Advanced Econometrics
1st year MPhil

Introduction to Potential Outcome Framework and Causal Inference

Martin Weidner

University of Oxford

2022-23, Hilary Term

"The credibility revolution" (some introductory comments)

Some interesting reading

[1] E.E. Leamer, *Let's Take the Con Out of Econometrics*, AER, 1983.

[2] J.D. Angrist & J.-S. Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics*, Journal of Economic Perspectives, 2010.

[3] E.E. Leamer, *Tantalus on the Road to Asymptopia*, Journal of Economic Perspectives, 2010.

[4] C.A. Sims, *But Economics Is Not an Experimental Science*, Journal of Economic Perspectives, 2010.

Some interesting listening (podcast episodes)

- From 2010: `https://www.econtalk.org/leamer-on-the-state-of-econometrics/`
- From 2014: `https://www.econtalk.org/joshua-angrist-on-econometrics-and-causation/`

Potential outcomes and randomized experiments

Textbook and review for causal inference:
Imbens and Rubin (2015); Imbens (2004)

Introduction

In everyday life, we make causal statements in a casual way.

- I should take aspirin to relieve my headache.
- My headache has gone because I took aspirin.
- Smoking causes cancer.
- She got a good job because she went to college.
- If I could have studied abroad, I would be able to speak English more fluently.

How do we define "cause and effect" mathematically?

Causality and Counterfactuals

- As these examples illustrate, we tend to draw a causal statement in a comparison between the observed/realized state and the counterfactual state.
- Already David Hume defined "cause and effect" using the concept of counterfactual.

Example

- Suppose I took aspirin and my headache went away.
- Consider the counterfactual outcome: what would have happened to me if I wouldn't have taken aspirin?
- An effect is the difference between what did happen (observed) with the actual treatment and what would have happened (counterfactual) if I would have had a different treatment.

Since we never observe the counterfactual, it is impossible to obtain causal evidence from the single observation!

Potential Outcome Framework

Neyman-Rubin's Causal Model: Proposed independently by Neyman (1923) and Rubin (1974).

Let $d \in \mathcal{D}$ be the treatment, or the policy that a planner can in principle manipulate.

Notation: Potential Outcomes

1. Define $\{Y(d) : d \in \mathcal{D}\}$ as a set of potential outcomes of an individual, where $Y(d)$ is the outcome that would be observed if the individual would have received treatment (been exposed to policy) $D = d$.
2. Her/his observed outcome Y satisfies

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

"Stable Unit Treatment Value Assumption" (SUTVA)

- Multiple units are needed to infer causality

Assumption: SUTVA

(i) The potential outcomes for any unit do not vary with the treatments assigned to other units.

(ii) For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

$\Rightarrow$ (i) rules out "spillovers", "peer effects", "general equilibrium effects", causal inference based on time-series data, etc

$\Rightarrow$ (ii) requires treatments to be clearly defined and uniform, e.g. the aspirin you are taking to relieve your headache should not be old (and therefore less effective).

Causal Estimands

- Let i represent unit, which could be a firm, an individual person, a household, etc. at a particular point in time.
- Individual causal effect is defined by $Y_i(d) - Y_i(d')$.
- Population level Average Treatment Effect (ATE):

$$\theta_{\text{ATE}} = \mathbb{E}\left[Y_i(d) - Y_i(d')\right]$$

- Sample ATE:

$$\theta_{\text{SATE}} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i(d) - Y_i(d')\right]$$

- ATE of subpopulation Ω is given by $\mathbb{E}\left[Y_i(d) - Y_i(d') \,\middle|\, i \in \Omega\right]$, for example:

$$\theta_{\text{CATE}}(x) = \mathbb{E}\left[Y_i(d) - Y_i(d') \,\middle|\, X_i = x\right]$$
$$\theta_{\text{ATT}}(x) = \mathbb{E}\left[Y_i(d) - Y_i(d') \,\middle|\, D_i = d\right]$$

Binary Treatment: Experimental Data

- From now on, we assume a binary treatment $d \in \{0, 1\}$.
- Given n units, we call $\{i : D_i = 1\}$ the treatment group and $\{i : D_i = 0\}$ the control group.
- Experimental data: the assignment mechanism of treatment is both known and controlled by the researcher.
- Let $\boldsymbol{Y}(d) = (Y_1(d), \ldots, Y_n(d))'$, $\boldsymbol{D} = (D_1, \ldots, D_n)'$, $\boldsymbol{X} = (X_1, \ldots, X_n)$.
- The assignment mechanism is defined by the probability law of $\boldsymbol{D}$ given $(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X})$.

