

A Practical Guide to the Program Evaluation Methods[†]Seiro Ito[‡]**Abstract**

This paper surveys the recent development of treatment effect literature. It is intended to give practical guidance for the applied researchers and policymakers. After showing the fundamental problem as that of missing counterfactual, we consider a variety of estimators according to the assumptions on exogeneity. First, we will see the benefits of randomized treatment assignment. Then, we will consider the case of randomization of eligibility, which is what is actually being randomized in most of the social experiments. We will point out the problems in randomized eligibility assignment, and some shortcoming of widely used intention-to-treat (ITT) estimator. Next, we will consider the average treatment effect (ATE) estimators based on exogenous treatment assignment. Noting exogeneity is a strong assumption, we will also consider the tests of exogeneity. Next, we will consider the bound-based method that can be applied when exogeneity fails and are left without instrumental variables. It is argued that one should use bound-based method more often rather than assuming unrealistic assumptions to get sharp conclusions. We will also consider the instrumental variable based method and its local average treatment estimator (LATE). We will see that LATE or IV estimators are valid only when the treatment effects are uniform across individuals, or when individuals participate to the program without thinking of the individual benefits from participation, which are both very unlikely. It is shown that, under heterogeneous impact, the treatment effect parameter (marginal treatment effect, MTE) differs across individuals with different propensity scores. With this motivation, we will study the treatment effect at given quantile rather than the mean impact, or the quantile treatment effect (QTE) estimators. An IV based estimator is shown to improve on LATE as it impose the distribution invariance, rather than treatment effect invariance, at each quantile. Then, we will consider the before-after data. Identification conditions of the widely popular difference-in-differences (DID) estimator is shown, and its limitation in the household context in developing countries. A novel changes-in-changes (CIC) estimator is also explained, while its limitation on the single index assumption is pointed out. The final section gives comparison over the methods.

Keywords: average treatment effect, randomized experiments, pitfalls in randomization, intention-to-treat estimator, bounds, exogenous treatment assignment, propensity score, matching, tests of exogeneity, instrumental variables, local average treatment effect, essential heterogeneity, marginal treatment effect, quantile treatment effect, difference-in-differences, changes-in-changes.

[†] Acknowledgements: This research is part of the research project on “Health Service and Poverty - Making Health Service More Accessible to the Poor” at the Institute of Developing Economies (IDE). The author would like to thank the members of research project for their comments.

[‡] Development Studies Center, Institute of Developing Economies.

Contents

Introduction	145
I Counterfactual and Treatment Effects	147
II A Selection Problem	149
III Randomized Experiments	150
III.1 Randomization of Treatment	150
III.2 Randomization of Eligibility	153
III.3 Pitfalls in Randomized Studies	154
IV Methods Based on Exogenous Treatment Assignment	159
IV.1 Regression Based Methods	162
IV.2 Matching Based Methods	164
IV.3 Propensity Score Based Methods	165
IV.4 Mixture of Methods	169
IV.5 Tests of Exogeneity	169
V Bound-Based Methods	170
V.1 Bounding the Conditional Probability	170
V.2 The Mixing Problem	171
VI Instrumental Variables Based Methods	174
VI.1 Instrumental Variable Estimator under Homogeneous Treatment Effects	175
VI.2 Instrumental Variable Estimator under Essential Heterogeneity	180
VII Quantile Treatment Effects	185
VII.1 Quantile Treatment Effects under Exogeneity	186
VII.2 IV Estimation of Quantile Treatment Effects	187
VIII Before-After Methods	189
VIII.1 Difference-in-Differences Estimation	189
VIII.2 Changes-in-Changes Estimation	192
Summary: Comparison of Estimation Methods	199

Introduction

What is program evaluation? Why do we need to do it? Plainly stated, program evaluation shows the impact of a program or a policy. It is given as the difference in the outcomes with and without the program. We need to evaluate a program because we want to know how much it had an impact on the outcome of interest.

The fundamental problem in program evaluation is that we cannot compare the outcomes with and without the program for the same individual. It is impossible for an individual to be both in and out of the program at the same time. So the problem we have is that of a missing counterfactual. Program evaluation literature seeks to find the ways to construct the missing counterfactual using statistical methods.

Naturally, the ways to construct the missing counterfactual depend on two things: the availability of information and data, and the statistical assumptions being met in the data. In what follows, we will see the hierarchy of statistical assumptions, from the strongest to the weakest: conditional treatment exogeneity, randomized eligibility, additive and time-invariant heterogeneity (fixed-effects), and presence of exogenous covariates. Depending on the assumptions being met, the possible choice of estimation method is determined.

Oddly enough, it is the strongest assumption, conditional treatment exogeneity, that is being employed most in the applied works. This is chosen probably out of the desire for ‘what needs to be done’. Evaluators are often tempted to assume too much than warranted in the data to draw sharp conclusions. This is because sharper conclusions are easier to read.^{*1} However, what one wants does not necessarily hold in reality, and one must adhere to the principle of ‘what can be done’. This principle calls for the knowledge of conditions that each estimator requires to correctly compute the impacts. This paper aims to serve as a practical guide for applied researchers and policymakers to the various estimators proposed in the treatment effect literature.

Having said the importance of statistical assumptions that hold in practice, it should be recognized that they are not something we must always take as given. They can be made to hold. When the epidemiologists run the clinical trials in the randomized, double-blind processes, they make sure that random treatment assignment (over the population of participating patients) holds. When the administrators of job training program under the Job Training Partnership Act (JTPA) randomly assigned the training eligibility, they tried make sure that eligibility

^{*1} Yet there is another factor which can be more important in practice: wider availability of statistical programs based on this assumption, a ‘supply side’ factor. For example, in \mathcal{R} , there are three different packages of programs that compute the propensity score matching estimator.

is exogenously given without any recourse to the ability of each individual. The efforts and initiative taken up by the Poverty Action Lab on social experiments are also intended to make sure that the eligibility is randomly assigned.

The statistical methods that can be used without calling for unrealistically strong assumptions depend on how well the evaluators are prepared. Preparedness can be understood in terms of: time and resources to be spent on, timing of evaluation, and implementability of program design suited for evaluation study. If a large scale survey can be done, it will equip us with the law of large numbers so the estimation will be more precise. If we have sufficient time, then we can wait until the outcome of interest has completed the change due to the program. If we can initiate the survey prior to implementation, then we can collect the baseline to be compared with the post intervention data, which gives more reliable control over the heterogeneity of individuals. If a blind experiment can be implemented, then, as we will see, we have less to worry about the biasedness in the estimates.

This suggests that one needs to at least plan ahead for evaluation. One needs to plan evaluation when they decide on program implementation. As we will see, having prior information buys us credible estimation even if we do not have a large number of explanatory variables (covariates), hence saves us with some money on collecting them. With stronger program implementability, one can randomize the eligibility to the point nobody would not decline to participate in the program if eligible, which will allow us more precise estimation. Or one can set and implement without exceptions an objective rule that is not related with the capacity of people, so we can safely assume that participants and nonparticipants are divided only by chance.

In the next section, we will articulate the nature of the evaluation problem. In section II, we will see how the naïve comparison between participants and nonparticipants bias the estimates, in the voluntary participation programs. In section III, we will see why randomized experiments are preferred, but also caution on the use of popular ITT estimator. In section IV, we will survey the most widely used set of cross-section estimators. We will study the tests of exogeneity assumption that the estimators are based on. In section V, we will briefly move away from point estimation and learn how we can bound the unknown parameter of interest with minimal set of assumptions. In section VI, we will cover the instrumental variables based methods, including LATE and IV estimators. It is pointed out that recent literature often argue against the use of IV based methods. Section VII provides a glance at the growing literature of quantile treatment effects, which can alleviate the shortcoming of IV estimators. In section VIII, we will consider the panel data models. Widely used DID estimator is shown and the plausibility of assumptions in the household context is discussed. We also give an illustration of the novel CIC estimator. In the last section, we will summarize and compare the various

techniques. We will omit the Bayesian and structural approaches because they are not practical for the applied researchers nor the policymakers. But this does not mean that one should be hesitant to use them, as they allow for richer heterogeneity and parameter identification.

I Counterfactual and Treatment Effects

The relationship we want to estimate is the effect of program compared with no intervention. The indicator for program intervention, or treatment statuses, is denoted as $D_i = 0, 1$ where 1 indicates with treatment and 0 otherwise. We will estimate D 's impact on the outcome of interest, y . The natural estimation is to get such an impact of a targeted person i :

$$\begin{aligned} \text{treatment effect of } D_i \text{ on } i &= (y_i \text{ when } D_i = 1) - (y_i \text{ when } D_i = 0), \\ &= (y_i | D_i = 1) - (y_i | D_i = 0), \end{aligned}$$

where we wrote $(y_i \text{ when } D_i = 1)$ as $(y_i | D_i = 1)$, and so forth. The symbol ' $|$ ' is used such as ' $y|x$ ' means ' y is conditioned on x ', or 'value of y when x is given.' If covariates (regressors) \mathbf{x} help explain the variations in y , then they should be included:

$$\text{treatment effect of } D_i \text{ on } i \text{ conditional on } \mathbf{x}_i = (y_i | D_i = 1, \mathbf{x}_i) - (y_i | D_i = 0, \mathbf{x}_i),$$

where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik-1})'$ is a $k \times 1$ vector of observables.

The fundamental problem in estimating the above is that we never observe $y_i | D_i = 1$ and $y_i | D_i = 0$ at the same time for the same individual i . So the challenge is to construct a suitable *counterfactual* of person i 's treatment status, that is, to construct what happened were i did (did not) get treated when i actually did not (did). For most of the time, constructing a counterfactual for each individual requires strong assumptions. They base on different sets of assumptions whose plausibility must be verified in the context of programs under question.

It is, however, possible to construct the *mean* of counterfactual over entire targeted population $\mathcal{E}[y_i | D_i, \mathbf{x}_i]$ under a set of reasonable assumptions, where $\mathcal{E}[\cdot]$ is an expectation operator.^{*2} The estimator we get using the population average is called the *average treatment effect (ATE)*:

$$ATE(\mathbf{x}) = \mathcal{E}[y_i | D_i = 1, \mathbf{x}_i] - \mathcal{E}[y_i | D_i = 0, \mathbf{x}_i].$$

There are several ways to construct the (means of) counterfactual. We can classify them into four categories: methods relying on randomized treatment assignment, methods based on observable treatment assignment rules, methods using before-after data, and methods based on instrumental variables.

^{*2} As we will later see, any statistic other than the mean, such as various quantiles, of population can be studied.

It may also be of an interest to know the mean treatment effect for those who took the treatment, or the *average treatment effect on the treated* which we denote as ATE_1 :

$$ATE_1 = \mathcal{E}[y_{1i}|D_i = 1, \mathbf{x}_i] - \mathcal{E}[y_{0i}|D_i = 1, \mathbf{x}_i]. \quad (1)$$

We can also consider the average treatment effect on the control ATE_0 as:

$$ATE_0 = \mathcal{E}[y_{1i}|D_i = 0, \mathbf{x}_i] - \mathcal{E}[y_{0i}|D_i = 0, \mathbf{x}_i]. \quad (2)$$

Although it may sound natural to pinning down the estimates of treatment effect parameters to the single numbers, there is another strand of thoughts that seeks to bound the estimates under weaker assumptions. This is appealing in the cases where the assumptions usually employed to point estimate the treatment effect parameters do not hold. We are often not as fortunate as the econometric theorists may think, and are left with data that do not satisfy the most popular set of assumptions. The bound-based approach allows us to get some information of the treatment effect, and with prior information, one can narrow down the bound to a reasonably width.

Not incidentally, it is rare to see in applied works the bound being used. This may be due to the fact that bound is not very popular among professional and nonprofessionals alike, and even if it is, the bound is sometimes too wide. However, this should not mean that we shall invoke the stronger assumptions to have a bound turned into a point estimate. As Manski (1995, 8) notes, one may have to develop tolerance for greater ambiguity in estimates, and may also have to face up with the hard fact that not all the questions can be answered credibly. A striking fact is that, even under a randomized experiment which is considered to be the gold standard in program evaluation, one can only get the bound of treatment effect parameter, and one sometimes has to invoke strong assumptions to get it pinned down to a single number. In this note, however, we will mostly focus on treatment effect parameters in numbers because most of the debate happens in the single valued parameter domain, not in the bound domain.

In what follows, we will assume that treatment will only affect those who are treated. This is called *stable unit treatment value assumption (SUTVA)*. It rules out externality in treatment such as deworming medicine studied by Miguel and Kremer (2004), or the repercussion on others through market mechanism, called the *general equilibrium effects* studied by Heckman, Lochner and Taber (1998). The latter can be considered as analogous to the price-taker assumption in microeconomics, which is often violated in a large scale intervention that affects prices. For example, in a large scale job training program, returns to skill may be lowered because of increased skill supply in the successful completion of the program, which may lower the treatment effects.

II A Selection Problem

Suppose that we want a consistent estimator of ATE. An estimator \hat{b} is said to be *consistent* if

$$\text{plim } \hat{b} = b.$$

This means that if we take plim or the *probability limit* (meaning if we increase the sample size to infinity), the estimator \hat{b} will coincide with the true value b .

When we run a regression with OLS

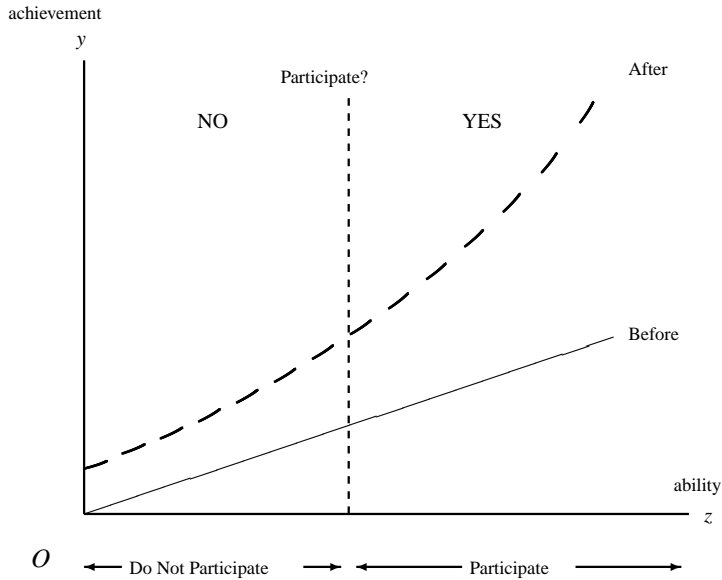
$$y_i = a + bx_i + \epsilon_i,$$

the necessary condition for estimated coefficients \hat{a} and \hat{b} to be consistent is that x_i and ϵ_i are uncorrelated.*³ Although the condition is simple, it is almost always unmet in the *observational data* where the value of x_i is chosen purposively by the agents under some optimizing process, rather than by a researcher who prefers to randomize the values of x_i for each i (for an estimation purpose, which is easily done in lab experiments of hard sciences that produce the *experimental data*). So all the estimation efforts boil down to devising a way to make these variables to become uncorrelated with each other (we say, we *orthogonalize* x_i and ϵ_i). For the simplest program evaluation where only participation affects y_i , we use $x_i = D_i$.

The reason for correlation between D_i and ϵ_i is clear. If the ‘ability’ (in benefiting from the program) among individuals is not uniform, then, under voluntary participation, the participants are more likely to have higher ability than the nonparticipants. Since we cannot observe ability, but it affects the outcome y_i nonetheless, it must be included in ϵ_i , for example, $\epsilon_i = c_i + e_i$ where c_i is ability of individual i known to i but unobservable to researchers and e_i is a random error term. Then, D_i and c_i must be positively correlated, as a higher value of c_i is more likely to be associated with $D_i = 1$ if the agents are rational, unless we explicitly include c_i in the regression. The bias of estimated \hat{b} , or deviation of expected value of \hat{b} from true b , is called a *selection bias*, as it originates from the fact that people voluntarily *self-select* themselves into the program.

*³ Sufficiency requires the model specification $y_i = a + bx_i + \epsilon_i$ to be correct, not, for example, $y_i = a + x_i^b + \epsilon_i$, $\sigma_\epsilon^2 < \infty$, and $x_i^2 \xrightarrow{P} \bar{x}^2 < \infty$.

Figure 1: Heterogenous Impacts and Self-Selected Program Participation



III Randomized Experiments

III.1 Randomization of Treatment

To develop a theory on treatment effect estimation, we start with the simplest setting and assume that the treatment assignment $D = 0, 1$ is random. When treatment status D is randomly assigned to individuals, the value of D is (statistically) independent of any variables:

$$\text{randomized trials} \iff \text{any variables} \perp\!\!\!\perp D.$$

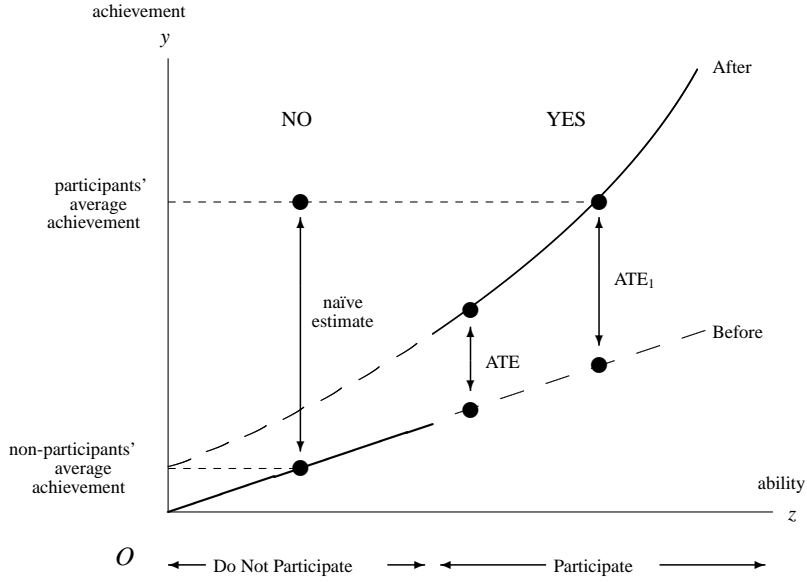
$a \perp\!\!\!\perp b$ means a is independent of b (and vice versa).

Let us consider the outcomes with and without treatment as random variables with different means. Denote the (stochastic) outcomes with treatment as y_1 and without treatment as y_0 . Consider a simple example where we assume y_D for $D = 0, 1$ is additively separable in a systematic part (mean) μ_D and a stochastic part u_D which varies from an individual to an individual and has a mean of zero. Under this simple setting, we can write:

$$\begin{aligned} y_{0i} &= \mu_0 + u_{0i}, \\ y_{1i} &= \mu_1 + u_{1i}. \end{aligned} \tag{3}$$

u_{Di} is an individual specific benefit for individual i for status D_i , and we have assumed that its means are zero's. Zero mean is a natural condition given we have purged all the systematic

Figure 2: Heterogenous Impacts and Upward Biasedness of Naïve Estimator



elements into μ_D .

Under this notation, if the treatment is randomly assigned, we see the first advantage of randomized treatment assignment:

$$\begin{aligned} ATE &\equiv \mathcal{E}[y_i|D_i = 1] - \mathcal{E}[y_i|D_i = 0], \\ &= \mathcal{E}[y_{1i}] - \mathcal{E}[y_{0i}] = \mathcal{E}[\mu_1 + u_{1i}] - \mathcal{E}[\mu_0 + u_{0i}] = \mu_1 - \mu_0, \end{aligned}$$

where we used $\mathcal{E}[u_{1i}] = \mathcal{E}[u_{0i}] = 0$. The sample analogue is:

$$\widehat{ATE} = \sum_{i=0}^{n_1} \frac{y_{1i}}{n_1} - \sum_{i=0}^{n_0} \frac{y_{0i}}{n_0}.$$

Alternatively, one can also estimate ATE by regressing y on 1 and D under a randomized trial. Note from the definition of y_{D_i} , we have:

$$y_i \equiv (1 - D_i)y_{0i} + D_i y_{1i}.$$

Plugging in the above into (3), we have:

$$y_i = (1 - D_i)(\mu_0 + u_{0i}) + D_i(\mu_1 + u_{1i}) = (\mu_0 + u_0) + D_i(\mu_1 - \mu_0) + D_i(u_{1i} - u_{0i}). \quad (4)$$

Taking $\mathcal{E}[y_i|D_i]$, we have:

$$\mathcal{E}[y_i|D_i] = \mu_0 + \mathcal{E}[u_0|D_i] + D_i(\mu_1 - \mu_0) + D_i\mathcal{E}[u_{1i} - u_{0i}|D_i] = \mu_0 + (\mu_1 - \mu_0)D_i, \quad (5)$$

where the last equality follows from $\mathcal{E}[u_{1i}|D_i] = \mathcal{E}[u_{1i}]$ and $\mathcal{E}[u_{0i}|D_i] = \mathcal{E}[u_{0i}]$ by random assignment of D_i , and we have $\mathcal{E}[u_{D_i}] = 0$ by assumption. Thus the regression parameter on D_i gives ATE.

If we take into account of the other factors $\tilde{\mathbf{x}}_i = (x_{1i}, \dots, x_{ik-1})$ that affect y , we can run OLS using all observations (treated and controls are used in the same regression):

$$y_i = a + \alpha D_i + \boldsymbol{\beta}' \tilde{\mathbf{x}}_i + \epsilon_i.$$

The estimated parameter $\hat{\alpha}$ is a consistent estimate of α or ATE, because $\epsilon_i \perp\!\!\!\perp D_i$ under randomized trials. Even if we do not include $\tilde{\mathbf{x}}_i$ explicitly and run a regression of y_i on 1 and D_i , the parameter on D_i gives a consistent estimate of $ATE(\mathbf{x})$. Omitting $\tilde{\mathbf{x}}_i$ in effect squeezes $\boldsymbol{\beta}' \tilde{\mathbf{x}}_i$ into the composite residual $u_i = \boldsymbol{\beta}' \tilde{\mathbf{x}}_i + \epsilon_i$. This does not affect consistency of $ATE(\mathbf{x})$ estimate $\hat{\alpha}$, because, under a randomized trial, D_i is uncorrelated with any variables.^{*4} So it gives:

$$\widehat{ATE}(\mathbf{x}) = \hat{\alpha}.$$

To be more rigorous on the conditional version of ATE, assume the separable model and take $\mathcal{E}[y_i|D_i, \mathbf{x}_i]$ on (4). Noting D_i is independent of any variables, including mean and stochastic part of y_{0i}, y_{1i} , we can take expectations of y_i conditional on \mathbf{x}_i separately from D_i . Thus taking an expectation conditional on D_i and \mathbf{x}_i of (4), which is equal to taking an expectation conditional on known \mathbf{x}_i with D_i , we have:

$$\begin{aligned} \mathcal{E}[y_i|D_i, \mathbf{x}_i] &= \mu_0 + \mathcal{E}[u_{0i}|D_i, \mathbf{x}_i] + (\mu_1 - \mu_0)D_i + \mathcal{E}[D_i(u_{1i} - u_{0i})|D_i, \mathbf{x}_i], \\ &= \mu_0 + \boldsymbol{\beta}'_0 \tilde{\mathbf{x}}_i + (\mu_1 - \mu_0)D_i + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' D_i \tilde{\mathbf{x}}_i, \end{aligned} \quad (6)$$

the first line follows because $\mathcal{E}[D_i|D_i, \mathbf{x}_i] = D_i$, and the second line follows because

$$\begin{aligned} \mathcal{E}[D_i(u_{1i} - u_{0i})|D_i, \mathbf{x}_i] &= D_i \mathcal{E}[u_{1i} - u_{0i}|D_i, \mathbf{x}_i], \\ &= D_i \mathcal{E}[u_{1i} - u_{0i}|\mathbf{x}_i], \\ &= D_i \mathcal{E}[\boldsymbol{\beta}'_1 \tilde{\mathbf{x}}_i + \epsilon_{1i} - \boldsymbol{\beta}'_0 \tilde{\mathbf{x}}_i - \epsilon_{0i}|\mathbf{x}_i]. \end{aligned}$$

So regressing y_i on 1, $\tilde{\mathbf{x}}_i$, D_i , $D_i \tilde{\mathbf{x}}_i$ gives an estimate of ATE as the coefficient on D_i when D_i is randomly assigned.

