Lecture 1

The Policy Evaluation Question and Randomization

Policy Evaluation
Nova SBE - Universidade Nova de Lisboa / IGC
Pedro C. Vicente
http://www.pedrovicente.org/

The problem of the counterfactual

- The main challenge of policy evaluation is to determine what would have happened to the beneficiaries if the program had not existed
    - e.g., one has to determine the per capita household income of beneficiaries in the absence of the intervention
- A beneficiary's outcome in the absence of the intervention is its **counterfactual**
- Ideally, one would like to compare how the same household would have fared with and without an intervention, or **treatment**
    - But, at a given point in time a household cannot be in the treatment and **control** groups at the same time
- The challenge of an impact assessment is to create a convincing comparison group

Figure 2.3 Evaluation Using a Before-and-After Comparison

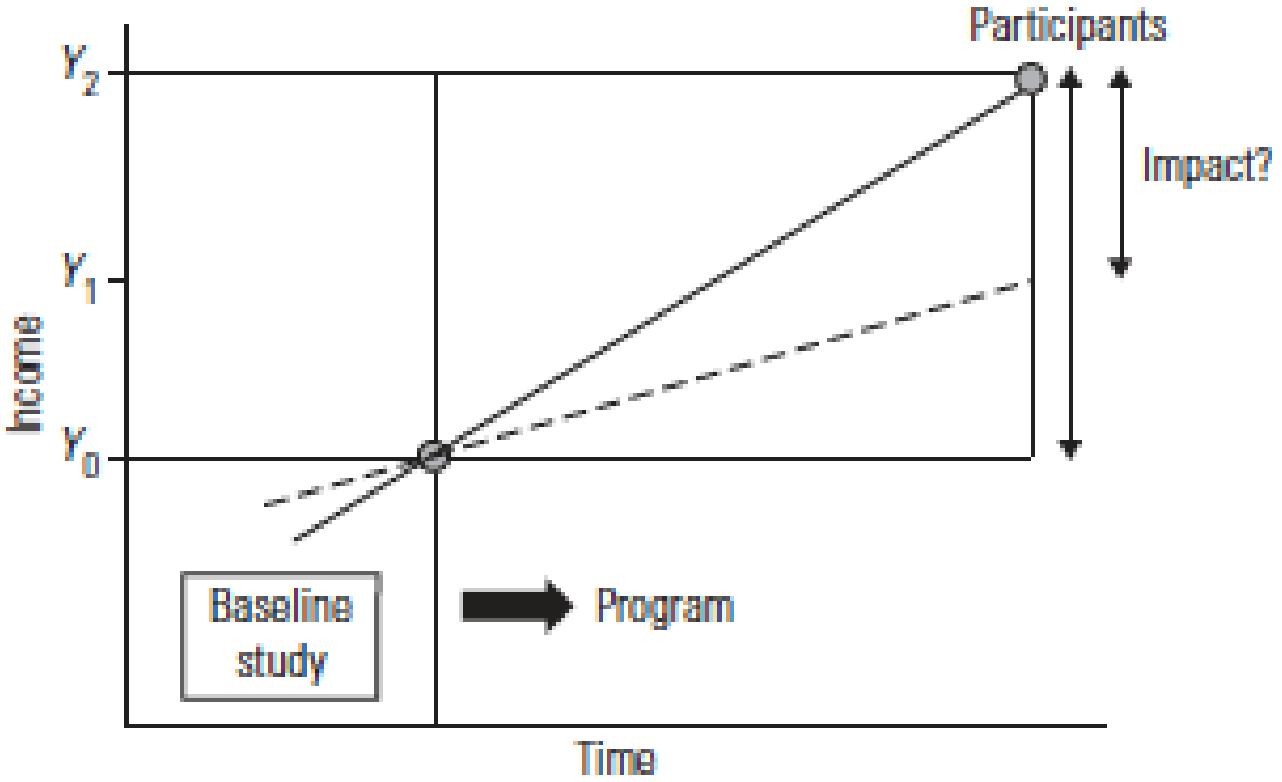
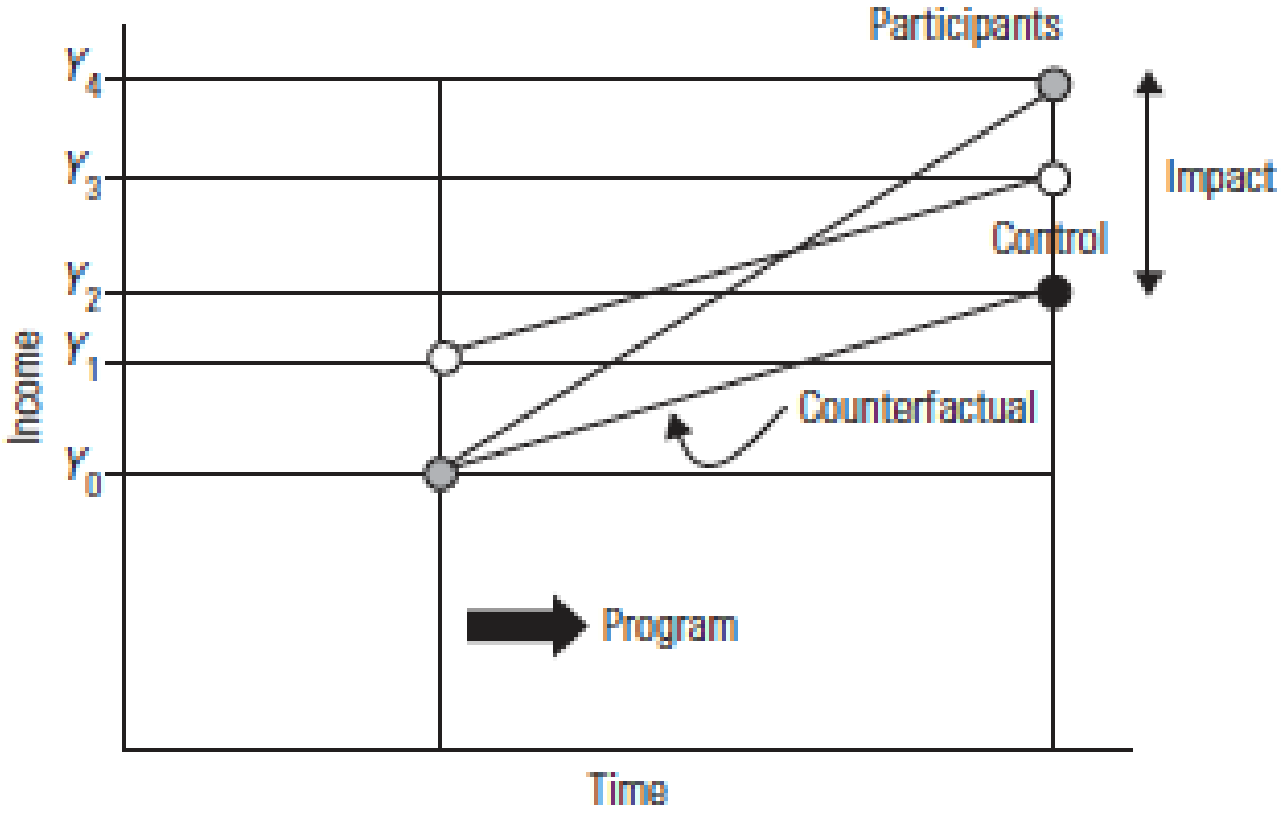


