

## Economics 696F, Causal Inference and Program Evaluation

### Lecture Note 1: “Econometrics without Error Terms”

Read articles: Holland (with discussion), Rubin 1974, Lalonde, Rosenbaum ch.2

$T$  = some treatment

For simplicity, assume binary, so  $T = 0, 1$ . (We will also consider multivalued/continuous treatments down the road.)

$Y$  = some outcome

Example:  $T = 0$  means HS education,  $T = 1$  means college education,  $Y$  is log hourly wage at age 40.

Suppose we want to investigate the relationship between  $T$  and  $Y$ , using data

$$(T_1, Y_1, T_2, Y_2, \dots, T_n, Y_n).$$

Assume  $(T_i, Y_i)$  is IID from some joint distribution.

We might start by writing down a linear model

$$Y_i = \beta_0 + \beta_1 T_i + \epsilon_i.$$

**Q:** What does  $\epsilon_i$  mean?

Obviously, the meaning of  $\epsilon_i$  will be closely related to the meaning we give  $\beta_0, \beta_1$ .

**Interpretation 1:** Let

$$\beta_0 := E[Y_i | T_i = 0].$$

(Here, “:=” stands for equality by definition)

$$\beta_1 := E[Y_i | T_i = 1] - E[Y_i | T_i = 0].$$

Then

$$E[Y_i | T_i] = \beta_0 + \beta_1 T_i.$$

Note that linearity is not restrictive, due to binary nature of  $T_i$ .

Define  $\epsilon_i := Y_i - \beta_0 - \beta_1 T_i$ . Then by construction

$$E[\epsilon_i | T_i] = 0.$$

Hence  $E[T_i \epsilon_i] = 0$ , and the OLS estimator  $\hat{\beta}$  is unbiased, consistent, etc.

Aside 1: Easy to show that

$$\begin{aligned}\hat{\beta}_0 &= \frac{\sum_i 1(T_i = 0)Y_i}{\sum_i 1(T_i = 0)} = E[\widehat{Y_i|T_i = 0}], \\ \hat{\beta}_1 &= \frac{\sum_i 1(T_i = 1)Y_i}{\sum_i 1(T_i = 1)} - \frac{\sum_i 1(T_i = 0)Y_i}{\sum_i 1(T_i = 0)} \\ &= E[\widehat{Y_i|T_i = 1}] - E[\widehat{Y_i|T_i = 0}].\end{aligned}$$

Aside 2: Suppose  $T$  is *not* binary. Then it might not be the case that  $E[Y_i|T_i]$  is linear. BUT, we can still define  $(\beta_0, \beta_1)$  as the solution to

$$\min_{\beta_0, \beta_1} E[(Y_i - \beta_0 - \beta_1 T_i)^2].$$

We can interpret  $\beta_0 + \beta_1 T_i$  as the “best linear predictor” of  $Y_i$  given  $T_i$ .<sup>1</sup> Can show that the solution is:

$$\beta_1 = \frac{Cov(T, Y)}{Var(T)},$$

$$\beta_0 = E(Y) - \beta_1 E(T).$$

Moreover, the OLS estimator  $\hat{\beta}$  is consistent for  $\beta$ .

Punchline: can always define  $\beta, \epsilon$  in such a way that  $\hat{\beta}$  is consistent for  $\beta$ .

**Interpretation 2:** disturbance term stands for “unobserved ability” or “motivation.”

To keep this interpretation distinct from interpretation 1, let’s introduce alternative notation:

$$Y_i = \gamma_0 + \gamma_1 T_i + v_i.$$

We suppose that unobserved ability might be correlated with treatment (e.g., high ability people tend to go to college).

Then

$$Cov(T_i, v_i) \neq 0.$$

So OLS will generally be biased, inconsistent for  $\gamma_0, \gamma_1$ .

---

<sup>1</sup>Equivalently, it is the projection of the random variable  $Y$  onto the space of linear functions of  $T$ .

We generally call parameters like  $\gamma_0, \gamma_1$  structural or causal parameters. It is important to note that unlike in interpretation 1, here the linearity and additivity assumptions may be quite important, even if  $T_i$  is binary.

There is a large body of work in econometrics on estimating structural/causal parameters under various assumptions.

**Important Point:** In general,  $\beta \neq \gamma, \epsilon \neq v$ . It may be the case that **both** interpretations are simultaneously valid.

A lot of people just write down the equation  $Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$ , without providing explicit assumptions on  $\epsilon_i$  (or equivalently, on the meaning of  $\beta_0, \beta_1$ ). This is ambiguous, and poor practice.

More generally, from the standpoint of scientific notation, it is not a good idea for convention to allow disturbance terms to be used to stand for two **distinct** concepts:

- an expectational residual;
- or a “structural” but unobserved explanatory factor.

A lot of confusion can be traced to ambiguity about the meaning of  $\epsilon_i$ .

Radical solution: get rid of the error term altogether.