Another advantage of randomization is that one can sample from the entire population. In terms of FIGURE 2, one can sample from entire support of ‘ability’ distribution. So one does not have to worry about differences in the values of covariates, because associated covariates can be considered as also being randomly selected. This will be an important upside in the propensity score based methods that require strong ignorability condition described in IV.3.

^{*4} Technically, this may affect the efficiency of estimate using the finite sample, so by including \mathbf{x}_i should give more precise estimate of α than the mean difference.

Note that, from ATE and ATE_1 we have:

$$\begin{aligned} ATE &= \int_{\mathbb{X}} \{\mathcal{E}[y_1 - y_0|D = 0, \mathbf{x}] \Pr[D = 0|\mathbf{x}] + \mathcal{E}[y_1 - y_0|D = 1, \mathbf{x}] \Pr[D = 1|\mathbf{x}]\} dF(\mathbf{x}), \\ &= \int_{\mathbb{X}} \mathcal{E}[y_1 - y_0|\mathbf{x}] dF(\mathbf{x}), \end{aligned}$$

because $\mathcal{E}[y_1 - y_0|D, \mathbf{x}] = \mathcal{E}[y_1 - y_0|\mathbf{x}]$ under conditional mean independence. Similarly for ATE_1 :

$$ATE_1 = \int_{\mathbb{X}_1} \mathcal{E}[y_1 - y_0|\mathbf{x}] dF(\mathbf{x}),$$

and if $\mathbb{X} \neq \mathbb{X}_1$, $ATE \neq ATE_1$.

III.2 Randomization of Eligibility

It should be noted that, under individual freedom, it is not the treatment status D_i that is being randomized, because it is difficult to force someone who are unwilling to be treated. An experimenter can only randomly assign the *eligibility* to participate in treatment. Eligibility is not equal to treatment, because some individuals can opt out (called an exclusion error or a type 1 error). Comparing the effects of the eligible group over the ineligible group gives the *intention-to-treat (ITT)* estimator. An ITT estimator can be of an interest to the policymakers who understand the inability to assign treatment at will. It gives the mean outcome difference when the treatment is offered and when it is not. Denoting the eligible with $z_i = 1$ and the ineligible with $z_i = 0$, an ITT estimator for mean impact is given by taking a difference between the mean outcome of the eligible and the ineligible:

$$\widehat{ITT} = \sum_{z_i=1} \frac{y_i}{n_1} - \sum_{z_i=0} \frac{y_i}{n_0},$$

where n_1 is the number of eligible and n_0 is the number of ineligible individuals.

The group of people who are influenced by eligibility belongs to unknown subpopulation, and the ITT estimator gives the weighted average of who participated and who opted out less the average of ineligible group, with the weights possibly being a function of unobservables. So the difference in the averages of two groups does not give ATE. This echoes with the criticism raised against the instrumental variable estimator of ATE which we will cover later on. It is also shown that the ITT estimator for the mean does not give ATE_1 either.

To see these points, let us consider an example. Suppose that there are two types of individuals, one who wishes to get treated if eligible (wishers), and one who wishes not to get treated even if eligible (nonwishers). The fraction of wishers in population is $w \in [0, 1]$. Non-wishers have the outcome of $y_i = a$, while wishers without treatment have $y_i = b$, and wishers

with treatment have $y_i = c$. So ATE_1 is given by $c - b$.^{*5} When an experimenter randomizes the eligibility, we assume that only the wishers with eligibility assigned will get treated. (We therefore assume that there is no one treated if ineligible or if being a nonwisher.) Then, the mean outcome for the eligible is a weighted average of nonwishers and eligible wishers, $(1 - w)a + wc$, while the mean outcome of the ineligible is a weighted average of nonwishers and ineligible wishers, $(1 - w)a + wb$. The ITT estimator gives:

$$\widehat{ITT} = w(c - b).$$

So it does not give ATE_1 , but ATE_1 multiplied with wisher proportion in the population.^{*6}

Note that the ITT estimator is increasing in wisher proportion w . So the ITT estimator may be sensitive to the popularity of treatment among the subjects, which poses inconvenience because the perception can be different from the objective facts. It is also problematic because popularity or wishers proportions can vary with regions or time, and it may also be a function of how much resources are spent on educating the public about the benefits of treatment. So w can be endogenous to both program placement and specificities of program operations. This means that a large ITT estimate may not hold in other areas under different population and different program administration. In short, ITT estimator certainly serves as a reference, but it may not be useful due to its lack of external validity.

III.3 Pitfalls in Randomized Studies

If it is admissible to restrict program implementation only to a group of regions, then one can use distant regions with similar characteristics to construct the counterfactual. However, in practice, there remains an ethical and political problem whether one can restrict program implementation to one group of regions when there is another, yet distant, group with similar characteristics hence the similar needs for intervention.

Then how feasible for an authority to randomize the eligibility? It may seem politically infeasible to randomize the eligibility across individuals. However, if the request for program is strong and the funding or logistical capacity is limited, sometimes it is perceived as fairer to

^{*5} ATE should not be the interest of experimenter under free individual will in this case, because no one from nonwishers would never, ever, get treated.

^{*6} Had the proportions of wishers differ between the eligible and the ineligible, then the ITT estimator does not give an interesting parameter. Denoting wisher proportion of the eligible as w_1 and the ineligible as w_0 ,

$$\widehat{ITT} = w_1(c - a) - w_0(b - a).$$

The proportions may differ if an experiment is done against two different areas. Note that we may not be able to estimate w_0 . If there is no one from the ineligible take the treatment, LATE, an instrumental variable estimator of treatment effects to be covered later, gives ATE_1 because it divides the ITT estimator with the difference in treated proportion among the eligible and the ineligible, or $w - 0 = w$.

randomize the program eligibility. This is mostly the case for NGOs or governments with limited administrative capacities. Examples include Progreso of Mexico which randomized over regions (to be the first to get the program), school voucher program in Columbia which used a lottery for students, Bolivian Social Fund that randomized over communities, deworming medicines and schooling inputs (flip charts, school meals) interventions that randomized over Kenyan schools in backward districts, and, (the converse case of too few demand of) military service draft in the US during the Viet Nam War.

In addition to ethical and political economy concerns, there may be bias induced in randomized studies due to lack of capacity on the part of experimenters. In their careful review of social experiments in the US, Heckman and Smith (1995) point that, in one labor market program, inability of experimenter to find a sufficient number of control has lead to an expansion of target population beyond the original plan which alters the composition of subject pool. They also note the possibility that experimenters may use the threats of termination on the individuals who are currently receiving other benefits not to drop out, so would-be drop outs or opt outs are included in experiments by forced compliance. Thus the operational aspects of randomized studies may affect the subject pools both in the treated and the control. Such *randomization bias* leads to different composition of subject pools than voluntary-based programs, so the estimated parameters may not be relevant for the ITT estimator.

Another possible bias Heckman and Smith (1995) suggest is *substitution bias*. This is the problem of contamination when there is an alternative program available for the control. Any treatment effect estimates are thus interpreted as the effects of treatment over whatever available substitutes to it, which are not the proper counterfactual. This is likely to be serious if the need for treatment under question is widely acknowledged and there is competition over implementation. NGOs in developing countries often compete with each other in achieving better outcomes, which is quite sound by itself. However, this may contaminate the control by providing the substitutes or inducing migration to other NGO's domain.^{*7} So it becomes crucial for successful implementation of randomized experiments that the experimenter holds monopoly power over the provision of services. This will narrow down the feasible area for a randomized study.

In an important paper, Manski (1996) shows three other types of problems that an evaluator may face in social experiments. The first problem is *partial compliance*. Note that the average treatment effect on the eligible can be written as:

$$\begin{aligned} \mathcal{E}[y_1 - y_0|z_i = 1] &= \{\mathcal{E}[y_1|D_i = 1, z_i = 1] - \mathcal{E}[y_0|D_i = 1, z_i = 1]\} \Pr[D_i = 1|z_i = 1] \\ &+ \{\mathcal{E}[y_1|D_i = 0, z_i = 1] - \mathcal{E}[y_0|D_i = 0, z_i = 1]\} \Pr[D_i = 0|z_i = 1]. \end{aligned}$$

^{*7} Even if each NGO segments the regions, such segmentation is endogenous and is likely to bias the estimates through nonrandom program placement.

Given that z_i is randomized, ATE on the eligible is just an ATE. In social experiments, we observe the outcomes of eligible compliers, so their mean value $\mathcal{E}[y_1|D_i = 1, z_i = 1]$ and their proportion $\Pr[D_i = 1|z_i = 1]$ among the eligible can be computed. It may be possible (but not in every social experiment) to observe the outcomes of eligible noncompliers, so we can get $\mathcal{E}[y_0|D_i = 0, z_i = 1]$ and their proportion $\Pr[D_i = 0|z_i = 1]$. But social experiments do not give the counterfactual outcomes of compliers $\mathcal{E}[y_0|D_i = 1, z_i = 1]$ and noncompliers $\mathcal{E}[y_1|D_i = 0, z_i = 1]$, which makes us impossible to compute ATE. This is due to the fact that experiments cannot provide the joint distribution of y_1 and y_0 because one cannot observe y_0 and y_1 at the same time. So one cannot obtain the joint distribution or conditional densities $f(y_1|y_0)$ and $f(y_0|y_1)$ which are required to compute $\mathcal{E}[y]$ under some policy.

If we assume that mean outcomes would be the same between compliers and noncompliers,

$$\begin{aligned}\mathcal{E}[y_0|D_i = 1, z_i = 1] &= \mathcal{E}[y_0|D_i = 0, z_i = 1], \\ \mathcal{E}[y_1|D_i = 0, z_i = 1] &= \mathcal{E}[y_1|D_i = 1, z_i = 1],\end{aligned}$$

or

$$\mathcal{E}[y_D|D_i = 1, z_i = 1] = \mathcal{E}[y_D|D_i = 0, z_i = 1] = \mathcal{E}[y_D|z_i = 1], \quad (7)$$

then ATE is identified.

$$\begin{aligned}\mathcal{E}[y_1 - y_0|z_i = 1] &= \{\mathcal{E}[y_1|D_i = 1, z_i = 1] - \mathcal{E}[y_0|D_i = 0, z_i = 1]\} \Pr[D_i = 1|z_i = 1] \\ &\quad + \{\mathcal{E}[y_1|D_i = 1, z_i = 1] - \mathcal{E}[y_0|D_i = 0, z_i = 1]\} \Pr[D_i = 0|z_i = 1], \\ &= \mathcal{E}[y_1|D_i = 1, z_i = 1] - \mathcal{E}[y_0|D_i = 0, z_i = 1].\end{aligned}$$

(7) is called the *exogenous compliance* assumption because the mean outcome of compliers and noncompliers are assumed to be the same once the treatment is assigned (or not assigned).

Imposing exogenous compliance is one way of dealing with partial compliance. However, it effectively trades wide credibility of estimate for stronger conclusions. As Manski (1996) notes, if credibility is a central concern in evaluation, it is not an attractive way in dealing with the problem of missing counterfactual. Another way to deal with it is computing the bounds on the estimates. Denote \underline{k}_D and \bar{k}_D as logical lower- and upper-bounds on the mean outcome y_D for eligible noncompliers. Then:

$$\begin{aligned}\mathcal{E}[y_1|D_i = 1, z_i = 1] \Pr[D_i = 1|z_i = 1] + \underline{k}_1 \Pr[D_i = 0|z_i = 1] \\ \leq \mathcal{E}[y_1|z_i = 1] \\ \leq \mathcal{E}[y_1|D_i = 1, z_i = 1] \Pr[D_i = 1|z_i = 1] \\ \quad + \bar{k}_1 \Pr[D_i = 0|z_i = 1],\end{aligned}$$

and

$$\begin{aligned}\mathcal{E}[y_0|D_i = 1, z_i = 1] \Pr[D_i = 1|z_i = 1] + \underline{k}_0 \Pr[D_i = 0|z_i = 1] \\ \leq \mathcal{E}[y_0|z_i = 1] \\ \leq \mathcal{E}[y_0|D_i = 1, z_i = 1] \Pr[D_i = 1|z_i = 1] \\ \quad + \bar{k}_0 \Pr[D_i = 0|z_i = 1].\end{aligned}$$

Then:

$$\begin{aligned}
& \left[\mathcal{E}[y_1 | D_i = 1, z_i = 1] - \bar{k}_0 \right] \Pr[D_i = 1 | z_i = 1] \\
& - \left[\mathcal{E}[y_0 | D_i = 0, z_i = 1] - \underline{k}_1 \right] \Pr[D_i = 0 | z_i = 1] \\
& \leq \mathcal{E}[y_1 - y_0 | z_i = 1] \\
& \leq \left[\mathcal{E}[y_1 | D_i = 1, z_i = 1] - \underline{k}_0 \right] \Pr[D_i = 1 | z_i = 1] \\
& - \left[\mathcal{E}[y_0 | D_i = 0, z_i = 1] - \bar{k}_1 \right] \Pr[D_i = 0 | z_i = 1].
\end{aligned}$$

or

$$(\mu_1 - \bar{k}_0)(1 - p) - (\mu_0 - \underline{k}_1)p \leq \mathcal{E}[y_1 - y_0] \leq (\mu_1 - \underline{k}_0)(1 - p) - (\mu_0 - \bar{k}_1)p, \quad (8)$$

where we suppressed the conditioning event $z_i = 1$ in $\mathcal{E}[y_1 - y_0]$ because eligibility is randomly assigned, and

$$\begin{aligned}
\mu_1 &= \mathcal{E}[y_1 | D_i = 1, z_i = 1], & \mu_0 &= \mathcal{E}[y_0 | D_i = 0, z_i = 1], \\
p &= \Pr[D_i = 0 | z_i = 1].
\end{aligned}$$

The above bound may not be narrow enough, so it may not give a useful answer to the question that policymakers ask. However, a weaker conclusion is a price of wider credibility as we maintained on not imposing strong assumptions as in (7). One sees that, if we assume to know the way this subpopulation of wishers represents the entire population, then one obtains the treatment effects in numbers. If, on the other hand, we would not want to assume such, then we only get the bounds.

The second problem in an experimental study is *stratification*, or experimentation on a subpopulation. This happens if the subjects are drawn from a particular subpopulation. Manski (1996) gives examples: clinical trials are often tested on the subpopulation of volunteers, Illinois Unemployment Insurance experiment is tested on people who are already on the unemployment insurance, Jobs Opportunities and Basic Skills of 1988 drew sample from Aid to Families with Dependent Children (AFDC) recipients. Denoting the indicator of subpopulation as $S = 1$, one needs the *exogenous stratification* assumption in order to obtain the population treatment effect from the subpopulation:

$$y_D \perp\!\!\!\perp S. \quad (9)$$

(9) is highly unlikely if S represents participants of a social program, because people who are under certain social program find it beneficial to subscribe. If S represents geographical stratification, (9) may hold in some cases, for example, randomization over implementation order. If S represents an income classes or social groups, again, (9) is implausible because the entitlements are usually different among different income classes and social groups. It is straightforward to show that assuming (9) on the subpopulation of wishers or particular strata

gives treatment effect parameter in numbers, while not doing so will give the bound on the parameter.

The third problem is *treatment variation*. This is a problem in scaling-up an experiment: an evaluator wants to estimate the treatment effect applied to the new population where some of the intended subpopulation may or may not get treated, while the evaluator has the knowledge of marginal distributions of y_D for both $D = 0, 1$ from the existing experiments. Despite randomized experiments give the marginals, it is rarely the case that the same implementation of assignment rule used in experiments applies when it came to be scaled-up. This happens if universal implementation of treatment is untenable due to budgetary and logistical constraints, or if there is partial compliance in the program. So the randomized experiments may not provide a sufficiently informative reference. This can be considered as the operational counterpart of randomization bias.

The problem, then, is to get the possible outcome distribution under partial compliance or different assignment rules. Manski (1997) calls it the *mixing problem*: finding $\mathcal{E}[y]$ under some policy that allows arbitrary partial compliance or nonuniform treatment, using the knowledge of marginals $\mathcal{E}[y|D]$ from fully complied experimental studies. Manski (1995), (1997) show that, under a binary treatment, one can construct the bound on $\Pr[y]$. Naturally, depending on the assumption one makes, the bounds differ. We will cover the bound based methods in detail in the later section.

So one must be careful when reading upon a claim that a randomized social program ‘is found to have an impact’ or ‘significantly affects the outcome’, because, even with the randomized studies, there may be bias. Even if there is no bias, one can only get an ITT estimator which has questionable external validity. Further, even if it has external validity, the actual, feasible policy implementation may be different from experiments.*⁸ Without prior information or further assumptions, the most robust statement can only be made with the bound, not the point estimates. Quoting from Manski (1996, 731):

My own research, whether based on experimental or nonexperimental data, reveals a preference to maintain weak assumptions to keep attention focused on treatment effects in populations in substantive interest. If that means one can only bound the treatment of interest, so be it.

While Manski (1996) stops at the conservative bound estimates, Heckman and Smith (1995) advocate for using nonexperimental estimation. In addition to the pitfalls in randomized studies suggested so far, they argue that randomized studies: do not provide insights into the mechanism behind the success/failure of program, are not easy to understand under the presence of randomization bias, cannot build upon the cumulative knowledge of the nonexperimental

*⁸ This applies equally to the nonexperimental estimators, though.

studies, cannot learn about the drop out or opt out processes, tend to suffer from lack of administrative supports.^{*9} Citing Lalonde (1986)'s influential study that compared nonexperimental estimators with an experimental estimator, they note that limited sample size, limited range of applicable nonexperimental estimators, and lack of model selection strategies usually performed in nonexperimental model checks. Heckman, Ichimura, Smith, and Todd (1998) use experimental data of JPTA and show that propensity score based methods and DID give results consistent with the experimental evidence. They also note the benefits of nonexperimental data when having participants and nonparticipants to be in the same labor market so one can identify parameters over entire support of propensity score, applying the same questionnaire to both groups, and including information on recent labor market experiences. This is in a sense a little odd because they are basing their benchmark on the experimental estimator, while Heckman is rather critical on its use in Heckman and Smith (1995).

A problem similar to partial compliance in randomized experiments is attrition. In practice, one can drop out after learning the net benefits of the treatment. This will pose a selectivity problem if the drop outs are experiencing or learning the unobservable disutility in the treatment. Chan and Hamilton (2006) use structural estimation to estimate the impacts of unobservable individual side effects in explaining the drop outs in a clinical trial. The identifying assumption of individual side effects bases on the fact that the trial was conducted in the double-blind process, so the assignment to particular treatment does not reflect the prior beliefs or the preferences over treatment choice, so drop out process reveals the heterogenous impact of each treatment. This condition, however, should not hold in the social experiments. So an ITT estimator may overestimate the impact if attrition is due to negative selection.

- Progresa (a conditional cash transfer program, see Skoufias, 2005): randomizing at the district levels, then enforce eligibility criterion on households.
- Kenyan school meal programs (Vermeersch, 2003), deworming projects (Miguel and Kremer, 2004): Randomizing at the school level.
- Angrist et al. (2002) examine randomized voucher assignments for private school tuitions in Columbia. They found increases in test scores and likelihood of finishing 8th grade, and reduction in repetitions.
- Banerjee et al. (2005) use randomized sample of schools for remedial education.

IV Methods Based on Exogenous Treatment Assignment

Although we used the independence of D_i with any other variables, (6) is actually derived under a weaker assumption. Suppose that:

$$\mathcal{E}[y_{Di}|D_i, \mathbf{x}_i] = \mathcal{E}[y_{Di}|\mathbf{x}_i] \quad \text{or} \quad (y_{0i}, y_{1i} \perp\!\!\!\perp D_i)|\mathbf{x}_i \quad \iff \quad (\epsilon_{0i}, \epsilon_{1i} \perp\!\!\!\perp D_i)|\mathbf{x}_i. \quad (10)$$

^{*9} The 'black-box'ness of randomized studies has also been pointed out in Ito (2006) in the context of development studies.

The first equality shows that even if $\text{corr}[D_i, y_{D_i}] \neq 0$ or outcome is allowed to depend on participation status D_i , we assume that participation is systematically explained by \mathbf{x}_i , meaning expected value of D_i is fully explained with \mathbf{x}_i , leaving only the random errors unexplained. So the value of D_i , the actual participation status of i , becomes redundant information in estimating the mean of y_{D_i} , the outcome of i under both treated $D_i = 1$ and untreated $D_i = 0$, once we condition on \mathbf{x}_i . This means that any systematic part in y_{D_i} that are correlated with D_i is fully explained by \mathbf{x}_i . This assumption is called under different names:

- *conditional mean independence* ($\mathcal{E}[y_{D_i}|D_i, \mathbf{x}_i] = \mathcal{E}[y_{D_i}|\mathbf{x}_i]$) or mean independence conditional on \mathbf{x}_i , as mean of y_{D_i} is independent of treatment D_i once we condition on \mathbf{x}_i , or,
- *ignorability of treatment* ($y_{0i}, y_{1i} \perp\!\!\!\perp D_i|\mathbf{x}_i$), as D_i does not play any role once we condition on \mathbf{x}_i , or,
- *selection on observables* \mathbf{x}_i ($\epsilon_{0i}, \epsilon_{1i} \perp\!\!\!\perp D_i|\mathbf{x}_i$), because the assumption that D_i is fully explained by \mathbf{x} , thus $D(\mathbf{x}_i)$, is equivalent to an assumption that the selection into the treatment is fully explained only by observables \mathbf{x}_i .

All of three point to the same statistical assumption in the context of estimating ATE.^{*10} We call this family of assumptions ‘exogenous treatment assignment’ because it implies $\mathcal{E}[D_i\epsilon_{D_i}] = 0$, as opposed to the endogenous treatment assignment which occurs when $\mathcal{E}[D_i\epsilon_{D_i}] \neq 0$.

$(\epsilon_{0i}, \epsilon_{1i} \perp\!\!\!\perp D_i)|\mathbf{x}_i$ means there can be some factors not captured in \mathbf{x}_i to be included in $\epsilon_{0i}, \epsilon_{1i}$, but they must be independent of treatment status D_i , such as errors unrelated to ‘ability.’ Then (4) becomes:

$$\begin{aligned}\mathcal{E}[y_i|D_i, \mathbf{x}_i] &= \mu_0 + \mathcal{E}[\epsilon_{0i}|D_i, \mathbf{x}_i] + (\mu_1 - \mu_0)\mathcal{E}[D_i|D_i, \mathbf{x}_i] + \mathcal{E}[D_i(\epsilon_{1i} - \epsilon_{0i})|D_i, \mathbf{x}_i], \\ &= \mu_0 + \mathcal{E}[\epsilon_{0i}|\mathbf{x}_i] + (\mu_1 - \mu_0)D_i + D_i(\mathcal{E}[\epsilon_{1i}|\mathbf{x}_i] - \mathcal{E}[\epsilon_{0i}|\mathbf{x}_i]),\end{aligned}$$

under (10), because $\mathcal{E}[\epsilon_{D_i}|D_i, \mathbf{x}_i] = \mathcal{E}[\epsilon_{D_i}|\mathbf{x}_i]$ and $\mathcal{E}[D_i\epsilon_{D_i}|D_i, \mathbf{x}_i] = D_i\mathcal{E}[\epsilon_{D_i}|D_i, \mathbf{x}_i] = D_i\mathcal{E}[\epsilon_{D_i}|\mathbf{x}_i]$ under conditional mean independence or $(\epsilon_{0i}, \epsilon_{1i} \perp\!\!\!\perp D_i)|\mathbf{x}_i$. Thus, if we let $\mathcal{E}[\epsilon_{D_i}|\mathbf{x}_i] = \tilde{\beta}'_D \tilde{\mathbf{x}}$, we get the same result as (6) under a weaker assumption of conditional mean independence.