Figure 2.2 Evaluation Using a With-and-Without Comparison



The problem of the selection bias

• The following equation presents the basic evaluation problem comparing outcomes *Y* across individuals *i*:

$$Y_i = a + bT_i + e_i$$

• *T* is a dummy equal to 1 for those who participate and 0 for those who do not participate

• ε is an error term reflecting unobserved characteristics that also affect the outcome

• The problem with estimating the above equation is that programs are placed according to the need of the communities and individuals, who in turn self-select to participate
• Treatment exposure is then subject to observed and unobserved characteristics of individuals
• Unobserved factors ε would then be correlated with T

$$\text{cov}(T, \varepsilon) \neq 0 \Rightarrow E(\varepsilon | T) \neq 0$$

• This implies the violation of one of the key assumptions of OLS (strict exogeneity) in obtaining unbiased estimates, namely of β

• A parenthesis on the relationships between independence, strict exogeneity (OLS assumption), and correlation, assuming $E(e) = 0$

$$(i)\ T \text{ ^ } e(independence) \Rightarrow \text{cov}(T, e) \equiv$$

$$\equiv E(Te) - E(T)E(e) = E(T)E(e) - E(T)E(e) = 0 \Leftrightarrow$$

$$\Leftrightarrow cor(T, e) \equiv \frac{\text{cov}(T, e)}{s_T s_e} = 0$$

$$(ii)\ T \text{ ^ } e(independence) \Rightarrow E(e|T) = E(e) \Leftrightarrow$$

$$\Leftrightarrow E(e|T) = 0\ (strict\ exogeneity)$$

$$(iii)\ E(e|T) = 0\ (strict\ exogeneity) \Rightarrow \text{cov}(T, e) = 0 \Leftrightarrow$$

$$\Leftrightarrow cor(T, e) = 0$$

• This problem can be represented in a conceptual framework
• Suppose we are evaluating an intervention (T) aimed at raising household income (Y)
• Define

$$Y_i(1), Y_i(0)$$

as the outcome of individual *i* in case he/she was treated (1) or not (0)
• This is defining

$$Y_i \equiv Y_i(1)T_i + Y_i(0)(1 - T_i)$$

• Then we can compare a group of individuals that was indeed treated with a group of individuals that was not treated, namely in terms of population means

$$D = E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]$$

- We can rearrange

$$D = E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]$$
$$\Leftrightarrow D = E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] +$$
$$+ E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 1]$$
$$\Leftrightarrow D = E[Y_i(1) - Y_i(0) \mid T_i = 1] +$$
$$+ [E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]]$$

