Impact Evaluation of Service Delivery Programs:

Methods and Concepts for Impact Assessments in Basic Education, Health, Water and Sanitation[*]

March 2008

Jakob Svensson[#] and Per Pettersson-Lidbom[¤]

[*] Paper prepared for the African Economic Research Consortium.

[#] IIES, Stockholm University, NHH and CEPR. Email: jakob.svensson@iies.su.se

[¤] Department of Economics, Stockholm University. Email: pp@ne.su.se

What is lacking among development practitioners today is not ideas, but an idea of whether or not the ideas work [Duflo, 2003].

1. Introduction

With limited resources and almost unlimited needs, impact evaluations ought to be an integral part of the policy formation process. The benefits of knowing which programs work and which do not extend far beyond any program or agency. A credible impact evaluation is also a global public good in the sense that it can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations in their ongoing search for effective ideas (Duflo, 2003). By credibly establishing impact, one can also counteract potential skepticisms about how funds are used and thereby build long-term support for international aid and development. It is also much easier to leverage resources when a project has been proven to work. With these facts in mind, it is surprising that impact evaluations are more an exception than a norm.

This paper is an introduction and, to some extent, a practical guide for researchers and practitioners interested in impact evaluation in education, health, water and sanitation. However, since the methods and concepts dealt with in this paper do not only apply to these sectors, the paper should be of more general interest. The paper is not a review of research using randomized or non-randomized evaluation methods, although we use past studies with a focus on sub-Saharan Africa to illustrate concepts and methods. Nor is it a paper that will in detail explore the properties of the various existing evaluation methods, although we provide references to in-depth studies.

The outline of the paper is as follows. In section 2, we discuss the question: What type of intervention/projects should one evaluate? We make a simple point, namely that the choice of which projects/interventions to evaluate should not be based on the ease at which the study can be implemented, but needs to be determined based on an overall assessment of how the sector works. Because some interventions are easier to evaluate than others, there is a risk that the researchers will pick projects to evaluate that are not necessarily of first-order importance. Section 3 discusses the issue of structural and reduced form relationships. The main section,

section 4, discusses the evaluation problem, or the selection bias problem, and reviews both randomized and non-randomized methods that have been developed to deal with this bias. We start by discussing the most credible design – the methodology of randomized evaluations. Thereafter, we discuss non-randomized methods, including the regression-control framework, matching methods, difference-in-differences and fixed effects methods, the instrumental variable approach and regression-discontinuity methods. Section 5 discusses inference issues, including problems associated with measurement errors and grouped data. Section 6 discusses data issues and power calculations and section 7, finally, concludes the paper.

2. Inputs or incentives: What should be the focus?

The deplorable state of publicly provided services in health, education, water and sanitations sectors in developing countries in general, and sub-Saharan Africa in particular, is evident from the data. For example, approximately 11 million children under-five die each year and almost half of these deaths occur in sub-Saharan Africa. More than half of these children – nearly 6 million – will die of diseases that could easily have been prevented or treated if the children had had access to a small set of proven, inexpensive services (Black et al, 2003; Jones et al, 2003).

Despite the tremendous progress in expanding enrollment and increasing years of schooling since 1960, 113 million children of primary school age are still not enrolled in school (Glewwe and Kremer, 2005). The problem once more appears to be most acute in sub-Saharan Africa. In 2000, the net enrollment rate in sub-Saharan Africa was 56 percent, compared to the average for the group of low-income and middle-income countries of 85% and 88%, respectively. Looking at the secondary school gross enrollment rate, sub-Saharan Africa comes out even worse. In 2000, the secondary school gross enrollment rate was 27 percent, almost half of that of South Asia – the region with the second lowest average. Maybe even more alarming, the quality of schooling in many developing countries is abysmal. As an example, a study on Ghanaian grade 6 students found the mean score on a very simple multiple-choice reading test to be similar to what one would expect from random guessing (Glewwe, 1999).

Evidence from water and sanitation sectors points in the same disturbing direction. For example, meeting the UN Millennium Development Goals of reducing the proportion of people without sustainable access to safe drinking water by half, will require providing over 900 million

people in rural areas of less developed countries with either household water connection or access to a constructed public water point within one kilometer (Kremer, et al 2006). In 2004, it was estimated that almost every other household (44% of the population) in sub-Saharan Africa lacked access to a sustainable water source – a small improvement from the 1990 figure of 51%.

What explains the dismal quality of the social services offered to the poor in developing countries? Clearly, inadequate funding is a plausible explanation. However, evidence presented in the 2004 World Development Report, and elsewhere, suggests that this is not the only reason. The provision of public services to poor people in developing countries is also constrained by weak incentives of service providers – schools and health clinics are not open when supposed to; teachers and health workers are frequently absent from schools and clinics and, when present, spend a significant amount of time not serving the intended beneficiaries; equipment, even when fully functioning, is not used; drugs and vaccines are misused; and public funds are expropriated (Bjorkman and Svensson, 2007).

Recent findings from rigorous, randomized evaluations in the education sector in Kenya and India also find little evidence that more resources on their own, with no changes in the way education is delivered, can improve the quality of education (Glewwe and Kremer, 2008). Thus, while there is still a great deal of value in identifying and evaluating the effects of increased supply or inputs, or the right mix of inputs, this alone will not get at the core of problem. As a consequence, attention is shifting towards understanding incentives and constraints facing both service providers and users. This involves studying both formal incentive schemes, like providing financial incentives to teachers or small in-kind incentives to mothers to get them to immunize their children, and demand driven approaches with an emphasis on popular participation, where the incentives are created through public pressure.

Overall, the focus on provider incentives seems promising, although the evidence to date is somewhat mixed. In education, where the bulk of the impact evaluation studies has been done, a number of studies document fairly large improvements in outcomes when modest incentives have been given to teachers. However, these examples involve financial incentives implemented by non-government organizations. When public officials have been involved in the implementation of the incentives scheme, things seem to work less well (Banerjee at al, 2008).