An illustration of conditional mean independence is:

$$\mathcal{E}[y_{0i}|D_i = 1, \mathbf{x}_i] = \mathcal{E}[y_{0i}|D_i = 0, \mathbf{x}_i] = \mathcal{E}[y_{0i}|\mathbf{x}_i].$$

The expected value of a hypothetical outcome y_{0i} (not being treated) of the individual who is actually treated is the same for that of the controls, once we condition on \mathbf{x}_i . So if we

^{*10} Strictly speaking, the latter two are conditional independence which can be applied to entire distribution while the first is mean independence which is only restricted to the mean.

control for the differences in \mathbf{x}_i , the (mean) outcome will be the same for the treated and the controls, meaning we have independence in mean conditional on \mathbf{x}_i . E.g., without FONCODES interventions, all the mean outcomes of villages will be the same between the treated and the control (for the latter we observe), if we base our expectation on observables \mathbf{x}_i . Under our assumption, this also holds true for y_{1i} , or $\mathcal{E}[y_{1i}|D_i = 1, \mathbf{x}_i] = \mathcal{E}[y_{1i}|D_i = 0, \mathbf{x}_i] = \mathcal{E}[y_{1i}|\mathbf{x}_i]$. Given the conditional expectations of y_{0i} , y_{1i} are the same for both treatment statuses $D_i = 0, 1$, the actual value of D_i becomes irrelevant in computing the conditional expectations.

As noted, under this assumption D_i becomes redundant once we condition on \mathbf{x}_i , we have:

$$\mathcal{E}[y_{0i}|D_i = 1, \mathbf{x}_i] = \mathcal{E}[y_{0i}|D_i = 0, \mathbf{x}_i]$$

indicating that the missing counterfactual $\mathcal{E}[y_{0i}|D_i = 1, \mathbf{x}_i]$ can be constructed from the controls by $\mathcal{E}[y_{0i}|D_i = 0, \mathbf{x}_i]$. Then, it is clear that if the above relationship is used for constructing the mean counterfactual for the treated $\mathcal{E}[y_{0i}|D_i = 1, \mathbf{x}_i]$, we are conditioning on $D_{1i} = 1$, or estimating ATE on the treated:

$$ATE_1 = \mathcal{E}[y_{1i}|D_i = 1, \mathbf{x}_i] - \mathcal{E}[y_{0i}|D_i = 1, \mathbf{x}_i].$$

This can be estimated under conditional mean independence:

$$ATE_1 = \mathcal{E}[y_{1i}|D_i = 1, \mathbf{x}_i] - \mathcal{E}[y_{0i}|D_i = 0, \mathbf{x}_i].$$

Also note that if we are going to estimate ATE on the controls ($D_i = 0$), then we use

$$\mathcal{E}[y_{1i}|D_i = 0, \mathbf{x}_i] = \mathcal{E}[y_{1i}|D_i = 1, \mathbf{x}_i],$$

thus we have ATE on the control ATE_0 as:

$$\begin{aligned} ATE_0 &= \mathcal{E}[y_{1i}|D_i = 0, \mathbf{x}_i] - \mathcal{E}[y_{0i}|D_i = 0, \mathbf{x}_i], \\ &= \mathcal{E}[y_{1i}|D_i = 1, \mathbf{x}_i] - \mathcal{E}[y_{0i}|D_i = 0, \mathbf{x}_i], \end{aligned}$$

which is exactly the same as ATE_1 . This is logical since, under exogenous treatment assignment, the opposite group can be used as the counterfactual after appropriate control of the covariates.

In estimating ATE_1 , we actually need only conditional mean independence for y_{0i} , not for y_{1i} (because we want the counterfactual of y_{1i}):

$$\mathcal{E}[y_{0i}|D_i, \mathbf{x}_i] = \mathcal{E}[y_{0i}|\mathbf{x}_i] \quad \text{or} \quad (y_{0i} \perp\!\!\!\perp D_i)|\mathbf{x}_i.$$

In estimating ATE_0 , we need likewise:

$$\mathcal{E}[y_{1i}|D_i, \mathbf{x}_i] = \mathcal{E}[y_{1i}|\mathbf{x}_i] \quad \text{or} \quad (y_{1i} \perp\!\!\!\perp D_i)|\mathbf{x}_i.$$

So we are invoking different statistical assumptions when estimating ATE_0 and ATE_1 , although the values of estimates are the same.

In practice, there are four methods that are available in estimating treatment effects under the assumption of exogenous treatment assignment. These differ in the way to compute the conditional means: regression, matching, propensity score, matching by blocking, and mixture methods. We will also consider the tests of exogeneity assumption. See Wooldridge (2002) and Imbens (2004) for more detailed presentation from which I draw heavily. I will turn to each of them in below.

IV.1 Regression Based Methods

Regression based method estimates the means of outcomes under no treatment μ_0 and under treatment μ_1 with regression, often using covariates \mathbf{x} . Thus

$$ATE(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}).$$

There are two ways to estimate means $\mu_1(\mathbf{x})$, $\mu_0(\mathbf{x})$.

- Parametric method:

$$y_i = c + \alpha D_i + \boldsymbol{\beta}' \mathbf{x}_i + \boldsymbol{\delta}' D_i (\mathbf{x}_i - \bar{\mathbf{x}}) + e_i,$$

with

$$ATE(\mathbf{x}_i) = \hat{\alpha} + \hat{\boldsymbol{\delta}}' (\mathbf{x}_i - \bar{\mathbf{x}})$$

\mathbf{x}_i term remains because we allow $\mathcal{E}[D_i \mathbf{x}_i] \neq \mathbf{0}$. This can be derived by assuming a linear function for $\mathcal{E}[\epsilon_{Di} | \mathbf{x}_i]$ in (6). Denoting $\mathcal{E}[\mathbf{x}_i] = \boldsymbol{\mu}_x$, take:

$$\mathcal{E}[\epsilon_{Di} | \mathbf{x}_i] = \boldsymbol{\beta}'_{Di} (\mathbf{x}_i - \boldsymbol{\mu}_x) \quad \text{with} \quad \mathcal{E}[\epsilon_{Di}] = \mathcal{E}_X[\mathcal{E}[\epsilon_{Di} | \mathbf{x}_i]] = 0.$$

We have subtracted $\boldsymbol{\mu}_x$ from \mathbf{x}_i for $\mathcal{E}[\epsilon_{Di}] = \mathcal{E}_X[\mathcal{E}[\epsilon_{Di} | \mathbf{x}_i]] = 0$ to hold. It can also be seen as the additional control for the difference in covariates distribution. Then

$$\begin{aligned} \mathcal{E}[y_i | D_i, \mathbf{x}_i] &= \mu_0 + \mathcal{E}[\epsilon_{0i} | \mathbf{x}_i] + (\mu_1 - \mu_0) D_i + D_i (\mathcal{E}[\epsilon_{1i} | \mathbf{x}_i] - \mathcal{E}[\epsilon_{0i} | \mathbf{x}_i]), \\ &= (\mu_0 - \boldsymbol{\beta}'_0 \boldsymbol{\mu}_x) + (\mu_1 - \mu_0) D_i + \boldsymbol{\beta}'_0 \mathbf{x}_i + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' D_i (\mathbf{x}_i - \boldsymbol{\mu}_x), \end{aligned} \quad (11)$$

so we have $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$, $c = \mu_0 - \boldsymbol{\beta}'_0 \boldsymbol{\mu}_x$, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. One can obtain flexibility by including polynomials in \mathbf{x}_i that are linear in parameters for $\mathcal{E}[\epsilon_{Di} | \mathbf{x}_i]$.

- *Regression discontinuity design* with a nonstochastic assignment rule $f(s_i)$ of treatment: an example is when we know the policy rule $D_i \stackrel{\text{def}}{=} f(s_i)$ for observable s_i . This is a special case of parametric method.

$$y_i = c + \alpha f(s_i) + \boldsymbol{\beta}' \mathbf{x}_i + \boldsymbol{\delta}' f(s_i) (\mathbf{x}_i - \bar{\mathbf{x}}) + e_i,$$

Because $f(s_i)$ is a deterministic function, and s_i is assumed to be orthogonal to e_i , hence $f(s_i)$ cannot be correlated with the error. A drawback is that any discrete changes in the outcome which may be due to other causes are attributed to the change in $f(s_i)$.

- Nonparametric method:

$$ATE(\mathbf{x}) = \frac{\sum_{i=1}^n \hat{\mu}_{1i}(\mathbf{x}_i) - \hat{\mu}_{0i}(\mathbf{x}_i)}{n}.$$

with $\mu_D(\mathbf{x}_i, e_{Di})$ being an unknown mean function to be estimated nonparametrically. Imbens (2004) points that regression based method relies on extrapolation of the control to get the counterfactual of the treated, thus there can be bias introduced in extrapolation if the distribution (support) of covariates are different. To control for the bias, he proposes to regress y_{0i} on \mathbf{x}_{0i} and $\bar{\mathbf{x}}_1 - \mathbf{x}_{0i}$ to include the average bias term $\delta'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$ in construction of counterfactual.

$$\mu_0(\mathbf{x}_i, e_{0i}) = \mu_0[\mathbf{x}_{0i}, \delta'(\bar{\mathbf{x}}_1 - \mathbf{x}_{0i}), e_{0i}].$$

One caution is that the means are not robust to outliers, thus if $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_0$ are very different, the predicted bias term can be sensitive to its specification. One can use kernel smoothing method, preferably with a local, not global, smoothing parameter, in estimating the conditional mean to appropriately control for the limited overlap in domain. However, one may have to cope with the dimensionality problem if the number of covariates is relatively large, a feature that is common in nonparametric regression.

As a (parametric) substitute of nonparametric estimation, one can use a known function with reasonable flexibility in μ_D , e.g., a choice of low-order polynomials in \mathbf{x}_i , $h(\mathbf{x}_i)$, e.g., $x_1^2 + x_1x_2 + x_2^2 + \dots$. Then

$$\hat{\mu}_{Di} = \hat{\beta}'_D \mathbf{x}_i + \hat{\gamma}'_D h(\mathbf{x}_i).$$

The reason for these estimators to be consistent is that, conditional on \mathbf{x} , means of y_D are independent of D , thus can be omitted from the two separate regressions of y_{Di} on \mathbf{x}_i .

- Angrist and Lavy (2004): Maimonides' Rule (class size < 40) of Babylonian Talmud enforced in Israel gives a deterministic rule of the class size as a function of number of potential students, and is likely to be uncorrelated with any other variables that affect school outcomes such as test scores. For example, area 1 with 40 potential students will have 2 classes and area 0 with 39 has only one. Provided that the areas with 40 and 39 potential pool of students can be similar in other respect, it gives a good case for comparison. Thus $s_{1i} = s_{\text{area } 1} = 40$ for all i in area 1 (treated), and $s_{0i} = s_{\text{area } 0} = 39$ for all i in area 0 (controls) and $f(40) = 1$ and $f(39) = 0$ is an indicator of treatment assignment in the class size experiment.
- Pitt and Khandker (1998): using the threshold of .5 acres as a deterministic eligibility (assignment) rule, they estimate the effects of group-based credit programs. They compare the outcomes of eligible households with ineligible households within the program village, conditional on other covariates \mathbf{x}_i and village fixed-effects that control for the placement endogeneity. Morduch (1999?) points that the .5 acre rule is not strictly enforced by the officials, thus regression discontinuity design fails in practice. See Armendariz-Aghion and Morduch (2005) for details.
- Ravallion and Wodon (2000) take community-level variables that central government use in allocating funds as instruments for treatment of Food-For-Education (FFE) program in Bangladesh. This follows as the central

government's allocation should not be correlated with the household-level variables, simply because they cannot observe them. This validates the use of central government allocation as instruments for household-level treatment. This means that the authors assume the treatment eligibility (that the village is allocated funds) is independent of any of household variables, as households cannot influence central government allocations. This also implies an assumption of independence of outcomes and participation statuses of eligible/ineligible households, conditional on village treatment eligibility z_i : after controlling for the fact that village is treated or nontreated, the decision on schooling/work should not differ between households across villages. That is, in the absence of this program, household behavior should be the same between the treated and the control villages. The estimation technique they used is not LATE, but probit with endogenous and censored variable. This is a generalization of Smith and Blundell (1986) and uses regression residuals of participation equation as a regressor in probits of work and schooling. Their estimation employs a clever strategy, however, they include household-level variables \mathbf{x}_i as explanatory variables in household's FFE receipt equation, $FFE_i = \gamma FFEV_i + \boldsymbol{\eta}' \mathbf{x}_i + v_i$, which can be correlated with household unobservables that bias estimates of $\hat{\eta}$ and residual \hat{v}_i . The estimate of γ may not be biased given eligibility is orthogonal to household-level factors. Although this is possibly done out of necessity that one needs the household-level variability that predicts household's FFE receipt, it invalidates the exogeneity test they perform in probits of work/school, as it relies on consistency of \hat{v}_i . Variables included in \mathbf{x}_i are household demography, marital status, religion, education, and land ownership. Some of those may be correlated with unobservable factors that influence FFE receipt, such as ability of members. \mathbf{x}_i should be confined to household-level exogenous variables, such as sex ratios (partly endogenous if there is a preference for balanced sex-ratio), land inheritance, and religion.

- Duflo (2001) also uses number of school primary construction in a district as the identification variable in explaining the education outcomes, based on the assumption that it is implemented across the board and is not correlated with individual-level variables. She compares education outcomes of cohorts prior to and after school construction period, for high-intensity (many school construction) areas and low-intensity (fewer school construction) areas, and found a significant increase in mean enrollment and mean wages for the post-intervention cohorts, especially for the high-intensity areas.

IV.2 Matching Based Methods

Matching based methods choose the counterfactual from the opposite treatment group that is close to the original reference observation. Closeness is determined by evaluating the distance between covariates \mathbf{x}_{1i} and \mathbf{x}_{0j} . Distance is computed with the choice of metric, e.g., Euclidean, Mahalanobis, etc. Even with the same distance metric, matching estimators can differ in the construction of counterfactual $\hat{\mu}_{0i}$. Once metric is chosen, researcher must decide on the number of matches for a given observation. Matches can be based on more than one counterfactual. It is always the case that there is no exact match; then one can use the nearest m neighbours, or can use the kernel estimator (smoothing over certain range of \mathbf{x}_i) of the control for i , etc. As noted earlier, kernel estimators have a problem of choosing the smoothing parameter, and a potential bias if the distribution of covariates differ significantly. There is little result on the optimal choice of metric, and is still under investigation (Imbens, 2004). The most popular estimator is propensity score matching estimator, which we will turn next.

One should note that matching based on metric and on propensity score differ in weighting on covariates. Propensity score matching estimator weights covariates according to the propensity score regression function. This gives efficiency in estimation provided that propen-

sity score estimates are consistent, in particular, exogeneity of covariates in selection equation, e.g., absence of measurement errors, omitted variables, fixed effects, etc. If not, it introduces bias. Thus metric-based matching estimators are more robust than propensity score matching estimator to the failure of exogeneity assumption.

IV.3 Propensity Score Based Methods

A propensity score G is a probability of being treated. Usually, it is estimated using logit or probit, by regressing treatment status on the set of covariates \mathbf{x}_i , given as $\hat{G}_i = G(\hat{\gamma}'\mathbf{x}_i)$, where $G(\cdot)$ is a logistic function for logit models and standard normal distribution function for probit models. For multiple treatment choices, one can use multinomial logit models. The merit of using the propensity score based methods is that one does not have to compare k -dimensional aspects of individuals to construct the counterfactual as in other matching estimators, but only one-dimensional \hat{G}_i (Rosenbaum and Rubin, 1983, see also the intuition given in Imbens, 2004). This may imply that the kitchen-sink regression of (11) may perform no worse than the propensity score based method (Wooldridge, 2002), and moreover, the former is a one-step procedure whereas the latter requires two-steps, losing efficiency. Another interpretation of propensity score based methods is that controlling for the propensity score can be seen analogously as controlling for the sampling weights in sampling theory. One controls for the probability of being selected into treatment, and use matched counterfactual.

- An additional assumption: $0 < G_i(\mathbf{x}_i) < 1$ for all i (called *strong ignorability of treatment* by Rosenbaum and Rubin, 1983). This is, in other words, there is a substantial overlap in covariates \mathbf{x}_i between the treated and the control, thus there is no point on the support of \mathbf{x}_i that only a single treatment status is observed. Then, Rosenbaum and Rubin (1983) show that ATE is given by:

$$\widehat{ATE}(\mathbf{x}_i) = n^{-1} \sum_{i=1}^n \frac{y_i(D_i - \hat{G}_i)}{\hat{G}_i(1 - \hat{G}_i)}$$

where \hat{G}_i is estimated propensity score of treatment. This follows from an application of the expectation operator on

$$(D - G)y = (D - G)[(1 - D)y_0 + Dy_1] = Dy_1 - G(1 - D)y_0 - GDy_1.$$

Taking an expectation on the above conditional on \mathbf{x} and D , which is allowed under the conditional mean independence assumption, gives:

$$\mathcal{E}[(D - G)y|D, \mathbf{x}] = D\mathcal{E}[y_1|\mathbf{x}] - G(1 - D)\mathcal{E}[y_0|\mathbf{x}] - GD\mathcal{E}[y_1|\mathbf{x}].$$

Then, taking an expectation over D , we have:

$$\begin{aligned}\mathcal{E}[(D - G)y|\mathbf{x}] &= \mathcal{E}_D [\mathcal{E}[(D - G)y|D, \mathbf{x}] | \mathbf{x}_i], \\ &= G\mathcal{E}[y_1|\mathbf{x}] - G(1 - G)\mathcal{E}[y_0|\mathbf{x}] - G^2\mathcal{E}[y_1|\mathbf{x}], \\ &= G(1 - G) \{ \mathcal{E}[y_1|\mathbf{x}] - \mathcal{E}[y_0|\mathbf{x}] \}.\end{aligned}$$

So

$$\mathcal{E} \left[\frac{(D - G)y}{G(1 - G)} \middle| \mathbf{x} \right] = \mathcal{E}[y_1|\mathbf{x}] - \mathcal{E}[y_0|\mathbf{x}] = ATE(\mathbf{x}).$$

Note that:

$$\begin{aligned}\mathcal{E}_x \left[\mathcal{E} \left[\frac{Dy}{G(\mathbf{x})} \middle| \mathbf{x} \right] \right] &= \mathcal{E}_x \left[\mathcal{E} \left[\frac{Dy_1}{G(\mathbf{x})} \middle| \mathbf{x} \right] \right] = \mathcal{E}_x \left[\frac{\mathcal{E}[D|\mathbf{x}]\mathcal{E}[y_1|\mathbf{x}]}{G(\mathbf{x})} \right], \\ &= \mathcal{E}_x \left[\frac{G(\mathbf{x})\mathcal{E}[y_1|\mathbf{x}]}{G(\mathbf{x})} \right] = \mathcal{E}_x [\mathcal{E}[y_1|\mathbf{x}]] = \mathcal{E}[y_1],\end{aligned}$$

where the second equality holds because $D \perp\!\!\!\perp y|\mathbf{x}$. Similarly,

$$\begin{aligned}\mathcal{E}_x \left[\mathcal{E} \left[\frac{(1 - D)y}{1 - G(\mathbf{x})} \middle| \mathbf{x} \right] \right] &= \mathcal{E}_x \left[\mathcal{E} \left[\frac{(1 - D)y_0}{1 - G(\mathbf{x})} \middle| \mathbf{x} \right] \right] = \mathcal{E}_x \left[\frac{(1 - \mathcal{E}[D|\mathbf{x}])\mathcal{E}[y_0|\mathbf{x}]}{1 - G(\mathbf{x})} \right], \\ &= \mathcal{E}_x \left[\frac{[1 - G(\mathbf{x})]\mathcal{E}[y_0|\mathbf{x}]}{1 - G(\mathbf{x})} \right] = \mathcal{E}_x [\mathcal{E}[y_0|\mathbf{x}]] = \mathcal{E}[y_0].\end{aligned}$$

Thus ATE is given by:

$$ATE = \mathcal{E} \left[\frac{Dy}{G(\mathbf{x})} - \frac{(1 - D)y}{1 - G(\mathbf{x})} \right],$$

and its sample analogue is:

$$\begin{aligned}\widehat{ATE} &= \frac{1}{n} \sum_{i=1}^n \left(\frac{D_i y_i}{G(\mathbf{x}_i)} - \frac{(1 - D_i) y_i}{1 - G(\mathbf{x}_i)} \right), \\ &= \frac{1}{n} \sum_{i=1}^n \frac{(D_i - G_i) y_i}{G_i(1 - G_i)},\end{aligned}$$

which gives the Rosenbaum and Rubin (1983)'s estimator. This is a propensity score weighted estimator: one uses the inverse of propensity score as weights to control for the 'sampling' probability. Hirano, Imbens, and Ridder (2003) use nonparametric estimation of propensity score, namely, logistic power series, or the power series of covariates to estimate the log odds ratio. They show that estimated parameter achieves the semi-parametric efficiency bound of Hahn (1998). As this estimator has weights that do not add up to 1, one reweights and:

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n (\omega_{1i} y_i - \omega_{0i} y_i),$$

where

$$\omega_{1i} = \frac{D_i}{G(\mathbf{x}_i)}, \quad \omega_{0i} = \frac{1-D_i}{1-G(\mathbf{x}_i)}.$$

$$\sum_{i=1}^n \frac{D_i}{G(\mathbf{x}_i)}, \quad \sum_{i=1}^n \frac{1-D_i}{1-G(\mathbf{x}_i)}$$

- Regression on propensity score. $\mathcal{E}[y_i|D_i = 1, \mathbf{x}_i] - \mathcal{E}[y_i|D_i = 0, \mathbf{x}_i]$ is uncorrelated with $\mathcal{V}[D_i|\mathbf{x}_i]$:

$$y_i = c + \alpha D_i + \beta \hat{G}_i + \delta D_i(\hat{G}_i - \bar{G}_i) + e_i.$$

This is, again, taking expectations on y_i conditional on $\mathbf{x}_i, G(\mathbf{x}_i)$, while assuming linear functions for $\mathcal{E}[y_0|G(\mathbf{x})]$ and $\mathcal{E}[\epsilon_{D_i}|G(\mathbf{x}_i)]$. Write

$$y = y_0 + (\mu_1 - \mu_0)D + D(\epsilon_1 - \epsilon_0).$$

Taking expectations given fixed D ,

$$\mathcal{E}[y|D, G(\mathbf{x})] = \mathcal{E}[y_0|G(\mathbf{x})] + (\mu_1 - \mu_0)D + D(\mathcal{E}[\epsilon_1|G(\mathbf{x})] - \mathcal{E}[\epsilon_0|G(\mathbf{x})]).$$

Assume linearity:

$$\mathcal{E}[y_0|G(\mathbf{x})] = \delta_0 + \delta_1 G(\mathbf{x}), \quad \mathcal{E}[\epsilon_D|G(\mathbf{x})] = \tilde{\delta}_D[G(\mathbf{x}) - \mu_G].$$

Then

$$\begin{aligned} \mathcal{E}[y_0|G(\mathbf{x})] + (\mu_1 - \mu_0)D + D(\mathcal{E}[\epsilon_1|G(\mathbf{x})] - \mathcal{E}[\epsilon_0|G(\mathbf{x})]) \\ = \delta_0 + (\mu_1 - \mu_0)D + \delta_1 G(\mathbf{x}) + \delta_2[G(\mathbf{x}) - \mu_G], \end{aligned} \quad (12)$$

where $\delta_2 = \tilde{\delta}_1 - \tilde{\delta}_0$. Thus the coefficient on D consistently estimates ATE.