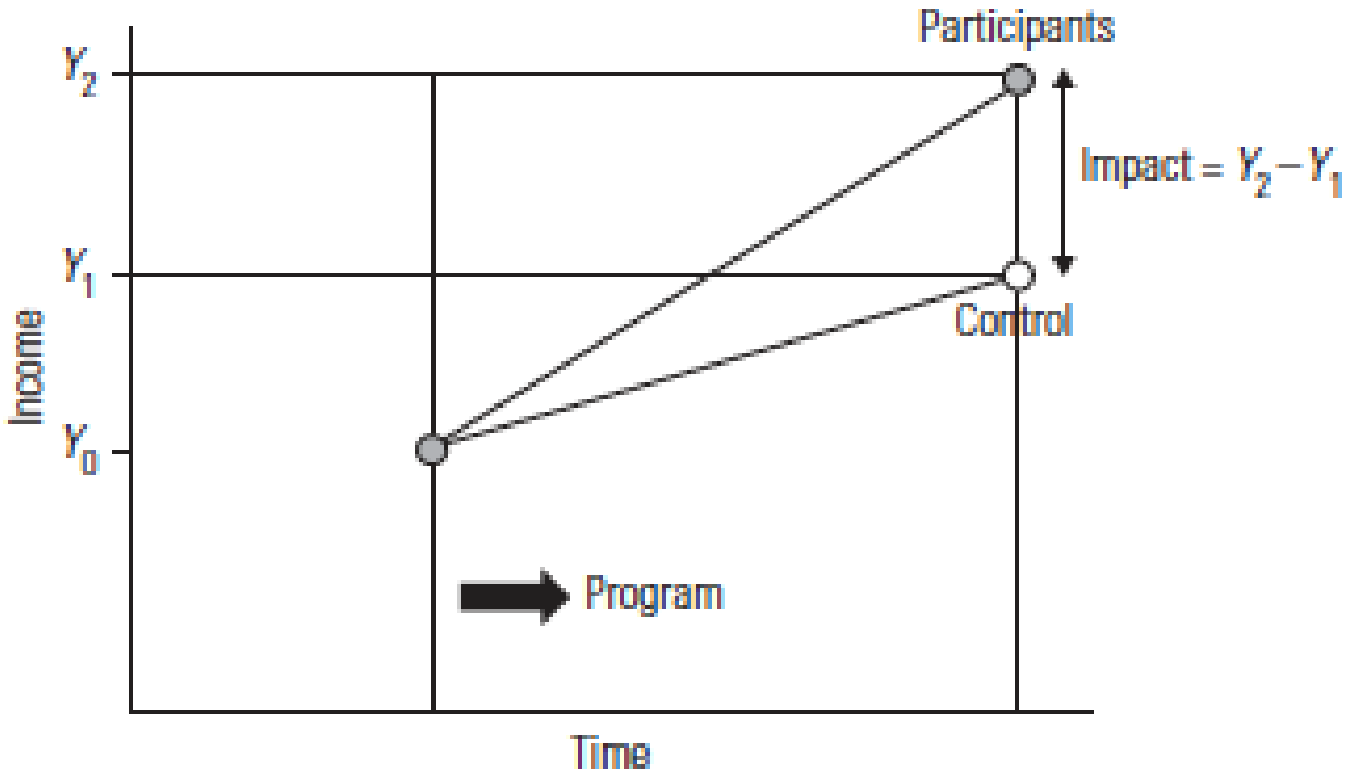
- We see the difference boils down to the sum of

$$TE = E[Y_i(1) - Y_i(0) \mid T_i = 1]$$
$$B = E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]$$

- *TE* is a treatment effect, and *B* is the selection bias

Randomization

Figure 3.1 The Ideal Experiment with an Equivalent Control Group



- Standard **randomization** procedure
  - In a first stage, a sample of potential participants is selected randomly from the relevant population; this is representative of the population with a certain sampling error; this stage ensures **external validity**
  - In a second stage, individuals are randomly assigned to treatment and comparison groups; this stage ensures **internal validity**

• The second stage guarantees that

$$E[Y_i(1) \mid T_i = 1] = E[Y_i(1) \mid T_i = 0]$$

$$E[Y_i(0) \mid T_i = 1] = E[Y_i(0) \mid T_i = 0]$$

• This implies that

$$B = E[Y_i(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] = 0$$

• And that

$$D =$$

$$= TT \circ E[Y_i(1) - Y_i(0) \mid T_i = 1] =$$

$$= TU \circ E[Y_i(1) - Y_i(0) \mid T_i = 0]$$

where *TT* is the treatment on the treated, and *TU* is the treatment on the untreated

• In addition to the second stage, the first stage guarantees that

$$ATE \equiv E[Y_i(1) - Y_i(0)] = TT = TU = D$$

where *ATE* is the average treatment effect, which is defined implicitly relative to a general population of interest

• The randomization procedure solves the policy evaluation question by equating all above treatment effects to *D*

• Coming back to our regression equation

$$Y_i = a + bT_i + e_i$$

• Given randomization, T and ε are independent, which means OLS gives an unbiased estimate of β

• Then

$$D = E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0] =$$
$$= a + b + E[e_i \mid T_i = 1] -$$
$$- a - E[e_i \mid T_i = 0] =$$
$$= b$$

where the last step employs independence between ε and T

Randomization in the real world

- J-PAL - Abdul Latif Jameel Poverty Action Lab

- IPA - Innovations for Poverty Action

• J-PAL is a research center based at the MIT and IPA is a nonprofit; both are dedicated to discovering what works to help the world's poor
• J-PAL and IPA design and evaluate programs in real contexts with real people
• J-PAL and IPA are committed to not just measuring the impact of a program, but also working with organizations to facilitate integration of research results into operations to ensure continuous improvement and the replication of successful ideas

• DIME – Development Impact Evaluation Initiative of the World Bank

• NOVAFRICA – Nova Africa Center for Business and Economic Development

- Different methods of randomization:
  - Oversubscription
    - If limited resources burden the program, implementation can be allocated randomly across a subset of eligible participants
  - Phase-in
    - Program is gradually phased-in across a set of eligible areas, so that controls represent eligible areas still waiting to receive the program
  - Within-group
    - Program is given to sub-groups within eligible areas – problem is spillovers
  - Encouragement
    - Instead of randomizing the treatment, researchers randomly assign subjects an announcement or incentive to participate in the program

- Concerns with randomization:
  - Ethics
    - Withholding a particular treatment from a random group of people and providing access to another may be simply unethical
  - External validity
    - A small-scale job training project may not affect overall wage rates, while a large-scale one may
  - Compliance
    - Arises when a fraction of the individuals who are offered the treatment do not take it
  - Spillover
    - Arises when the treatment reaches the control group (contamination)

