Counterfactuals and Causal Inference in the Social Sciences

Professor Florencia Torche

Spring 2017

May 10, 2017

Which of the following statements is true?

1) Propensity score matching is a substantial improvement over regression (or any other parametric model).

OR

2) Propensity score matching is subject to the same untestable assumption as regression analysis: Conditional on observed confounders, treatment allocation is independent of potential outcomes ("conditional ignorability").

Sensitivity analysis

- How strongly should an unmeasured confounder affect selection into treatment in order to annul the conclusions about a causal effect from a matching analysis?

- **Rosenbaum Bounds**: Assumes that, conditional on observed covariates, individuals differ on the basis of an unobserved confounder so that treated cases have an *odds of receiving the treatment* that is up to $\Gamma >= 1$ greater than the odds for control cases.
  - If $\Gamma = 1$, treated cases are equally likely to receive the treatment as control cases (no hidden bias).
  - If $\Gamma = 1.5$, treated cases are 50 percent more likely to receive the treatment than observationally identical control cases due to an unobserved confounder.

Sensitivity analysis (2)

- For each value of hidden bias measured by Γ, the sensitivity analysis computes end points of the bounds for the significance level (p-value) of the test of the null hypothesis that the treatment has no effect using the Wilcoxon signed rank test for matched pairs.
- Γ = 1 describes the situation of no hidden bias. In such situation, a single p-value is obtained -- the p-value for a randomized experiment.
- For Γ > 1 one obtains an interval of p-values reflecting uncertainty due to hidden bias, which could result in overestimating or underestimating the treatment effect.
- As Γ increases, the interval becomes larger and eventually becomes uninformative, including both large and small p-values, among them, p-values larger than a threshold to determine statistical significance (e.g. p=.05).

Example Lee (2010) “The early socioeconomic effects of teenage childbearing: A propensity score matching approach”

Table 4: **A sensitivity analysis using the Rosenbaum bounds of the causal effects of teenage childbearing**

	Γ	*p*-critical
A. Educational Attainment		
Dropout	1.6	0.002
	1.7	0.007
	1.8	0.023
	1.9	0.057
College attendance	1.7	0.002
	1.8	0.010
	1.9	0.039
	2.0	0.110

Γ captures the increase in the odds of receiving the treatment among the treatment cases compared with the control cases.

Sensitivity analysis (3)

- Hidden bias depends on two parameters: The association between unobserved confounder and treatment allocation (Γ) and the association between unobserved confounder and outcome (what Harding [2003] calls Δ).

- The Rosenbaum bounds strategy provides a very conservative test of hidden bias, because it is based on the assumption of nearly perfect association between the unobserved confounder and the outcome (**Δ~1) (Rosenbaum 2002 chapter 4).**

- The value of Γ associated with a p value >.05 indicates that the confidence interval for the treatment effect would include zero if an unobserved confounder caused the odds ratio of treatment assignment to differ between treatment and control groups by Γ, *and* if the confounder's effect on the outcome was so strong as to almost perfectly determine whether the outcome would be higher for the treatment over control case in each pair of matched cases.

Rosenbaum bounds in stata

- rbounds for continuous outcome
- mhbounds for dichotomous outcome

Appendix: Wilcoxon signed rank test for matched pairs

- Non-parametric test for matched data (treatment – control)
- Examines whether population medians between the two distributions are different from zero.
- Analogous to paired t-test when distribution of differences between pairs are (severely) non-normally distributed.

Appendix: Wilcoxon signed rank test for matched pairs: Example

i	T	C	T-C	|T-C|	Ordered |T-C|	Ranks	Signed ranks
1	85	75	10	10	5	1	-1
2	70	50	20	20	10	3	-3
3	40	50	-10	10	10	3	3
4	65	40	25	25	10	3	3
5	80	20	60	60	15	5	5
6	75	65	10	10	20	6	6
7	55	40	15	15	25	7	7
8	20	25	-5	5	60	8	8

H0=Median difference=0

HA=Median difference>0 (one-sided test)

Test statistic W is the smaller of W+ (sum of the positive ranks) and W- (sum of the negative ranks). (Smaller values of W are less likely under the null!).