- A simple (one-to-one) *propensity score matching estimator*:

$$\widehat{ATE} = \sum_{i=1}^{n_1} \frac{\hat{u}_{1i, \hat{G}_i} - \hat{u}_{0i, \hat{G}_i}}{n_1}.$$

The procedure is to estimate propensity scores and obtain predicted propensity scores for all individuals, then, for a given i in the treated, choose the individuals with the closest propensity score as \hat{G}_i to form a pair in $\hat{u}_{1i, \hat{G}_i} = y_{1i, \hat{G}_i} - \hat{\beta}' \mathbf{x}_{i, \hat{G}_i} = \hat{\alpha} + \hat{\epsilon}_{1i, \hat{G}_i}$ and $\hat{u}_{0i, \hat{G}_i} = y_{0i, \hat{G}_i} - \hat{\beta}' \mathbf{x}_{i, \hat{G}_i} = \hat{\epsilon}_{0i, \hat{G}_i}$. This follows since, under the conditional mean independence assumption,

$$\mathcal{E}[y_{D_i}|D_i, \mathbf{x}_i, G(\mathbf{x}_i)] = \mathcal{E}[y_{D_i}|\mathbf{x}_i, G(\mathbf{x}_i)] = \begin{cases} \alpha + \beta' \mathbf{x}_{i, \hat{G}_i} & \text{for } D_i = 1 \\ \beta' \mathbf{x}_{i, \hat{G}_i} & \text{for } D_i = 0 \end{cases}$$

There is an important caveat in the propensity score matching estimators that have been pointed out by Abadie and Imbens (2006). Since there will not likely to be perfect matches between the observed i and the counterfactual j over the k -dimensional vector \mathbf{x}_i and \mathbf{x}_j , we should expect a bias in matching. Under regularity conditions, the bias induced by imperfect

matches is shown to be of order $O_p(N^{-\frac{1}{k}})$, and in the case of \widehat{ATE}_1 , the bias will be of order $O_p(N_1^{-\frac{r}{k}})$ where $r \geq 1$ is the relative (*vis-a-vis* the treated) speed of increase in the number of controls used as the sample size increases.^{*11} As the standard bootstrapped covariance estimates may not be consistent (Abadie and Imbens, 2005), they provide a consistent covariance matrix estimator. It is also shown that, the smaller the number k of continuous covariates used in propensity score estimation, less the bias in estimated treatment effects. This follows because the use of greater number of covariates introduces greater biasedness, while the discrete covariates have a greater chance of having perfect matches. Also, the greater the number of matches per observation (in ATE), or, the greater the number of controls (in \widehat{ATE}_1), more efficient the estimates will be. This result poses a potentially serious problem in application. One may want to estimate propensity score as efficiently as possible, so throwing in as many covariates as possible. But Abadie and Imbens (2006)'s results tell us that it is likely that we are increasing the bias. This is similar to efficiency-bias trade off by overfitting within data that we see in forecasting.

- Jalan and Ravallion (2003) uses nearest 5 neighbours to the treated household/child i . Denoting the health of treated child i as h_{1i} , the estimated counterfactual \hat{h}_{0i} is given by:

$$\hat{h}_{0i} = \sum_{j=1}^5 W_{ij} h_{0ij}, \quad \sum_{j=1}^5 W_{ij} = 1.$$

W_{ij} is the weight obtained from other procedure. The nearest neighbor of i is defined as the observation j of the controls that minimizes the squared odds ratio difference $\left[\frac{\hat{G}(x_{1i})}{1-\hat{G}(x_{1i})} - \frac{\hat{G}(x_{0j})}{1-\hat{G}(x_{0j})} \right]^2$. The four closest to j are easily found. Matches were only accepted if the squared odds ratio difference is less than 0.001 (an absolute difference in odds less than 0.032). They used 2 levels of matching: villages and households. They used nearest 5 neighbours for the household matching and nearest single neighbour for the village matching. They discarded 62 out of 324 villages with piped water for no close match, 650 out of 9000 households with piped water. The ATE estimator is given as:

$$\widehat{ATE}(\mathbf{x}_i) = \sum_{i=1}^{n_1} \omega_i \left(h_{1i} - \sum_{j=1}^5 W_{ij} h_{0ij} \right) = \sum_{i=1}^{n_1} \omega_i (h_{1i} - \hat{h}_{0i}),$$

where ω_i is sampling weight that sums to 1 in the case they oversampled some of i (if it is pure random sampling, $\omega_i = \frac{1}{n_1}$ for all i). They also incorporated other covariates \mathbf{x}_i that may affect health. They first run a regression (using only the controls to avoid any contamination from the treatment):

$$h_{0i} = \beta'_0 \mathbf{x}_{0i} + u_{0i},$$

then estimate ATE as:

$$\widehat{ATE} = \sum_{i=1}^{n_1} \omega_i \left((h_{1i} - \hat{\beta}'_0 \mathbf{x}_{1i}) - \sum_{j=1}^5 W_{ij} (h_{0ij} - \hat{\beta}'_0 \mathbf{x}_{0ij}) \right) = \sum_{i=1}^{n_1} \omega_i (\tilde{h}_{1i} - \tilde{h}_{0i}),$$

where

$$\tilde{h}_{1i} = h_{1i} - \hat{\beta}'_0 \hat{\mathbf{x}}_{1i}, \quad \tilde{h}_{0i} = \sum_{j=1}^5 W_{ij} (h_{0ij} - \hat{\beta}'_0 \mathbf{x}_{0ij}).$$

^{*11} Precisely, it is $\frac{N^r}{N_0} \xrightarrow{a} \theta \in [0, \infty)$.

- Heckman, Ichimura and Todd (1997) use local linear kernel weights in W_{ij} that uses all observation in the controls. Another popular weight which uses all observations in the control is the kernel of some density function. Local linear matching is more efficient at the boundary points (propensity scores close to 0 or 1).

IV.4 Mixture of Methods

A simple yet promising approach is to use matching and regression. This will control for the difference in the distribution of covariates. Suppose that, using some metric or propensity score matching, one obtain a matched pair of y_{0i} and y_{1i} . Since we observe the treated and the counterfactual for it is extrapolated with the control, we get some imputed matched observation y_{0i} . Then, simple matching estimator is given by $y_{1i} - y_{0i} = \alpha + e_i$. Adding to it some difference in covariates to control for the bias, we estimate:

$$y_{1i} - y_{0i} = \alpha + \beta'(\mathbf{x}_{1i} - \mathbf{x}_{0i}) + e_i,$$

where \mathbf{x}_{0i} is matched treated covariates for \mathbf{x}_{1i} . While Imbens (2004) suggests using $\bar{\mathbf{x}}_1$, this uses matched \mathbf{x}_{1i} to gain efficiency (?). One can alternatively estimate

$$y_{0i} = \beta' \mathbf{x}_{0i} + e_i,$$

to obtain imputed mean $\hat{\mu}_0(\mathbf{x}_{1i})$ using treated group's covariates \mathbf{x}_{1i} and estimate:

$$ATE_1(\mathbf{x}) = \sum_{i=1}^{n_1} (y_{1i} - \hat{\mu}_0(\mathbf{x}_{1i})).$$

IV.5 Tests of Exogeneity

Imbens (2004) provides two tests of exogeneity (unconfoundedness). First is to test if the ineligible and the opt-outs have the same characteristics for the outcome y_i :

$$y_i \perp\!\!\!\perp 1(z_i = 1) | \mathbf{x}_i, z_i D_i = 0.$$

If the opt-outs have the different distributional features with the ineligibles, one cannot use the opt-outs as the control. Eligibility z_i can be defined by availability of the social program. In the above, conditioning on $z_i D_i = 0$ which holds for $D_i = 0$ with $z_i = 1$ and $z_i = 0$, one tests whether being eligible (opt-outs) is statistically independent with the outcome y_i . A simple test would be to regress y_i on covariates \mathbf{x}_i and eligibility z_i , and test if the coefficient on eligibility is significantly different from zero. One can use higher moments or quantiles of y_i for further examination.

The second test Imbens (2004) proposes is similar in spirit. It tests if the treatment effect can be observed in lagged outcome, i.e., outcome prior to the intervention. If the treatment

assignment is exogenous, one should not logically expect any correlation between lagged outcomes and treatment status. If the difference in treatment status significantly affects the lagged outcome, then the exogeneity assumption is likely to be violated, because outcomes tend to be serially correlated, i.e., $\text{cov}[y_{it}, y_{it-1}] \neq 0$. It tests:

$$y_{it-1} \perp\!\!\!\perp D_i | \mathbf{x}_{it}, y_{it-2}, \dots, y_{it-s}, z_i.$$

If the coefficient on D_i is not significantly different from zero, it is plausible that unconfoundedness is not violated. One can use, for example, a vector of proxy variables $y_{it-2}, \dots, y_{it-s}$ for lagged outcomes as a substitute. With a sufficient number of lags, power of the test can be reasonably high. One needs, however, exchangeability and weak stationarity in i, s so the conditional density $y_{it-1} | y_{it-2}, \dots, y_{it-s}$ does not depend on i, s . Stationarity can be tested with other data sets. Another point to be noted that, although this test gives some insights, one can use panel estimator if lagged covariates \mathbf{x}_{it-1} are available. So one can use this test to see if DID rather than methods based on exogeneity is necessary to estimate ATE.

Another test of exogeneity is given by Heckman and Vytlacil (2005) in the context of IV estimator. Note:

$$\begin{aligned} \mathcal{E}[y|G(\mathbf{x}_i) = p] &= \mathcal{E}[y_0 + D(ATE + u_1 - u_0)|G(\mathbf{x}_i) = p], \\ &= \mathcal{E}[y_0|G(\mathbf{x}_i) = p] + \mathcal{E}[D\mathcal{E}[ATE + u_1 - u_0|G(\mathbf{x}_i) = p, D = 1]|G(\mathbf{x}_i) = p], \\ &= \mathcal{E}[y_0|G(\mathbf{x}_i) = p] + p\mathcal{E}[ATE + u_1 - u_0|G(\mathbf{x}_i) = p, D = 1]. \end{aligned}$$

Thus with $p > p'$:

$$\begin{aligned} \mathcal{E}[y|G(\mathbf{x}_i) = p] - \mathcal{E}[y|G(\mathbf{x}_i) = p'] \\ = (p - p')ATE + p\mathcal{E}[u_1 - u_0|G(\mathbf{x}_i) = p, D = 1] - p'\mathcal{E}[u_1 - u_0|G(\mathbf{x}_i) = p', D = 1], \end{aligned}$$

or

$$\begin{aligned} \frac{\mathcal{E}[y|G(\mathbf{x}_i) = p] - \mathcal{E}[y|G(\mathbf{x}_i) = p']}{p - p'} \\ = ATE \\ + \frac{p\mathcal{E}[u_1 - u_0|G(\mathbf{x}_i) = p, D = 1] - p'\mathcal{E}[u_1 - u_0|G(\mathbf{x}_i) = p', D = 1]}{p - p'}. \end{aligned}$$

Hence $\mathcal{E}[y|G(\mathbf{x}_i) = p]$ is nonlinear in p if $\mathcal{E}[u_1 - u_0|G(\mathbf{x}_i) = p, D = 1]$ is not uniform over p . So one can visually inspect to see if linearity holds by plotting $\mathcal{E}[y|G(\mathbf{x}_i) = p]$ against p .

V Bound-Based Methods

V.1 Bounding the Conditional Probability

When the exogenous eligibility assignment assumption does not hold, as often does not in observational data, what should, or can, a researcher do? In the context of lacking the valid

instrumental variables, a popular way in dealing with the failure of exogeneity (exclusion, if in the IV context) assumption is to estimate as if the assumption holds, and verbally state the direction of possible bias. But is this the best we can do?

Another way to deal with the failure of exogeneity is to bound the estimate. Under the bounds analysis, Manski (1995), (1996) show that one cannot pin down the estimate to a single number, but one can nevertheless bound the possible value of estimate. For example, note:

$$\begin{aligned}\Pr[y_0|\mathbf{x}] &= \Pr[y_0|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}] + \Pr[y_0|\mathbf{x}, D = 1] \Pr[D = 1|\mathbf{x}], \\ \Pr[y_1|\mathbf{x}] &= \Pr[y_1|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}] + \Pr[y_1|\mathbf{x}, D = 1] \Pr[D = 1|\mathbf{x}].\end{aligned}$$

Let us focus on y_0 . $\Pr[y_0|\mathbf{x}, D = 1]$ is counterfactual distribution, thus cannot be observed. Nevertheless, one knows the lower and upper bounds of it, namely, 0 and 1. Thus

$$\Pr[y_0|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}] \leq \Pr[y_0|\mathbf{x}] \leq \Pr[y_0|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}] + \Pr[D = 1|\mathbf{x}].$$

Analogously, we have:

$$\Pr[y_1|\mathbf{x}, D = 1] \Pr[D = 1|\mathbf{x}] \leq \Pr[y_1|\mathbf{x}] \leq \Pr[y_1|\mathbf{x}, D = 1] \Pr[D = 1|\mathbf{x}] + \Pr[D = 0|\mathbf{x}].$$

Then, $\Pr[y_1|\mathbf{x}] - \Pr[y_0|\mathbf{x}]$ is bounded with:

$$\begin{aligned}\Pr[y_1|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}] - (\Pr[y_0|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}] + \Pr[D = 1|\mathbf{x}]) \\ \leq \Pr[y_1|\mathbf{x}] - \Pr[y_0|\mathbf{x}] \\ \leq (\Pr[y_1|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}] + \Pr[D = 0|\mathbf{x}]) - \Pr[y_0|\mathbf{x}, D = 0] \Pr[D = 0|\mathbf{x}].\end{aligned}\tag{13}$$

(13) is what Manski (1995) calls as the worst case scenario, because this is the widest bound which must be satisfied for all binary treatment policies. The width of the bound is 1. Still, this is better than the case without data where the bound can be anywhere between -1 and 1 , or in the width of 2. If the exogeneity assumption is suspicious and if we do not have instruments, then we must base our analysis on (13) for credibility.

V.2 The Mixing Problem

An innovation of Manski (1995) is that he considers, under a binary treatment, the distribution of outcomes y_m under an *arbitrary policy* m , given the knowledge of $\Pr[y_0|\mathbf{x}]$ and $\Pr[y_1|\mathbf{x}]$. This unspecified policy assigns individuals to treatment with an unspecified assignment rule D_m . Thus:

$$y_m \equiv D_m y_1 + (1 - D_m) y_0.$$

The *mixing problem* Manski considers is defined as: what can we know about $\Pr[y_m|\mathbf{x}]$ with the knowledge of $\Pr[y_0|\mathbf{x}]$ and $\Pr[y_1|\mathbf{x}]$?

It turns out that one can bound the probability of an arbitrary policy y_m that falls into some outcome set B , or $\Pr[y_m \in B|\mathbf{x}]$. Since y_m is a convex combination of y_1 and y_0 , we have

$(y_1 \in B) \cap (y_0 \in B) \implies y_m \in B$, and $(y_1 \notin B) \cap (y_0 \notin B) \implies y_m \notin B$. These are the trivial cases that give $\Pr[y_m \in B|\mathbf{x}] = 1$ and $\Pr[y_m \in B|\mathbf{x}] = 0$, respectively, and we do not have to worry about them.

Instead, we consider the two polar cases. A treatment policy minimizes $\Pr[y_m \in B|\mathbf{x}]$ if the assignment rule D_m follows:

$$\begin{aligned} (y_1 \notin B) \cap (y_0 \in B) &\implies D_m = 1, \\ (y_1 \in B) \cap (y_0 \notin B) &\implies D_m = 0. \end{aligned} \tag{14}$$

Then the smallest possible value of $\Pr[y_m \in B|\mathbf{x}]$ when it is minimized by policy is $\Pr[(y_1 \in B) \cap (y_0 \in B)|\mathbf{x}]$. Another treatment policy maximizes $\Pr[y_m \in B|\mathbf{x}]$ if the assignment rule D_m follows:

$$\begin{aligned} (y_1 \notin B) \cap (y_0 \in B) &\implies D_m = 0, \\ (y_1 \in B) \cap (y_0 \notin B) &\implies D_m = 1. \end{aligned} \tag{15}$$

Then the largest possible value of $\Pr[y_m \in B|\mathbf{x}]$ when it is maximized by policy is $\Pr[(y_1 \in B) \cup (y_0 \in B)|\mathbf{x}]$. So:

$$\Pr[(y_1 \in B) \cap (y_0 \in B)|\mathbf{x}] \leq \Pr[y_m \in B|\mathbf{x}] \leq \Pr[(y_1 \in B) \cup (y_0 \in B)|\mathbf{x}].$$

Unfortunately, we do not know these bounding values, so the lower bound must be substituted with the smallest possible value that is consistent with the marginals $\Pr[y_0|\mathbf{x}]$ and $\Pr[y_1|\mathbf{x}]$, and so does the upper bound which has to be replaced with the largest possible value consistent with $\Pr[y_0|\mathbf{x}]$ and $\Pr[y_1|\mathbf{x}]$.

The lower bound is given by rearranging

$$\Pr[(y_1 \in B) \cup (y_0 \in B)|\mathbf{x}] = \Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}] - \Pr[(y_1 \in B) \cap (y_0 \in B)|\mathbf{x}] \leq 1,$$

or

$$\Pr[(y_1 \in B) \cap (y_0 \in B)|\mathbf{x}] \geq \max \{ \Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}] - 1, 0 \},$$

where the maximum operator is necessary because $\Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}]$ can be less than 1. The largest possible value for the upper bound is given when there is no overlap between y_1 and y_0 over B , so $\Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}]$. Then:

$$\begin{aligned} \max \{ \Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}] - 1, 0 \} \\ \leq \Pr[y_m \in B|\mathbf{x}] \\ \leq \Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}]. \end{aligned} \tag{16}$$

This is the worst case bound in the mixing problem. It can be seen that if $\Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}] < 1$, the width of bound is smaller than 1. When we compare it with the width of 1 in (13), we see that knowledge of the marginals provides the possibility of narrowing the width. The improvement will be greater for the set B such that $\Pr[y_1 \in B|\mathbf{x}] + \Pr[y_0 \in B|\mathbf{x}]$ is smaller.

CONTINGENCY TABLE OF HECKMAN AND SMITH (1995)
UNTREATED

		UNTREATED		
		$y_0 = 1$	$y_0 = 0$	
TREATED	$y_1 = 1$	$\Pr[y_1 = 1, y_0 = 1]$	$\Pr[y_1 = 1, y_0 = 0]$	$\Pr[y_1 = 1]$
	$y_1 = 0$	$\Pr[y_1 = 0, y_0 = 1]$	$\Pr[y_1 = 0, y_0 = 0]$	$\Pr[y_1 = 0]$
		$\Pr[y_0 = 1]$	$\Pr[y_0 = 0]$	

A binary outcome example of Manski (1997) gives $\mathcal{E}[y_1] = \Pr[y_1 = 1] = 0.67$ and $\mathcal{E}[y_0] = \Pr[y_0 = 1] = 0.49$. An outcome is graduation, denoted with $y = 1$ if graduated and $y = 0$ otherwise. y_1 indicates the outcome under treatment and y_0 the outcome under no treatment. Then, what is the contribution of program on y ? Or what are the joint probabilities $\Pr[y_0, y_1]$? There is an infinite number of combinations that are consistent with $\mathcal{E}[y_1] = \Pr[y_1 = 1] = 0.67$ and $\mathcal{E}[y_0] = \Pr[y_0 = 1] = 0.49$. There are several possibilities:

- Widest bound (Hoeffding-Frechet bounds). If treatment is assigned to attain the highest feasible graduation rate that it is assigned only to individuals with $y_1 = 1, y_0 = 0$, and no treatment is given to individuals with $y_1 = 0, y_0 = 1$ or $y_1 = 0, y_0 = 0$, then graduation probability is $1 - \Pr[y_1 = 0, y_0 = 0]$. Conversely, if the program is to achieve the lowest feasible graduation rate that it gives treatment to individuals with $y_1 = 0, y_0 = 1$, and no treatment is given to individuals with $y_1 = 1, y_0 = 0$, then graduation probability is $\Pr[y_1 = 1, y_0 = 1]$. Then we want the bounds that minimize both $\Pr[y_1 = 0, y_0 = 0]$ and $\Pr[y_1 = 1, y_0 = 1]$ to get the widest bound. This follows because the upperbound is given by $1 - \Pr[y_1 = 0, y_0 = 0]$ and the lower bound is given by $\Pr[y_1 = 1, y_0 = 1]$, so minimizing $\Pr[y_1 = 0, y_0 = 0]$ and $\Pr[y_1 = 1, y_0 = 1]$ gives the most conservative (or widest) bounds on graduation probability $\Pr[y = 1]$. Thus

$$\begin{aligned} \Pr[y_1 = 0, y_0 = 0] &= 0 & \Pr[y_1 = 0, y_0 = 1] &= .33 \\ \Pr[y_1 = 1, y_0 = 0] &= .51 & \Pr[y_1 = 1, y_0 = 1] &= .16 \end{aligned}$$

are the joint probability distribution. So the highest graduation rate consistent with experimental evidence is 1 and the lowest is .16, or [.16, 1].

- If treatment does not harm graduation, or $\Pr[y_1 = 0, y_0 = 1] = 0$, then

$$\Pr[y_1 = 0] = \Pr[y_1 = 0, y_0 = 0] + \Pr[y_1 = 0, y_0 = 1] = \Pr[y_1 = 0, y_0 = 0] = .33,$$

and

$$\Pr[y_0 = 1] = \Pr[y_1 = 0, y_0 = 1] + \Pr[y_1 = 1, y_0 = 1] = .49.$$

Then $\Pr[y_1 = 1]$ is .67 so the interval is [.49, .67].

- If y_D are independent,

$$\Pr[y_1 = 0, y_0 = 0] = \Pr[y_1 = 0] \Pr[y_0 = 0] = .17,$$

$$\Pr[y_1 = 1, y_0 = 1] = \Pr[y_1 = 1] \Pr[y_0 = 1] = .33.$$

The bounds on $\Pr[y = 1]$ is [.33, .83].

The widest bound seems too large, however, this is the bound that is consistent with the widest credibility. The widest bound can be narrowed if we add more assumptions. □

As seen in the example above, one can consider a variety of restrictions to narrow the worst case bounds in (16). An interesting restriction among them is the case when we know the control proportion $\Pr[D_m = 0|\mathbf{x}] = p$. Then:

$$\Pr[y_D] = \Pr[y_D|D_m = 1](1 - p) + \Pr[y_D|D_m = 0]p.$$

This may not be realistic because it is almost impossible to predict p . But it will be helpful in the policy debate if we can see what bound will be obtained for different values of p .

