Australian National University

EE468: Integrated Public Economics, Development and Political Economics

Lecture 2: Empirical Methods for Development Policy Analysis

24 January 2014

Sommarat Chantarat (sommarat.chantarat@anu.edu.au)

- Overview
- Econometrics Vs. Randomised control experiments
- A closer look at policy impact evaluation

Causal inference for development policy analysis

- Causal inference to answer development questions
  - What constitute admissible evidence in evaluating development problems?
  - How do we determine what policies work and why they (not work)?
  - How credible are the underlying assumptions for making prediction?
- Economic theory can be used to create theoretical predictions on causal relationships between households and economic environments and to make ex-ante evaluation of policy impacts

  x → household models → $y = f(x|Z)$
- Today, we will learn how to identify and test these theoretical predictions using empirical data

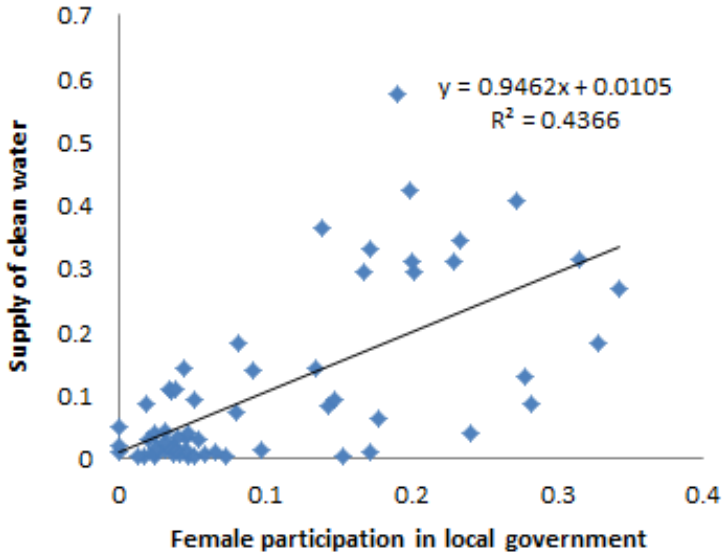
  x → y? using econometric analysis/experimental techniques
- Understanding different methods for identifying causality is an important tool kit for studying development problems and policies!

Causation Vs. Correlation: A threat for policy inference

- If x and y move together, this does not imply that x causes y → identifying direction of causality is the most important challenge in causal inference

- Ex) A study of determinant of availability of clean water in rural areas

$$Clean\ water_v = a + b \cdot women\ participation\ in\ local\ gov_v + e_v$$



- What does this evidence imply? → increasing women participation in local government will help improving available of clean water?
- Is this simply a result of reverse causality? In rural area, women's responsibility is to fletch water and so without available water supply, they could be busy fletching water and have no time to participate in local gov…

- Without correctly establishing the direction of causality, analysis will be at risk of describing reverse reality and making bad policy inference!

What we learn from basic econometrics

- Back to the basic → least squared regression (a review!)

$y = \beta' x + u$ where y: dependent (outcome) variable
x: scalar or vector of independent (explanatory) variables
u: error term (omitted vars and measurement errors in y, x)

With key assumptions: $E(u) = 0, cov(x, u) = 0$ (exogeneity), we know that

- $\hat{\beta} = \frac{cov(x,y)}{var(x)}$ is BLUE of the true population β
- $\hat{\beta}$ is unbiased → $E(\beta) = E\left[\frac{cov(x,y)}{var(x)}\right] = E\left[\frac{cov(x,\beta' x+u)}{var(x)}\right] = E\left[\frac{cov(x,\beta' x)}{var(x)} + \frac{cov(x,u)}{var(x)}\right] = \beta$

What we learn from basic econometrics

- With endogeneity → $cov(x, u) \neq 0$

$y = \beta' x + u$ where y: dependent (outcome) variable
x: scalar or vector of independent (explanatory) variables
u: error term (omitted vars and measurement errors in y, x)

We can show that

- $\hat{\beta}$ is biased → $E(\beta) = E\left[\frac{cov(x,y)}{var(x)}\right] = E\left[\frac{cov(x,\beta' x+u)}{var(x)}\right] = E\left[\frac{cov(x,\beta' x)}{var(x)} + \frac{cov(x,u)}{var(x)}\right]$

$$= \beta + \frac{cov(x,u)}{var(x)} \neq \beta$$

And so with endogeneity, the OLS estimated $\hat{\beta}$ would under or over estimate the true β by the magnitude of $\frac{cov(x,u)}{var(x)}$

Identification problems in causal inference

- What are the key sources of endogeneity?