W+ =32, W-=4

Test statistic=-4, Critical value (one-tailed, α=.05)=5, so reject the null.

Non-binary treatments

- Area of continuous research.
- Simple solution: Suppose one has a control and two treatments.
- Same as 3 treatments (T = 0; 1; 2, does not matter that 0 is “control").
- If outcomes unordered, you can do pairwise comparison: Compare effect of 0 to 1, 0 to 2, and 1 to 2 separately, obtaining balance in each analysis.
- Similar to multinomial logit approach.

Non-binary treatments

- Approach suggested by Imbens, 2000 “The role of the Propensity score in Estimating Dose-Response Functions”. *Biometrika* 87(3):706–710, based on propensity score weighting.
- First predict the probability of belonging to each of the j treatment groups using multinomial regression (logit/probit).
- Each unit will have j probabilities attached to them, one for each of the treatment groups.
- Expected values of the outcome for each treatment condition can be calculated by weighting the outcomes of people in that group using the inverse of each unit’s probability of being in that treatment condition.

Continuous treatment

- Approach suggested by Imai and Van Dyk 2004. “Causal Inference with General Treatment Regimes: Generalizing the Propensity Score” *Journal of the American Statistical Association* 99 (467): 854-866.
- Calculate propensity scores by running a standard linear regression (or the like) of the treatment variable t on confounders and obtain predicted values for t (θ,predicted xb values), which I & VD call “propensity function”.
- Stratify the data based on θ.
- Within each stratum, regress Y on t and θ.
- Calculate weighted average of within-strata estimates of effect of t to obtain average effects.

Propensity score methods using multilevel data

- PSM: Arpino and Mealli, 2011 “The specification of the propensity score in multilevel studies” *Computational Statistics and Data Analysis* 55: 1770–1780.
- PSW: Li, Zaslavsky, Landrum. 2013. “Propensity Score Weighting with Multilevel Data” *Statistics in Medicine* 32(19): 3373–3387.
- Zubizarreta & Keele. 2017. “Optimal Multilevel Matching in Clustered Observational Studies” *Journal of the American Statistical Association*, in press.

Instrumental Variables (IVs)

Instrumental Variables

- Research question: Effect of military service (in Vietnam) on later earnings.
- Military service is correlated with multiple factors that have an independent effect on earnings –personality, motivation, ability, socioeconomic resources, risk aversion, etc.
- Some of these factors are unobserved/unobservable.
- Military services is likely endogenous because of the unobserved selectivity of those who serve.
- PS-based methods unlike to address this issue.

Instrumental variable

- We could use an instrument (Z), a variable correlated with military service (T) but uncorrelated with unobserved factors shaping earnings (i.e. uncorrelated wit the error term in the equation predicting earnings).

- Instrument needs to be a source of exogenous variation in military service. A plausible source is Vietnam era draft lotteries, in which priority for induction was determined by Random Sequence Numbers (RSN).

- RSN provides a “natural experiment” – event that occur in the natural or social world, which is allocated at “as random”.

Instrumental variable: Formalization

- OLS Regression: $Y = \beta_0 + \beta_1 T + \varepsilon$

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

  $\beta_1 = Cov(Y,T)/Cov(T,T)$

- But if X is correlated with ε, then β_1 is biased….

- If Z is correlated with T but is not correlated with Y (via ε), then:

  $Y = \beta_0 + \beta_{IV} Z + \varepsilon$

  $Cov\ (Y,T) = \beta_{IV} Cov(Z,T)$

$$\beta_{IV} = \frac{Cov\ (Y,Z)}{Cov(T,Z)} = \frac{Cov(Y,Z)/Var(Z)}{Cov(T,Z)/Var(Z)} = \text{IV estimator}$$

IV Estimation

$$\beta_{IV} = \frac{\text{Cov}(Y,Z)}{\text{Cov}(T,Z)} = \frac{\text{Cov}(Y,Z)/\text{Var}(Z)}{\text{Cov}(T,Z)/\text{Var}(Z)} = \frac{\text{Reduced Form}}{\text{First stage}}$$

- The IV estimator is the ratio of two covariances/ regression coefficients.