- Intent-to-treat
  - Despite efforts to randomize the program intervention ex-ante, actual program participation may not be entirely random
    - Treatment may affect control units
    - Treated individuals may not participate in the program after all
    - Treated individuals may drop out (selective attrition)
  - Focusing on intent-to-treat (ITT) effects or instrumenting actual program participation by the randomized assignment strategy (IV) are the possibilities
  - Example: call *Z* the variable that is randomly assigned (encouragement)

$$D = E[Y_i(Z_i = 1) - Y_i(Z_i = 0) \mid Z_i = 1] +$$
$$+[E[Y_i(Z_i = 0) \mid Z_i = 1] - E[Y_i(Z_i = 0) \mid Z_i = 0]] =$$
$$= E[Y_i(Z_i = 1) - Y_i(Z_i = 0) \mid Z_i = 1] = ITT \neq TE$$

Sample size, design, and power of experiments

• Take the simple regression framework:

$$Y_i = a + bT_i + e_i$$

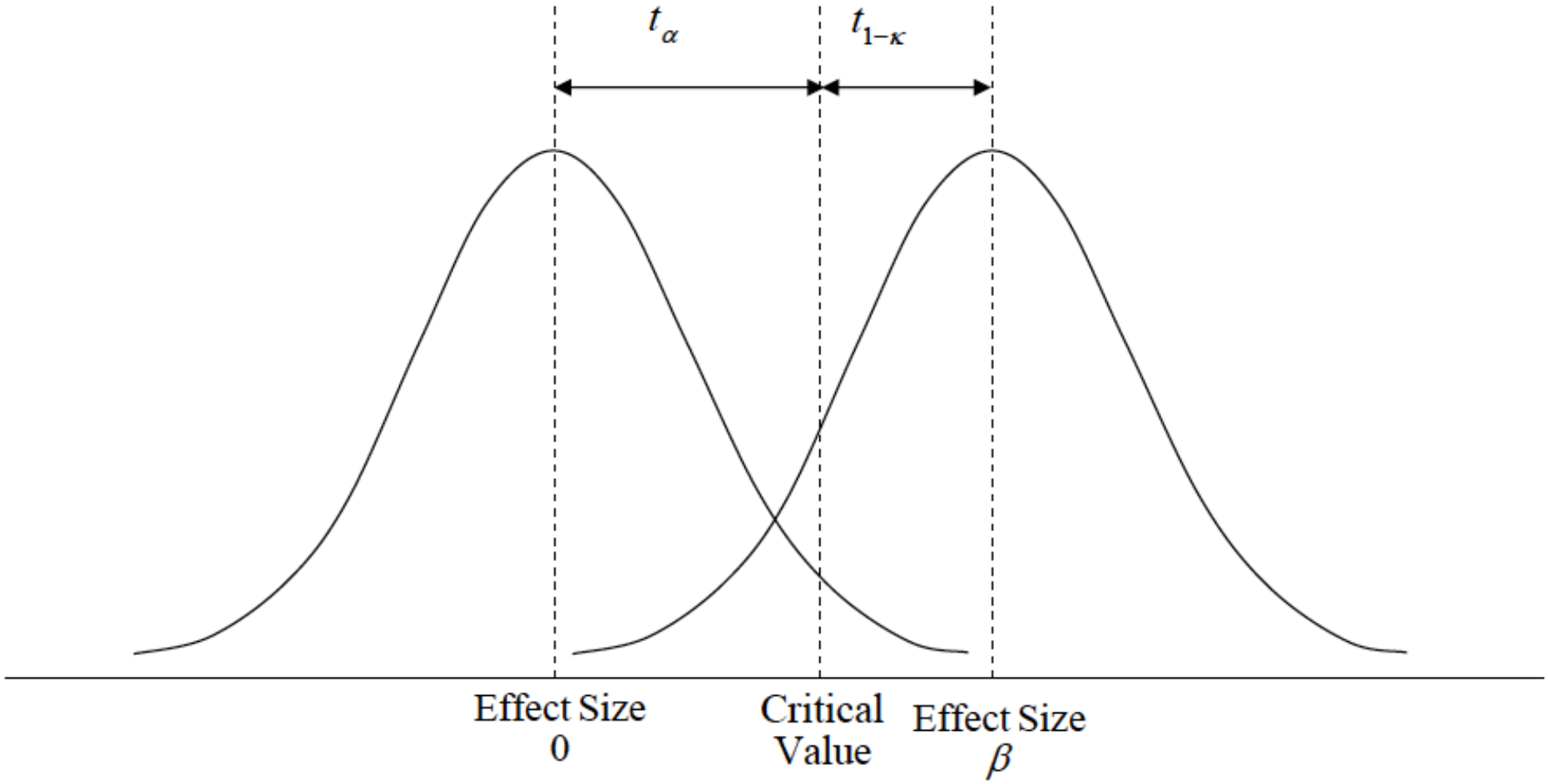
• Assume only one treatment, that a proportion P of the sample is treated, that observations are iid with variance σ^2

• Then the variance of β^{OLS} is given by

$$\frac{1}{P(1-P)}\frac{s^2}{N}$$

• We are generally interested in testing the hypothesis H0 that the effect of the program is equal to zero against the alternative that it is not

• The significance level of a test represents the probability of a type I error (the probability we reject the hypothesis when it is true)


t_{α}
$t_{1-\kappa}$
Effect Size
0
Critical
Value
Effect Size
β

• The distribution to the left is the distribution of β^{OLS} under H0
• For a given *significance level* α, H0 is rejected if $|\beta^{OLS}|$ falls to the right of the critical value

$$\left|b^{OLS}\right| > t_a * SE_{b^{OLS}}$$

where t_α is obtained from the standard *t*-distribution ($t_{\alpha/2}$ for a two-sided test)
• The distribution to the right shows the distribution of β^{OLS} under true impact β
• The *power* of the test for a true effect size β is the fraction of the area under this curve that falls to the right of the critical value t_α, i.e., the probability that we reject H0 when it is false
• To achieve power κ, it must be that

$$b > (t_{1-\kappa} + t_a) * SE_{b^{OLS}}$$

where $t_{1-\kappa}$ is obtained from the standard *t*-distribution

• The *minimum detectable effect* (MDE) size for a given power (κ), significance level (α), sample size (N), and portion of subjects allocated to treatment (P) is then given by

$$MDE = (t_{1-\kappa} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{s^2}{N}}$$

for a single sided test (where t_{α} is replaced for $t_{\alpha/2}$ for a two-sided test)