Under interpretation 1, it’s really enough to write

$$E[Y_i|T_i] = \beta_0 + \beta_1 T_i.$$

or

$$(\beta_0, \beta_1) := \arg \min_{\beta_0, \beta_1} E [(Y_i - \beta_0 - \beta_1 T_i)^2].$$

These expressions don’t require  $\epsilon_i$ , yet they are more precise.

It seems harder to get rid of disturbance terms in the structural interpretation 2, but it turns out to be possible, and often pretty fruitful.

## Potential Outcomes Approach

Why do we even care about interpretation 2? Presumably, we have in mind that  $\gamma_0, \gamma_1$  represent the causal or structural relationship between  $T$ , and  $Y$ , and not mere association.

Example: it might be that college educated workers earn higher wages on average than high school graduates (so  $\beta_1 > 0$ ). But part of that difference might reflect higher innate ability or motivation, rather than a real effect of college education.

So we want a general way to define and notate cause-and-effect relationships.

Define:

$$\begin{aligned} Y_i(0) &= \text{Potential outcome under treatment 0 ('control')} \\ Y_i(1) &= \text{Potential outcome under treatment 1 ('treatment')} \end{aligned}$$

Important implicit assumption:

**Stable Unit Treatment Value Assumption:** (Neyman, 1923, Rubin, 1980). potential outcome for  $i$  does not depend on the treatments received by other units.

Causal effect for individual  $i$ :  $Y_i(1) - Y_i(0)$ .

Often, we focus on the average treatment effect:

$$ATE := E[Y_i(1) - Y_i(0)] = E(Y_i(1)) - E(Y_i(0)).$$

Observed outcome:

$$Y_i := (1 - T_i)Y_i(0) + T_iY_i(1).$$

so we only observe one of the two potential outcomes. The other potential outcome is counterfactual.

Connection to linear structural model  $Y_i = \gamma_0 + \gamma_1 T_i + v_i$ : let

$$\begin{aligned} Y_i(0) &:= \gamma_0 + v_i \\ Y_i(1) &:= \gamma_0 + \gamma_1 + v_i \end{aligned}$$

So, the linear structural model is a special case of the general potential outcome model, with the restriction that

$$Y_i(1) - Y_i(0) = \gamma_1 \quad \text{for all } i.$$

We call this the constant treatment effect assumption.

In general, the potential outcome model allows for heterogeneous treatment effects across individuals. However, since we only observe one of the two potential outcomes for any individual, there is a limit on how much we can learn about the distribution of individual-level treatment effects.

## Interpretation 1 vs. 2

Recall we defined

$$\beta_1 = E[Y_i|T_i = 1] - E[Y_i|T_i = 0].$$

Does  $\beta_1 = ATE$ ?

In general, no.

Consider

$$\begin{aligned} E[Y_i|T_i = 1] &= E[(1 - T_i)Y_i(0) + T_iY_i(1)|T_i = 1] \\ &= E[Y_i(1)|T_i = 1] \\ &\neq E[Y_i(1)] \text{ in general} \end{aligned}$$

Likewise,

$$E[Y_i|T_i = 0] = E[Y_i(0)|T_i = 0] \neq E[Y_i(0)].$$

Example: Suppose that for all  $i$ ,  $Y_i(1) = Y_i(0)$ . So  $ATE = 0$ . There is no causal effect of college. However, if  $T_i$  is positively correlated with  $Y_i(0) = Y_i(1)$  (more motivated students go to college), then we would have

$$E[Y_i(1)|T_i = 1] > E[Y_i(1)],$$

$$E[Y_i(0)|T_i = 0] < E[Y_i(0)].$$

Hence,  $\beta_1 > 0$ .

We could have told much the same story using  $\epsilon_i$  and  $v_i$ , but this way avoids notational confusion, and every variable is at least potentially observable. It is worth getting used to this notation.

### Aside: some philosophy of causality

There is a rich literature in analytic philosophy on causality. There have been two main approaches to defining causality:

- **Regularity approaches:** Hume: “we 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 second.” (from *An Enquiry Concerning Human Understanding*, section VII.)
- **Counterfactual approaches:** Hume: “Or, in other words, where, if the first object had not been, the second never had existed.” (from *An Enquiry Concerning Human Understanding*, section VII.)

Regularity approach: a minimal constant conjunction between the two objects. Suppes: a probabilistic association between the two objects, which cannot be explained

away by other factors. This is the basic idea behind Granger causality, a concept used in time series.

Difficulty: what are the other factors? If we limit ourselves to only observable factors, this is somewhat unsatisfying. If factors are not all observable, then what?

Counterfactual approach: Lewis (1973) proposes to imagine a range of possible worlds. Given two propositions  $A$  and  $C$ , a counterfactual  $A \rightsquigarrow C$  is true if the “closest” hypothetical world (to our own) such that  $A$  holds, has  $C$  holding.

Connection to our notation: consider the effect of college for a single individual  $j$ . Suppose that in our world,  $T_j = 1$  (so individual goes to college). Then the closest world with  $T_j = 1$  is our world, and the associated counterfactual  $Y_j(1)$  is simply the observed outcome.

To define  $Y_j(0)$ , look for the closest world which has  $T_j = 0$ , and take the outcome in that world.