Letting Ψ be the set of all possible distributions on \mathbb{Y} , and ψ be its element, then:

$$\Pr[y_1|D_m = 1] \in \Psi_1(p) \equiv \Psi \cup \left\{ \frac{\Pr[y_1] - p\psi}{1-p} : \psi \in \Psi \right\}, \quad (17)$$

and

$$\Pr[y_0|D_m = 0] \in \Psi_0(p) \equiv \Psi \cup \left\{ \frac{\Pr[y_0] - (1-p)\psi}{p} : \psi \in \Psi \right\}. \quad (18)$$

Given that $\psi \in [0, 1]$, we immediately have the bound on each probability:

$$\max \left\{ 0, \frac{\Pr[y_1 \in B] - p}{1-p} \right\} \leq \Pr[y_1 \in B|D_m = 1] \leq \min \left\{ 1, \frac{\Pr[y_1 \in B]}{1-p} \right\}, \quad (19)$$

and

$$\max \left\{ 0, \frac{\Pr[y_0 \in B] - (1-p)}{p} \right\} \leq \Pr[y_0 \in B|D_m = 1] \leq \min \left\{ 1, \frac{\Pr[y_0 \in B]}{p} \right\}. \quad (20)$$

Since distribution of y_m is $(1-p, p)$ mixture of $\Pr[y_1|D_m = 1]$, $\Pr[y_0|D_m = 0]$, we have:

$$\Pr[y_m] \in \{(1-p)\psi_1 + p\psi_0 : (\psi_1, \psi_0) \in \Psi_1(p) \times \Psi_0(p)\}. \quad (21)$$

Thus, combining (19), (20), (21), we have:

$$\begin{aligned} \max \{0, \Pr[y_1 \in B] - p\} + \max \{0, \Pr[y_0 \in B] - (1-p)\} \\ \leq \Pr[y_m \in B] \\ \leq \min \{1-p, \Pr[y_1 \in B]\} + \min \{p, \Pr[y_0 \in B]\}. \end{aligned} \quad (22)$$

The maximum of lower bound is $\Pr[y_1 \in B] + \Pr[y_0 \in B] - 1$, but can take the intermediate values of $\Pr[y_1 \in B] - p$ or $\Pr[y_0 \in B] - (1-p)$. The minimum of the upper bound is lowered if $1-p > \Pr[y_1 \in B]$ or $p > \Pr[y_0 \in B]$. Thus the knowledge of p provides a possibility of narrowing the bound in (16).

It should be noted that, despite its usefulness, we may not be able to use (16) nor (22) because experiments may only offer an ITT estimator, not the marginals. Then one must base the analysis on (13) which has the regrettably large width of 1. Another thing to note is that, despite being critical on randomized experiments, the mixing problem needs the marginals hence the perfectly conducted randomized experiments. So it should not be understood that the use of (16) or (22) can serve as a substitute to a randomized experiment, but rather it shows another way of utilizing the experimental evidence.

VI Instrumental Variables Based Methods

The instrumental variable estimator, or the *local average treatment effect (LATE)*, is an ATE for the population whose participation eligibility is changed from 0 to 1 (and among them, from a nonparticipant status to a participant status). The key assumption is that the change in

status is induced only by the change in eligibility from 0 to 1, not by any other variables, and such a change is orthogonal to (or independent of) any factors that affect outcomes.^{*12}

Assumption: eligibility z_i of participation is randomly assigned.

$$(y_{1i}, y_{0i}, D_{1i}, D_{0i}) \perp\!\!\!\perp z_i,$$

or equivalently,

$$y_{Di}, D_i(z_i) \perp\!\!\!\perp z_i.$$

This means that eligibility is assigned independent of the possible outcomes y_{1i}, y_{0i} , nor the likely responses D_{1i}, D_{0i} of individuals to the assigned value z_i . Conditional mean independence of y_{Di} and D_i given \mathbf{x}_i is not necessary. Under this setting, an authority randomly assigns each individuals the eligibility z to participate in the program, and the individuals who are eligible $z = 1$ decide on participation D_{1i} . D_{1i} is participation status if the person is eligible $z_i = 1$, D_{0i} is the participation status of the ineligible person $z_i = 0$. We naturally expect $D_{0i} = 0$ and $D_{1i} = 1$, nevertheless, there can be some *targeting errors* that:

- Type 1 error (exclusion error): eligible individuals do not participate, $D_{1i} = 0$.
- Type 2 error (leakage error^{*13}): ineligible individuals participate, $D_{0i} = 1$.

In the case of type 1 error, one may not want to call it an error given that the individuals voluntarily opt out. But in the context of poverty reduction, an eligible individual is a low income individual, and the scheme which prompts them to opt out can be considered as having targeting errors. To summarize, there can be four cases as in **Table 1**.

VI.1 Instrumental Variable Estimator under Homogeneous Treatment Effects

The most popular IV-based estimator is the Wald estimator:

$$\widehat{LATE} = \frac{\bar{y}_{z_i=1} - \bar{y}_{z_i=0}}{\bar{D}_1 - \bar{D}_0}.$$

In the example of **Table 1**, $LATE = \frac{1600-1300}{.8-.2} = 500$.

Noting

$$y_i = \{1 - D(z_i)\}y_{0i} + D(z_i)y_{1i},$$

^{*12} The former is called relevancy and the latter is called validity of instrument z for D . See below.

^{*13} Should be better termed as an inclusion error.

Table 1: Eligibility and Actual Participation Status

Classification of Treatment and Eligibility Statuses

		TREATMENT		D_z
		$D = 0$	$D = 1$	
ELIGIBILITY	$z = 0$	not targeted	leakage (type 2 error)	D_{0i}
	$z = 1$	exclusion (type 1 error)	targeted and complied	D_{1i}
y_D		y_0	y_1	

A Numerical Example

		TREATMENT		\bar{D}_z	\bar{y}_z
		$D = 0$	$D = 1$		
ELIGIBILITY	$z = 0$	80	20	0.20	1300
	$z = 1$	50	200	0.80	1600
\bar{y}_D		1200	1800		

we have

$$\begin{aligned}
 & \mathcal{E}[y_i|z_i = z] - \mathcal{E}[y_i|z_i = z'] \\
 &= \mathcal{E}[\{1 - D(z)\}y_{0i} + D(z)y_{1i}|z_i = z] - \mathcal{E}[\{1 - D(z')\}y_{0i} + D(z')y_{1i}|z_i = z'], \\
 &= \mathcal{E}[\{1 - D(z)\}y_{0i} + D(z)y_{1i}] - \mathcal{E}[\{1 - D(z')\}y_{0i} + D(z')y_{1i}], \\
 &= \mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})],
 \end{aligned} \tag{23}$$

where the second to last line follows from the assumption that $y_{Di}, D_i(z_i) \perp\!\!\!\perp z_i$.^{*14} Thence

$$\begin{aligned}
 \mathcal{E}[y_i|z_i = z] - \mathcal{E}[y_i|z_i = z'] &= \mathcal{E}[y_{1i} - y_{0i}|D(z) - D(z') = 1] \Pr[D(z) - D(z') = 1] \\
 &\quad + \mathcal{E}[y_{1i} - y_{0i}|D(z) - D(z') = -1] \Pr[D(z) - D(z') = -1].
 \end{aligned}$$

^{*14} Wooldridge's derivation uses:

$$D_i = 1(z_i = z)D(z) + \{1 - 1(z_i = z)\}D(z') = D(z') + 1(z_i = z)\{D(z) - D(z')\}.$$

Plugging in $y_i = y_{0i} + D_i(y_{1i} - y_{0i})$, we have:

$$\begin{aligned}
 y_i &= y_{0i} + [D(z') + 1(z_i = z)\{D(z) - D(z')\}](y_{1i} - y_{0i}), \\
 &= y_{0i} + D(z')(y_{1i} - y_{0i}) + 1(z_i = z)\{D(z) - D(z')\}(y_{1i} - y_{0i}).
 \end{aligned}$$

Taking expectations conditional on $z_i = z$ (meaning, with $z_i = z$ imposed), we have:

$$\begin{aligned}
 \mathcal{E}[y_i|z_i = z] &= \mathcal{E}[y_{0i}|z_i = z] + \mathcal{E}[D(z')(y_{1i} - y_{0i})|z_i = z] + \mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})|z_i = z], \\
 &= \mathcal{E}[y_{0i}] + \mathcal{E}[D(z')(y_{1i} - y_{0i})] + \mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})],
 \end{aligned}$$

where the last equality follows because $D(z_i), y_{Di} \perp\!\!\!\perp z_i$. Taking expectations conditional on $z_i = z'$, we have:

$$\begin{aligned}
 \mathcal{E}[y_i|z_i = z'] &= \mathcal{E}[y_{0i}|z_i = z'] + \mathcal{E}[D(z')(y_{1i} - y_{0i})|z_i = z'], \\
 &= \mathcal{E}[y_{0i}] + \mathcal{E}[D(z')(y_{1i} - y_{0i})].
 \end{aligned}$$

Then we will get (23):

$$\mathcal{E}[y_i|z_i = z] - \mathcal{E}[y_i|z_i = z'] = \mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})].$$

Under monotonicity, $D(z) - D(z') \geq 0$ then $D(z) - D(z') = 1$ if $D(z) = 1, D(z') = 0$. So:

$$\begin{aligned} \Pr[D(z) - D(z') = 1] &= \Pr[D(z) = 1, D(z') = 0] = \Pr[D = 1|z_i = z] + \Pr[D = 0|z_i = z'] \\ &= \Pr[D = 1|z_i = z] - \Pr[D = 1|z_i = z']. \end{aligned}$$

Thus we have:

$$\mathcal{E}[y_{1i} - y_{0i}|D(z) - D(z') = 1] = \frac{\mathcal{E}[y_i|z_i = z] - \mathcal{E}[y_i|z_i = z']}{\Pr[D = 1|z_i = z] - \Pr[D = 1|z_i = z']} \quad (25)$$

(25) shows that LHS depends on the particular values of z_i , therefore nor does the RHS. Heckman (1997) argues that, given that it depends on the particular values of z_i and is ATE of unknown subpopulation, LATE does not give an answer to an interesting policy question. Although his exposition is correct, one can always check the characteristics of subpopulation whose choices are affected by the value of z_i , and it gives the net effect on the economy of certain intervention. This is still an interesting question being answered.

Then:

$$\begin{aligned} \mathcal{E}[y_i|z_i] &= \mathcal{E}[y_{0i}|z_i] + \mathcal{E}[D(z_i)(y_{1i} - y_{0i})|z_i] + z_i \mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})|z_i], \\ &= \mathcal{E}[y_{0i}] + \mathcal{E}[D(z_i)(y_{1i} - y_{0i})] + z_i \mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})], \end{aligned}$$

because $y_{Di}, D(z_i) \perp\!\!\!\perp z_i$. Then, taking differences:

$$\mathcal{E}[y_i|z_i = z] - \mathcal{E}[y_i|z_i = z'] = (z - z') \mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})]. \quad (26)$$

It will be clear that LATE derived in this fashion relies on particular values of z_i . Expanding the expectations,

$$\mathcal{E}[\{D(z) - D(z')\}(y_{1i} - y_{0i})] = \mathcal{E}[y_{1i} - y_{0i}|D(z) - D(z') = 1] \Pr[D(z) - D(z') = 1],$$

as we assume monotonicity. Noting that $\Pr[D(z) - D(z') = 1] = \Pr[D = 1|z_i = z] - \Pr[D = 1|z_i = z']$, we have:

$$\mathcal{E}[y_{1i} - y_{0i}|D_i(z) - D_i(z') = 1] = \frac{1}{z - z'} \frac{\mathcal{E}[y_i|z_i = z] - \mathcal{E}[y_i|z_i = z']}{\Pr[D = 1|z_i = z] - \Pr[D = 1|z_i = z']}, \quad (27)$$

(27) is sometimes called the *Wald estimator*. It is the same as (25) with $z - z' = 1$. Note that the Wald estimator only exploits the variation induced by changes in z , $D|z$, that are random by assumption, which ensures zero correlation between $D|z$ with unobservables. So it is ATE of unknown subpopulation whose choice is changed by variation in IV. It is also noteworthy that LATE is a function of particular z_i value by construction, even after dividing it with $z - z'$. This is in contrast with the IV estimator which is not a function of z_i as we will examine below.

Consider an instrumental variables estimator on D_i :

$$y_i = c + \alpha D_i + e_i,$$

An instrumental variable estimator using eligibility z_i as an instrument for participation gives:

$$\hat{\alpha}_{IV} = (\mathbf{d}'\mathbf{z})^{-1}\mathbf{y}'\mathbf{z} = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (D_i - \bar{D})(z_i - \bar{z})},$$

where $\mathbf{d} = (D_{1i}, \dots, D_{ni})'$, $\mathbf{z} = (z_1, \dots, z_n)'$, $\mathbf{y} = (y_1, \dots, y_n)'$. In the population moment terms,

$$\alpha_{IV} = \frac{\text{cov}[y, z]}{\text{cov}[D, z]} = \alpha \frac{\text{cov}[D, z]}{\text{cov}[D, z]} + \frac{\text{cov}[e, z]}{\text{cov}[D, z]} = \alpha,$$

if

$$\text{cov}[e, z] = 0,$$

which is called the *validity* requirement of an instrumental variable, while another requirement

$$\text{cov}[D, z] \neq 0,$$

is called the *relevancy* requirement.

Although the IV estimator is attractive due to its ability to give the unbiased ATE estimate under weak assumptions, there are at least two drawbacks. First is that it measures the average treatment effects of undefined subpopulation. It is ATE of subpopulation, or marginal population, whose choices were changed due to changes in eligibility. If it is the school meals program, it is the marginal households in the sense that they have undernourished children who will change their schooling choices due to school meals. It is not likely to be the ATE, or the treatment effect averaged over the entire population. However, this may not be a weakness if we want to know the effects on the subpopulation who have undernourished children in their home, or the subpopulation who are in need of public assistance to have their children stay in schools. In addition, they can always compare the characteristics of marginal population with others to see on which population the policy is working. The second drawback is that we can rarely find the valid IVs. So most often we opt to randomize the eligibility,^{*15} which then is subjected to the same operational criticisms for randomized trials. One may wish to compare randomized estimator and IV estimator and see how the estimates change. These should become closer to each other if the share of so-called ‘unknown subpopulation’ becomes larger. Oreopoulos (2006) shows the UK and Northern Irish case study on the effects of compulsory schooling law where a significant portion of population was affected by such legislation. The estimated marginal returns to schooling is similar to the US estimates of Angrist and Krueger (1990) where only a small fraction of population was affected.

^{*15} Most researchers use randomized eligibility for instruments. See Bitler, Gelbach, and Hoynes (2006) who use data of Connecticut’s Jobs First program, and Abadie, Angrist, and Imbens (2002) who use JTPA data.

- Angrist (1990) uses the Vietnam War draft lottery numbers as randomly assigned eligibility for military service, which assigns the service to individuals with low lottery numbers, and estimates its effect on subsequent incomes. They found the veteran status to affect negatively their incomes.
- Angrist and Krueger (1990) use birth date as eligibility: in some US states, one cannot drop out if he/she is below 16 in August 30. So students born in September 1st and August 30 has a difference of 1 year of compulsory education, whose assignment should be random given the birth dates are random. They divided birth dates into 4 quarters, and compared with the first (beginning from Sep 1) and the last three quarters using Wald estimates. They found LATE to be significantly positive.
- Angrist et al. (2002) use randomized voucher assignment as instruments for using the vouchers, as only 90% of households used them. They found significant increase in grade attainment, test scores, likelihood of finishing 8th grade, and a reduction in grade repetitions.
- Banerjee et al. (2005) use Wald estimator for the effects of remedial education on test scores. Randomization at the school level for getting remedial education teachers. They find significant improvement on scores, especially for underachievers, and significant cost-effectiveness over other interventions.

VI.2 Instrumental Variable Estimator under Essential Heterogeneity

Heckman (1997), Heckman and Vytlacil (2005), (2006), and other Heckman papers show an important result that LATE and IV estimators are valid only when the treatment effect is the same for all individuals, or when the individuals do not take into account the treatment effects when participating. This is seen by introducing the *essential heterogeneity* that for different individuals, treatment effect α_i is generally different:

$$y_i = c + \mu_i(\mathbf{x}_i) + \alpha_i D_i + e_i, \quad (28)$$

We assume a separable (between mean and disturbance) model:

$$y_{0i} = \mu_0(\mathbf{x}_i) + u_{0i}, \quad y_{1i} = \mu_1(\mathbf{x}_i) + u_{1i},$$

so

$$y_{1i} - y_{0i} = \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i) + u_{1i} - u_{0i} = \alpha + u_{1i} - u_{0i},$$

where $\alpha = \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)$ is ATE. Plugging this into

$$y_i = y_{0i} + (y_{1i} - y_{0i})D_i,$$

gives (28):

$$y_i = \mu_0(\mathbf{x}_i) + (\alpha + u_{1i} - u_{0i})D_i + u_{0i} = \mu_0(\mathbf{x}_i) + \alpha D_i + u_{0i} + (u_{1i} - u_{0i})D_i.$$

where we denoted $\mu_0(\mathbf{x}_i) = c + \mu(\tilde{\mathbf{x}}_i)$ (with $\tilde{\mathbf{x}}_i$ does not include intercept) and $u_{0i} = e_i$. So the coefficient α_i on D_i is:

$$\alpha_i = \alpha + u_{1i} - u_{0i}.$$

As can be seen, the treatment effect model under essential heterogeneity can be cast in a random coefficient framework. It can also be cast in the fixed-effect framework, if the perceived

individual gain from participation $u_{1i} - u_{0i}$ is a function of the unobserved individual fixed effect c_i .

In the population moment terms, IV estimator of ATE gives:

$$\begin{aligned}\alpha_{IV} &= \frac{\text{cov}[y_i, z_i]}{\text{cov}[D_i, z_i]} = \frac{\text{cov}[\alpha D_i, z_i]}{\text{cov}[D_i, z_i]} + \frac{\text{cov}[u_{0i} + (u_{1i} - u_{0i})D_i, z_i]}{\text{cov}[D_i, z_i]}, \\ &= \alpha + \frac{\text{cov}[(u_{1i} - u_{0i})D_i, z_i]}{\text{cov}[D_i, z_i]}.\end{aligned}$$

So consistency of IV estimator rests on $\frac{\text{cov}[(u_{1i} - u_{0i})D_i, z_i]}{\text{cov}[D_i, z_i]} = 0$, or that $(u_{1i} - u_{0i})D_i$ is not a function of z_i . Denoting the individual gain as $\Delta u_i \equiv u_{1i} - u_{0i}$, the denominator can be rewritten as:

$$\begin{aligned}\text{cov}[\Delta u_i D_i, z_i] &= \mathcal{E}[\Delta u_i D_i z_i] - \mathcal{E}[\Delta u_i D_i] \mathcal{E}[z_i], \\ &= \mathcal{E}_z [\mathcal{E}[\Delta u_i D_i | z_i] z_i] - \mathcal{E}[\Delta u_i D_i] \mathcal{E}[z_i], \\ &= \mathcal{E}_z [\mathcal{E}[\Delta u_i D_i | z_i] z_i] - \mathcal{E}_z [\mathcal{E}[\Delta u_i D_i | z_i]] \mathcal{E}[z_i].\end{aligned}$$

This can be zero if $\mathcal{E}[\Delta u_i D_i | z_i] = \mathcal{E}[\Delta u_i D_i]$ by the second line, or if $\mathcal{E}[\Delta u_i D_i | z_i] = 0$ by the third line. Since $\mathcal{E}[\Delta u_i D_i | z_i] = \mathcal{E}[\Delta u_i D_i(z_i) | z_i]$, we should more likely to see stronger correlation between individual benefit $\Delta u_i = u_{1i} - u_{0i}$ and participation under $z_i = z$, which is consistent with our assumption of monotonicity that $D_i(z) \geq D_i(z')$. So the first condition should not hold. The second condition also does not hold if $\mathcal{E}[\Delta u_i | D_i, z_i] \neq 0$ because:

$$\mathcal{E}[\Delta u_i D_i | z_i] = \mathcal{E}_{D|z} [\mathcal{E}_{\Delta u D, z} [\Delta u_i | D_i, z_i] D_i | z_i].$$

Heckman (1997) argues that it is unlikely that $\mathcal{E}[\Delta u_i | D_i, z_i] = 0$, given individuals make participation decisions based on individual gains. Condition that $\mathcal{E}[\Delta u_i | D_i, z_i] \neq 0$ is highly plausible, and under this, we should have $\text{cov}[(u_{1i} - u_{0i})D_i, z_i] \neq 0$. This covariance will be zero if $\Delta u_i \perp\!\!\!\perp D_i | z_i$, or weaker $\mathcal{E}[\Delta u_i | D_i, z_i] = \mathcal{E}[\Delta u_i | z_i]$ suffices. The latter is analogous to the ‘ignorability of treatment’ and ‘selection on observables’ only that z_i plays the role of covariates \mathbf{x}_i in exogenous treatment assignment case.

So it is crucial that $\mathcal{E}[\Delta u_i | D_i, z_i] = \mathcal{E}[\Delta u_i | z_i]$ for an IV estimator to be consistent under essential heterogeneity. Yet another way of describing how this condition means, or to understand the consistency requirement of IV estimator, is to note that $y_{D_i} = \mu_D(\mathbf{x}_i) + u_{D_i}$ where $\mu_D(\mathbf{x}_i)$ is a function only of \mathbf{x}_i :

$$\mathcal{E}[y_{1i} - y_{0i} | \mathbf{x}_i, z_i, D_i = 1], = \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i) + \mathcal{E}[u_{1i} - u_{0i} | \mathbf{x}_i, z_i, D_i = 1].$$

As argued, Heckman (1997) points that the third term would not be zero. He shows that the conditional density of Δu_i given \mathbf{x}_i, z_i , and D_i , $f(\Delta u_i | \mathbf{x}_i, z_i, D_i = 1)$, which we use in computing $\mathcal{E}[\Delta u_i | \mathbf{x}_i, z_i, D_i = 1]$, will be dependent on D_i . Using D^* as a latent index variable for program

participation such that $D = 1$ iff $D^* \geq 0$, and suppressing \mathbf{x}_i for notational simplicity, we have:

$$\begin{aligned}
\Pr[D_i = 1|z_i, \Delta u_i]f(\Delta u_i|z_i)f(z_i) &= \int 1(D^* \geq 0)f(D^*|z_i, \Delta u_i)f(\Delta u_i|z_i)f(z_i)dD^*, \\
&= \int 1(D^* \geq 0)f(D^*, z_i, \Delta u_i)dD^*, \\
&= \int 1(D^* \geq 0)f(\Delta u_i|D^*, z_i)f(D^*|z_i)f(z_i)dD^*, \\
&= \int_0^\infty f(\Delta u_i|D = 1, z_i)f(z_i)f(D^*|z_i)dD^*, \\
&= f(\Delta u_i|D = 1, z_i)f(z_i) \int_0^\infty f(D^*|z_i)dD^*, \\
&= f(\Delta u_i|D_i = 1, z_i)f(z_i) \Pr[D_i = 1|z_i].
\end{aligned}$$

Third to the last equality follows because $D = 1$ for $D^* \geq 0$, so we can condition on $D = 1$. By the Bayes' rule,

$$f(\Delta u_i|z_i, D_i = 1) = \frac{\Pr[D_i = 1|z_i, \Delta u_i]f(\Delta u_i|z_i)}{\Pr[D_i = 1|z_i]}. \quad (29)$$

Thus for $f(\Delta u_i|z_i, D_i = 1) = f(\Delta u_i|z_i)$, we need:

$$\Pr[D_i = 1|z_i, \Delta u_i] = \Pr[D_i = 1|z_i].$$

Only under this case, an IV estimator gives a consistent ATE estimate. This is unlikely to hold since individuals make participation decisions based on the individual gains Δu_i .^{*16} Heckman (1997, 449) points out that zero correlation between Δu_i and D_i is 'a behavioral assumption', and thus 'cannot be settled by a statistical analysis.'