In a standard regression $y = \beta' x + u$, $u(q)$ where q captures observed and unobserved but omitted variables and other measurement errors in y, x)

Exogeneity assumption → q is uncorrelated with x

$y = \beta' x + u$ → $cov\big(x, u(q)\big) = 0$

q

→ so the only effect of x on y is direct effect via $\beta' x$

Endogeneity → q contains "unobserved heterogeneity" that seem to be correlated with both x and y (e.g., omitted variables, selection bias, reverse causality) → $cov\big(x, u(q)\big) \neq 0$

$y = \beta' x + u$ → $cov\big(x, u(q)\big) \neq 0$

q

→ so the estimated effects of x on y would include both direct effect of x and the influence of q through x

→ $\mathrm{E}(\hat{\beta}) = \beta + \frac{cov(x,u(q))}{var(x)}$

Identification problems in causal inference

- **Ex 1) Estimating return to education:**

$$earning_i = a + b \cdot schooling_i + c \cdot experience_i + u_i$$

With potential unobserved heterogeneity like ability, health status, location, etc.

→ To what extend would positive $\hat{b}$ results from the fact that high income, ability, tend to attend school? NOT the direct return to education itself

- **Ex 2) Estimating local demand for food:**

$$fooddemand_i = a + b \cdot price_i + c \cdot HH_i + u_i \quad ; HH_i \text{ vector of hh characteristics}$$

Downward sloping demand might not be identified with $\hat{b}$ with potential reverse causality of demand → price

Identification problems in causal inference

- **Ex 3) Estimating impacts of fertilizer on farm productivity:**

  Deaton and Benjamin use 1985 survey in Cote'd Ivory to estimate

$$productivity_i = \underset{(SE)}{5.62} + \underset{(18.5)}{0.15} fertilizer_i + \underset{(0.07)}{0.06} insecticide_i + \underset{(2.51)}{0.53} size_i \quad (4.32)$$

Yet, only 10% of farmers in the area adopted fertilizer. What went wrong? Did this overstate real impact? To what extend would the estimated impact (0.15*) reflect the fact that progressive, high-yield farmers (who likely to get credit) tend to use fertilizer?

- **Ex 4) Estimating impacts of health clinics on community's health outcomes:**

$$healthoutcome_c = a + b \cdot T_c + c \cdot Com_c + u_c$$

where $T_c = 1$ with clinic, $= 0$ without and Com_c is community characteristics

Would $\hat{b}$ over or under-estimate the real impact if health clinics tend to be built in villages with poor health outcomes?

This is the classic "selection bias" in program evaluation! → What might happen to the policy recommendation with selection bias?

Solving endogeneity using econometrics: Panel Data

- If we can observe the same individuals in more than one periods, one can construct individual, hh, location fixed effect variables to control for some of "unobserved heterogeneities"

- **In estimating return to schooling with panel data**, we can instead estimate:

$$y_{it} = a + b \cdot s_{it} + c \cdot e_{it} + \delta_i + u_{it}$$

where δ_i controls for other observed and unobserved individual-specific characteristics. Rewrite above by $-y_{it-1}$ from both sides, we get

$$y_{it} - y_{it-1} = (a - a) + b(s_{it} - s_{it-1}) + c(e_{it} - e_{it-1}) + (\delta_i - \delta_i) + (u_{it} - u_{it-1})$$

$$\Delta y_i = b\Delta s_i + c\Delta e_i + \theta_i \quad \text{where} \quad \theta_i = u_{it} - u_{it-1}$$

Would $\text{cov}(\Delta s_i, \theta_i) = 0$ now?