The demand driven approach has been subject to less scrutiny. Bjorkman and Svensson (2007) is an exception and documents large effects on both utilization and health outcomes from

a community-based monitoring project in primary health in Uganda. Other studies, however, document much smaller effects (see, for instance, Olken, 2007, Banerjee at al, 2008). Taken together, these findings stress the need to better understand if and under what conditions demand driven approaches to strengthen providers' incentives to serve the poor may work. More generally, the findings suggest the need of focusing impact evaluation not only on the last link in the service delivery chain; i.e., using variation across service providers and users to estimate the impact of various programs and interventions. After all, a country's ability to improve service delivery outcomes is not only (and sometimes not even primarily) determined by what happens at the school or health clinic level, but by the behavior of different actors and agencies involved in the design and implementation of education policy. And since the implementation of social service delivery in developing countries is often plagued by inefficiencies and corruption, interventions that focus on improving governance in general and governance of social services in particular may be a cost-effective way of improving service delivery outcomes (Reinikka and Svensson, 2007).

3. Methodological issues

This section lays out a simple framework to help us think about structural forms, reduced forms, and causal relationships in social sectors like education, health, water and sanitation. Without much loss of generality, we structure the discussion around primary education and education policies.[1]

Consider a household, or specifically the parents of a child, with a utility function

$$U = U(C, S, A), \tag{3.1}$$

where C is a vector consumption of goods and services, including leisure, at different time periods; S is a vector of each child's years of schooling, and A is a measure of learning for each child.

A natural starting point for economists is to consider a household that maximizes (3.1), subject to a budget constraint, a production function for learning, and the function linking learning to future labor income. To simplify, we assume that each household only has one child, so we can treat A and S as scalars, and only one school to choose from.

[1] See Glewwe and Kremer for a more thorough discussion of these issues and Glewwe (2005) for a similar exposition focusing on child health.

The production function for learning is

$$A = A(S, q, cc, hc, I), \tag{3.2}$$

where q is a vector of school and teacher characteristics, cc is a vector of child characteristics (like innate ability), hc is a vector of household characteristics (like parents' education), and I is a vector of school inputs under the control of parents. We use capital letters to denote endogenous variables, or choice variables, and small letters to denote exogenous variables. If we assume, somewhat unrealistically, that parents cannot influence school or teacher characteristics, we can treat q as exogenous. In I we include factors such as purchases of textbooks by parents, private tutoring and child health.

The (inter-temporal) budget constraint tells us that the household's income (parental income, the income generated from home production by the child, and transfers from the child when working as an adult) cannot exceed the household's expenditure (which depends on the quantity of goods consumed, the quantity of schooling, and the prices of these goods, inputs and schooling).

We can close the model by specifying an equation that relates the child's cognitive skills to her income Y when working as an adult

$$Y = Y(A, cc, hc). \tag{3.3}$$

This is the simplest set-up, to which we could add (as a constraint) an agricultural production function, and possibly a credit constraint. The set-up could also be extended by considering the household's choice conditional on the gender of the child, and/or introducing bargaining between household members.

Maximizing (3.1), subject to the budget constraint, (3.2) and (3.3), yields solutions for the quantity of schooling S and the parents' financial involvement in education I.

$$S = \Pi(q, cc, hc, p), \tag{3.4}$$

and

$$I = \Omega(q, cc, hc, p), \tag{3.5}$$

where p is a vector of prices (for schooling, inputs and other goods and services).

Inserting (3.4) and (3.5) into (3.2), we have

$$A = \Phi(q, cc, hc, p). \tag{3.6}$$

Equations (3.4)-(3.6) constitute causal relationships. That is, they inform us about the causal effects of changes in the exogenous variables in vectors q, cc, hc, and p on the quantity of

education (3.4), parents' financial involvement in schooling (3.5) and learning outcomes (3.6). The equations also constitute reduced form relationships. That is, they inform us how, through its effect on the endogenous variables, a change in some element in q or p in the end affects the endogenous variable of interest.

To illustrate this, compare equation (3.2) with equation (3.6). The former depicts the structural relationship between A and the various determinants. Consider a change in one element of q – call it q_1 (the provision of textbooks for example). Equation (3.2) then gives us the partial derivative, i.e. the change in A due to q_1 holding all other variables constant. A change in the same element in (3.6), on the other hand, gives us the total derivative, i.e. it allows for changes in S and I in response to the change in q_1.

For a policymaker, the reduced form estimates, or the total derivative, is typically of most interest since it informs the policymaker of how changes in q_1 actually influence A. Note, though, that this information alone may not be sufficient to evaluate the welfare effects of the policy change. For example, if publicly provided textbooks and textbooks supplied by parents (which will show up as changes in I) are substitutes, the total effect on learning may be small. Only observing estimates from (3.6), it would then be concluded that the provision of textbooks has little impact (although it would be correct to conclude that publicly provided textbooks have no effect on learning in this context). However, one would not be able to tell why this is the case. It could be because the provision of textbooks has a minor effect on learning, i.e. that both the partial derivative, dA/dq_1, from (3.2) and the total derivate, dA/dq_1, from (3.6) are small. It could also be the case that dA/dq_1 in (3.2) is large and positive, but that parents reduce their own supply of textbooks in response to the intervention, i.e. that $dI/dq_1 < 0$ in 3.5, leaving the total number of textbooks per student roughly unchanged. So while the intervention may have had little impact on learning, it presumably had a positive effect on parents' welfare.

Lack of knowledge about the structural relationship, or at least lack of knowledge about other key endogenous variables like I in this model, also makes it more difficult to extrapolate from a policy experiment because the behavioral response may vary across space and time. This is important to keep in mind, given that impact evaluations almost exclusively focus on policy parameters, i.e. the reduced form estimates.

4. Impact Evaluation: Empirical Methods

4.1 The Evaluation Problem

An impact evaluation attempts to address a causal question about the relationship between the variable (or policy) *T* and outcome *Y*. With no loss of generality, think of *T* as a binary variable indicating whether the individual participated (*T = 1*) or not (*T = 0*) in a program we want to evaluate, and *Y* as the outcome of interest. For example, if we want to evaluate a program that freely distributes insecticide treated bed nets to different communities, *T = 1* for a community (or individuals in a community) that benefits from the free distribution of bed nets and *T = 0* for communities which have not received bed nets. The outcome variable *Y* could then be a measure of the under-five mortality rate in the community.

Assume now that we have *N* units (this could be individuals, households, communities, or service delivery units like schools or clinics) and let *i* be an index for the unit in the population. In the above example, *i* would then indicate a specific community. Assume further that some units have participated in the program in question, or been exposed to treatment, and some have not. Let Y_i be the observed outcome and let Y_{i1} be the potential outcome of unit *i* in case of treatment ($T_i = 1$) and Y_{i0} be the potential outcome of unit *i* in case of no treatment ($T_i = 0$). We can now write the observed outcome Y_i for each unit in terms of potential outcomes as[2]

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0} = Y_{i0} + (Y_{i1} - Y_{i0}) T_i . \tag{4.1}$$

This expression requires thinking in terms of counterfactuals. We must be able to imagine what might have happened to someone who participated in the program if he/she had not participated and vice versa. In other words, we imagine two worlds for each unit, one world where the unit received treatment and one where it did not.