Heckman and Vytlacil (2005) show that the LATE estimator is a weighted average of what they call the *marginal treatment effect*, which is a treatment effect conditioned on $G = p$, which is given by:

$$MTE \stackrel{\text{def}}{=} \mathcal{E}[y_1 - y_0|u_D = p] = \frac{\partial \mathcal{E}[y|G(D=1|z)=p]}{\partial p}.$$

Note that this is a differential version (limit case by differentiating from the right) of Wald estimator evaluated at p . This is derived as follows. Assume a latent variable D^* that determines the participation:

$$D^* = \mu_D(\mathbf{z}) - v \begin{cases} \geq \\ < \end{cases} 0 \iff D = \begin{cases} 1, \\ 0. \end{cases}$$

^{*16} IV validity does not hold generally even $f(\Delta u_i|z_i) = f(\Delta u_i)$, unless another condition holds: $\Pr[D_i = 1|z_i, \Delta u_i] = \Pr[D_i = 1|z_i]$ or D_i is independent of Δu_i conditional on z_i or $D_i \perp \Delta u_i|z_i$. (29) shows that even if the instrument validity holds for the marginal density $f(\Delta u_i|z_i) = f(\Delta u_i)$, the density of Δu_i conditional on D_i and z_i , $f(\Delta u_i|z_i, D_i = 1)$, still generally is a function of, hence correlated with, D_i :

$$\frac{\Pr[D_i = 1|z_i, \Delta u_i]f(\Delta u_i)}{\Pr[D_i = 1|z_i]} = f(\Delta u_i|z_i, D_i = 1).$$

We have assumed a separability between regressors and disturbance term in the above. The above inequality can be rewritten by using the distribution function F_V , a monotonic transformation, as:

$$\mu_D(\mathbf{Z}) \geq V \iff F_V[\mu_D(\mathbf{Z})] \geq F_V(V) \iff G(\mathbf{Z}) \geq U_D,$$

where we denoted $G(\mathbf{Z}) = F_V[\mu_D(\mathbf{Z})]$ and $U_D = F_V(V)$. Note by construction $U_D \stackrel{d}{\sim} \mathbb{U}[0, 1]$.

$$\begin{aligned} \mathcal{E}[y|\mathbf{Z} = \mathbf{z}] &= \mathcal{E}[y|G(\mathbf{Z}) = p], \\ &= \mathcal{E}[Dy_1 + (1 - D)y_0|G(\mathbf{Z}) = p], \\ &= \mathcal{E}[y_0] + \mathcal{E}[D(y_1 - y_0)|G(\mathbf{Z}) = p], \\ &= \mathcal{E}[y_0] + p\mathcal{E}[y_1 - y_0|D = 1], \\ &= \mathcal{E}[y_0] + \int_0^p \mathcal{E}[y_1 - y_0|U_D = u_D]du_D. \end{aligned}$$

So

$$\frac{\partial \mathcal{E}[y|G(D=1|z)=p]}{\partial p} = \mathcal{E}[y_1 - y_0|U_D = p].$$

They show that LATE estimator is a weighted average of MTE:

$$\alpha_{IV} = \int_0^1 \omega(x, u_D) MTE(x, u_D) du_D,$$

with

$$\begin{aligned} \omega(x, u_D) &= \frac{\mathcal{E}[J(\mathbf{Z}) - \mathcal{E}[J(\mathbf{Z})|\mathbf{X} = \mathbf{x}, G(\mathbf{Z}) \geq u_D] \Pr[G(\mathbf{Z}) \geq u_D | \mathbf{X} = \mathbf{x}]]}{\text{cov}[J(\mathbf{Z}), G(\mathbf{Z}) | \mathbf{X} = \mathbf{x}]}, \\ &= \frac{\int (j - \mathcal{E}[J(\mathbf{Z})|\mathbf{X} = \mathbf{x}]) \int_{u_D}^1 f_{G,J}(g, j | \mathbf{X} = \mathbf{x}) dg dj}{\text{cov}[J(\mathbf{Z}), G(\mathbf{Z}) | \mathbf{X} = \mathbf{x}]}, \\ &= \frac{\int \left[\int_{u_D}^1 f_{G|J}(g|J(\mathbf{Z}) = j, \mathbf{X} = \mathbf{x}) dg \right] (j - \mathcal{E}[J(\mathbf{Z})|\mathbf{X} = \mathbf{x}]) f_J(j|\mathbf{X} = \mathbf{x}) dj}{\text{cov}[J(\mathbf{Z}), G(\mathbf{Z}) | \mathbf{X} = \mathbf{x}]}, \\ &= \frac{\int \Pr[G(\mathbf{Z}) \geq u_D | J(\mathbf{Z}) = j, \mathbf{X} = \mathbf{x}] (j - \mathcal{E}[J(\mathbf{Z})|\mathbf{X} = \mathbf{x}]) f_J(j|\mathbf{X} = \mathbf{x}) dj}{\text{cov}[J(\mathbf{Z}), G(\mathbf{Z}) | \mathbf{X} = \mathbf{x}]}, \end{aligned}$$

They also show that ATE is another form of weighted average of MTE. Weights used in LATE are generally different from those used in ATE, so they discredit LATE as not being meaningful for a treatment effect parameter. They argue that LATE is only meaningful for policy effect parameter which measures the *net* effect of the policy, rather than the *gross* effect such as the treatment effect. They also show that the weights used in LATE estimator, although sum to 1, are negative at some u_D if $J(\mathbf{Z})$, any function constructed with a vector of IVs \mathbf{Z} that is used for IV estimation, is nonmonotonic in (or being negatively correlated with) propensity score: if the probability $\Pr[G(\mathbf{Z}) \geq u_D | J(\mathbf{Z}) = j, \mathbf{X} = \mathbf{x}]$ is negatively correlated with $J(\mathbf{Z})$ for a certain value of u_D , the numerator of $\omega(x, u_D)$ becomes negative. This happens if there is

essential heterogeneity in participation that people may enter or exit the program to a given change in \mathbf{Z} .^{*17} One thus needs *uniformity* (or ‘monotonicity’ in Imbens and Angrist (1994) sense) in propensity score that a greater value of $J(\mathbf{Z})$ leads to a greater propensity score $G(\mathbf{Z})$ for everyone. If the propensity score is used as instruments $J(\mathbf{Z}) = G(\mathbf{Z})$, then all the weights will be positive.

The issue that instrumental variables give an aggregate of different estimates of different subpopulation was well recognized, because it was pointed out by Heckman and Robb (1985), (1986). A similar, yet not using MTE, result is found in Imbens and Angrist (1994, Theorem 2). An analogous proposition has been derived ahead of time by Angrist, Graddy, and Imbens (2000, Theorem 1, 2) in a continuous treatment intensity $D \in \mathbb{R}_+$ case that shows an IV estimator in simultaneous equations models is a weighted average of MTEs (although they do not use the term ‘MTE’).

$$\begin{aligned}\alpha_{IV}^{z^k, z^l} &= \int_0^\infty \mathcal{E} \left[\frac{\partial y(D, \mathbf{Z})}{\partial \tilde{D}} \Big| D(\mathbf{z}^k) \leq \tilde{D} \leq D(\mathbf{z}^l) \right] \omega(\tilde{D}) d\tilde{D}, \\ &= \int_{D(\mathbf{z}^k)}^{D(\mathbf{z}^l)} \omega(\tilde{D}) MTE(\tilde{D}, \mathbf{Z}) d\tilde{D},\end{aligned}$$

with the weights being given by the ratio of probability of particular treatment intensity \tilde{D} to the total sum of probability over entire support of treatment intensity, or:

$$\omega(\tilde{D}) = \frac{\Pr[D(\mathbf{z}^k) \leq \tilde{D} \leq D(\mathbf{z}^l)]}{\int_0^\infty \Pr[D(\mathbf{z}^k) \leq r \leq D(\mathbf{z}^l)] dr} = \frac{\Pr[D(\mathbf{z}^k) \leq \tilde{D} \leq D(\mathbf{z}^l)]}{\int_{D(\mathbf{z}^k)}^{D(\mathbf{z}^l)} \Pr[D(\mathbf{z}^k) \leq r \leq D(\mathbf{z}^l)] dr},$$

where \mathbf{z}^k and \mathbf{z}^l are indexed as $D(\mathbf{z}^k) \leq D(\mathbf{z}^l)$ with $k \neq l$. Angrist, Graddy, and Imbens (2000, Theorem 2) also notes that, under many instrument values $\{\mathbf{z}^1, \dots, \mathbf{z}^K\}$,

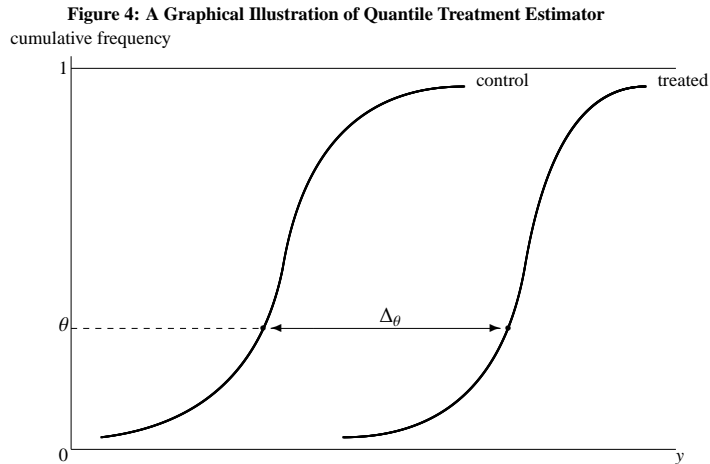
$$\alpha_{IV}^{z^k, z^{k+1}} = \sum_{k=1}^{K-1} \lambda_{k+1} \alpha_{IV}^{z^k, z^{k+1}}$$

with weights

$$\lambda_{k+1} = \frac{[G(\mathbf{z}^{k+1}) - G(\mathbf{z}^k)] \sum_{l=k}^K f_z(\mathbf{z}^l) [J(\mathbf{z}^l) - \mathcal{E}[J(\mathbf{Z})]]}{\sum_{m=1}^{K-1} [G(\mathbf{z}^{m+1}) - G(\mathbf{z}^m)] \sum_{l=m}^K f_z(\mathbf{z}^l) [J(\mathbf{z}^l) - \mathcal{E}[J(\mathbf{Z})]]},$$

where $f_z(\mathbf{Z})$ is the probability mass function of \mathbf{Z} . λ_{k+1} can be negative if the order of instruments $\mathbf{z}^k, \mathbf{z}^l$ are such that, for some \mathbf{z}^k and \mathbf{z}^{k+1} , $J(\mathbf{Z})$ and $G(\mathbf{Z})$ are nonmonotonic. This weight is discrete version of Heckman and Vytlačil (2005)’s weight and the condition for nonnegativity is exactly the same. However, what the IV estimate averages over is IV estimates of subpopulation, not MTEs as in Heckman and Vytlačil (2005).

^{*17} This also happens even under treatment effect homogeneity when participation decision is nonseparable in IVs and disturbance $D^* = \mu_D(\mathbf{Z}, V)$.



Given both sides recognize the problem from the beginning, it seems as if the controversy is taking too much toll in terms of excessive arguments, and it might have become a source of confusion. They are pretty much in accord that IV estimator is of limited usefulness when there is essential heterogeneity. Chamberlain (1986), Heckman and Robb (1985), (1986), Imbens and Angrist (1994), among others, all point to the fact that IV is not suited under essential heterogeneity, simply because the instrument loses validity. Most fruitful debate may be found in pointing out that independence is too strong for mean statistics but conditional mean independence suffices (Heckman, 1999, 831), and that IV estimators and LATE estimators are all weighted averages of MTE (Heckman and Vytlacil, 2005).

As the standard IV estimate gives the weighted average of subpopulations, it is natural to consider the IV estimation of QTE, or IV estimation of nonseparable functions, which we will turn next.

VII Quantile Treatment Effects

Most of regression and propensity score based methods estimate the mean impact. However, as Abadie, Angrist, and Imbens (2002) and Bitler, Gelbach, and Hoynes (2006) show, they may miss some important heterogeneity of the impacts over the entire population. Quantile treatment effect (QTE) estimates the impact on the quantile of the population. Suppose that we are interested in θ quantile. Then, θ quantile treatment effect is given by:

$$\Delta_{\theta}(\mathbf{x}) = q_{\theta}(\mathbf{x}|D = 1) - q_{\theta}(\mathbf{x}|D = 0),$$

where $q_\theta(\mathbf{x}|D)$ is an outcome function at quantile θ given covariates \mathbf{x} and treatment status D (Doksum, 1974). That is, we compare the θ quantile of the treated and θ quantile of the control. With appropriate construction of the counterfactual (the control), there is nothing wrong in directly comparing the quantiles. Bitler, Gelbach, and Hoynes (2006) simply compare the response at the specified quantile of treated and control distributions in randomized experiments.

Average treatment effect is related by:

$$ATE(\mathbf{x}) = \int_0^1 \Delta_\theta(\mathbf{x})d\theta.$$

Quantile regression is a well-established estimation technique and its computation can be done using standard statistical programs such as \mathcal{R} and its package `quantreg`, or using the algorithm of Buchinsky (1998).^{*18}

VII.1 Quantile Treatment Effects under Exogeneity

Firpo (2007) shows the efficient semiparametric QTE estimator and its \sqrt{n} convergence to asymptotic normality. QTE, or overall quantile treatment effects (OQTE) in his terminology, is identified in a two-step procedure. First, one estimates the propensity score nonparametrically.

^{*18} Quantile regression is derived as follows. Consider

$$\begin{aligned} \min_m \left[\sum_{y_i < m} (1 - \theta)|y_i - m| + \sum_{y_i \geq m} \theta|y_i - m| \right] &= \min_m \left[\sum_{y_i < m} -(1 - \theta)(y_i - m) + \sum_{y_i \geq m} \theta(y_i - m) \right], \\ &= \min_m \left[\sum_{i=1}^n \rho_\theta(y_i - m) \right], \end{aligned}$$

where $\rho_\theta(a) = a \cdot (\theta - 1[a < 0])$ is called a *check function*, thus

$$\rho_\theta(a) = [a|a < 0] \cdot (\theta - 1) + [a|a > 0]\theta = (1 - \theta)|a| \cdot 1[a < 0] + \theta|a| \cdot 1[a > 0] > 0.$$

So the problem is:

$$\min_m \mathcal{E} [\rho_\theta(y_i - m)] = \min_m \left[\int_{-\infty}^m -(1 - \theta)(y - m)f(y)dy + \int_m^{\infty} \theta(y - m)f(y)dy \right]$$

FOC is:

$$(1 - \theta) \cdot [- (y - m^*)f(y)] \Big|_{y=m^*} + \int_{-\infty}^{m^*} (1 - \theta)f(y)dy - \theta \cdot [(y - m^*)f(y)] \Big|_{y=m^*} - \int_{m^*}^{\infty} \theta f(y)dy = 0,$$

thus

$$-\theta \int_{m^*}^{\infty} f(y)dy + (1 - \theta) \int_{-\infty}^{m^*} f(y)dy = 0,$$

or

$$\frac{\Pr[y \leq m^*]}{\theta} = \frac{\Pr[y > m^*]}{1 - \theta}.$$

Then we see that $m^* = \text{quantile}_\theta(y)$, where $\text{quantile}_\theta(y)$ is y such that $\Pr[y \leq m^*] = \theta$. (y such that lower-tail probability is equal to θ .) $\theta = \frac{1}{2}$ where m^* is median, a median LAD estimator, is the special case of θ quantile.

The approach he proposes is to follow Hirano, Imbens, and Ridder (2003) and estimate the propensity score nonparametrically. Second, using the estimated propensity score $\hat{G}(\mathbf{x}_i)$, one solves:

$$\min_{\{q_\theta\}} \sum_{i=1}^n \hat{\omega}_{ig} \rho_\theta(y_i - q_\theta)$$

where the weights are

$$\hat{\omega}_{i0} = \frac{1}{n} \frac{1-D_i}{1-\hat{G}(\mathbf{x}_i)}, \quad \hat{\omega}_{i1} = \frac{1}{n} \frac{D_i}{\hat{G}(\mathbf{x}_i)}.$$

VII.2 IV Estimation of Quantile Treatment Effects

Using the monotonicity assumption, Abadie, Angrist, and Imbens (2002) propose a two-step, IV procedure for binary treatment that applies the weighted quantile regression technique of Newey and Powell (1990). The instrument z is randomly assigned eligibility that explains the participation status D_z , with the IV ‘monotonicity’ assumption $D_1 > D_0$. It solves, for a given θ :

$$\operatorname{argmin}_{\{\alpha_\theta, \beta_\theta\}} \mathcal{E}[\rho_\theta(y - \alpha_\theta D - \beta'_\theta \mathbf{x} | D_1 > D_0)]$$

Their two-step procedure requires in the first stage to estimate the propensity score which is used to construct the complier finding function:

$$\kappa = 1 - \frac{D(1-z)}{1-\pi_0(\mathbf{X})} - \frac{(1-D)z}{\pi_0(\mathbf{X})}, \quad \pi_0(\mathbf{x}) = \Pr[z = 1 | \mathbf{x}].$$

Note that $\kappa = 1$ if $Z = D = 0$ or $Z = D = 1$, and $\kappa < 0$ if $Z = 0, D = 1$ or $Z = 1, D = 0$, hence it finds the complier with $\kappa = 1$. Abadie (2003) has shown that

$$\mathcal{E}[\rho_\theta(y - \alpha_\theta D - \beta'_\theta \mathbf{x} | D_1 > D_0)] = \frac{1}{\Pr[D_1 > D_0]} \mathcal{E}[\kappa \rho_\theta(y - \alpha_\theta D - \beta'_\theta \mathbf{x})].$$

Then one can use the estimated $\hat{\kappa}$ as weights in weighted quantile regression.

$$\operatorname{argmin}_{\{\alpha_\theta, \beta_\theta\}} \mathcal{E}[\hat{\kappa} \rho_\theta(y - \alpha_\theta D - \beta'_\theta \mathbf{x})].$$

Since this is an example of M-estimators, it is straightforward to derive the robust covariance matrix for inference.

Chernozhukov and Hansen (2005) show the moment conditions that can be used in nonparametric estimation of quantile treatment response functions, which they call the *instrumental variable quantile regression (IVQR)* model:

$$\Pr[y \leq q_\theta(\mathbf{x}|D) | \mathbf{Z}] = \theta \quad \text{and} \quad \Pr[y < q_\theta(\mathbf{x}|D) | \mathbf{Z}] = \theta.$$

A sample analogue is given by a vector of empirical moment conditions:

$$\frac{1}{n} \sum_{i=1}^n [1[y_i \leq \alpha_\theta D_i + \beta'_\theta \mathbf{x}_i] - \theta] \begin{bmatrix} \mathbf{x}_i \\ \mathbf{z}_i \end{bmatrix} = o_p(n^{-\frac{1}{2}}).$$

A distinguished feature of their methods is that IV monotonicity assumption is replaced with *rank similarity* condition of an index that captures the heterogeneity in outcomes. Instead of IV monotonicity, they assume the net outcome gain for a given ‘ability’ in the treatment is known *ex ante* up to a distribution that is common to everyone. Under this assumption, the estimated QTE has an interpretation as the treatment effect holding the distribution of unobservable, or the mean of unobservable, fixed. Unfortunately, one needs a similar condition as IV monotonicity for global identification of parameters.*¹⁹

Another feature of IVQR is that one does not require independence between the instrument and selection equation disturbance terms, unlike other IV estimators or other estimators relying on selection on observables. IVQR assume that decisions are explained by a general function $D(\mathbf{x}_i, z_i, \mathbf{v}_i)$ where \mathbf{v}_i is a random vector, allowing an arbitrary relationship between z_i and \mathbf{v}_i in selection. This is helpful when IVs are measured with errors hence become correlated with selection errors as in Hausman (1977). Another instance is the EXAMPLE 2 of Imbens and Angrist (1994) that eligibility assignment z_i may be random yet with which official the applicants must work on application may affect the participation decision D_i . This works analogous to the measurement error problem: $z_i = 1$ for eligible applicants may actually be less than 1 if assigned to an obnoxious or difficult official. Their method also allows estimation over entire distribution of compliers and allows discrete and continuous treatment variables, unlike Abadie, Angrist, and Imbens (2002) who consider only the binary treatment case.

Chernozhukov and Hansen (2004a) provide an application of their methodology to effects of pension plan on wealth accumulation. Despite being flexible and not requiring IV monotonicity, Chernozhukov and Hansen (2005)’s method requires an untestable assumption of rank similarity in the unobservable. Chernozhukov and Hansen (2004a) use Abadie, Angrist, and Imbens (2002)’s estimator to check the robustness thereby indirectly asserting the plausibility of rank similarity condition in their application. Chernozhukov and Hansen (2004b) illustrate a simple, two-stage computation procedure to estimate the linear quantile function using instruments for an endogenous regressor D_i for a given τ :

1. Define a grid of $\{\alpha_j, j = 1, \dots, J\}$. Choose the IV function $\phi(\mathbf{x}_i, \mathbf{z}_i|\tau)$. Choose weights $v_i(\tau)$. Recommended choices are projection of D_i on $\mathbf{x}_i, \mathbf{z}_i$ for $\phi(\mathbf{x}_i, \mathbf{z}_i|\tau)$, and $v_i(\tau) = 1$.

*¹⁹ Chernozhukov and Hansen (2005)’s THEOREM 2 shows that global identification condition required for uniqueness of parameter estimates is monotone likelihood ratio condition that likelihood ratio $\frac{f_{Y_1}}{f_{Y_0}}$ is increasing in z_i . This indicates that eligibility increases the likelihood of being treated relative to untreated, which is similar but different from IV monotonicity, that is based on distribution function rather than density function, used in LATE.

Then run weighted quantile regression to estimate $\beta_\tau(\alpha_j)$ and $\gamma_\tau(\alpha_j)$:

$$\frac{1}{n} \sum_{i=1}^n \rho_\tau \left(y_i - \alpha_j D_i - \beta' \mathbf{x}_i - \gamma \phi(\mathbf{x}_i, \mathbf{z}_i | \tau) \right) v_i(\tau).$$

Store $\hat{\gamma}_\tau(\alpha_j)$ for each α_j .

2. For each $\hat{\gamma}_\tau(\alpha_j)$, find α_j such that

$$\min_{\{\alpha_j\}} \sqrt{\frac{\hat{\gamma}_\tau^2(\alpha_j)}{a(\mathbf{x}_i, \mathbf{z}_i | \tau)}}$$

with $a(\mathbf{x}_i, \mathbf{z}_i | \tau) = \mathcal{E}[\phi(\mathbf{x}_i, \mathbf{z}_i | \tau)^2]$.

VIII Before-After Methods

VIII.1 Difference-in-Differences Estimation

In words, the assumption employed are: omitted variables are fixed variables, either in level or in first-difference. Conditional mean independence is not necessary, but one needs the homogeneous treatment effect for all individuals.

Difference-in-differences estimator is:

$$\widehat{ATE}(\mathbf{x}) = \sum_{i=1}^{n_1} \frac{\Delta \hat{y}_{1it}(\Delta \mathbf{x}_{it})}{n_1} - \sum_{i=1}^{n_0} \frac{\Delta \hat{y}_{0it}(\Delta \mathbf{x}_{it})}{n_0}.$$

with

$$\epsilon_{it} = c_i + e_{it},$$

where e_{it} is a random error with mean zero satisfies conditional mean independence $(c_i, v_{it}) \perp (e_{it}, e_{it-1}) | \mathbf{x}_{it}, \mathbf{x}_{it-1}$. $\Delta \hat{y}_{Dit}(\Delta \mathbf{x}_{it})$ is predicted Δy_{Dit} in the regression of Δy_{Dit} on $\Delta \mathbf{x}_{it}$.