While panel data could resolve some sources of endogeneity, we still cannot solve the potential reverse causality!

Solving endogeneity using econometrics: IV

- With $y = \beta' x + u$ and $cov(x, u) \neq 0$, the idea is to use exogenous "instrument variable (IV) z to separate good variation of x (that is uncorrelated with u) from the bad one (correlated with u) and then only impute causal relationship $x \rightarrow y$ from the good variation of x only
- This could work if we can find an instrument variable z such that

  (1) Relevance → $cov(z, x) \neq 0$ (z explains great variations in x)

  (2) Exogeneity → $cov(z, u) = 0$ (z has no direct relationship with y)
- IV estimation is then done through two stage least square (2SLS):

  First: Decompose the endogenous variation of x into good one explained by z

  $x = \alpha' z + e$

  Second: Using the good variation of x imputed by z to estimate causality

  $y = \beta_{iv}{}' \hat{x}(z) + \epsilon$ where $\hat{x}(z) = \hat{\alpha}' z$ from the first regression
- We can show that $\hat{\beta}_{iv} = \frac{cov(z,y)}{cov(z,x)}$ is now unbiased and consistent if z satisfies (1),(2)

$$E(\hat{\beta}_{iv}) = \left[\frac{cov(z,\beta' x+u)}{cov(z,x)}\right] = \beta \frac{cov(z,x)}{cov(z,x)} + \frac{cov(z,u)}{cov(z,x)} = \beta \text{ if } cov(z,x) \neq 0 \text{ and } cov(z,u) = 0$$

Solving endogeneity using econometrics: IV

- In estimating return to education:

$$lnearning_i = a + b \cdot schooling_i + c \cdot experience_i + u_i$$

What might be a good IV? Card (1995) uses proximity to school

- IV can be innovatively drawn from **natural experiments** (taking advantage of existing natural or man-made randomness) e.g., date of birth, random rationing into school, etc.

- In estimating impacts of fertilizer on farm productivity:

$$lnproductivity_i = a + b \cdot fertilizer_i + c \cdot insecticide_i + d \cdot size_i + u_i$$

Can you think of some potential good IV?

- The key challenges in using IV approach thus are
  - Causality inevitably revolve around the quality of the instrument z and the defense of the instruments chosen is as much an art as a science!
  - It is just not easy to find suitable instruments!

Solving endogeneity using randomised experiments

- The idea is to use well-crafted randomisation to create "exogenous variations" in x so that causal relationship can be identified in a confined, representative sample

- Randomisation involves randomly assigning different realisations of x to different analytical units (individuals, hhs, communities, etc.) in the sample such that each unit has equal probability of being assigned different variations of x

- In estimating demand for food:

$demand_i = a + b \cdot p_i + c \cdot HH_i + u_i$ where $cov(p_i, u_i) \neq 0$ with reverse causality

How about randomly distributing variations of price (discount coupons with different discount rates) to randomly chosen households from targeted population and so create exogenous variable, ex),$coupon_i = (150\%, 120\%, 100\%, 80\%, 50\%)$ and re-estimate

$$demand_i = a + b \cdot coupon_i + c \cdot HH_i + u_i$$

By randomisation → $cov(coupon_i, u_i) = 0$ and so $\hat{b}$ is now unbiased impact of price on demand for food!

Solving endogeneity using randomised experiments

- In estimating impacts of fertilizer on farm productivity

$$productivity_i = a + b \cdot fertilizer_i + c \cdot HH_i + c \cdot farm_i + u_i$$

How about randomly making fertilizer available for some households but not the others?

$$productivity_i = a + b \cdot givenfertilizer_i + c \cdot HH_i + c \cdot farm_i + u_i$$

where $givenfertilizer_i = \begin{cases} 1 \; if \; household \; is \; given \; fertilizer \\ 0 \; if \; household \; is \; not \; given \; fertilizer \end{cases}$

By randomisation → $cov(givenfertilizer_i, u_i) = 0$ and so $\hat{b}$ is now unbiased estimator of direct impact of fertilizer on productivity!