In reality, we do not observe what would happen to *i* under both *T* and *C* simultaneously. Is it possible to get around this problem by comparing outcomes for those units that received and did not receive treatment? That is, is it possible to make a causal statement about impact by comparing average effects in the two groups? In general, the answer is no. To see this, note that the differences in averages (or, formally, expectations), using the expression for potential outcomes, is

$$E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0] = E[Y_{i1} - Y_{i0} \mid T_i = 1] + E[Y_{i0} \mid T_i = 1] - E[Y_{i0} \mid T_i = 0]. \tag{4.2}$$

[2] Holland (1986).

The difference in expectations consists of two terms. The first term, $E[Y_{i1}-Y_{i0} \mid T_i=1]$, is the *average treatment effect on the treated*. This term captures the average difference in outcome between those who have been treated $E[Y_{i1} \mid T_i=1]$, and what would have happened to them had they not been treated, $E[Y_{i0} \mid T_i=1]$. The second term is the selection effect. The selection effect is the differences in (counterfactual) outcomes in the no treatment case between those that did and did not receive treatment.

To see what this implies, consider once more the example of insecticide treated bed nets with Y_i being morbidity from malaria in the community. If the program focused on providing bed nets in communities where the threat of malaria is particularly severe, or in communities where few households owned bed nets, the selection effect would result in a bias. That is, absent treatment, there would be differences in outcomes between those treated and those not treated. In the following sections, we will discuss several empirical methods that have been developed to overcome this selection bias.

4.2 Randomized evaluations

One method for solving the selection problem, i.e. where the selection bias has been entirely removed, is when the N units are randomly assigned to receive treatment.[3] In this case, the N units are randomly divided into two groups: the treatment group with N_T units and the comparison group with N-N_T units. Since treatment has been randomly assigned, units assigned to the treatment and comparison groups differ in expectations only through their exposure to treatment; that is, $E[Y_{i0} \mid T_i=1] - E[Y_{i0} \mid T_i=0] = 0$.

As the sample size increases, the difference in empirical means

$$(4.3) \quad \frac{1}{N_T}\sum_j^{N_T} E[Y_i \mid T_i=1] - \frac{1}{N-N_T}\sum_j^{N-N_T} E[Y_i \mid T_i=0] = \hat{E}[Y_i \mid T_i=1] - \hat{E}[Y_i \mid T_i=0]$$

will converge to

$$(4.4) \quad E[Y_{i1} \mid T_i=1] - E[Y_{i0} \mid T_i=0].$$

[3] The method, design, and various other methodological and practical issues with respect to randomized evaluations are discussed in detail in Duflo, Glennerster and Kremer (2007).

Further, if the potential outcomes of a unit are unrelated to the treatment status of any other unit[4], it follows that $E[Y_{i0} \mid T_i=0] = E[Y_{i0} \mid T_i=1]$, and the difference in empirical means further simplifies to

$$E[Y_{i1} - Y_{i0} \mid T_i = 1] = E[Y_{i1} - Y_{i0}], \tag{4.5}$$

the causal parameter of interest for treatment.

The role of random assignment can be restated in terms of a regression model with a single explanatory variable

$$Y_i = a + bT_i + e_i, \tag{4.6}$$

where T_i is a dummy for assignment to the treatment group and e_i is an error term. Equation (4.6) can be estimated with OLS. The OLS estimator of ß, $\hat{b}_{OLS}$, is

$$\hat{b}_{OLS} = \hat{E}[Y_i \mid T_i = 1] - \hat{E}[Y_i \mid T_i = 0], \tag{4.7}$$

that is, the difference in empirical means.

When a randomized evaluation is correctly designed and implemented, it provides an unbiased estimate of the impact of the program in question. For this reason, a randomized experiment is usually considered to be the gold standard for studying causal effects. When implementing a randomized experiment in the field in a developing country, however, it may not be possible to fully replicate the simple example above. In the following sub-sections, we will discuss the most common problems and design issues that may arise. Before that, we will briefly discuss the issue of how randomized evaluations can be introduced in the field.

4.2.1 When are randomized evaluations appropriate/possible?

To date, most randomized evaluations have been implemented during the pilot phase of a program, often involving an NGO partner, although one of the most well-known randomized evaluations, PROGRESSA (now called Oportunidades) was conducted by the Mexican government. From the financier's perspective, such a pilot phase evaluation will not only tell if and to what extent the project works, but can also counteract potential skepticisms about how funds are being used. Below, we briefly discuss several settings for when randomization is potentially possible.

[4] This is the so-called Stable-Unit-Treatment Value Assumption (SUTVA) from Rubin (1978).

Randomization can also be introduced when there are limited resources or implementation capacity. In this case, the program may in need of being phased in over a number of years anyway. With high demand across many potential recipients, a fair method for determining the order in which the recipients receive the project is to randomize who receives the project in the first phase. The later recipients can then, in the initial phase, constitute the comparison group (as in Miguel and Kremer, 2004). Similarly, if demand exceeds supply, a fair way of rationing resources is to select those who will receive the program through a lottery, as was done in Karlan and Zinman's (2008) evaluation of the impact of expanded consumer credit in South Africa.

In some cases, for ethical reasons, for example, it may not be possible to get cooperation from participants in the comparison group, even if they will benefit from the project later on. In this case, it is still possible to introduce an element of randomization by providing the program to some subgroup in each area. Banerjee et al's (2007) evaluation of a remedial education program in India is an example of this. In that case, all schools participated in the project. However, the evaluators randomly assigned the remedial education to different grades in different schools, thus ensuring a treatment and comparison group for each grade.

Even programs that are available to all units in a study area could, in some cases, be evaluated using a randomized design. For example, if the take-up of a program is not universal, i.e. if everyone does not participate (or does not participate fully), one could randomly choose who to encourage to participate. An important difference from the above example is that the so-called encouragement design does not involve randomizing over treatment itself; instead the evaluators randomly assign subjects an encouragement to receive treatment. An example of this approach is Duflo, Kremer, and Robinson's (2006) evaluation of different interventions to understand the adoption of fertilizers in Western Kenya.

