Development

ECON 8830
Anant Nyshadham

Projections & Regressions

Linear Projections

- If we have many potentially related (jointly distributed) variables
  - Outcome of interest Y
  - Explanatory variable of interest X
  - Additional potential confounders A, B, C
- We are interested in how much of Y is explained ***incrementally*** by X accounting for any confounding co-variation with A, B, C
- A projection is a decomposition of the variation in variable into the independent (orthogonal) planes or spaces of other variables.
- Each of the independent sources of variation in the full set of variables Y,X,A,B,C is given a plane that is separated by right angles from each of the other planes.

Linear Projections

- Projections are analytical/theoretical representations and are true by construction
- We can always represent one variable as a projection on other variables

$$Y = \varrho + \lambda_x X + \lambda_a A + \lambda_b B + \lambda_c C + \psi$$

- If A contributes nothing to Y, $\lambda_a=0$
- If Y,X,A,B,C are *jointly normally distributed* ψ is independent from X,A,B,C and the linear projection fully explains the relationship between Y and X,A,B,C

Regression

- Regressions are the data (empirical) analogue to projections
- A regression of Y on X,A,B,C separates the ***observed*** variation in Y into the orthogonal planes of ***observed*** variation in X,A,B,C

$$Y = \alpha + \Upsilon_x X + \Upsilon_a A + \Upsilon_b B + \Upsilon_c C + \varepsilon$$

- Υ's measure the ***observed*** covariance between Y and regressor (X,A,B,C) divided by the variance of the regressor

Partition Regression

- A regression of Y on X,A,B,C will yield the same Υx as a regression of $\cancel{Y}$ on $\cancel{X}$

$$Y = \alpha + \Upsilon_x X + \Upsilon_a A + \Upsilon_b B + \Upsilon_c C + \varepsilon$$

$$Y = \delta + \kappa_a A + \kappa_b B + \kappa_c C + \cancel{Y} \quad \rightarrow$$
$$\cancel{Y} = Y - (\delta + \kappa_a A + \kappa_b B + \kappa_c C)$$

$$X = \tau + \eta_a A + \eta_b B + \eta_c C + \cancel{X} \quad \rightarrow$$
$$\cancel{X} = X - (\tau + \eta_a A + \eta_b B + \eta_c C)$$

$$\cancel{Y} = \Upsilon_x \cancel{X} + \varepsilon$$

Fixed (Group) Effects

- Operation: include as controls a set of dummy variables that spans a dimension of variation
  - Omit one dummy if general constant is estimated
    - 1 dummy for gender
    - 11 dummies for month
- Concept: Assigns varying intercept (constant) to individual groups or time periods
  - Effectively demeans variables within cells of variation (e.g., by month or gender)
  - Ensures that coefficient of interest does not reflect these course differences across groups or time

Causality

Causality

- Program X was implemented; because of X, outcome Y happened
  - If this is true, we can say with confidence that if we implement X in a similar setting, we would expect Y to happen again
- Causal estimates measure the "true effect" of policy interventions:
  - Compare Y in a world with X versus an otherwise identical world without X

Causality and policy evaluation

- Causal estimates allow us to determine *which policies work and which do not*
  - How effective is policy X?
  - What are the measurable benefits per unit cost of policy X?
    - What are the benefits (per cost) of alternative interventions?
    - Thus, how "comparatively effective" is X?

Causal chains:

Subsidies/extension programs

↓

Adoption of better farm inputs & new technologies

↓

Higher agricultural yields

Examples of Policy Evaluations

- Can loans and subsidies encourage the use of chemical fertilizer?
  - Which policy works better – loans or subsidies? [comparative effectiveness]
  - Does the timing of loans/subsidies matter? (i.e. seasonal variation in liquidity)
- Are matching grants effective in increasing adoption of high-yielding hybrid varieties?

Causal impact v. correlations

- *Causal impact* is not the same as *association*!
- Example: What is the impact of mechanized agricultural inputs on yield?
- Suppose we had data on these two variables, and the correlation > 0
  - Does this imply that a policy of subsidizing mechanized agriculture will increase yields?