Randomized Experiment

Let $p_i(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X}) := \Pr(D_i = 1 | \boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X})$ Consider assignment mechanisms that satisfy the following properties:

- probabilistic: for all $1 \leq i \leq n$,

$$0 < p_i(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X}) < 1.$$

- unconfounded: for all $1 \leq i \leq n$,

$$p_i(\boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X}) = e_i(\boldsymbol{X}).$$

Theorem

If an experimental design is probabilistic and unconfounded and $e_i = e_i(\boldsymbol{X})$ is known, then θ_{SATE} can be unbiasedly estimated.

Proof

Remember that $e_i = \Pr(D_i = 1 | \boldsymbol{Y}(1), \boldsymbol{Y}(0), \boldsymbol{X})$ is known here.

Let

$$\widehat{\theta}_{\text{SATE}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{D_i Y_i}{e_i} - \frac{(1 - D_i) Y_i}{1 - e_i} \right].$$

We obtain that

$$\mathbb{E}\left[\widehat{\theta}_{\text{SATE}} \,\middle|\, \boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X} \right] = \theta_{\text{SATE}}.$$

Examples of such assignment mechanisms

- Bernoulli assignment: flip a coin (the probability of i being treated can depend on her/his characteristic X_i)
- Completely randomized experiment: randomly select n_t (prespecified constant) treated units out of n.
- Stratified randomized experiment: Partition the sample into subsamples (strata) according to $\boldsymbol{X}$, and conduct completely randomized experiment in each stratum.
- Paired randomized experiment: Each stratum contains only two units.

$\Rightarrow$ For all these mechanisms we have that $\widehat{\theta}_{\text{SATE}}$ is unbiased but

$$\text{Var}\left[\widehat{\theta}_{\text{SATE}} \,\middle|\, \boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}\right]$$

will be different for different assignment mechanisms.

Natural Experiment: Cholera in London

The earliest and one of the most successful natural experiment took place in London.

- In 1848 and 1853-54, there were cholera epidemics in London. Meanwhile, cholera was believed to be caused by "bad air." The microbiology was not yet developed and existence of germ was not widely accepted.
- A physician John Snow claimed that cholera was caused by contaminated water.
- He found the three water companies supplied water to the central London. Southwark and Vauxhall had the intake pipe in the sewage-polluted sections of the Thames. By contrast, Lambeth moved its intake pipe upstream of Thames due to the Metropolis Water Act in 1852, which prohibits any water company to extract water from the tidal reaches of the Thames.

Natural Experiment: Cholera in London

Snow collected data and tabulated cholera death and source of water.

	No. Houses	Death	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

- The death rate of S & V is 9 times higher than Lambeth!
- This data can be analyzed as if they had resulted from a randomized experiment.

Identification and estimation under unconfoudedness

Causal Inference for Observational Studies

- Experimental studies are not always possible, and we often have to rely on observational data, in which assignment of treatment is not under the control of researchers.
- From now on, we assume the data $\{(Y_i, D_i, X_i) : i = 1, \ldots, n\}$ consists of n i.i.d. observations drawn from a population.
- $X_i \in \mathbb{R}^K$: a vector of pre-treatment covariates (characteristics).
- Whenever observational data are used, we always have to worry about confounding factors (confounders): factors that are associated with both an assignment (choice) of treatment and outcomes.

Simpson's paradox

Kidney stone data in Charig, Webb, Payne and Wickham (1986): success rates (the number of successes/patients treated)

Treatment	Overall	small stones	large stones
Open surgery	78% (273/350)	93% (81/87)	73% (192/263)
Needle surgery	83% (289/350)	87% (234/270)	69% (55/80)

$\Rightarrow$ Which treatment is more effective?

Unconfoundedness

Assumption: Unconfoundedness

An assignment mechanism satisfies ***unconfoundedness*** *(= selection on observables) if*

$$(Y(1), Y(0)) \perp D|X$$

$\Rightarrow$ i.e. X captures all the confounders; among the units that are observationally identical in terms of pretreatment observable characteristics X, treatment is randomly assigned (chosen).