- Randomised experiments have been widely used especially for policy impact evaluation … to resolve the classic "selection bias", let's now look more into this!

A closer look at policy impact evaluation

- The impact on outcome y_i of individual i who participate in the program ($T_i = 1$):

  $$\Delta_i = (y_i|T_i = 1) - (y_i|T_i = 0)$$

  The classic evaluation problem → we in fact cannot observe Δ_i since we cannot observe for the same person in both states (participate and not participate in the program) at the same point in time!

- Rigorous ex-post impact evaluation thus focuses on estimating "average treatment effect (ATE)" taken from observed outcomes of a larger set of sampled population

- $$ATE = E(\Delta_i) = \underbrace{E(y_i|T_i = 1)}_{\text{Mean observed outcomes of participants (treatment group)}} - \underbrace{E(y_j|T_j = 0)}_{\text{Mean observed outcomes of counterfactual (control group)}}$$

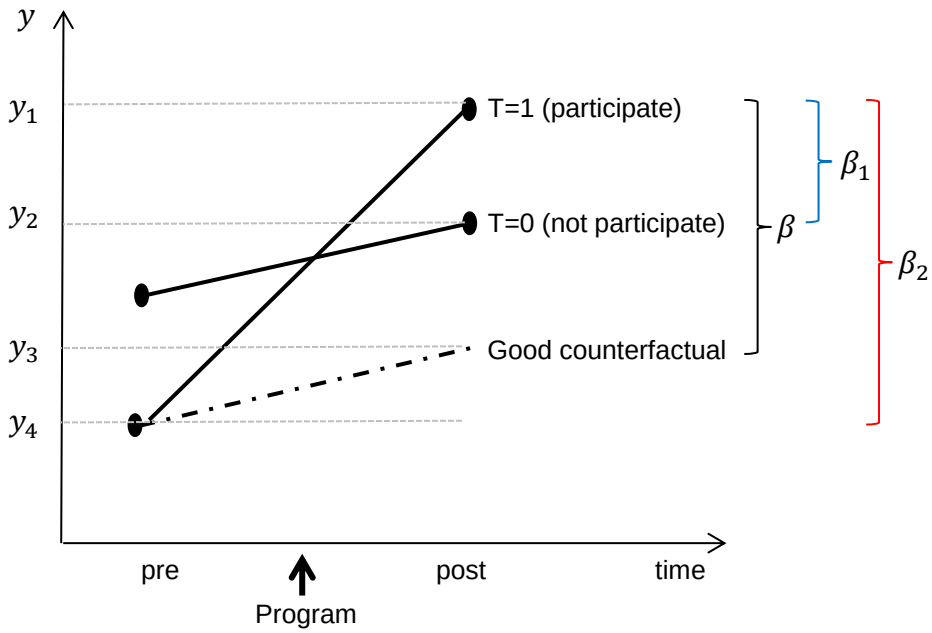
- The key is therefore to find appropriate "counterfactual (control group)" to capture mean outcome of participating group if they have not participated in the program

A closer look at policy impact evaluation

- What constitute good choice of counterfactual (control group)?
  - Both T and C groups should be identical in the absence of the program
  - Both cannot be differently exposed to other environments during evaluation

  And so we can be sure that the observed difference in the mean outcomes between the two groups is due solely to the program!

- Ex) Outcomes of participating in a microcredit program



Common but counterfeit counterfactual

- Pre and post comparison?

$\beta_2 = y_1 - y_4$ could overstate impacts as there might be others affecting how outcomes of the groups change over time

- Takers and non-takers comparison?