A convenient way is to embed in the regression. Participation process is allowed to include the fixed-effect c_i and idiosyncratic error v_{it} that should not be correlated with other idiosyncratic error e_{it} , and we will write it as:

$$D_{it} = D_{it}(c_i, v_{it}, \mathbf{x}_{it}).$$

Under this, we allow for a linear time trend γt :

$$y_{it} = c + \alpha D_{it}(c_i, v_{it}, \mathbf{x}_{it}) + \gamma t + \beta' \mathbf{x}_{it} + (c_i + e_{it}).$$

Note, when program is introduced in time t , it is $D_{it} = 1$ and $D_{it-1} = 0$ for the treated, thus

$$\Delta D_{it}(c_i, v_{it}, \mathbf{x}_{it}) = D_{it}(c_i, v_{it}, \mathbf{x}_{it}) - D_{it-1} = 1 \quad \text{for the treated,}$$

and $\Delta D_{it}(c_i, v_{it}, \mathbf{x}_{it}) = 0$ for the controls. Note $\Delta t = t - (t - 1) = 1$. Then,

$$\begin{aligned}\Delta y_{it} &= \gamma + \alpha \Delta D_{it}(c_i, v_{it}, \mathbf{x}_{it}) + \beta'_1 \Delta D_{it}(c_i, v_{it}, \mathbf{x}_{it}) \Delta \mathbf{x}_{it} + \beta'_0 [1 - \Delta D_{it}(c_i, v_{it}, \mathbf{x}_{it})] \Delta \mathbf{x}_{it} + \Delta e_{it}, \\ &= \gamma + \alpha D_{it}(c_i, v_{it}, \mathbf{x}_{it}) + \beta'_1 D_{it}(c_i, v_{it}, \mathbf{x}_{it}) \Delta \mathbf{x}_{it} + \beta'_0 [1 - D_{it}(c_i, v_{it}, \mathbf{x}_{it})] \Delta \mathbf{x}_{it} + \Delta e_{it}.\end{aligned}$$

OLS gives consistent estimates if $(c_i, v_{it}) \perp (e_{it}, e_{it-1}) | \mathbf{x}_{it}, \mathbf{x}_{it-1}$.

Further, if

$$\beta_1 = \beta_0,$$

which is frequently assumed in the simplest applications of DID estimator, then:

$$\Delta y_{1it} - \Delta y_{0it} = \alpha + \Delta e_{1it} - \Delta e_{0it}.$$

Taking expectations, we have:

$$\mathcal{E}[\Delta y_{1it}] - \mathcal{E}[\Delta y_{0it}] = \alpha = ATE.$$

In this case, ATE estimator is;

$$\widehat{ATE} = \sum_{i=1}^{n_1} \frac{\Delta y_{1it}}{n_1} - \sum_{i=1}^{n_0} \frac{\Delta y_{0it}}{n_0}.$$

If we omit some variables of $\Delta \mathbf{x}_{it}$ from the regression when $\beta_1 \neq \beta_0$, we can still estimate ATE consistently if changes in \mathbf{x}_{it} are uncorrelated with c_i . This holds, for example, when c_i is constant through time and its effect on \mathbf{x}_{it} is also constant through time, thus differencing eliminates them.

The identification condition $(c_i, v_{it}) \perp (e_{1it}, e_{1it-1}) | \mathbf{x}_{it}, \mathbf{x}_{it-1}$ precludes correlation between v_{it} and (e_{it}, e_{it-1}) through common, unobservable time-varying shocks. For example, a health shock realized in t or $t - 1$ to the family member may prompt a person to participate the social program while the shock may affect the outcome of interest y_{it} in a time-varying way.^{*20} Another example is that a weather shock which is omitted in the regression affects both participation D_{it} and y_{it} . This can be partially resolved for the weather shocks at the village level by including the village dummies in the regression.^{*21} Another important caveat is that, when there is a serial correlation (residuals are correlated through time), there may be a substantial bias in estimated standard errors but using the heteroskedasticity-consistent covariance matrix reduces it. See Bertrand, Duflo, and Mullainathan (2004) for details.

^{*20} If its effects on outcomes are time-invariant, then we can use the fixed-effect model.

^{*21} Including the residual of participation equation, as done in Smith and Blundell manner, does not work because one cannot consistently estimate participation due to the presence of fixed effects. Ravallion and Wodon (2000) makes a mistake of including the endogenous variables of households. If we have three periods of data, using \mathbf{x}_{it-2} as instruments that is correlated with participation $D(c_i, v_{it}, \mathbf{x}_{it})$ but not with the changes in outcome Δy_{it} , unfortunately, does not work, because \mathbf{x}_{it-2} and c_i can be correlated.

Difference-in-difference-in-differences estimator (robust to fixed-growth-effects):

$$ATE(\mathbf{x}) = \sum_{i=1}^{n_1} \frac{\Delta^2 \hat{y}_i (\Delta^2 \mathbf{x}_i | D_i = 1)}{n_1} - \sum_{i=1}^{n_0} \frac{\Delta^2 \hat{y}_i (\Delta^2 \mathbf{x}_i | D_i = 0)}{n_0}.$$

Robust to an heterogenous fixed-growth-effect selection bias $(\gamma_i + \gamma)t$. Redefine the errors as

$$u_{it} = c_i + \gamma_i t + \eta_{it}, \quad e_{it} = \gamma_i t + \eta_{it},$$

where η_{it} are idiosyncratic errors that satisfy the conditional mean independence with c_i, v_{it}, γ_i given covariates, or $(c_i, v_{it}, \gamma_i) \perp (\eta_{1it}, \eta_{1it-1}, \eta_{1it-2}) | \mathbf{x}_{it}, \mathbf{x}_{it-1}, \mathbf{x}_{it-2}$. Then:

$$y_{it} = c + \alpha D_{it} + (\gamma_i + \gamma)t + \boldsymbol{\beta}' \mathbf{x}_{it} + (c_i + \eta_{it}).$$

We assume $D_{it} = D_{it}(\mathbf{x}_{it}, c_i, \gamma_i)$. Taking a first-difference:

$$\Delta y_{it} = \gamma + \alpha \Delta D_{it} + \boldsymbol{\beta}' \Delta \mathbf{x}_{it} + (\gamma_i + \Delta \eta_{it}),$$

in which ΔD_{it} is positively correlated with γ_i of the composite error term $\gamma_i + \Delta \eta_{it}$. Taking a second-difference $\Delta^2 \mathbf{x}_{it} \stackrel{\text{def}}{=} \Delta \mathbf{x}_{i,t} - \Delta \mathbf{x}_{i,t-1} = (\mathbf{x}_{i,t} - \mathbf{x}_{i,t-1}) - (\mathbf{x}_{i,t-1} - \mathbf{x}_{i,t-2})$:

$$\begin{aligned} \Delta^2 y_{it} &= \alpha \Delta^2 D_{it} + \boldsymbol{\beta}' \Delta^2 \mathbf{x}_{it} + \Delta^2 \eta_{it}, \\ &= \alpha \Delta^2 D_{it} + \boldsymbol{\beta}' \Delta^2 \mathbf{x}_{it} + \Delta^2 \eta_{it}, \end{aligned}$$

which purges the individual trending term γ_i from the error, so $D_{it} \perp \Delta^2 \eta_{it}$, thus it is robust to heterogeneous individuals with different growth rates in Δy_{it} . The identification condition is $(c_i, v_{it}, \gamma_i) \perp (\eta_{1it}, \eta_{1it-1}, \eta_{1it-2}) | \mathbf{x}_{it}, \mathbf{x}_{it-1}, \mathbf{x}_{it-2}$. The same argument follows for the rest as the first-difference case, only that it takes at least 3 periods, with 2 periods prior to the intervention to implement this estimator.

Abadie (2005) discusses the nonlinear, semiparametric estimation of DID estimator. Denote $y_{i,t}$ as outcome value of period t . It shows one can estimate DID nonparametrically with

$$\mathcal{E} \left[y_{1,1} - y_{0,1} | \mathbf{x}, D = 1 \right] = \mathcal{E} \left[h(D, \mathbf{x})(y_{1,1} - y_{0,1}) | \mathbf{x} \right], \quad h(D, \mathbf{x}) = \frac{1}{G(\mathbf{x})} \frac{D-G(\mathbf{x})}{1-G(\mathbf{x})},$$

because

$$\begin{aligned} \mathcal{E} \left[\frac{1}{G(\mathbf{x})} \frac{D-G(\mathbf{x})}{1-G(\mathbf{x})} (y_{1,1} - y_{0,1}) | \mathbf{x} \right] &= \mathcal{E} \left[\frac{1}{G(\mathbf{x})} \frac{D-G(\mathbf{x})}{1-G(\mathbf{x})} (y_{1,1} - y_{0,1}) | \mathbf{x}, D = 1 \right] \cdot G(\mathbf{x}) \\ &\quad + \mathcal{E} \left[\frac{1}{G(\mathbf{x})} \frac{D-G(\mathbf{x})}{1-G(\mathbf{x})} (y_{1,1} - y_{0,1}) | \mathbf{x}, D = 0 \right] \cdot [1 - G(\mathbf{x})], \\ &= \mathcal{E} \left[y_{1,1} - y_{0,1} | \mathbf{x}, D = 1 \right] - \mathcal{E} \left[y_{1,1} - y_{0,1} | \mathbf{x}, D = 0 \right], \\ &= \mathcal{E} \left[y_{1,1} - y_{1,0} | \mathbf{x}, D = 1 \right] - \mathcal{E} \left[y_{0,1} - y_{0,0} | \mathbf{x}, D = 0 \right], \\ &= \mathcal{E} \left[y_{1,1} - y_{1,0} | \mathbf{x}, D = 1 \right] - \mathcal{E} \left[y_{0,1} - y_{0,0} | \mathbf{x}, D = 1 \right], \\ &\equiv \text{average treatment effect on the treated} \\ &= \mathcal{E} \left[y_{1,1} - y_{0,1} | \mathbf{x}, D = 1 \right], \end{aligned}$$

the fourth equality follows under the DID identifying assumption $\mathcal{E}[y_{0,1} - y_{0,0} | \mathbf{x}, D = 1] = \mathcal{E}[y_{0,1} - y_{0,0} | \mathbf{x}, D = 0]$. Noting that linearity assumption in standard DID to be restrictive, he has shown a semiparametric way to approximate the unknown function $\mathcal{E}[y_{i,1} - y_{i,0} | \mathbf{x}_i, D_i = 1]$. It is well known that when \mathbf{x}_i has relatively large dimension, it poses a problem in nonparametric estimation. Approximation is given by solving

$$\operatorname{argmin}_{\{\gamma\}} \sum_{i=1}^n \left[\hat{G}(\mathbf{x}_i) \left\{ \hat{h}(D_i, \mathbf{x}_i)(y_{i,1} - y_{i,0}) - g(\mathbf{x}_{ki}; \gamma) \right\}^2 \right]$$

where $g(\cdot)$ is an approximating function of choice such as polynomial in \mathbf{x}_i , and $\mathbf{x}_{ki} \in \mathbb{X}_k$, $\mathbf{x}_i \in \mathbb{X}$, and $\mathbb{X}_k \subseteq \mathbb{X}$. Thus we take:

$$\mathcal{E}[y_{i,1} - y_{i,0} | \mathbf{x}_i, D_i = 1] \simeq g(\mathbf{x}_{ki}; \gamma).$$

- DID: Operation Blackboard (Chinn, 2005). Effects of reallocation of teachers from large schools to small schools with single teacher on school outcomes. Estimation is at the state level using the number of both schools.
- Banerjee et al. (2005) use DID in estimating the effects of remedial education on test scores.
- Duflo (2001): DID between high- and low-intensity groups.

VIII.2 Changes-in-Changes Estimation

In their seminal and important paper, Athey and Imbens (2006) proposed an entirely new approach to program evaluation with before-after data. Unlike DID, which estimates the mean of treatment effects under fixed-effect and constant treatment effect assumptions, they show how one can derive the entire counterfactual distribution, both no-treatment for the treated and with-treatment for the control, under arbitrary treatment effect heterogeneity. The assumptions they used in deriving the changes-in-changes (CIC) estimator for continuous y are:

1. A single index $u_i \in \mathbb{U}$ that explains the differences in outcomes $y_{ig,t}$, given group g , time t , and covariates \mathbf{x}_i . Call U ability.
2. A common fixed outcome mapping $h : U \times T \rightarrow Y$ with strict monotonicity in U , $\frac{\partial h}{\partial U} > 0$ (does not have to be differentiable, strictly speaking). Commonality plays a big role here, because it ensures that there is no fundamental differences between the treated and the control with the same outcome Y . So it is implied that there is no difference in economic environment between the groups except for the treatment. This may be seen as another way of addressing the ignorability of treatment.
3. Distribution of U is assumed to be fixed through time within a group (or more precisely, independent across time within a group), or $U \perp\!\!\!\perp T | G$. But the same individual i does not have to have the same value of u in different periods, only the distribution of them to be unchanged. So any changes in outcome distribution is interpreted as changes in functional forms of $h(u, t)$.

4. There exists a substantial common support $\mathbb{U}_1 \subseteq \mathbb{U}_0$.

Denote the random variable Y_{gt} as outcome of group g at time t . Group $g = 1$ is the treated, and $g = 0$ is the control. When we denote the counterfactual outcome, we will make the (hypothetical) treatment status D explicit as $Y_{D,gt}$. Program is implemented at time $t = 1$. Then, the counterfactual of the treated is denoted by $Y_{0,11}$.

Athey and Imbens (2006)'s main theorem is:

$$F_{Y_{0,11}}(y) = F_{Y_{10}} \left[F_{Y_{01}}^{-1} \left[F_{Y_{01}}(y) \right] \right]. \quad (30)$$

This shows how the quantile of $Y_{0,11}$ can be computed.*22

Proof is relatively simple. Note:

$$\begin{aligned} F_{Y_{gt}^N}(y) &= \Pr[h(U, t) \leq y | g, t] = \Pr[U \leq h^{-1}(y; t) | g, t], \\ &= \Pr[U \leq h^{-1}(y; t) | g] = \Pr[U_g \leq h^{-1}(y; t)], \\ &= \Pr[h^{-1}(y; t)]. \end{aligned}$$

The second line follows because we assume $U \perp\!\!\!\perp T | G$. Then:

$$\begin{aligned} F_{Y_{gt}^N}(y) &= F_{U,g}[h^{-1}(y; t)] \\ &\iff \\ F_{Y_{gt}^N}[h(U, T)] &= F_{U,g}[h^{-1}(h(U, T); t)] \end{aligned} \quad (31)$$

So

$$h(U, T) = F_{Y_{gt}^N}^{-1} \left[F_{U,g}(u) \right]. \quad (32)$$

For $g = 0, t = 0$,

$$h(U, 0) = F_{Y_{00}^{-1}} \left[F_{U,0}(u) \right]. \quad (33)$$

Apply (31) for $g = 0, t = 1$, then $F_{Y_{01}}[h(U, 1)] = F_{U,0}[h^{-1}(h(U, 1); 1)]$. Then

$$F_{U,0}^{-1} \left[F_{Y_{01}}[h(U, 1)] \right] = h^{-1}(y; 1). \quad (34)$$

From (33), $h(U, 0) = F_{Y_{00}^{-1}} \left[F_{U,0}(u) \right]$ so

$$h \left[h^{-1}(y; 1), 0 \right] = F_{Y_{00}^{-1}} \left[F_{U,0} \left(h^{-1}(y; 1) \right) \right] = F_{Y_{00}^{-1}} \left[F_{Y_{01}}^{-1}(y) \right], \quad (35)$$

*22 We first choose $y_{0,11}$, the counterfactual outcome value in period 1, because we want to know at given value of y how much its quantile would be in the counterfactual distribution. Then we find the quantile of $y_{0,11}$ in $F_{Y_{10}}$, the period 1 control distribution. We use the actual control group's distribution $F_{Y_{10}}$ because this is the 'placebo' outcome distribution we want to compare with, and it is assumed that the only difference between the groups is the treatment. This quantile θ gives the how much the latent index u_0 should be had the observation $y_{0,11}$ been in the control group, which is counterfactual. For this θ we can get its period 0 value with $F_{Y_{00}}^{-1}(\theta)$, which we denote with y_{00} . This is justified because the latent index's ranking is assumed not to change over time, or we are interested in the changes of outcome at the given quantile thus want to find the corresponding quantile's period 0 value. Then, we can use $F_{Y_{10}}$ to get the quantile of y_{00} to see what quantile y_{00} should have been had it been included in the treated group. This is justified because of the common production function assumption with a single latent index, so for $y_{00} = y_{10}, u_0 = u_1$. The obtained quantile is the counterfactual quantile of $y_{0,11}$.

where the last line used (34). Applying (31) with $g = 1, t = 0$, we have $h(U, 0) = F_{Y_{10}}^{-1} [F_{U,1}(u)]$, so

$$F_{Y_{10}} [h(U, 0)] = F_{U,1}(u). \quad (36)$$

Then

$$\begin{aligned} F_{Y_{11}^N}(y) &= F_{U,1} [h^{-1}(y; 1)], \\ &= F_{Y_{10}} [h [h^{-1}(y; 1), 0]], && \text{[by (36)]} \\ &= F_{Y_{10}} [F_{Y_{01}}^{-1} [F_{Y_{01}}(y)]] && \text{[by (35)]} \end{aligned} \quad \blacksquare$$

To state in the simplest way, we will estimate the treatment effect for the quantile θ of the treated:

$$\Delta_\theta = (y_{11}^\theta - y_{10}^\theta) - (y_{01}^{\tilde{\theta}} - y_{00}^{\tilde{\theta}}), \quad y_{10}^\theta = y_{00}^{\tilde{\theta}}, \quad \theta \neq \tilde{\theta} \text{ generally.}$$

So

$$\Delta_\theta = y_{11}^\theta - y_{01}^{\tilde{\theta}}.$$

Under the common outcome mapping assumption, the counterfactual of y_{10}^θ is $y_{00}^{\tilde{\theta}}$ in control with $y_{10}^\theta = y_{00}^{\tilde{\theta}}$. y_{10}^θ and $y_{00}^{\tilde{\theta}}$ are observations with the same value of u_i . Then, tracking the change through time for $y_{00}^{\tilde{\theta}}$, the change at quantile $\tilde{\theta}$, should give the counterfactual for y_{11}^θ , because the rank of u_i are assumed to be invariant through time. For all y_{10} (all θ) on the common support with y_{00} , this should give the entire distribution of counterfactual for y_{11} . One can see that it is assuming that group affiliation does not matter in outcome. It is assumed that if the unobservable ability u_i is the same, then the outcome in both periods should be the same.*²³

The cookbook method for deriving the counterfactual, continuous outcome distribution for the treated is:

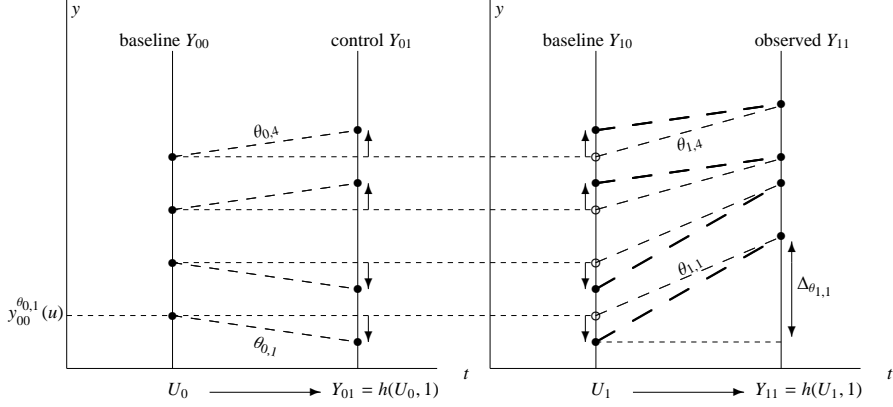
1. Choose y_{10} in the baseline of the treated. For this value y_{10} , find the quantile q_0 in the baseline of the control.

*²³ One can rewrite (30) as:

$$F_{Y_{01}}^{-1} [F_{Y_{00}} [Y_{10}(Y_{0,11} = y)]] = y, \quad (37)$$

Using this, we can extrapolate Y_{10} to Y_{01} . Take any value of Y_{10} , say y_{10} . This outcome y_{10} corresponds to a specific value u_1 via $y_{10} = h(u, 0)$. We can also find the value of u_0 if the chosen value y_{10} is to be hypothetically found in the control group, because $h(U, 0)$ is common across the groups hence $U_1 = U_0$ for the same $Y_{10} = Y_{00}$. Note that even with $U_1 = U_0$, corresponding quantile θ_1 and θ_0 need not to be equal because the distribution $F_{Y_{10}}$ and $F_{Y_{00}}$ (hence distributions of U_1 and U_0 , following common production function assumption and monotonicity assumption) are different. Nevertheless, we can find θ_0 if for the chosen y_{10} is to be found in $F_{Y_{00}}$. Using the transition (inverse) mapping $F_{Y_{01}}^{-1} : \Theta \rightarrow Y_{01}$ at quantile θ_0 , we obtain the outcome that we should observe in period 1 for such period 0 quantile θ_0 . This gives the counterfactual period 1 outcome for the chosen y_{10} , or $y_{0,11}$.

Figure 5: Changes-in-Changes Algorithm for $\mathbb{Y}_{10} \subseteq \mathbb{Y}_{00}$



Note: For a given y_{10} and its associated quantile θ_1 , we find the same value in control $y_{00} = y_{10}$ and find its quantile θ_0 . Then the counterfactual increment in Y_1 in the absence of treatment is $\Delta y_{0,11} = y_{01}^{\theta_0} - y_{00}^{\theta_0}$, thus CIC estimate is $\Delta \theta_1 = y_{11}^{\theta_1} - y_{10}^{\theta_1} - \Delta y_{0,11} = (y_{11}^{\theta_1} - y_{10}^{\theta_1}) - (y_{01}^{\theta_0} - y_{00}^{\theta_0})$.

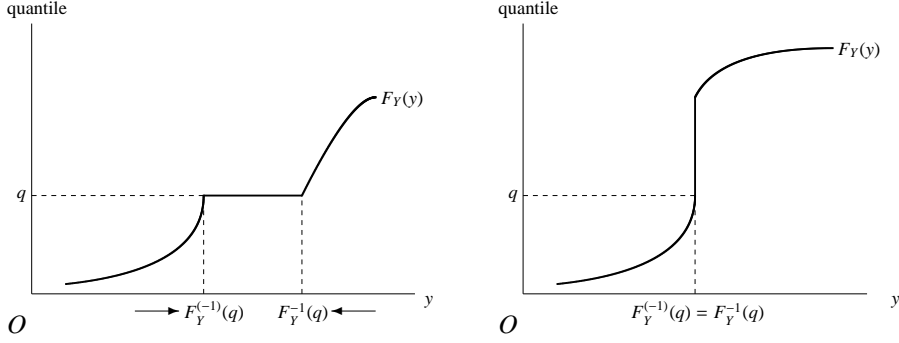
2. For the quantile q_0 , find the outcome y_{01} in the second-period of the control. This is the counterfactual outcome for the treated with baseline outcome y_{10} in the absence of treatment.
3. Repeat 1. and 2. for entire support of $\mathbb{U}_1 \cap \mathbb{U}_0$.

Since it gives you the entire distribution (on the common support), it gives the quantile treatment effects for any quantile provided that there is a common support. It can be used in reverse to obtain the counterfactual distribution for the control, e.g., take y_{00} and find the quantile from the treated, and so on. It is also shown that it can be used for multi-period, and multi-groups.

One can incorporate covariates by estimating $\Delta y_{ig,t} = \gamma'_g \Delta \mathbf{x}_{ig,t} + u_{ig,t}$ using the fixed-effect model before the exercises. Get the fixed-effect estimate $\hat{c}_{ig} = \sum_{t=1}^{T_i} \frac{\hat{u}_{ig,t}}{T_i}$. Then use \hat{c}_{ig} in place of $y_{ig,0}$. The recommended procedure of Athey and Imbens (2006) is to estimate $y_{i,t} = \delta' \mathbf{d}_{i,t} + \beta' \mathbf{x}_{i,t} + u_{i,t}$ where $\mathbf{d}_{i,t} = [gt, g(1-t), (1-g)t, (1-g)(1-t)]'$ is a group-time dummy interaction vector, and $\mathbf{x}_{i,t}$ is without an intercept. Then construct group-time-inclusive residual measure $\hat{y}_{i,t} = y_{i,t} - \hat{\beta}' \mathbf{x}_{i,t} = \hat{\delta}' \mathbf{d}_{i,t} + \hat{u}_{i,t}$ for CIC estimation. This procedure accounts for cohort fixed-effects, but not individual fixed-effects.