Propensity scores and overlap

- Define the propensity score:

$$e(x) := \Pr(D = 1|X = x),\ x \in \mathcal{X} \subset \mathbb{R}^K.$$

Assumption: Overlap

An assignment mechanism satisfies ***overlap*** *if*

$$0 < e(x) < 1,\ \forall x \in \mathcal{X}.$$

Identification of θ_{ATE}

Theorem:

Under unconfoundedness and overlap, θ_{ATE} is nonparametrically identified:

$$\theta_{\text{ATE}} = \int_{\mathcal{X}} [\mathbb{E}(Y|D=1,X) - \mathbb{E}(Y|D=0,X)]\, dF_X(x).$$

- The expression for θ_{ATE} given in the theorem shows that it is identified, because $\mathbb{E}(Y|D=1,X)$, $\mathbb{E}(Y|D=0,X)$ and $F_X(x)$ are all identified from the data on (Y,D,X).
- Equivalently, we can show that θ_{ATE} is identified by "recycling" the unbiasedness proof of $\widehat{\theta}_{\text{SATE}}$ above:

$$\theta_{\text{ATE}} = \mathbb{E}\left[\frac{DY}{e(X)} - \frac{(1-D)Y}{1-e(X)}\right]$$

Remarks on unconfoundedness

How should we choose a set of covariates to make the unconfoundedness assumption credible?

- Unconfoundedness is not testable, and often a controversial assumption.
- We want to make sure that the covariates are determined and measured before treatment assignment.
- Exploiting available background knowledge. E.g., caseworkers assign job training programs according to a known index that depends on observable characteristics (previous earning, education, sex, etc.), with probabilistic rationing due to a capacity constraint (e.g., first come first serve).
- Many estimation procedures for average treatment effects rely on unconfoundedness. We discuss some of them in the following.

Many different approaches to estimation:

1. Regression approach:
   - 1.1 Regressions on covariates
   - 1.2 Regressions on $e(x)$/propensity score blocking
2. Matching
3. Propensity score weighting
4. Doubly robust estimation

1.1. Regressions on covariates

Typical parameters of interest:

$$\theta_{\text{ATE}} = \int_{\mathcal{X}} [\mathbb{E}(Y|D=1, X=x) - \mathbb{E}(Y|D=0, X=x)] dF_X(x)$$

$$\theta_{\text{ATT}} = \int_{\mathcal{X}} [\mathbb{E}(Y|D=1, X=x) - \mathbb{E}(Y|D=0, X=x)] dF_{X|D=1}(x)$$

- Estimate $\mathbb{E}(Y|D=1, X)$ and $\mathbb{E}(Y|D=0, X)$ by regressions (OLS/WLS).
- Take the sample average of $\widehat{\tau}(x_i) := \widehat{E}(Y|D=1, X=x_i) - \widehat{E}(Y|D=0, X=x_i)$.
- Simple to implement but estimates can be very sensitive to the chosen regression specification, see e.g. LaLonde (1986).

1.2. Regressions on propensity score

Theorem (Rosenbaum and Rubin 1983)

Under unconfoundedness we have:

$$(Y(1), Y(0)) \perp D \,\big|\, e(X)$$

- The theorem motivates regressions on the estimated propensity scores
  1. Estimate $e(x)$ by a (flexible) binary regression.
  2. Run regressions to estimate $\mathbb{E}(Y|D=d, e(X))$ using the estimated propensity scores.
  3. Take the sample average of $\widehat{\tau}(x_i) := \widehat{E}(Y|D=1, e=\widehat{e}(x_i)) - \widehat{E}(Y|D=0, e=\widehat{e}(x_i))$.
- Propensity score blocking method estimate $\mathbb{E}(Y|D=d, e(X))$ by piecewise constants.

2. Matching Estimation I

- The idea of matching estimators goes back to William G. Cochran's work in 1960's, e.g. Cochran (1968).
- Given unconfoundedness, estimation of causal effects can be done by comparing the outcomes of the individuals who look identical in terms of X (or $e(X)$), but differ in their treatment status.
- For each unit in the treatment (control) group, we impute the missing potential outcomes $Y(0)$ (resp. $Y(1)$) using the observed outcomes of the control (resp. treated) units that are similar in terms of X.
- Can be viewed as a nonparametric way of implementing regression approach.