• Note that the above equation implicitly defines the sample size N required to achieve a given level of power, given the MDE and the level of significance, as well as the portion of treated subjects

• When the significance level (α) decreases (more exigent, lower probability of type I error), t_{α} increases, so that MDE increases

• When power (κ) increases (more exigent, higher probability that we reject H0 when it is false), $t_{1-\kappa}$ increases, so that MDE increases

• Check the other parameters (min MDE): decrease σ^2, increase N, and divide sample equally between treatment and control

- There are extensions of the above power calculations for cases when:
  - The randomization happens across groups of individuals while the data are available at the individual level
    - Crucially, in this case individuals in the same group may be subject to common shocks, which means their outcomes may be correlated

$$MDE = \frac{(t_{1-k} + t_a)}{\sqrt{P(1-P)J}} * \sqrt{r + \frac{1-r}{n}s}$$

where J is number of groups (clusters), n is number of units within cluster, and ρ is intra-cluster correlation

  - There is imperfect compliance

$$MDE = (t_{1-k} + t_a) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{s^2}{N}} \frac{1}{c-s}$$

where c is the share of treated subjects who actually get the treatment, and s is the share of control who actually get it

- Control variables
  - In a simple randomized experiment, controlling for baseline values of covariates likely to predict the outcome does not affect the expected value of an estimator of β, but it can reduce its variance
  - Note that controlling for covariates affected by the treatment would bias the estimate of the treatment effect by capturing part of its impact
  - Information on covariates should therefore be collected at the baseline
- Block randomization
  - Blocks of units that share some observable conditions (strata) are formed, and randomization is performed within each block
  - Very much like controlling ex-post, but more efficient

Inference issues

- Grouped data
  - Problem: standard errors need to take into account possible correlation in the outcomes between members of the same group
  - Possible solution: use the cluster-correlated Huber-White covariance matrix estimator
    - Only when the number of groups randomized is large enough – when the number of clusters is small, standard errors too small, leading to over-rejection of null hypothesis of no effect
- Multiple outcomes
  - Problem: the probability of null rejection on one outcome is high
  - Possible solution: aggregation
- Sub-groups and covariates
  - Problem: relevant sub-groups and covariates sample-driven
  - Possible solution: pre-plan

Heckman's general selection framework

- We follow here Heckman and Vytlacil (2005, Econometrica)
- Selection model with two potential outcomes

$$Y_0 = \mu_0(X, U_0) = \mu_0(X) + U_0$$

$$Y_1 = \mu_1(X, U_1) = \mu_1(X) + U_1$$

$$D^* = \mu_D(Z) - U_D$$

$$D = \begin{cases} 1 \; if \; D^* \geq 0 \\ 0 \; otherwise \end{cases}$$

$$Y = Y_1 D + Y_0 (1 - D)$$

• Main additional assumptions

$$(i)\ \pi_D(Z)\ \textit{nondegenerate conditional on } X$$

$$(ii)\ (U_1, U_0, U_D) \perp Z \mid X$$

$$(iii)\ U_D\ \textit{distributed continuously}$$

• The first two are standard IV assumptions (we will come back to this later)
• The third is for simpler exposition
• There are other technical assumptions that I omit here for simplicity (refer to the paper)

• Treatment effect definitions

$$(i)\ \Delta^{ATE}(x) \equiv E(\Delta \mid X = x),\ where\ \Delta = Y_1 - Y_0$$

$$(ii)\ \Delta^{TT}(x) \equiv E(\Delta \mid X = x, D = 1)$$

$$(iii)\ \Delta^{TU}(x) \equiv E(\Delta \mid X = x, D = 0)$$

• The paper argues that all treatment effects in the literature are functions of the following treatment effect

$$(iv)\ \Delta^{MTE}(x, u_D) \equiv E(\Delta \mid X = x, U_D = u_D)$$

• This is the Marginal Treatment Effect

  • A small value of u_D is likely to produce participation, a large value is likely not to produce participation

Example: class sizes

Krueger (QJE, 1999)

• This paper provides an econometric analysis of a large-scale randomized experiment on class size conducted in the US, the Tennessee Student/Teacher Achievement Ratio experiment (STAR)

• Project STAR was a longitudinal study in which kindergarten students and their teachers were randomly assigned to one of three groups beginning in the 1985–1986 school year:

- Small classes (13–17 students per teacher)
- Regular-size classes (22–25 students)
- Regular/aide classes (22–25 students) which also included a full-time teacher's aide

• After their initial assignment, the design called for students to remain in the same class type for four years

• Over all four years, the sample had 11,600 students from 80 schools

• Each school was required to have at least one of each class-size type, and random assignment took place within schools

• Students were given standardized tests at the end of each year

- Deviations from the ideal experimental design:
  - Students in regular-size classes were randomly assigned again between classes with and without full-time aides at the beginning of the first grade
    - This means a different set of classmates (compared to the students in small-size classes who had the same set of classmates)
  - Approximately 10% of students switched between small and regular classes between grades (behavioral problems, and parental complaints)
  - Class sizes varied more than intended due to relocation of parents
  - Sample attrition was common: half of students who were present in kindergarten were missing in at least one subsequent year, as some students may have switched to another school upon knowing their class-type assignments