As this argument shows, incorporating covariates brings in the problem of consistently estimating the individual residuals. This is not possible for repeated cross-section data when individual fixed effects are present. Usefulness of CIC that repeated cross-section suffices will be limited under individual fixed-effects when we incorporate covariates.

Figure 6: Two Inverse Mapping



Note: Inf is the greatest lower bound of the set, and sup is least upper bound of the set.

$$F_Y^{-1} \equiv \inf\{y \in \mathbb{Y} | F_Y(y) \geq q\},$$

$$F_Y^{(-1)} \equiv \sup\{y \in \mathbb{Y} | F_Y(y) \leq q\}.$$

For discrete outcomes, one observes masses at certain outcome values, thus one needs to modify the strict monotonicity assumption to weak monotonicity of h in u . Accordingly, one can identify the (lower- and upper-) bounds of counterfactual distribution. Before doing so, one must properly define the inverse of the distribution function. Athey and Imbens (2006) define two inverse mapping, F_Y^{-1} , $F_Y^{(-1)}$:

$$F_Y^{-1} \equiv \inf\{y \in \mathbb{Y} | F_Y(y) > q\},$$

$$F_Y^{(-1)} \equiv \sup\{y \in \mathbb{Y} | F_Y(y) < q\}.$$

As **FIGURE 6** shows, two inverse mapping agree when $F(y)$ is dense. For the flat segment, we have:

$$F_Y^{(-1)}(q) < F_Y^{-1}(q),$$

and

$$F_Y[F_Y^{(-1)}(q)] < q.$$

Hence for all $q \in [0, 1]$:

$$F_Y[F_Y^{(-1)}(q)] \leq q \leq F_Y[F_Y^{-1}(q)].$$

Then their second main theorem (counterpart for the discrete outcome cases):

$$F_{Y_{10}}[F_{Y_{00}}^{(-1)}(F_{Y_{01}}(y))] \leq F_{Y_{11}^N}(y) \leq F_{Y_{10}}[F_{Y_{00}}^{-1}(F_{Y_{01}}(y))]. \quad (38)$$

This proof is also relatively simple. Given $\mathbb{U}_1 \subseteq \mathbb{U}_0$, normalizing $U_0 \sim \mathcal{U}[0, 1]$. Then:

$$F_{Y_{0t}}(y) = \Pr[h(U_0, t) \leq y] = \sup\{u : h(u, t) = y\}. \quad (39)$$

Probability is equal to the value of u_0 because u_0 is uniform on $[0, 1]$, and it is supremum of u such that $h(u, t) = y$ by definition. Then:

$$\begin{aligned} F_{Y_{1r}^N}(y) &= \Pr[Y_{1r}^N \leq y] = \Pr[h(U_1, t) \leq y] = \Pr[U_1 \leq \sup\{u : h(u, t) = y\}], \\ &= \Pr[U_1 \leq F_{Y_{0r}}(y)]. \end{aligned}$$

This gives

$$F_{Y_{10}}[F_{Y_{00}}^{(-1)}(F_{Y_{01}}(y))] = \Pr[U_1 \leq F_{Y_{00}}[F_{Y_{00}}^{(-1)}(F_{Y_{01}}(y))]].$$

Using $F_Y[F_Y^{(-1)}(q)] \leq F_Y[F_Y^{-1}(q)]$ for all $q \in [0, 1]$, we have:

$$\begin{aligned} \Pr[U_1 \leq F_{Y_{00}}[F_{Y_{00}}^{(-1)}(F_{Y_{01}}(y))]] &\leq \Pr[U_1 \leq F_{Y_{00}}[F_{Y_{00}}^{-1}(F_{Y_{01}}(y))]] = \Pr[U_1 \leq F_{Y_{01}}[F_{Y_{00}}^{-1}(F_{Y_{00}}(y))]] \\ &= \Pr[U_1 \leq F_{Y_{01}}(y)] = F_{Y_{11}^N}(y) \end{aligned} \quad (40)$$

Similarly, $F_Y[F_Y^{-1}(q)] \geq q$,

$$\begin{aligned} F_{Y_{10}}[F_{Y_{00}}^{-1}(F_{Y_{01}}(y))] &= \Pr[U_1 \leq F_{Y_{00}}[F_{Y_{00}}^{-1}(F_{Y_{01}}(y))]] \\ &\geq \Pr[U_1 \leq F_{Y_{00}}(y)] = \Pr[U_1 \leq F_{Y_{01}}(y)] = F_{Y_{11}^N}(y) \end{aligned} \quad (41) \quad \blacksquare$$

Note that from (39) we have $F_{Y_{00}}(y) = \sup\{u : h(u, 0) \leq y\}$. Since we normalized $U_0 \sim \mathcal{U}[0, 1]$ we have:

$$F_{U_0}(u) = u.$$

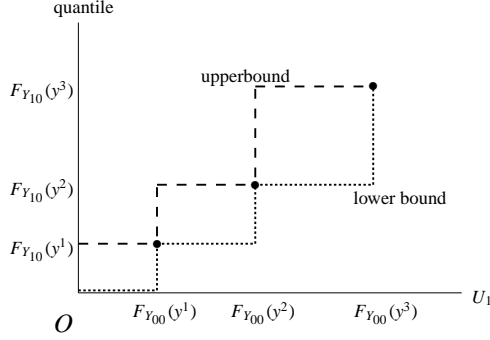
Then the value of $F_{Y_{00}}(y)$ gives the corresponding U value u . For $Y_{10} = Y_{00} = y$, corresponding value u is the same under the common mapping assumption. Then for $Y_{10} = Y_{00} = y$, distribution $F_{U_1}(u)$ is identified for the particular point u in the support of U_1 , which is equal to the value of $F_{Y_{10}}(y)$. Given $\mathbb{U}_1 \subseteq \mathbb{U}_0$, distribution over \mathbb{U}_1 is identified by $U_1 = F_{Y_{00}}(Y_{00})$, but only at points with $Y_{10} = Y_{00} = y$. Once points for $Y_{10} = Y_{00} = y$ are identified, they define the natural lower bound and upperbound for distribution. **FIGURE 7** gives an illustration. There are three masses at $\{y^1, y^2, y^3\}$ such that $Y_{00} = Y_{10}$ and corresponding frequencies in $F_{Y_{00}}(y)$ and $F_{Y_{10}}(y)$. With normalization, we have three $F_{Y_{00}}(y)$ as realization of U_1 , and corresponding $F_{Y_{10}}(y)$ as its frequencies. Lower- and upper-bounds are defined naturally.

In the discrete case, one needs to add exogeneity of covariates and weak monotonicity of $h(u, \mathbf{x}, t)$ in covariates \mathbf{x} to assumptions. Then covariates will help narrowing the bounds. This happens because for any $u^k(t, \mathbf{x}_i)$ and $u^k(t, \mathbf{x}_j)$ that are associated with the same outcome $y^{<k>}$ with $\mathbf{x}_i \neq \mathbf{x}_j$, in general we have $u^k(t, \mathbf{x}_i) \neq u^k(t, \mathbf{x}_j)$ thus adding more points in support of u .

Note that DID is a special case of CIC that the former imposes linearity and common treatment effect over entire support. CIC relaxes them by introducing two new assumptions, common outcome mapping and time-invariance of rank of ability. If it is a panel, then one can track each individual to obtain transitional mapping $Y_{D0} \rightarrow Y_{D1}$, then time invariance of rank is not necessary.

CIC is different from quantile DID (QDID). For a given y_{10} , the former fixes the period 0 outcome with $y_{10} = y_{00}^\theta$ to define the quantile θ of counterfactual observation, and obtain

Figure 7: Point Identification with Bounds in Discrete CIC



Note: Three values of $Y_{00} = Y_{10} = \{y^1, y^2, y^3\}$ are observed, with corresponding $F_{Y_{00}}(y)$ and $F_{Y_{10}}(y)$. By normalization $F_{U_0}(u) = u$. For $Y_{00} = Y_{10}$, underlying U are the same for both groups, $U_0 = U_1$. Noting $F_{Y_{00}}(y) = \sup\{u : h(u, 0) = y\}$ we have $F_{Y_{00}}(y^i) = u^i$. Thus points on support of U_1 are identified u^1, u^2, u^3 , and its corresponding quantiles $F_{U_1}(u^i) = F_{Y_{10}}(y^i)$. Lower- and upper-bounds are defined naturally.

an intertemporal change at the quantile θ in the control by taking $y_{01}^\theta - y_{00}^\theta$, treating it as the counterfactual change in time for y_{10} . The latter fixes the quantile $\hat{\theta}$ such that $q(\hat{y}_{10}|\mathbb{Y}_1) = \hat{\theta}$, then define the counterfactual observation as the quantile satisfying $q(\hat{y}_{00}|\mathbb{Y}_0) = \hat{\theta}$. Then the counterfactual intertemporal change for y_{10} is $\hat{y}_{10} - \hat{y}_{00} = y_{01}^{\hat{\theta}} - y_{00}^{\hat{\theta}}$. This, however, ‘does not make sense’ because it ignores the difference in outcome distribution between the treated and the control. It is not immediate why we want to compare the same quantile of different groups, if the distributions are not identical. Two distributions become identical if the treatment assignment is random, hence QDID becomes relevant.

As one can see, this does not require the same individual i to have the same specific u_i , only the rank to be the same across periods. So one can estimate CIC using the repeated cross section data, provided that each period sample is representative. However, one may be hesitant to attribute all the outcome differences solely to the ‘ability’ U , even after taking into account the covariates, because one should expect idiosyncratic shocks v_{git} to play some role. This should not be a problem if the errors U_{gi} and V_{git} are additive and if we are concerned only with the mean impact, or DID, as the mean of such idiosyncratic shocks can be normalized to be zero.*²⁴ So the interpretation is that CIC identifies the shock-inclusive treatment effects

*²⁴ It is impossible to integrate out V_{git} because we cannot identify the joint distribution of U_i and V_{git} . We only identify the transformed distribution of a single random variable $Y_{gt} = h(U_g, V_{gt}|G = g, T = t)$. With two unobservables U_i, V_{git} and one observed outcome, one cannot back out U_i and V_{git} without functional form restrictions. Consider an example $Y_t = V_{gt}U_g$. One generally needs the changes in variables by defining $Y_t = V_{gt}U_g, X_t = V_{gt}$, derive a joint distributions for $f_{X_t, Y_t}(x_t, y_t)$ and integrate X_t out to get $f_{Y_t}(y_t)$. To integrate out X_t , we need to know $f_{X_t, Y_t}(x_t, y_t)$ as a function of X_t and Y_t . To be concrete, assume $V_{gt} \sim \text{beta}(a, b)$, $U_g \sim \text{beta}(a + b, c)$. Then one can show $f_Y(y) = \text{beta}(a, b + c)$. But the problem is that, even the parametric

if one compares quantile-by-quantile, not at the mean. This is the result of the single index assumption that one can capture the difference in outcomes only with U . This limitation is the same with other QTE estimators which we will cover later on. In contrast, DID estimator averages out the idiosyncratic shocks if the shocks are separable, because it imposes linearity and constant treatment effect assumptions. If we impose linearity on $h(u, t)$ and assume constant treatment effects in CIC, then one can average out the idiosyncratic shocks by taking averages over certain range of quantiles.^{*25} So CIC is a nonlinear generalization of DID.

Summary: Comparison of Estimation Methods

In evaluating the program, practitioners may ask one of the following questions:

1. Which method is most appropriate for the given data at hand?
2. Given the program implementation cycle and the resource constraints, which estimator should we choose? How do we collect data for the chosen estimator?

Unfortunately, in most of the time, they ask the first question. This is because the evaluators usually do not have sufficient budget nor time, and they must deal with the problem with given data at hand. It is typically *ex post* data that they have, so they must rely on cross-sectional variations in outcomes and covariates to explain treatment effects. They thus need a broad range of covariates and a large number of potential control pool. If these conditions are not met, it is not likely that, whichever the method one uses, the estimates will give reliable answer to the question. So one must be content with the bound-based method.

Even if a broad range of covariates and a large potential control pool are available in *ex post* data, one still has to test the plausibility of exogeneity assumption. This is done by finding ineligible and opt-outs, comparing lagged outcomes by treatment status, or estimate propensity scores G and inspect linearity of treatment effect in G . We have pointed out that the first two may be demanding in terms of data requirement, and the linearity test may be the best option.

If exogeneity is rejected using an eye-ball test, one can infer the direction of selectivity bias. If G is nonlinear and increasing in the treatment effect, there will likely to be an upward bias. Without any credible instruments, one cannot resolve the problem of self-selection. Again, one

assumptions are correct, we will only know a and $b + c$, but not b and c separately, so one cannot deduce $a + b$ and c .

^{*25} Taking the mean for a certain range over quantile $[u_a, u_b]$, then the separable idiosyncratic shock term will be equal to its population mean of zero. But this will not give us a consistent estimator for 'average' quantile treatment effect for the chosen range, because quantile function can be nonlinear and so the average of quantile estimators may not be the quantile estimator for the average over the range, or $E_{u_a, u_b} [q(\theta)] \neq q(E_{u_a, u_b} [\theta])$. Even if $h(u, t)$ is linear, $q(\cdot)$ needs not to be linear, hence the equality does not hold.

must be content with the bound-based method in this case.

Although very unlikely, if exogeneity is not rejected one can use several estimators to get ATE. One caution is that covariate-conditioned $ATE_1(\mathbf{x})$ is equal to covariate-conditioned $ATE(\mathbf{x})$ under conditional mean independence, but unconditional ATE_1 is not the same as ATE if covariate support \mathbb{X}_1 for the treated is a strict subset of entire support \mathbb{X} . One can use regression-based estimators, parametric or nonparametric, propensity score based estimators, matching-based estimators. The literature has not established the finite sample property of these estimators, which makes us hard to choose from.

If an instrument is available, one can use IV estimator or Wald estimator. These have advantages that one can handle self-selectivity of the treatment sample. Such an estimator gives LATE, not ATE. It is ATE of people whose treatment status is changed with the (in)eligibility, hence local. It is further shown that, in the presence of heterogenous treatment effects, IV and Wald estimators may give weighted average of marginal treatment effects with unknown weights. Thus applicability of IV-based estimator, despite being convenient in dealing with endogeneity, is rather limited. Unfortunately, there is no clear consensus on the solution for essential heterogeneity so far.

If one is fortunate enough to decide on sampling design before the intervention, one can invoke on either randomization or before-after data collection, or even both. Randomization will give the ITT estimator, but it is sensitive to placement and operational spscificities, and its external validity is in question. Collecting the baseline always helps, as it gives more information on the population. But the identifying assumption of most widely used DID estimator is strong that pre-program condition cannot be correlated with participation decision. In the meantime, it may be useful to extend the bound-based approach to the panel data setting where one can control the additive individual effects.

REFERENCES

- Abadie, Alberto (2003)**, "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, Vol. 113, No.2 (), pp.231-263.
- _____ (2005), "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, Vol. 72, No. (), pp.1-19.
- _____, **Joshua Angrist, and Guido Imbens (2002)**, "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, Vol. 70, Issue 1 (January), pp.91-117.
- _____ and **Guido Imbens (2002)**, "On the Failure of Bootstrap for Matching Estimators," *mimeo*, Harvard University
- _____ and _____ (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, Vol. 74, Issue 1 (January), pp. 235-267.
- Angrist, Joshua D. (1990)**, "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, Vol. 80, No. 3 (June), pp. 313-336.
- _____, **Eric Bettinger, Erik Bloom, Elizabeth King and Michael Kremer (2002)**, "Vouchers for Private Schooling

- in Columbia: Evidence from Randomized Natural Experiment,” *American Economic Review*, No. 92, No.5 (December), pp. 1535-1558.
- , **Kathryn Graddy, and Guido W. Imbens (2000)**, “The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish,” *Review of Economic Studies*, Vol. 67, No. 3 (July), pp. 499-527.
- , and **Alan B. Krueger (1990)**, “Does Compulsory School Attendance Affect Schooling and Earnings?,” *Quarterly Journal of Economics*, Vol. 106, No. 4 (November), pp. 979-1014.
- , and **Victor Lavy (1999)**, “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement,” *Quarterly Journal of Economics*, Vol. 114, No. 2 (May), pp. 533-575.
- Athey, Susan, and Guido Imbens (2006)**, “Identification and Inference in Nonlinear Difference-In-Differences Models,” *Econometrica*, Vol. 74, Issue 2 (March), pp. 431-497.
- Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden (2005)**, “Remedying Education: Evidence from Two Randomized Experiments in India,” *Working Paper*, MIT.
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004)**, “How Much Should We Trust Differences-In-Differences Estimates?” *Quarterly Journal of Economics*, Vol. 119, Issue 1 (February), pp.249-275.
- Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes (2006)**, “What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments,” *American Economic Review*, No. 96, No.4 (September), pp. 988-1012.
- Buchinsky, Moshe (1998)**, “Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research,” *Journal of Human Resources*, Vol. 33, No. 1 (Winter), pp.88-126.
- Card, David, and Lara D. Shore-Sheppard (2004)**, “Using Discontinuous Eligibility Rules to Identify the Effects of the Federal Medicaid Expansions on Low-Income Children,” *Review of Economics and Statistics*, Vol. 86, No. 3 (), pp.752-766.
- Chamberlain, Gary (1986)**, “Asymptotic Efficiency in Semi-parametric Models with Censoring,” *Journal of Econometrics*, Vol.32, Issue 2 (July), pp.189-218.
- Chan, Tat Y., and Barton H. Hamilton (2006)**, “Learning, Private Information, and the Economic Evaluation of Randomized Experiments,” *Journal of Political Economy*, Vol. 114, No. 6 (), pp.998-1040.
- Chinn, Aimee (2005)**, “Can Redistributing Teachers Across Schools Raise Educational Attainment? Evidence from Operation Blackboard in India,” *Journal of Development Economics*, Vol.78, No.2 (), pp.384-405.
- Chernozhukov, Victor, and Christian Hansen (2004a)**, “The Effects of 401(k) Participation on the Wealth Distribution: An Instrumental Quantile Regression Analysis,” *Review of Economics and Statistics*, Vol. 86, Issue 3 (August), pp. 735-751.
- and ——— (2004b), “Inference on the Instrumental Quantile Regression Process for Structural and Treatment Effect Models,” *working paper*, MIT.
- and ——— (2005), “An IV Model of Quantile Treatment Effects,” *Econometrica*, Vol. 73, Issue 1 (January), pp. 245-261.
- Doksum, Kjell (1974)**, “Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case,” *Annals of Statistics*, Vol. 2, No. 2 (March), pp.267-277.
- Duflo, Esther (2001)**, “Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment,” *American Economic Review*, No. 91, No.4 (September), pp. 795-813.
- Firpo, Sergio (2007)**, “Efficient Semiparametric Estimation of Quantile Treatment Effects,” *Econometrica*, forthcoming.
- Galiani, Sebastian, Paul Gertler, and Ernesto Schargrotsky (2002)**, “Water for Life: The Impact of the Privatization of Water Services on Child Mortality,” *mimeograph*, UC Berkeley.
- Hahn, Jingyong (1998)**, “On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, Vol. 66, Issue 2 (March), pp. 315-332.
- Hausman, Jerry A. (1977)**, “Errors in Variables in Simultaneous Equation Models,” *Journal of Econometrics*, Vol. 5, Issue (), pp. 389-401.
- Heckman, James J. (1997)**, “Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations,” *Journal of Human Resources*, Vol. 32, No. 3 (Summer), pp.441-462.
- (1999), “Instrumental Variables: Response to Angrist and Imbens,” *Journal of Human Resources*, Vol. 34,

No. 4 (Autumn), pp.828-837.

- Heckman, James J., Hidehiko Ichimura, Jefferey A. Smith, and Petra E. Todd (1998)**, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, Vol. 66, No. 5 (September), pp.1017-1098.
- Heckman, James J., Hidehiko Ichimura, and Petra E. Todd (1997)**, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, Vol. 64, No. 4 (October), pp.605-654.
- Heckman, James J., and Richard Robb (1985)**, "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, New York: Wiley.
- _____ and _____ (1986), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in Howard Wainer ed., *Drawing Inferences from Self-Selected Samples*, Berlin: Springer-Verlag.
- Heckman, James J., and Jefferey A. Smith (1995)**, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, Vol. 9, No. 2 (Spring), pp.85-110.
- Heckman, James J., and Edward Vytlacil (2005)**, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, Vol. 73, Issue 3 (May), pp.669-738.
- _____ and _____ (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economic and Statistics*, Vol. 88, No. 3 (August), pp.389-432.
- Hirano, Keisuke, Guido W. Imbens, and Geert Ridder (2003)**, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, Vol. 71, Issue 4 (July), pp. 1161-1189.
- Imbens, Guido W. (2004)**, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economic and Statistics*, Vol. 86, No. 1 (February), pp. 4-29.
- Imbens, Guido W., and Joshua D. Angrist (1994)**, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 62, Issue 2 (March), pp. 467-475.
- Jalan, Jyotsna and Martin Ravallion (2003)**, "Does Piped Water Reduce Diarrhea for Children in Rural India?" *Journal of Econometrics*, Vol. 112, No. 0, pp.153-173.
- Lalonde, Robert J. (1986)**, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, Vol. 76, Issue 4 (September), pp.604-620.
- Leibowitz, Arleen. (1990)**, "The Response of Births to Changes in Health Care Costs," *Journal of Human Resources*, Vol. 25, Issue 4 (Autumn), pp.697-711.
- Manski, Charles F. (1995)**, *Identification Problems in Social Sciences*, Harvard University Press, Cambridge.
- _____ (1996), "Learning about Treatment Effects from Experiments with Random Assignment of Treatments," *Journal of Human Resources*, Vol. 31, No. 4 (Autumn), pp.709-733.
- _____ (1997), "The Mixing Problem in Program Evaluation," *Review of Economic Studies*, Vol. 64, No. 4 (October), pp. 537-553.
- Miguel, Edward, and Michael Kremer (2004)**, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, No. 72, Issue 1 (January), pp. 159-217.
- Newey, Whitney K., and James L. Powell (1990)**, "Efficient Estimation of Linear and Type I Censored Regression Models Under Conditional Quantile Restrictions," *Econometric Theory*, Vol. 6, Issue 3 (September), pp. 295-317.
- Oreopoulos, Philip (2006)**, "Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter," *American Economic Review*, Vol. 96, No. 1 (March), pp.152-175.
- _____ and **Shahidur R. Khandker (1998)**, "The Impact of Group-Based Credit Programs on Poor Households in Bangladesh: Does the Gender of Participants Matter?" *Journal of Political Economy*, Vol. 105, No. 5 (October), pp.958-996.
- Ravallion, Martin, and Quentin Wodon (2000)**, "Does Child Labour Displace Schooling? Evidence on Behavioural Responses to an Enrollment Subsidy," *Economic Journal*, No. 110, March, pp. C158-C175.
- Rosenbaum, Paul R. and Donald B. Rubin (1983)**, "The Central Role of Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Vol. 70, No. 1 (April), pp.41-55.
- Rosenzweig, Mark R. and Kenneth I. Wolpin (1984)**, "Evaluating the Effects of Optimally Distributed Public Programs: Child Health and Family Planning Interventions," *American Economic Review*, Vol. 76, Issue 3 (June),

pp.470-482.

_____ and _____ (2000), "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*, Vol. 38, No. 4 (December), pp.827-874.

Stock, James H., Joathan H. Wright, and Motohiro Yogo (2002), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, Vol. 20, No. 4 (October), pp. 518-529.

Vermeersch, Christel (2003), "School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation," *manuscript*, University of Oxford.

Wooldridge, Jeffrey M. (2002), *Econometric Analyses of Cross-Section and Panel Data*, MIT Press.