4.2.2. Design and implementation issues

While still representing a small fraction of all impact evaluations, randomized experiments have become a popular method in development economics. Unlike social experiments in the US, many of these experiments have been implemented with a fairly small budget by working with local partners. In this and the following sub-sections, we will discuss the most common design issues and problems that may arise when implementing randomized experiments in the field.

Level of randomization. In the evaluation of new drugs and vaccines, treatment is in general randomly allocated to individual subjects. However, many social experiments and field trials in medicine are not randomized to individuals but to intact groups or clusters (could be a village, or a service delivery unit with the corresponding catchment population) – a so-called cluster-randomized experiment. In some cases, randomization is introduced at the cluster level by necessity. For example, due to their nature, some interventions must be implemented at the community level (like water and sanitation schemes or an intervention that focuses on processes such as community monitoring) and therefore, there is no room for randomization within the cluster. Cluster-randomized evaluations may also be preferred for logistical convenience or to avoid resentment among treatment and control subjects or towards the implementation organization. However, in other cases, there might be a choice between randomizing at the individual and the group level. In these cases, a couple of factors need to be taken into account:

(a) The level of randomization potentially has large implications for the power of the experiment (as discussed in section 6). Individual-level randomization requires a smaller sample size to detect an effect with a given level of power and a statistical significance level.

(b) Spillovers from treatment to comparison groups can bias the estimation of treatment effects. In such cases, the randomization should occur at a level capturing these effects.[5]

(c) Randomizing at the group level can be one way of addressing concerns with partial compliance.

Control variables. When using data from a randomized experiment, other factors need not be controlled for. By construction, these controls, call them X_i, are uncorrelated with the treatment indicator T_i and thus, will not affect the estimate of $ß$. Nevertheless, the inclusion of control variables may generate more precise estimates of the treatment effect $ß$, since the inclusion of controls reduces the variance of the error term.[6] This effect will be larger the more explanatory power the control variables have. One implication of this is that for outcome variables that are persistent (like test scores), controlling for baseline outcomes Y_{it-1}, may greatly improve the precision of the treatment effect. Importantly, one must only control for

[5] Miguel and Kremer's (2004) evaluation of deworming drugs in Western Kenya is an example where these concerns led the researchers to randomize at the school level rather than at the individual level.

[6] The variance of $ß_{OLS}$ is $s^2(X'X)^{-1}$, where X is a matrix of all covariates including the treatment indicator. As noted in Duflo et al (2006), the inclusion of control variables may increase $(X'X)^{-1}$ and thereby increase the variance of $ß_{OLS}$.

pretreatment variables. Controlling for variables that are affected by the experiment will bias the estimate of the treatment effect.

Stratification. Stratification, that is, the use of pretreatment characteristics to stratify the sample, can be used to improve the precision of the estimates. This involves decomposing the full sample into smaller subgroups that share similar characteristics (for instance, being located in the same geographical area). As discussed in Duflo et al (2006) and Cox and Reid (2000), stratification will improve precision to the extent that the variable used for decomposing the sample explains the variation in the treatment of interest. In many social experiments, the sample size is relatively small. In this scenario, while randomization ensures that treatment and control groups will be similar in expectations, stratification will ensure this to also be true in practice. The regression equation X can be adjusted by adding the variables used to stratify the sample as additional controls.[7]

Factorial designs. Many programs involve more than one component. For instance, Miguel and Kremer's (2004) deworming intervention involved both the distribution of deworming pills and health education. Likewise, Bjorkman and Svensson's (2007) study of a community-monitoring project in the primary health care sector in Uganda involved both the provision of baseline information and encouragement of participation. When there is uncertainty about the fact that either component, in itself, may make a large difference, it makes sense to first evaluate the combined package and later follow up with studies trying to disentangle the effects. The combined effect is also typically what interests policymakers, at least as long as the combined package it possible to scale up. If the budget so allows, it is also possible to test both the individual components and the combined package at the same time. This would involve testing different treatments simultaneously.

Baseline survey.[8] In principle, since randomization ensures that the treatment and control groups are similar in expectations, there is no need for a baseline survey. Nevertheless, there are several reasons why a baseline provides a potentially high value.

(a) The inclusion of baseline controls will typically generate more precise estimates of the treatment effect, thereby reducing the required sample size. Baseline controls may also be needed in order to stratify the sample. On the other hand, a baseline survey has budgetary

[7] See Duflo et al. (2006) and Imbens et al. (2006) for a more thorough discussion of the issue of stratification and control variables.

[8] Data issues and power calculations are discussed in section 6.

implications and in some cases other data, including administrative data, may substitute for a baseline survey.

(b) A baseline survey expands the possibilities to study heterogeneous effects by looking at the interaction between initial conditions and the impact of the program.

(c) A baseline survey allows the evaluator to test if the randomization was properly conducted by looking at differences between the treatment and control groups in baseline characteristics. If the randomization is successful, there should be no difference between treatment and control groups in baseline characteristics (or at least no systematic differences).

(d) A baseline survey also allows the evaluator to refine the data collection process.

4.2.3. Potential Problems

In practice, there is a number of potential problems with experiments. These can be divided into problems with internal or external validity. A statistical study is said to be internally valid if the statistical inference about causal effects is valid for the population studied, i.e., there is no correlation between treatment and the error term, while a statistical study is said to be externally valid if its inference and conclusions can be generalized from the population and setting studied to other populations and settings.

4.2.3.1. Threats to internal validity

Failure to randomize. Random assignment to treatment and control groups is the fundamental feature of a randomized experiment that makes it possible to estimate the causal effect. If treatment is not assigned randomly, but is instead partly based on the characteristics or the preferences of the subjects, the experimental outcomes will reflect both the effect of the treatment and the effect of nonrandom assignment. In general, nonrandom assignment leads to biased inference.

Failure to follow treatment protocol. In actual experiments, subjects not always do what they are told. Therefore, even if the treatment assigned is random, the treatment actually received might not be random. The failure of subjects to completely follow the randomized treatment protocol is called partial compliance with the treatment protocol. With partial compliance, the

treatment and control groups are no longer random samples from the larger population from which the subjects were originally drawn: instead the treatment and control groups have an element of self-selection. Failure to follow the treatment protocol leads to biased inference. Problems with partial compliance could be handled by using the variables that are being randomly manipulated (initial assignment) when analyzing outcomes. One can then either look at the reduced form (when non-compliance is part of the issue studied) or use initial assignment as an instrument. The latter estimator is often labeled Intention-To-Treat estimator (ITT).