The "evaluation problem" (1 of 2)

- The *effect* of mechanized farm implements on yields for farmer X can be expressed as:

[Yield if farmer X used mechanized inputs]

minus

[Yield if farmer X did not use mechanized inputs…]

…at the same moment in time

The “evaluation problem” (2 of 2)

- The fundamental problem is constructing a *counterfactual*
  - We can never observe *both states of the world* at the same time
- The goal of empirical evaluation is to find a valid proxy for what would have happened to farmer X *had he not* adopted the intervention
  - This often involves finding someone (or a group of people) who “looks like” farmer X but who did not adopt, and comparing outcomes for the two

The search for a counterfactual

- The treated group and the counterfactual (or "control") group should be statistically identical on observable dimensions, except that the treated group benefited from the intervention
- If so, then we reason that the only cause for differences in outcomes between treated and untreated is the intervention
- <u>Example</u>: subsidy for chemical fertilizer adoption

Common Issues 1

- Observe treatment and control groups of farmers before and after intervention
- Compare yields of treatment group farmers before and after intervention, find yields went ***down***
- Compare yields of treatment group to yields of control group after treatment intervention, find treatment yields are higher?
- What is "true" effect of intervention?

Common Issues 1

- Time-varying unobservables
  - What else changes for treatment and control groups during intervention time?
    - Rainfall or temperature shocks? Pest infestation?
  - Are changes same across both groups?
    - If yes, we can compare ***changes*** across groups (differencing)
    - If no, cannot separate effect of intervention from effects of time-varying unobservables
- Must make reasonable assumption

Common Issues 2

- Compare yields for adopters of chemical fertilizer to non-adopters after subsidy program, find yields of adopters are ***lower***
- Key problem is *selection*: who chooses to adopt?
  - Those who choose to adopt might have worse soil (need fertilizer more)
  - Non-adopters might be participating in other programs
- Possible solution: matching
  - Requires ability to predict unobserved returns to adoption using observed characteristics

Common Issues 3

- Compare yields for farmers who were *eligible* for subsidies to those who were *ineligible*
- Key concern is that the determinants of eligibility might be correlated with the effectiveness of the intervention
- Potential solutions:
  - Experimental variation
  - Exploit discontinuity in eligiblity rule, if one exists

Conclusions

- To identify effective interventions and compare alternatives, we need to be able to estimate causal effects
- Important to construct a valid counterfactual: a group that would behave the same as the treated group would have in the absence of the intervention
- Invalid counterfactuals (in general):
  - Before and after: time-varying variables
  - Participants vs. non-participants: characteristics
- Options: Choice of method depends on program design, operational considerations, and the question

Endogeneity

True Model

- Suppose true model of yields is:
  - Y = a + bX + cZ + e
    - a, b, and c are parameters to be estimated; e is error term
    - WHAT DOES b REPRESENT IN TERMS OF POLICY? Why do we care to estimate it?
- Do not observe Z
- Can only estimate:
  - Y = a + bX + e
- What happens to estimate of b, ~~b~~?

Original Example

- $Y = a + bX + cZ + e$
  - Y is agricultural yield (total production / area)
  - X is use of HYV (=1 if used HYV in last season, 0 otherwise)
  - Z is a vector of soil characteristics (esp. suitability for planting HYV)
- Uninteresting case: c=0
- Two important cases
  - Z is not known by farmer
  - Z is known by farmer (and affects X)

Irrelevant Z

- Suppose that c=0
  - Soil characteristics have no effects on yields (not really believable!)
  - Profit-maximizing farmer would thus not base his choice of X on Z
- Estimates of a and b will be unaffected by omission of soil quality (Z)
  - Y = a + bX + cZ + e, c=0
  - Y = a + bX + e (estimated model ***is* the** true model)
  - Thus linear regression will give us ~~a~~ = a, ~~b~~ = b

Exogenous X

- Farmer does not know soil quality (Z)
  - Thus Z does not affect farmer's choice of X
- Suppose HYV adoption makes yield (Y) very large if Z=1, but very small if Z=0
  - Y will depend on ***both*** X and Z
  - Farmer cannot act on relationship between X and Z; therefore, X will not depend on Z!
- Estimate of b is unaffected:
  - $\cancel{b} = b$; $\cancel{a} = [E(Y - bX)] = a + c\,\cancel{Z}$
    - ($\cancel{Z}$ is average soil quality in sample)