2. Matching Estimation II

M-nearest neighbor covariate matching: For each i in the sample,

$$\widehat{Y}_i(0) = \begin{cases} Y_i & \text{if } D_i = 0, \\ \frac{1}{M}\sum_{j\in J_M(i)} Y_j & \text{if } D_i = 1, \end{cases}$$

$$\widehat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in J_M(i)} Y_j & \text{if } D_i = 0, \\ Y_i & \text{if } D_i = 1, \end{cases},$$

$$\widehat{\tau}_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i(1) - \widehat{Y}_i(0)\right), \quad \widehat{\tau}_{\text{ATT}} = \frac{1}{n_t}\sum_{i:D_i=1}\left(Y_i - \widehat{Y}_i(0)\right),$$

where $J_M(i)$ is the subset of $\{1,\ldots,n\}$ with size M (fixed), specifying the M-nearest units in the opposite group of i.

3. Propensity score weighting methods

- We already discussed the identity:

$$\theta_{\text{ATE}} = E\left[\frac{DY}{e(X)} - \frac{(1-D)Y}{1-e(X)}\right].$$

$\Rightarrow$ A propensity score weighting method estimates θ_{ATE} by a sample analogue:

$$\widehat{\theta}_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i Y_i}{\widehat{e}(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{e}(X_i)}\right]$$

where $\widehat{e}(X_i)$ is obtained by e.g. a logistic regression of D_i in X_i.

4. Doubly robust estimation of ATE

(an example of "Neyman orthogonality", which we discuss later)

- Let $\mu_d(x) = \mathbb{E}(Y|D=d, X=x)$. You showed in your problem set that the formula

$$\theta_{\text{ATE}} = \mathbb{E}\left[\frac{(Y-\mu_1(X))D}{e(X)} - \frac{(Y-\mu_0(X))(1-D)}{1-e(X)} + (\mu_1(X)-\mu_0(X))\right]$$

  is robust towards misspecification of either $e(x)$ or $\mu_d(x)$

- The corresponding doubly robust (DR) estimator reads

$$\widehat{\theta}^{\text{DR}}_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(X_i,\widehat{\beta}) - \mu_0(X_i,\widehat{\beta}) + \frac{D_i(Y_i-\mu_1(X_i,\widehat{\beta}))}{e(X_i,\widehat{\gamma})} - \frac{(1-D_i)(Y_i-\mu_0(X_i,\widehat{\beta}))}{1-e(X_i,\widehat{\gamma})}\right]$$

  where $\mu_d(X_i,\beta)$ and $e(x,\gamma)$ are (non-) parametric specifications and $\widehat{\beta}$ and $\widehat{\gamma}$ denote the corresponding estimates.

$\Rightarrow$ A consequence of the above robustness is that

$$\left|\widehat{\theta}^{\text{DR}}_{\text{ATE}} - \theta_{\text{ATE}}\right| = O_P\left(\frac{1}{n}\right) + O_P\left(\left\|\widehat{\beta}-\beta\right\| \cdot \|\widehat{\gamma}-\gamma\|\right)$$

LATE interpretation of IV estimation

Linear IV

- Recall the standard linear IV model,

$$Y_i = \alpha + \beta D_i + \epsilon_i,$$
$$D_i = \gamma_0 + \gamma_1 Z_i + \nu_i,$$

  if the instrumental variable Z satisfies the instrument exclusion restriction, $\mathbb{E}(Z_i \epsilon_i) = 0$, and the rank condition, $\gamma_1 \neq 0$, then the two-stage least squares (2SLS) estimator, $\widehat{\beta}_{\text{2SLS}}$, consistently estimates β.

- If we interpret the outcome equation as a structural equation with homogeneous effect, $Y_i(1) - Y_i(0) = \beta$, then 2SLS consistently estimates ATE.

$\Rightarrow$ **Question**:

If the treatment effects are heterogeneous, what is $\widehat{\beta}_{\text{2SLS}}$ estimating?

Example: draft lottery in Vietnam war (Angrist 1990; Angrist and Krueger 1992)

Estimation of the effect of serving in military on earnings. A simple comparison of the means do not tell the causal effects since those choosing to serve in the military are likely to be different from those who don't.

- Y_i: post-Vietnam war earnings (weekly wage) in 1980-84 obtained from CPS.
- D_i: Indicator for Vietnam war veteran status.
- Z_i: Binary indicator for whether or not the draft lottery draws i's birthdate in high places in the draft eligibility list.