Attrition. People may move during the experiment. If people who leave have particular characteristics systematically related to the outcome, then there is attrition bias (non response bias). Attrition also reduces statistical power. The way of minimizing problems of attrition is to ensure that all (most) participants in the two groups are tracked during data collection.

Experimental effects. In experiments with human subjects, the mere fact that the subjects are in an experiment can change their behavior, a phenomenon sometimes called the Hawthorne effect.

4.2.3.2. Threats to external validity

A statistical study is said to be externally valid if its inference and conclusions can be generalized from the population and setting studied to other populations and settings.

Non-representative sample. The population studied and the population of interest must be sufficiently similar to justify generalizing the experimental results. An example of when a non-representative sample might arise is when the experimental participants are volunteers. Even if the volunteers are randomly assigned to treatment and control groups, these volunteers might be more motivated than the overall population and, for them, the treatment could have a greater effect. More generally, non-randomly selecting the sample from the greater population of interest can compromise the ability to generalize the results from the population studied to the population of interest.

Non-representative program or policy. The policy program of interest must also be sufficiently similar to the program studied to permit generalizing the results. One important feature is that the program in a small-scale, tightly monitored experiment could be quite different from the program actually implemented. Another difference between an experimental program and an actual program is its duration: the experimental program only lasts for the length of the

experiment, while the actual program under consideration might be available for longer periods of time.

General equilibrium effects. An issue related to scale and duration concerns what economists call "general equilibrium" effects. Turning a small, temporary experimental program into a widespread, permanent program might change the economic environment sufficiently so that the results from the experiment cannot be generalized. Phrased in econometric terms, an internally valid small experiment might correctly measure a causal effect, holding constant the market or policy environment. General equilibrium effects mean that these other factors are not, in fact, held constant when the program is broadly implemented.

Treatment vs. eligibility effects. Another potential threat to external validity arises because, in economics and social programs more generally, participation in the actual (non-experimental) program is voluntary. Thus, an experimental study that measures the effect of the program on randomly selected members of the population will not, in general, provide an unbiased estimator of the program effect when the recipients of the actual implemented program are permitted to decide whether or not to participate

4.2.3.3. Threats to power

Small samples. Because experiments are difficult to administer, samples are often small, which makes it difficult to obtain significant results. It is important to compute power calculations before starting an experiment (what is the sample size required to be able to discriminate an effect of a given size from 0?) and adjust the sample size accordingly (or possibly abandon the evaluation).[9]

Experiment design and power of the experiment. When the unit of randomization is a group (e.g. a school), we may need to collect data on a very large number of individuals to get significant results, if the outcomes are strongly correlated within groups (see section 6).

4.3. Review of non-experimental methods used in economics

This section will review the issues in causal modeling primarily using the framework adapted in economics, i.e., modeling by the use of regression equations with explicitly represented error

[9] See section 6.

terms. When boiled down to essentials, the economic framework is observationally equivalent to the potential outcomes model used in the previous section to describe randomized trials. The reason why we do not use the potential outcomes notation is that most empirical work only uses non-experimental or observational data. Moreover, many studies have multi-valued or continuous treatments for which regression methods are well suited since regression coefficients have an "average derivative" interpretation. The remainder of section 4.3. is structured as follows.

Section 4.3.1. describes the most common method used in empirical work, the regression-control framework, while the closely related matching method is discussed in section 4.3.2. These two methods share the same identifying assumption since they both use selection on observables, i.e., all relevant confounding factors are known and precisely measured.

In sections 4.3.3. and 4.3.4., difference-in-differences and fixed effects methods are described. These methods share the same basic identifying assumption since both approaches assume there to be omitted confounding variables that cannot be observed. Nevertheless, since these omitted confounders are supposed to be time invariant, data with a time dimension can be used to control for the unobserved but time invariant confounders.

Section 4.3.5. discusses the instrumental variable approach that uses the assumption of the existence of a variable, i.e., an instrument, which is assumed to be unrelated to any unmeasured confounding factors to estimate the causal relationship of interest.

In section 4.3.6., we discuss regression-discontinuity methods which can be seen either as a selection-on-observable approach as in the case of the sharp RD design or an IV approach as in the fuzzy RD case.

4.3.1 Regression-control method

Most impact evaluation studies are based on non-experimental data and, at least historically, a regression-control framework has been the most common method used to deal with selection biases (omitted variable biases) with observational data. Typically, a multiple regression of the following form is estimated

(4.8) $$Y_i = ?_0 + ?_1 W_{1i} + ?_2 W_{2i} + \ldots + ?_K W_{Ki} + \nu_i ,$$

where Y_i is some outcome of interest of unit i, W_1, W_2 ,…, W_K are independent variables and ν is the regression error. In impact assessment work, typically, only one regressor in equation (4.8) is

usually of direct interest while the others regressors are best considered as controling for confounding factors, i.e., they have no causal interpretations. Thus, it makes sense to restate equation (4.8) as

(4.9) $$Y_i = ?_0 + b X_i + ?_2 W_{2i} + ?_3 W_{3i} + \ldots + ?_K W_{Ki} + v_i ,$$

where X is the regressor of direct interest (corresponding to T above) and parameter b is the causal effect.

To fix the ideas, we assume that the causal effect b is the same for everybody in the population, $b_i = b$ for all i, and that X only takes two values (below we will discuss how the analysis is affected by these assumptions). In other words, either one is "treated" (X=1) or one is not treated (X=0). Causal inference can be made if the following assumption is valid

(4.10) $$E(v \mid X, W_1, W_2, \ldots, W_K) = E(v \mid W_1, W_2, \ldots, W_K) .$$

This assumption is called conditional independence (CI) or "selection on observables".[10] If this assumptions is valid, then the OLS estimator of the causal effect will be unbiased and consistent, i.e., plim $\hat{b} = b$. It is useful to consider three cases when the CI assumption holds.

(i) When the classical least square assumption holds, i.e., E(u | X, W_1, W_2,..,W_K)=0.

(ii) If X is randomly assigned.

(iii) If X is randomly assigned conditional on the W's.

It is noteworthy that the CI is weaker than the zero conditional mean assumption since the W's are allowed to be correlated with the error. A regression-control framework might therefore provide an unbiased and consistent estimate of b if the CI holds.