Endogenous X

- Farmer *knows* soil quality (Z) and takes it into account when choosing to adopt HYV (X)
  - Farmer wants to maximize yield
  - Suppose soil quality can be of two types
    - good for HYV (Z=1)
    - bad for HYV (Z=0)
  - Extreme Case: Farmer chooses to adopt (X=1) only when soil is good for HYV (i.e. Z=1); and thus X=0 if Z=0
- Estimate of b will be biased in this case:
  - $\hat{b} = b + c$; $\hat{a} = a$

Endogenous Z (cont.)

- A less extreme, more believable case:
  - Suppose farmer *more likely* to use HYV (X=1) if his soil is suitable for it (Z=1)
  - Bias will then depend on degree of dependency between X and Z
  - b <= ~~b~~ <= b + c ; a <= ~~a~~ <= a + c ~~Z~~
- If we observe soil quality (Z), or know exact relationship between Z and X, can still get estimate of true b!
- But this is not common…in general, we don't know Z or its exact relationship to X
- What can we do?

Overcoming Endogeneity

- Induce variation in X which is void of relationship with Z (<u>randomization</u>)
- Remove effects of static unobserved Z by comparing two groups over time (<u>differencing</u>)
- Use other ***observed*** characteristics to fully predict portion of X which depends on unobserved Z (<u>matching</u>)
- Exploit discontinuity in relationship between X and Z by comparing observations within bandwidth of discontinuity (<u>discontinuity</u>)

Workshop examples

- Effects of formal sector healthcare on health outcomes
- Effects of school fee subsidies on enrollment
- Effects of access to credit on self enterprise
- Effects of nutrition on farm labor productivity

Methods

- Regression Analysis / Decomposition
- Difference in Differences
- Instrumental Variables
- Regression Discontinuity
- Structural Estimation

The Goal

- **Establish Causality**
  - We did X (or X happened), and because of it, Y happened.
- **Why?**
  - Policy: if we do X again, we can expect Y to happen; if we want Y to happen, perhaps we should do X.
  - Generalizability: if X happens in another context or a different time, we can expect Y to happen

Getting to Causality

- In a more research-friendly universe, we’d be able to observe a single person (call him Fred) in both states of the world at the same time: with the treatment and without the treatment.

“counterfactual comparison”

$$Y_{\text{treated Fred}} - Y_{\text{untreated Fred}}$$

Getting to Causality

- In the real world, finding this "counterfactual" is impossible.
  - We cannot see the same person at the same time in two different states.
- Should we get more people? Some with the treatment and some without.
- Should we measure Y for Fred before and after he is treated?

Getting to Causality

- With more people, we can calculate Average (treated)-Average(untreated).
  - But what if there are underlying differences between the treated and untreated that also impact their Y's?
- With multiple measurements of Y for Fred with different values of X (treated and untreated), we can calculate $Y_{\text{treated Fred}} - Y_{\text{untreated Fred}}$
  - But what if other things changed for Fred during the same time that impacted his Y?

Randomized Experiment

- If we randomize the treatment, on average, treatment and control groups should be the same in all respects, and there won't be underlying differences that cause "bias."
- Check that it's true for all observables.
- Hope that it's therefore true for all unobservables.
- This technique is called *randomization* and is the most common strategy for establishing causality in the sciences.

Randomization

Randomize who gets treated.
Check if it came out OK.

$$\overline{Y}_T - \overline{Y}_C$$

Basically, that's it.