Krueger – results:

• Adjustments for school effects, attrition, re-randomization after kindergarten, non-random transitions, and variability in actual class size do not overturn the main findings on STAR:

  • Students in small classes scored higher on standardized tests than students in regular-size classes

  • The provision of a full-time teacher aide has only a modest effect on student achievement, although this effect may be attenuated because of the frequent availability of part-time aides in regular classes

• Interestingly, at least for the early grades, the analysis suggests that the main benefit of attending a small class seems to arise by the end of the initial year a student attends a small class (perhaps a socialization effect)

TABLE II

P-VALUES FOR TESTS OF WITHIN-SCHOOL DIFFERENCES AMONG SMALL, REGULAR, AND REGULAR/AIDE CLASSES

Variable	Grade entered STAR program			
	K	1	2	3
1. Free lunch	.46	.29	.58	.18
2. White/Asian	.66	.28	.15	.21
3. Age	.38	.12	.48	.40
4. Attrition rate	.01	.07	.58	NA
5. Actual class size	.00	.00	.00	.00
6. Percentile score	.00	.00	.46	.00

Each p-value is for an F-test of the null hypothesis that assignment to a small, regular, or regular/aide class has no effect on the outcome variable in that grade, conditional on school of attendance.

All rows except 4 pertain to the first grade in which the student entered the STAR program. The attrition rate in row 4 measures whether the student ever left the sample after initially being observed.

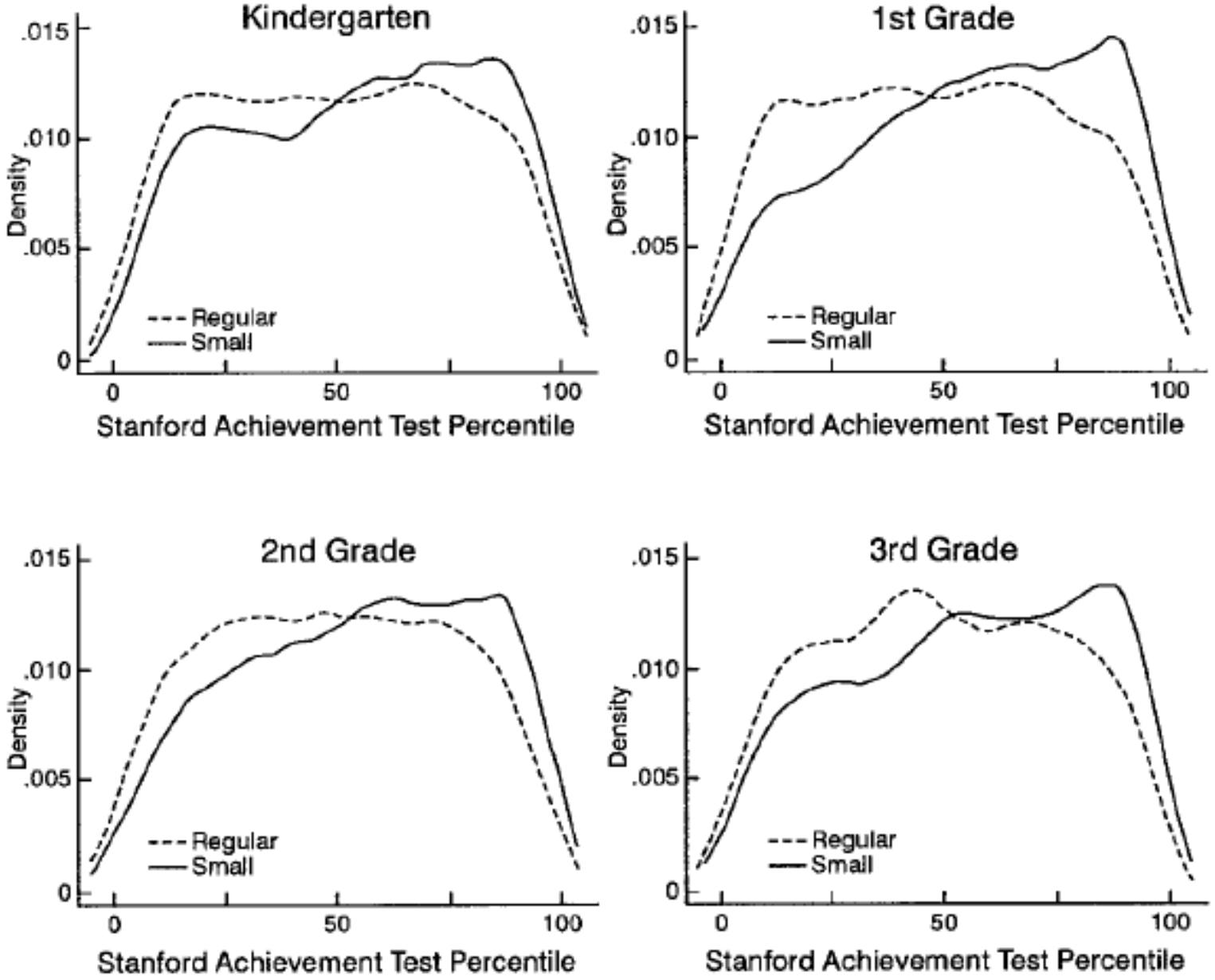


FIGURE I

Distribution of Test Percentile Scores by Class Size and Grade

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/A_{cs} + \beta_3 X_{ics} + \alpha_s + \varepsilon_{ics}$$