There is an alternative method for estimating the causal effect b in regression (2) that is useful in understanding in what way a regression-control framework can potentially solve any selection or omitted variable bias. This alternative method is based on a partitioned regression originally derived by Frisch and Waugh (1933), where the estimate of b can be computed in two steps. The first step is to regress X on all other control variables W_1, W_2,.., W_K and obtain the residual $\hat{u}$. The second step involves regressing Y on the residual $\hat{u}$, which yields the estimate of b since

(4.11) $$\hat{b} = \frac{Cov(Y_i, \hat{u}_i)}{V(\hat{u}_i)}.$$

[10] The assumption is sometimes called ignorability of treatment (given the observed covariates W) or unconfoundedness.

This regression formula gives a demonstration of parameter $\hat{b}$ as having a partial effect interpretation since all observable confounding variables have been netted out from the variation in X. In other words, it is what is "left over" in the variation in X that constitutes the identifying variation of b. This is also the reason why the identifying assumption is called selection-on-observables.

The concepts of internal and external validity discussed in section 4.2.3 provide a useful framework for evaluating whether a regression-control study is useful for answering a specific causal question of interest. An empirical result is said to be internally valid if the estimated regression coefficient is unbiased and consistent, while it is externally valid if the results can be generalized to other populations than that being studied.

A key question in any regression-control study is whether the results are internally valid, i.e., whether the CI assumption is plausible. The CI assumption clearly makes sense when there is an actual random assignment conditional on W. Even without a random assignment, however, CI might make sense if we know a great deal about the process generating the treatment. However, in many applications, the regressor of interest is not randomly assigned and we do not have any detailed knowledge about the process that actually determines the treatment. Thus, the choice of control variables is crucial, but which are the potential confounding factors that should be included in the population model? Economic theory does usually not specify what other variables should be held constant in order to isolate the primary effect of interest. For example, when we look at the impact of education on individual earnings, what else should be held constant? IQ, work effort, occupational choice, and family background etc. If we do not correctly include all relevant factors in the OLS regression, the CI assumption is in general not satisfied and therefore, the OLS estimator will be biased and inconsistent. In short, there will be selection or an omitted variable bias.

We can use the omitted variable bias (OVB) formula to describe the direction of the bias. The OVB formula describes the relationship between the regression estimates in models with sets of control variables. Suppose that the true regression can be written as

$$Y_i = b_0 + bX_i + ?W_i + v_i \,, \tag{4.12}$$

but (4.12) is estimated without the variable W. Since the OLS estimator for the regression equation without W is $\text{Cov}(Y_i, X_i)/\text{Var}(X_i)$, we can derive the OVB formula by plugging 4.12 into the OLS formula, yielding

(4.13) $$\frac{Cov(Y_i, X_i)}{V(X_i)} = ß + ?p_{wx},$$

where g_{wx} is the vector of coefficients from the regression of the elements of W_i on X_i. In words, this OVB formula states that the OLS estimate from the short regression equals the long regression (the true regression) plus the effect of omitted variables times the regression of omitted variables on included variables.

The OVB can now be used to get a sense of the likely consequences of omitting a variable for the direction of the bias of the OLS coefficient. To give a concrete example, let us once more consider the evaluation of a program that freely distributes insecticide treated bed nets to different communities. Assume that the bed nets were distributed in such a way that children with more educated parents received more bed nets. If one were to find a correlation between, say, the under-five mortality rate and bed nets, this does not mean that bed nets are causally related to child mortality since more educated families may be better at protecting their children from getting malaria independent of whether they received a bed net. In this case, one must control for the educational level of parents to estimate the causal effect of bed nets on child mortality.

The OVB formula also suggests that a simple approach to detect potential with a regression-control strategy is to check whether the regression results are highly sensitive to changes in the set of control variables. If the regression results are sensitive to changes in the set of control variables, there is reason to wonder whether there might be unobserved covariates that would change the estimates even further.[11] Thus, controlling for an insufficient number of factors may cause bias.

Less known is the fact that controlling for too many factors may also give rise to a bias if these variables are outcome variables themselves. For example, if wage and ability (as measured by IQ, for example) are both caused by education, then controlling for IQ in an OLS regression of wage on education will lead to a downward bias in the OLS coefficient of education. Intuitively, the ability variable picks up some of the causal effect of education, namely the increase in wages which is due to the effect of education on ability, which in itself affects wages. To avoid controlling for outcome variables, variables measured before the treatment was determined are generally valid control variables.

[11] See Altonji et al (2005) for a formal framework using this idea.

See Angrist and Krueger (1999) and Angrist and Pischke (2008) for discussions about the regression control framework.

4.3.2 Matching

A related approach to the regression-control framework is the matching approach. The key identifying assumption in both methods is selection on observables. An attractive feature of matching methods is that they are typically accompanied by an explicit statement of the CI assumption required to give matching estimates a causal interpretation. In contrast, work based on the regression-control framework typically does not explicitly state and discuss the CI assumption. Nevertheless, we have just seen that the causal interpretation of a regression coefficient is based on exactly the same CI assumption. Thus, since both methods depend on the knowledge that all confounding factors are known and quantified, one may therefore ask whether or to what extent matching differs from regression-control analysis.

In the matching approach, treatment effects are constructed by matching subjects with the same covariates, while a regression analysis uses a linear model for the effects of covariates. In practice, however, regression estimates can be understood as a type of weighted matching estimator as discussed by Angrist (1998). Thus, the difference between a regression-control analysis and a matching approach will typically not be of any major empirical importance. For this reason, the matching approach will only be discussed very briefly.

When the covariates take on many values, it becomes difficult to find good matches for each possible value of the covariates. A possible solution in this case is to match subjects in the control and treatment groups on the propensity score, i.e., the conditional probability of treatment given control variables. Rosenbaum and Rubin (1983) show that conditioning on the propensity score eliminates the omitted variable bias. The dimensionality of the matching problem is therefore reduced since the propensity score is scalar. However, there are many details to be filled in when implementing propensity score matching methods, such as how to model the propensity score and how to do inference. Since these procedures have not yet been standardized, there is a nontrivial chance that the results are sensitive to the precise implementation although the same data and covariates are being used. Therefore, matching is

most suitable when the covariates are few and discrete, since matches will be prefect. Moreover, propensity score methods are also exclusively used when the treatment is binary.[12]

For a further discussion of matching and the relationship to regressions, see Angrist and Kreuger (1999), Angrist and Piscke (2008) and Imbens (2004). A recent example of when this approach is used is Levinson et al's (2008) evaluation of the impact of HIV on labor market participation in South Africa.