2SLS estimate

Tabulate D and Z.

	Z=1	Z=0
D=1	954	1,507
D=0	2,280	6,896

- First stage:

$$D_i = 0.18 + 0.12 Z_i + \widehat{\epsilon}_i.$$

- Reduced-form regression for Y:

$$Y_i = 237 - 4.55 Z_i + \widehat{v}_i.$$

$\Rightarrow \widehat{\beta}_{\text{2SLS}} = \frac{-4.55}{0.12} = -40$

Local Average Treatment Effects

Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) analyze the question in the potential outcome framework.

- Define potential outcomes and potential treatment selection:

$$\left(Y_i\left(z,d\right) : z \in \{1,0\},\ d \in \{1,0\}\right) \text{ and } \left(D_i\left(z\right) : z \in \{1,0\}\right).$$

Assumption: Random Assignment (RA)

$$\left(Y(1,1), Y(1,0), Y(0,1), Y(0,0), D(1), D(0)\right) \perp Z$$

$\Rightarrow$ IV is randomly assigned independent of any unobserved heterogeneity determining one's outcome and/or treatment selection.

Exclusion Restriction

Assumption: Exclusion Restriction (ER)

$$Y(1,d) = Y(0,d) =: Y(d)$$

(with probability one)

$\Rightarrow$ IV does not directly affect the outcome.

Instrument Monotonicity

Assumption: Instrument Monotonicity (IM, no-defier)

$$D(1) \geq D(0)$$

(with probability one)

$\Rightarrow$ IV affects individuals treatment selection only in one direction.

- Define selection types:

$$T_i = \begin{cases} at, & \text{always-taker: } D_i(1) = D_i(0) = 1, \\ c, & \text{complier: } D_i(1) = 1,\ D_i(0) = 0, \\ nt, & \text{never-taker: } D_i(1) = D_i(0) = 0, \\ df, & \text{defier: } D_i(1) = 0,\ D_i(0) = 1. \end{cases}$$

Local Average Treatment Effects

Under weak regularity conditions (WLLN applicable and $\Pr(D=1|Z=1) - \Pr(D=1|Z=0) \neq 0$) we have

$$\widehat{\beta}_{\text{2SLS}} \to_p \beta_* := \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(D|Z=1) - \mathbb{E}(D|Z=0)}$$

LATE Theorem (Imbens and Angrist 1994)

Assume RA, ER, IM, and $\Pr(D=1|Z=1) - \Pr(D=1|Z=0) > 0$. Then,

$$\beta_* = \mathbb{E}\left[Y(1) - Y(0) \,\middle|\, T = c\right]$$

Remarks/Implications

- $\widehat{\beta}_{\text{2SLS}}$ estimates compliers effect (LATE) instead of ATE.
- Since compliers are defined in terms of $(D(1), D(0))$, we cannot identify who are the compliers in the data.
- The complier subpopulation is specific to the instrument used; different IVs estimate different LATEs.
- With heterogeneous treatment effects, we cannot identify ATE in the IV setting described above.

Extensions to non-binary instrument/treatment and additional covariates

- Multi-valued ordered D. Angrist and Imbens (1995)
- Multi-valued unordered D. Heckman, Urzua and Vytlacil (2008)
- Continuous D in simultaneous equation models (demand and supply): Angrist, Graddy and Imbens (2000)
- Multi-valued/continuous Z. Imbens and Angrist (1994), Heckman and Vytlacil (2005)

Bibliography I

Angrist, J. and A. B. Krueger (1992). Estimating the payoff to schooling using the vietnam-era draft lottery.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The american economic review*, 313–336.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies 67*(3), 499–527.

Angrist, J. D. and G. W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association 90*(430), 431–442.

Bibliography II

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Charig, C. R., D. R. Webb, S. R. Payne, and J. E. Wickham (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed) 292*(6524), 879–882.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295–313.

Heckman, J. J., S. Urzua, and E. Vytlacil (2008). Instrumental variables in models with multiple outcomes: The general unordered case. *Annales d'Economie et de Statistique*, 151–174.

Bibliography III

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica 73*(3), 669–738.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics 86*(1), 4–29.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Bibliography IV

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences 10*, 1–51.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.