TABLE V

OLS AND REDUCED-FORM ESTIMATES OF EFFECT OF CLASS-SIZE ASSIGNMENT ON AVERAGE PERCENTILE OF STANFORD ACHIEVEMENT TEST

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Kindergarten								
Small class	4.82	5.37	5.36	5.37	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)	(2.19)	(1.25)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44	—	—	8.35	8.44
			(1.35)	(1.36)			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39	—	—	4.48	4.39
			(.63)	(.63)			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07	—	—	−13.15	−13.07
			(.77)	(.77)			(.77)	(.77)
White teacher	—	—	—	−.57	—	—	—	−.57
				(2.10)				(2.10)
Teacher experience	—	—	—	.26	—	—	—	.26
				(.10)				(.10)
Master's degree	—	—	—	−.51	—	—	—	−.51
				(1.06)				(1.06)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31	.01	.25	.31	.31
B. First grade								
Small class	8.57	8.43	7.91	7.40	7.54	7.17	6.79	6.37
	(1.97)	(1.21)	(1.17)	(1.18)	(1.76)	(1.14)	(1.10)	(1.11)
Regular/aide class	3.44	2.22	2.23	1.78	1.92	1.69	1.64	1.48
	(2.05)	(1.00)	(0.98)	(0.98)	(1.12)	(0.80)	(0.76)	(0.76)
White/Asian (1 = yes)	—	—	6.97	6.97	—	—	6.86	6.85
			(1.18)	(1.19)			(1.18)	(1.18)
Girl (1 = yes)	—	—	3.80	3.85	—	—	3.76	3.82
			(.56)	(.56)			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61	—	—	−13.65	−13.77
			(.87)	(.87)			(.88)	(.87)
White teacher	—	—	—	−4.28	—	—	—	−4.40
				(1.96)				(1.97)
Male teacher	—	—	—	11.82	—	—	—	13.06
				(3.33)				(3.38)
Teacher experience	—	—	—	.05	—	—	—	.06
				(0.06)				(.06)
Master's degree	—	—	—	.48	—	—	—	.63
				(1.07)				(1.09)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30	.01	.23	.29	.30

TABLE V
(CONTINUED)

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
C. Second grade								
Small class	5.93	6.33	5.83	5.79	5.31	5.52	5.27	5.26
	(1.97)	(1.29)	(1.23)	(1.23)	(1.70)	(1.16)	(1.10)	(1.10)
Regular/aide class	1.97	1.88	1.64	1.58	.47	1.44	1.16	1.18
	(2.05)	(1.10)	(1.07)	(1.06)	(1.23)	(0.87)	(0.81)	(0.81)
White/Asian (1 = yes)	—	—	6.35	6.36	—	—	6.27	6.29
			(1.20)	(1.19)			(1.21)	(1.20)
Girl (1 = yes)	—	—	3.48	3.45	—	—	3.48	3.44
			(.60)	(.60)			(.60)	(.60)
Free lunch (1 = yes)	—	—	−13.61	−13.61	—	—	−13.75	−13.77
			(.72)	(.72)			(.73)	(.73)
White teacher	—	—	—	.39	—	—	—	.43
				(1.75)				(1.76)
Male teacher	—	—	—	1.32	—	—	—	.82
				(3.96)				(4.23)
Teacher experience	—	—	—	.10	—	—	—	.10
				(.06)				(.07)
Master's degree	—	—	—	−1.06	—	—	—	−1.16
				(1.06)				(1.05)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.22	.28	.28	.01	.21	.28	.28
D. Third grade								
Small class	5.32	5.58	5.01	5.00	5.51	5.42	5.30	5.24
	(1.91)	(1.22)	(1.19)	(1.19)	(1.46)	(1.08)	(1.03)	(1.04)
Regular/aide class	−.22	−.16	−.33	−.75	−.30	.12	.13	−.10
	(1.95)	(1.12)	(1.11)	(1.07)	(1.17)	(0.85)	(0.81)	(0.78)
White/Asian (1 = yes)	—	—	6.12	6.11	—	—	5.97	5.96
			(1.45)	(1.44)			(1.44)	(1.43)
Girl (1 = yes)	—	—	4.16	4.16	—	—	4.17	4.18
			(.66)	(.65)			(.66)	(.66)
Free lunch (1 = yes)	—	—	−13.02	−12.96	—	—	−13.21	−13.16
			(.81)	(.81)			(.82)	(.81)
White teacher	—	—	—	.64	—	—	—	.19
				(1.75)				(1.75)
Male teacher	—	—	—	−7.42	—	—	—	−6.83
				(2.80)				(2.76)
Teacher experience	—	—	—	.04	—	—	—	.03
				(.06)				(.06)
Master's degree	—	—	—	1.10	—	—	—	.88
				(1.15)				(1.15)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.17	.22	.23	.01	.16	.22	.22

All models include constants. Robust standard errors that allow for correlated residuals among students in the same class are in parentheses. Sample size is 5861 for kindergarten, 6452 for first grade, 5950 for second grade, and 6109 for third grade.

$$Y_{ig} = \beta_0 + \beta_1 S_{io} + \beta_2 REG/A_{io} + \beta_3 N_{ig}^{S} + \beta_4 N_{ig}^{A} + \beta_5 X_{ig} + \alpha_g + \alpha_f + \alpha_s + \varepsilon_{ig}$$