4.3.3 Differences-in-differences

The difference-in-differences (DD) approach is a method for estimating the effect of policy interventions or other sharp changes in the economic environment. DD methods are used in problems with multiple subpopulations, where some subpopulations are subject to a policy intervention or treatment and others not. Outcomes are measured in each group before and after the policy intervention. To account for changes over time unrelated to the intervention, the change experienced by the group subject to the intervention (referred to as the treatment group) is adjusted by the change experienced by the group not subject to treatment (the control group). The underlying assumption is that the time trend in the control group is an adequate proxy for the time trend that would have occurred in the treatment group, in the absence of the policy intervention (i.e., the parallel trend assumption).

The DD method is useful for evaluating policy changes in environments where important underlying time trends may be present. The DD approach has been popular for evaluating government policy changes that take place in some administrative units, such as states, but not in neighboring units. An illustrative example is Duflo's (2001) study on the impact on schooling and labor market outcomes of a school construction program in Indonesia. She basically compares educational attainment for cohorts born before and after the school-construction program. The treatment group is the cohorts born in regions with a large number of newly built schools, while the control group is cohorts born in regions with a small number of newly built schools. Not surprisingly, she finds that the educational attainment has increased more in the treatment group relative to the control group. Moreover, she also finds that wages have increased more in the treatment group relative to the control group, which she solely attributes to the increased educational attainment in the treatment group due to the school construction program.

[12] Although it can be adapted to multi-valued treatments (e.g., Imbens 2000).

The DD design requires two years of data in the form of pooled cross sections, i.e. a new random sample is taken from the population each year. Let $\bar{Y}^{treatment,before}$ denote the sample average of Y_i for those in the treatment group before they have received treatment and let $\bar{Y}^{treatment,after}$ be the sample average of Y_i for those in the treatment group after they have received treatment. Similarly, let $\bar{Y}^{controlbefore}$ be the sample average of Y_i for those in the control group before they have received treatment while $\bar{Y}^{control,after}$ is the sample average of Y_i for those in the control group after they have received treatment. The differences-in-differences estimator is the average change in Y for those in the treatment group, minus the average change in Y for those in the control group

$$(4.14)\quad \hat{b}^{DD} = (\bar{Y}^{treatment.,after} - \bar{Y}^{treatment.,before}) - (\bar{Y}^{control.,after} - \bar{Y}^{control,before}) = \Delta\bar{Y}^{tretment} - \Delta\bar{Y}^{control},$$

where $\Delta\bar{Y}^{tretment}$ is the average change in y in the treatment group and $\Delta\bar{Y}^{control}$ is the average change in Y for the control group. The idea behind the difference-in-differences estimator is to correct the simple before and after difference for the treatment group by subtracting the simple before and after difference for the control group. Since the control group should reflect the counterfactual outcome for the treatment group, the DD-estimate is an unbiased estimate of the causal effect if, absent the treatment, the average change in Y (i.e., $\Delta\bar{Y}$) would have been the same for treatment and control groups. This is known as the "parallel trend" assumption.

To test whether $\hat{b}^{DD}$ is statistically different from zero, we can use a regression analysis. The DD estimator can be written in regression notation as

$$(4.15)\qquad Y_{igt} = \beta_1 + \beta_2 D_t + \beta_3 X_g + \beta_4 X_g \times D_t + u_{igt},$$

where D is a dummy variable equal to one in the post-policy intervention period and zero in the pre-policy intervention period, x is a binary treatment indicator equal to one if the subject is in the treatment group and zero if she is in the control group. The DD estimator is $\hat{\delta}_4$, since $\{E(Y \mid X=1, D=1) - E(Y \mid X=1, D=0)\} - \{E(Y \mid X=0, D=1) - E(Y \mid X=0, D=0)\} = \{(\beta_1+\beta_2+\beta_3+\beta_4) - (\beta_1+\beta_3)\} - \{(\beta_1+\beta_2) - \beta_1\} = \beta_4$. The advantage of formulating the DD estimator in this way is that it makes clear that the key identifying assumption is that there is no interaction between the time effect and the treatment group except for the treatment under study, i.e., $E(u \mid X \times D)=0$. In other words, the time effect D captures the way in which both the control and treatments groups are influenced by time and the fixed group effect X captures any fixed unmeasured differences between treatment and control groups. Thus, by comparing the time changes in the means for the

treatment and control groups, both group-specific and time-specific effects are allowed for in the DD method. The incorporation of the influences of other variables W_1, W_2,…, W_K is straightforward in the DD approach

(4.16) $$Y_{igt} = ?_1 + ?_2 D_t + ?_3 X_g + ?_4 X_g \acute{} D_t + p_1 W_{1igt} + p_2 W_{2igt} + \ldots + p_3 W_{Kigt} + u_{igt}.$$

This regression provides a simple way of adjusting for observable differences between the observations in the different groups. That is, the *W* variables account for the possibility that the groups have systematically different characteristics before and after the policy change i.e., they take into account compositional bias due to changes in the sample before and after the treatment.

One of the main pitfalls of the DD approach is the possibility of an interaction (besides the treatment) between treatment group and time (i.e., omitted interactions), implying that E(u | $X \times D$) ? 0. The DD approach is most plausible when the control group is very similar to the treatment group, so that interactions are less likely. It is useful to examine the size and significance of the estimated time effect $\hat{q}_2$ and group effect $\hat{q}_3$ for an indication of the comparability of the groups. These coefficients should be close to zero. For example, if the time effect is large in absolute value, it suggests that the period-to-period changes in *Y* are not unusual, since the time effect picks up the effect of omitted variables and trends in *Y*. A sizeable time effect suggests that the effects from these sources vary substantially across treatment and control groups i.e., there are likely to be omitted interactions. As previously mentioned, a situation favorable to the DD design is one where, both before and after, the control group has a distribution of outcomes close to that for the treatment group in the before period.

When the average levels of outcome *Y* are very different for control and treatment groups before the treatment, the magnitude or even the sign of the DD effect is very sensitive to the functional form posited. Suppose that you look at the effect on child mortality rates of providing bed nets. In one place, the mortality rate (under five) falls from say 140 to 100 while in another place it falls from 100 to 70. Because of the dramatic difference in pre-mortality levels (140 vs. 100), it is difficult to assess whether the treatment was effective. The DD estimate in levels would be (140 - 100) - (100 - 70) = 10, which suggests a positive effect of providing bed nets, while the DD in logs would be [log(140) - log(100)] - [log(100) - log(70)] = log(1.4) - log(1.43) < 0, which suggests that bed nets have a negative effect on child mortality.

