w\right] = \frac{\partial m\left(d,w\right)}{\partial d} + \frac{\partial}{\partial d}\mathbb{E}\left[U|d,w\right] = \frac{\partial m\left(d,w\right)}{\partial d},$$

where the second equality holds if $\frac{\partial}{\partial d}\mathbb{E}\left[u|d,w\right]=\mathbb{E}\left[\frac{\partial}{\partial d}u|d,w\right]$. At a first glance, it seems that the mean independence assumption $\mathbb{E}\left[U|d,w\right]=0$, which is weaker than CIA, implies the equivalence between $\mathrm{ATE}\left(d,w\right)$ and $\partial m\left(d,w\right)/\partial d$ here. However, such slight weakening is achieved by at the cost of assuming $h\left(D,W,U\right)$ follows the additive separable form $m\left(D,W\right)+U$. There is a tradeoff between the restrictiveness of the assumption and the generality of the function form.

The *structural approach* here models the economic mechanism, hopefully guided by economic theory. In recent years, the belief in this econometric tradition is encroached by the more intuitive and straightforward *reduced-form* approach, which documents stylized facts when suitable economic theory is not available. The most prominent reduced-form approach is the potential outcome framework, as in the first part of the lecture notes.

1.3 Summary

There are constant debates about the pros and cons of the two approaches; see *Journal of Economic Perspectives* Vol. 24, No. 2 Spring 2010. In macroeconomics, the so-called Phillips curve, attributed to A.W. Phillips about the negative correlation between inflation and unemployment, is a stylized fact learned from the reduced-form approach. The Lucas critique (Lucas, 1976) exposed its lack of microfoundation and advocated modeling deep parameters that are invariant to policy changes. The latter is a structural approach. Ironically, more than forty years has passed since the Lucas critique, equations with little microfoundation still dominate the analytical apparatus of central bankers. For the structural approach, the devotion to a DGP that no one has ever seen is hardly justifiable. For the reduced-form approach, *ad hoc* regressions are likely falling short of external validity; what is worse those most crucial economic issues are not to be experimented with. You can go to poor villages in a forgotten concern of the world to randomly distribute deworming pills, but there is no way you can persuade the PBOC to experiment with interest rates.

When we care about the structural causality concerning some treatment d to the dependent variable y, under CIA we can find equivalence between ATE and the partial derivative of CEF. All analyses are conducted in population. We have not touched the sample yet.

Historical notes: Regressions and conditional expectations are concepts from statistics and they are imported to econometrics in early time. Researchers at the Cowles Commission (now Cowles Foundation for Research in Economics) — Jacob Marschak (1898–1977), Tjalling Koopmans (1910–1985, Nobel Prize 1975), Trygve Haavelmo (1911–1999, Nobel Prize 1989) and their colleagues — were trailblazers of the econometric structural approach.

The potential outcome framework is not peculiar to economics. It is widely used in other fields such as biostatistics and medical studies. It was initiated by Jerzy Neyman (1894–1981) and extended by Donald B. Rubin (1943–).

`Zhentao Shi. January 10, 2026`