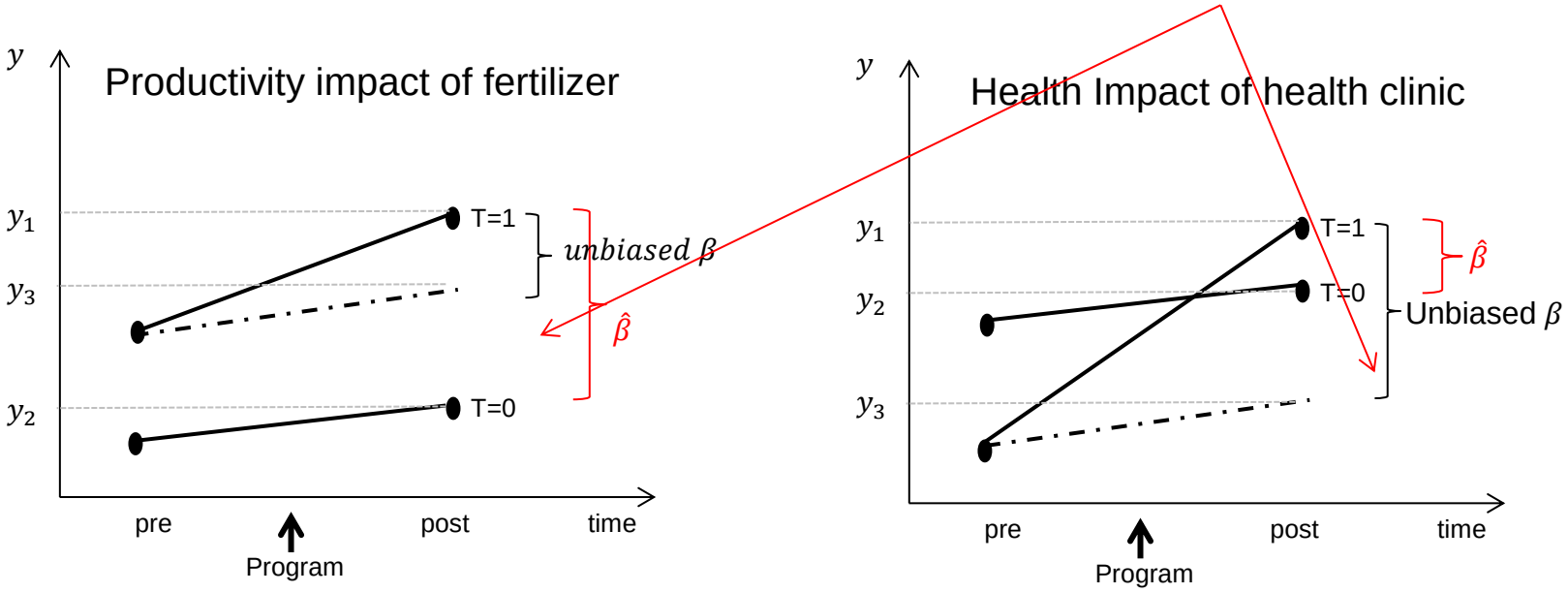
$\beta_1 = y_1 - y_2$ could understate impacts as participants and non-participants may be different in the absence of program

A closer look at policy impact evaluation

- **Selection problem in takers and non-takers comparison**: As we see before, program impact can be estimated from econometric model:

  $y_i = \beta T_i + \alpha HH_i + u_i$; $T_i = 1$ if participate and $= 0$ if not

  As selection into the program can be based on observed and/or unobserved characteristics that also correlates with y_i → $E(\hat{\beta}) = \beta + selection\ bias$