- (Under certain assumptions necessary for identification): The IV estimator is the ratio of two causal effects.

Intent to Treat (ITT) Effect

- The IV estimator can be seen as the ratio of two intent-to-treat effects (ITT):

$$\frac{E(Y|Z=1) - E(Y|Z=0)}{E(T|Z=1) - E(T|Z=0)} =$$

Effect of being assigned into treatment on outcome

Effect of being assigned into treatment on receiving the treatment

Where the instrument is the variable that assigns individuals into treatment.

IV Estimation

Wald Estimator:

$$\frac{\text{Cov}(Y,Z)}{\text{Cov}(T,Z)} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(T|Z=1) - E(T|Z=0)}$$

- Wald estimator is the simplest IV estimator uses a single binary (0/1) instrument to estimate a model with one endogenous regressor and no covariates.

- This can be easily extended to include covariates.

IV Estimation

Two-Stage Least Square Estimator (2SLS, TSLS):

- First stage: Regress *T* on *Z* to obtain $\hat{T}$
- Second stage: Regress Y on $\hat{T}$
- The coefficient on $\hat{T}$ is the IV estimate
- Standard errors won't be correct, but that can be addressed.

Assumptions for IV estimation: Covariance between treatment and instrument.

1) Non-zero covariance between the instrument and the treatment: Z is correlated with T ("first stage"), and correlation is "sufficiently large".

$$E[T_i(1) - T_i(0)] \sim= 0$$

Assumptions for IV estimation: Random assignment

2) Instrument is as good as randomly assigned (i.e. independent of potential outcomes conditional on covariates).

$$\Pr(Z=c) = \Pr(Z=c')$$

Assumptions for IV estimation: Exclusion restriction

3) Any effect of Z on Y must be via an effect of Z on T. Instrument only affects outcome through the treatment ("no alternative pathways of influence").

$$Y(Z,T) = Y(Z',T) \text{ for all } Z, Z' \text{ and for all } T$$

If the treatment is not different even if the treatment assignment (instrument) is different, then the outcome should not be different either.

The treatment assignment (instrument) must be unrelated with potential outcomes once treatment received is taken into account.

Assumptions for IV estimation: Monotonicity

4) Monotonicity:

$$T_i(1) \geq T_i(0)$$

The instrument has a positive or zero OR a negative or zero effect on the treatment for all individuals. It cannot have a positive effect on some individuals and a negative effect on others.

While the instrument may have no effect on some individuals, all of those who are affected are affected in the same way.

Assumptions for IV estimation

5) Stable Unit Treatment Value Assumption (SUTVA):

- *No-interference:* Outcome of individual *i* does not depend on the treatment assignment of other individuals, or on the treatment status of other individuals.

$$\text{If } Z_i = Z'_i \text{ then } T_i(Z) = T_i(Z')$$
$$\text{If } Z_i = Z'_i \text{ and } T_i = T'_i \text{ then } Y_i(Z,T) = Y_i(Z',T')$$

- Treatment assigned to one unit is the same as the treatment assigned to another unit (i.e., there are no “versions” or “types” of treatment that are not recorded in our treatment assignment vector).

Reminder: SUTVA formally

$$y_i(T)\begin{bmatrix} T_1 \\ T_2 \\ . \\ . \\ T_i \\ . \\ . \\ T_{n-1} \\ T_n \end{bmatrix} = y_i(T')\begin{bmatrix} T'_1 \\ T'_2 \\ . \\ . \\ T_i \\ . \\ . \\ T'_{n-1} \\ T'_n \end{bmatrix}$$

where T and T' denote two treatment assignment regimes for all units in the population.