TABLE IX

ESTIMATES OF POOLED MODELS

DEPENDENT VARIABLE: AVERAGE PERCENTILE RANKING ON SAT TEST

COEFFICIENT ESTIMATES WITH ROBUST STANDARD ERRORS IN PARENTHESES

Variable	(1)	(2)	(3)
Initial class small (1 = yes)	2.87	3.16	2.99
	(.83)	(.80)	(.80)
Initial class regular/aide (1 = yes)	.29	.49	.58
	(.69)	(.67)	(.67)
Cumulative years in small class	1.19	1.05	.65
	(.39)	(.38)	(.39)
Cumulative years in reg/aide class	.37	.25	.14
	(.39)	(.37)	(.37)
Fraction of classmates in class previous year	—	—	.60
			(1.03)
Average fraction of classmates together previous year	—	—	−.46
			(1.52)
Fraction of classmates on free lunch	—	—	−2.73
			(1.62)
Fraction of classmates who attended kindergarten	—	—	6.85
			(1.67)
Student and teacher characteristics	No	Yes	Yes
3 current grade dummies; 3 dummies indicating first grade appeared in sample; school effects	Yes	Yes	Yes
R^2	.18	.23	.23
Sample size	25,249	24,350	24,349

Student and teacher characteristics are as follows: student race, gender, and free lunch status; and teacher race, gender, experience, and master's degree or higher status. OLS estimates are reported, with robust standard errors that adjust for a possible correlation of residuals for the same student over time in parentheses.

Example: textbooks for schools

Glewwe, Kremer, and Moulin (AEJ, 2008)

- A randomized evaluation in rural Kenya looks at whether providing textbooks increases test scores for the average student
- 25 treatment and 25 control schools
- Results:
    - No significant results for the average student
    - Disaggregating the results by students' initial academic achievement suggests a potential explanation for the lack of an overall impact
        - Textbooks increased scores for students with high initial academic achievement and increased the probability that the students who had made it to the selective final year of primary school would go on to secondary school
- Many pupils could not read the textbooks, which are written in English, most students' third language
- The results are consistent with the hypothesis that the Kenyan educational system are oriented to the academically strongest

Table 1: Differences in Normalized Pre-Test Scores between Textbook Schools and 25-School Comparison Group

Subject	English		Math		Science		All subjects combined	
Grade	Grades with texts (3-7)	All grades (3-8)	Grades with texts (3, 5, 7)	All grades (3-8)	Grades with texts (8)	All grades (3-8)	Grades with texts	All grades
Difference between textbook schools and comparison schools	0.046 *(0.105)*	0.033 *(0.101)*	0.056 *(0.090)*	0.054 *(0.085)*	0.173 *(0.105)*	-0.017 *(0.088)*	0.061 *(0.091)*	0.023 *(0.087)*
Observations	8,516	9,332	5,069	9,302	816	9,276	14,401	27,910

Notes: Each column represents a regression of pre-test scores from January of year 1 on a constant and a dummy variable for being in a textbook school, with school random effects. The sample consists of all students from the 25 textbook schools and the 25-school comparison group who took the pre-test in January of year 1.

Columns (1) – (6) combine different grades and include dummy variables for each grade. Columns (7) and (8) combine subjects and grades and have dummy variables for each grade/subject combination. Columns (1), (3), (5) and (7) exclude grade/subject combinations that did not receive textbooks.

Standard errors in parentheses.

Table 4: Impact of Textbook Program on Normalized Test Scores

Dependent Variable	Normalized test score [a b]	Normalized test score [b]	Normalized test score minus pretest score [c]	Normalized test score minus pretest score [c]
	(1)	(2)	(3)	(4)
Textbook school	0.023 (0.087)	0.020 (0.104)	0.018 (0.053)	-0.046 (0.071)
Received a textbook				
Region and sex dummies	YES	YES	YES	YES
Years exposed to textbooks	1	2	1	2
Grades	3-8	4-7	3-8	4-7
Observations	24,132	12,663	11,321	7,354

Notes: * significant at 10%, ** significant at 5%, *** significant at 1%. Standard errors in parentheses.

[a] Running the same regressions for individual subjects English, math, and science (not shown in this Table), yields similar results, with the coefficients on textbooks never statistically significantly different from zero.

[b] Sample includes all children enrolled in January of year 1 who took the relevant October/November test

[c] Sample includes all children who were enrolled in January of year 1 and took the relevant October/November test as well as the pre-test in January of year 1.

Table 9: Normalized Test Scores by Quintile of Pre-Test Scores

Years exposed	*Quintile 1*	*Quintile 2*	*Quintile 3*	*Quintile 4*	*Quintile 5*
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
1	-0.049	-0.021	0.032	0.142*	0.218**
	(0.064)	*(0.069)*	*(0.073)*	*(0.079)*	*(0.096)*
2	-0.077	-0.109	-0.089	0.021	0.173
	(0.081)	*(0.094)*	*(0.104)*	*(0.101)*	*(0.131)*

Notes: * significant at 10%, ** significant at 5%, *** significant at 1%

Each row represents five random effects regressions, one for each quintile (based on pre-test scores from January of year 1), of post-test scores on a dummy variable indicating whether a child is in a textbook school and on dummy variables for region and sex. The sample consists of all children enrolled in January of year 1 who took both the pre-test in year 1 and the relevant post-test. All results are aggregated over all grade/subject combinations that received textbooks.

What have we learnt?

- Evaluation question demands a counterfactual
- Selection bias is the fundamental problem
- Randomization makes the selection bias disappear
- Randomization has problems too: ethics, external validity, compliance, spillovers
- Power calculations are important but require common sense
- Inference issues: standard errors allowing for intra-cluster correlation, pre-plan
- Examples

Problem Set 1

Exercise: Follow the instructions in chapter 12 of Khandker et al (2009). You should produce a do file and a log file, which should be commented to show that you understood the results. These should be emailed to Matilde Grácio (grader): only one email per group, please.

Note: Matilde should be able to run the do file on her computer given the original datafile and produce the raw log file.

Due date: Friday, February 18.