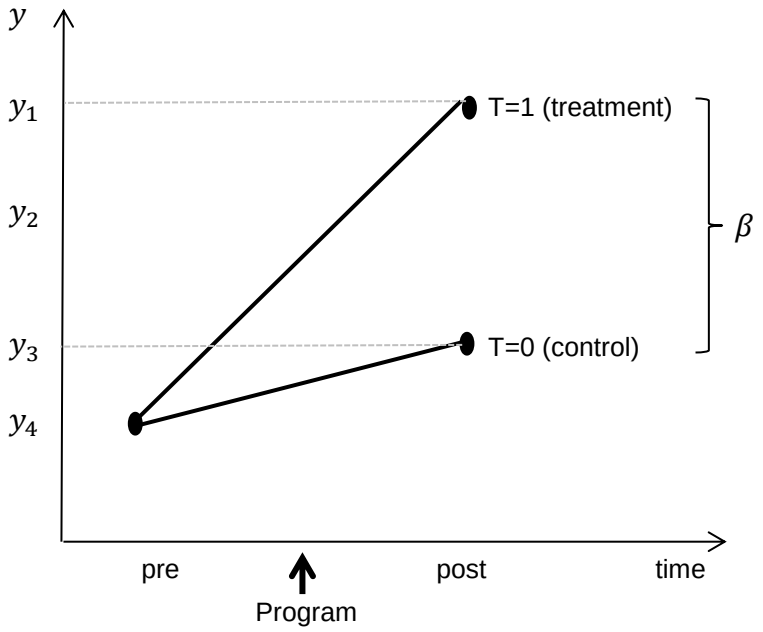


- Rigorous Impact evaluation thus need to isolate impact of the program from the potential selection bias and from impacts of other factors

Randomised control experiments for impact evaluation

- RCEs can be used to generate exogenous variations in program participation by randomly assigning eligible units (individual, hh, school, or village) into treatment and control groups with all units having equal opportunity of selection for treatment

- **By construction, RCEs eliminate various sources of selection bias**
  - Both treatment and control groups should have similar pre-program outcomes and be exposed to same environments over time and only differ in program



For pure randomisation;

$$ATE = E(y_i|T_i = 1) - E(y_j|T_j = 0)$$

We can also estimate ATE from

$$y_i = \beta T_i + \alpha HH_i + u_i$$

where $T_i = 1$ if treatment, =0 if control

$\hat{\beta}$ is thus an unbiased estimate of ATE

Randomised experiments for impact evaluation

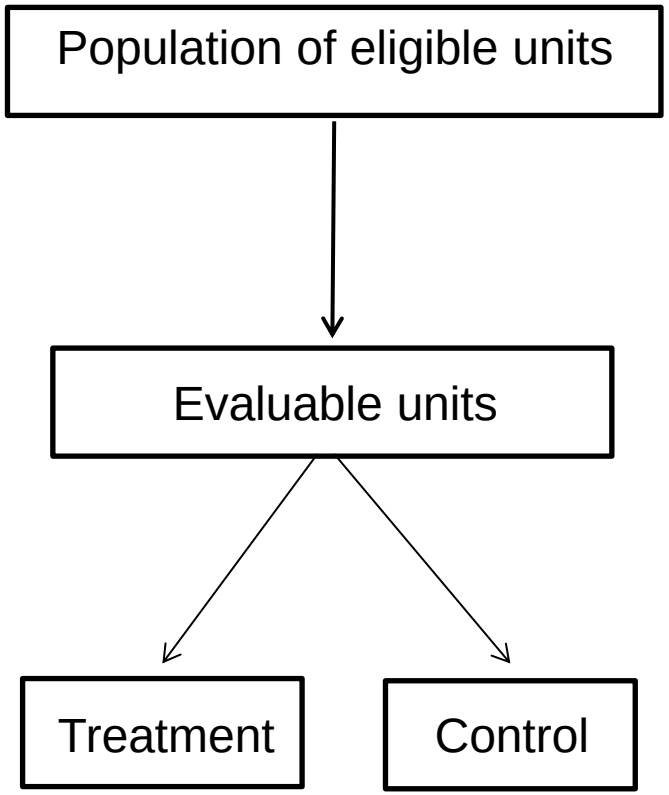
- **Three steps in implementing a randomised experiment**

1) Define randomised units (ind, hh, school, village) depending on intervention

   - too large – too small sample size?
   - too small – risk of spillover effects

2) Select evaluable sample from targeted population units (power calculation, cost)

3) Randomise evaluable units into treatment and control (public lottery, dice, computer generation)



- Spatial vs. temporally (by randomised program phased in)
- Two types: randomise treatment vs. randomise encouragement to take treatment
- Not everyone assigned to treatment will participate – compliance largely imperfect