- Let Yi(T) denote the outcome for unit i when the units are assigned according to T and Yi(T') denote the outcome for unit i when the units are assigned according to T'.
- SUTVA requires that Yi(T) = Yi(T') for all versions of T.
- The outcome of unit i depends only on the treatment assignment of i, and not on the assignment of other units.

IV – What is the effect being estimated? (1)

- The effect of the treatment is not necessarily homogeneous across the population.
- The IV estimator only captures the effect of the treatment among those whose treatment status changes because of the instrument.

- Example:
  - Effect of military service among men who served because they received a low lottery number, but not among men who would have served anyway (regardless of the draft) and those who did not serve in spite of having a low number.
  - Effect of attending Catholic school among those who attended because they received a randomly-assigned voucher but not among those who attend Catholic school regardless of the voucher or those who do not attend even if they receive the voucher.

IV – What is the effect being estimated (2)

		T(0)	
		0	1
T(1)	0	Never-taker	Defier
	1	Complier	Always-taker

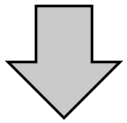
Where T(Z) is the treatment status T={0,1} given a particular state of the instrument Z={0,1}.

Y(T(Z))=Status of the outcome given treatment status and instrument status.

IV – What is the effect being estimated? (3)

Monotonicity assumes there are no "defiers"

		T(0)	
		0	1
T(1)	0	Never-taker	Defier
	1	Complier	Always-taker



- Effect captured by an IV estimator is effect among those who will receive the treatment if exposed to the instrument but will not receive the treatment if not exposed to the instrument.

- Called, appropriately, Local Average Treatment Effect (LATE) or Complier Average Causal Effect (CACE).

IV – What is the effect being estimated?(4)

- ATE: Weighted average of the effect of the treatment among compliers, never takers, and always takers. If it is reasonable to think that the effect is the same across these three groups, then LATE=ATE.

- ATT: Weighted average of the effect of the treatment among compliers and always-takers. If it is reasonable to think that the effect is the same across these two groups, then LATE=ATT.

How to find a good instrument?

- The main challenge in an IV analysis is finding a credible instrument; i.e., a variable Z that affects T but not Y (other than via T).

- Common sources of instruments include
  - Nature: geography, weather, biology in which a truly random source of variation influences T (no possible reverse causation)
  - History: things determined a long time ago, which were possibly endogenous contemporaneously, but which no longer plausibly influence Y
  - Institutions: formal or informal rules that influence the assignment of T in a way unrelated to Y
- Good IVs based on deep knowledge of the processes shaping T and Y.
- Sometimes, a good IV does not exist.

Checking the validity of the instrument

- Non-zero covariance between instrument and treatment: The IV is "sufficiently" correlated with the endogenous treatment X.
  - "Weak instrument" can be tested empirically using a F-test on the instrument in the first stage.
  - A rule of thumb is that the F should be above 10 (very preliminary rule of thumb… there are currently more precise tests).

Checking the validity of the instrument

- Exclusion restriction: The IV is uncorrelated with the outcome Y other than through T.
  - Cannot be tested directly if model is exactly identified.
  - With multiple instruments, we can test the validity of over-identifying restrictions, but test based on the *assumption* that one of the instruments is valid.
  - Example:

    y = b0 + b1z1 + b2z2 + e

    We can estimate the model using only z1 as instrument, then check whether z2 and estimated residual $\hat{e}$ are correlated (H0=uncorrelated).

    If they are, z2 is not a valid instrument *under the assumption* that z1 is a valid instrument.

    But failure to reject H0 may be because estimates are very imprecise

    Rejection of H0 may be because there is effect heterogeneity and instruments capture different LATEs.

Other ways addressing plausibility of exclusion restriction

We can't test directly. But we may be able to indirectly:

1. If possible, look for an association between the instrument and post-instrument outcomes that could be considered to be alternate paths but occur before the treatment or would not be expected to be affected by the treatment.

2. Look for an association between the instrument(s) and outcome(s) in samples where there is no reason for such a relationship.