Quasi-Experiment

- What do we do if we cannot randomize treatment?
  - Treatment has already occurred in the past
  - Random assignment would be unethical
  - Treatment is too grandiose or expensive
- Compare individuals with varying treatment who are otherwise as identical as possible.
  - Exploit what we know about treatment assignment
    - Regression Discontinuity, Instrumental Variables
  - Account for any non-random differences
    - Observables: Multivariate Regression, Matching
    - Unobservables: Diff-in-Diff, Control Function
- These techniques are considered "quasi-experimental"

Example Papers

- Impacts of
  - salt iodization on education and labor outcomes
  - temperature and lighting on worker productivity
  - health care on health outcomes and household enterprise activity
  - scholarships on college outcomes
  - health insurance on criminal activity
  - soft skills training on worker productivity and retention
  - managerial quality on worker productivity dynamics

Treatment Assignment

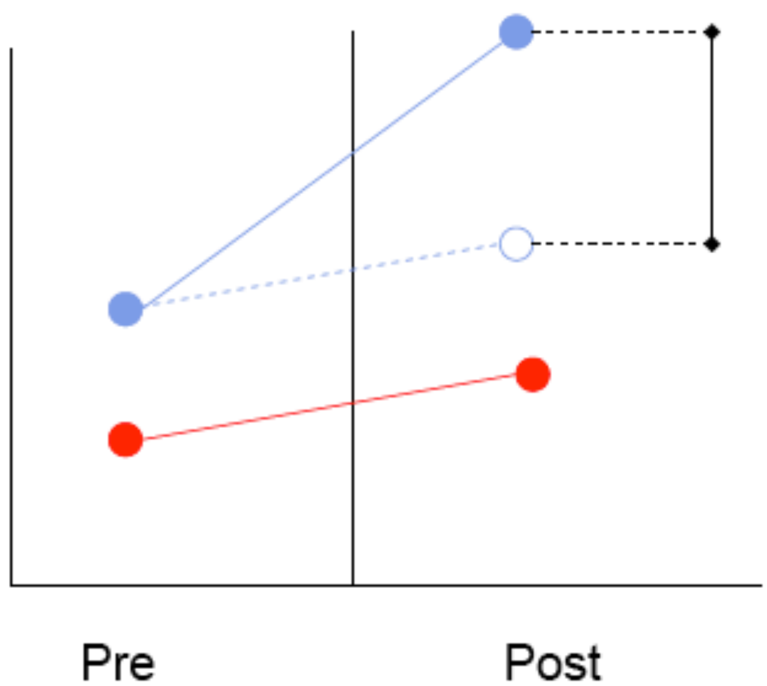
- Treatment is often *clearly* not random.
  - Many health improvements and infrastructural changes coincided with salt iodization
  - Seasonal garment styles and buying patterns are correlated with temperature
  - Sicker people seek out formal health care
  - Smarter kids and needier kids get scholarships.
  - Prevalence of crime and health conditions are both increasing in poverty
  - Workers who engage in extra-training are also more likely to put forth more effort at work
  - Production teams with better supervisors and faster learning workers might get assigned different tasks

Differencing

- If we can see treated and untreated groups before and after, we can compare the **CHANGES** in Y for treated before and after treatment to coincident changes for the untreated and
  - High and low goiter states before and after iodization
  - Factories with and without LED during high and low temperatures in the same day, month, year, etc
- Assume changes in everything else are common to both treated and untreated groups

Identifying Assumption

- Whatever happened to the control group over time is what would have happened to the treatment group in the absence of the program.



Effect of program difference-in-difference (taking into account pre-existing differences between T & C and general time trend).

Instrumental Variables

- If we know of some factor Z that at least partially determines treatment X without directly impacting outcome Y, we can use Z as a predictor (instrument) of treatment X that can bypass any confounders.
  - Ease of accessing health care predicts health care utilization but not incidence and severity of sickness
- 2 key requirements
  - Z must adequately predict X (testable)
  - Z must not impact Y except through X (assumed)

Regression Discontinuity

- If we know the exact assignment rule, we can use this rule to construct instrument Z for treatment X.
  - Merit-based tuition subsidies given based on GPA and SAT/ACT cutoffs
  - Subsidized health care provided to those below wealth cutoff
- Compare those just above cutoff to those just below cutoff
- Assume at tiny increments of eligibility all else is equivalent across treated and untreated

Matching

- Match each treated participant to one or more untreated participant based on observable characteristics.
- Assumes no selection on unobservables
- Condense all observables into one "propensity score," match on that score.

Matching

- After matching treated to most similar untreated, subtract the means, calculate average difference

$$\frac{Y_{Jon(T)} - Y_{John(C)} + Y_{Jim(T)} - Y_{Tim(C)}}{2}$$