DD estimates are more reliable when you compare outcomes just before and after the policy change, because the identifying assumption (parallel trends) is more likely to hold over a short time period. With a long time period, many other things are likely to happen and confound the treatment effect. However, for policy purposes, it is often more interesting to know the medium or long-term effect of a policy change. In any case, one must be cautious in extrapolating short-term responses to long-term responses.

Another important concern for the validity using a DD approach is whether the program is implemented based on pre-existing differences in outcomes. For example, it is common to compare wage gains among participants and non-participants in training programs to evaluate the effect of training on earnings. However, Ashenfelter and Card (1985) note that training participants often experience a dip in earnings just before they enter the program (which is presumably why they did enter the program in the first place). Since wages have a natural tendency to mean reversion, this leads to an upward bias of the DD estimate of the treatment effect.

Endogenous change in policy due to a governmental response to variables associated with past or expected future outcome is another threat to internal validity in the DD design. For example, a few years of very high infant mortality rates due to unusual circumstances, say draught, may stimulate some sort of policy intervention. A subsequent reduction in child mortality rates after unusual years should not be taken to indicate that policy intervention was effective, if a drop would have been expected anyway. The way to avoid the problems of endogenous change in policy is to know the circumstances surrounding the change.

A test for whether the results are likely to be internally valid; i.e. a check of whether the parallel trend assumption is likely to hold, is to use data in periods before the policy intervention and compute a DD estimate by comparing, say period t-1 with period t. If this DD estimate is nonzero, given that there was no policy intervention between period t and period t-1, it suggests that the original estimate will not capture the causal effect of the treatment since the treatment group and the control group did not have parallel trends in the outcome before the policy intervention. More generally, when data are available for many years, it is very useful to plot the series of average outcomes for treatment and control groups and see whether trends are parallel and whether there is a sudden change for the treatment group just after the reform.

The DD approach can be strengthened by the use of additional control groups since they reduce the importance of biases or random variation in a single control group. If the DD estimate with the alternative control group is different from the DD estimate with the original control group, the original DD estimate is likely to be biased. For example, Duflo (2001) uses more than one control group in her study of the school construction program in Indonesia.

The use of additional outcome measures is another important robustness check. The idea is to replace Y by another outcome that is not supposed to be affected by the treatment. If the DD estimate using the other outcome is non-zero, it is likely that the DD for the original outcome is also biased. The DD approach is also strengthened by the presence of several distinct groups that are subject to the treatment. Especially useful are treatment groups in different settings such as different time periods or treatment groups receiving treatments of differential intensities.

See Meyer (1995), Angrist and Kreuger (1999) and Angrist and Pischke (2008) for more discussions about the DD approach.

4.3.4 Fixed effects methods

A related approach to DD is the fixed effect approach where the data is in the form of a panel instead of repeated cross sections. Panel data consists of observations on the same units in two time periods or more. If the data set contains observations on variables X and Y, the data is denoted (X_{it}, Y_{it}) where i=1,2…, N and t=1, 2, …,T. Suppose that we have the following population model

(4.17) $$Y_{it} = ß_0 + ßX_{it} + c_i + u_{it},$$

where c_i is an unobservable random variable that is time-constant. The variable c_i captures all unobserved, time-constant factors that affect Y_{it}. Without loss of generality, we set the coefficient on c_i equal to one since c_i is unobserved and virtually never has a natural unit of measurement (i.e., it would be meaningless to try to estimate its partial effect). An unobserved time constant variable is called an unobserved effect in panel data analysis. In the case of panel data on individuals, the unobserved effect can be interpreted as capturing features of the individual, such as cognitive ability, motivation or early family upbringing that are given and do not change over time. In the case of panel data on service providers (e.g. health clinics or schools), c_i contains unobserved provider characteristics, such as managerial quality or structure, which can be considered as being (roughly) constant at least over the period of study. The unobserved effect c_i

is also referred to as a fixed effect or unobserved heterogeneity (or individual heterogeneity, firm heterogeneity, city heterogeneity, and so on). If the unobserved effect is correlated with the regressor of interest, i.e., Cov(X_i, c_i)?0, this will lead to an omitted variable bias in the OLS estimate (also called heterogeneity bias). One possible solution to the omitted variable problem is to find a suitable proxy variable for c_i, but this will not be a convincing solution since it is typically difficult to get a good measure of the unobserved effect.

In such a case, panel data offers a much more compelling solution since we can eliminate any influence of c_i on $Y_{it,}$ even without being able to observe and measure these time constant factors. Hence, panel data allows us to control for any time-constant omitted variables that may otherwise lead to an omitted variable bias in a pure cross-section analysis. Nevertheless, using panel data comes at a price, namely that only variables causing time variance can be used in the empirical analysis.

The simplest kind of panel data is to have two years of data for some cross section of units. Call the two periods $t=1$ and $t=2$. These years need not be adjacent, but $t=1$ corresponds to the pre-treatment year. Suppose that at $t=1$, no units have received treatment while at $t=2$, some units are in the control group and others in the treatment group. Let $?Y_{it}$ be the change of the value of Y_{it} from $t=1$ to $t=2$, that is, $?Y_{it} = Y_{i2} - Y_{i1}$. We can now obtain a panel data version of the DD estimator by estimating

(4.18) $$?Y_{it} = ß_0 + ß?X_{it} + u_{it} = ß_0 + ßX_{it} + u_{it},$$

where we have used the fact that $?X_{it} = X_{it}$ since treatment only takes place in period 2. The OLS estimator of ß will therefore be similar to the DD estimator in the previous section, since it is the difference in group means of $?Y$. However, there is an important difference between the DD estimators with panel data as compared to the DD estimators with repeated cross-section data. The important difference is that we can difference the outcome across the same cross-sectional units, which allows us to control for unobserved heterogeneity across units, whereas in the pooled cross-section data case we can only control for unobserved heterogeneity across groups.

It is also possible to generalize this simple DD panel data estimator to the case of more than two periods, say T periods, where subjects get treatments in any period

(4.19) $$Y_{it} = c_i + ßX_{it} + l_t + v_{it},$$ i=1,2…, N and t=1, 2, …,T,

where c_i is an unobserved fixed effect and l_t is a fixed-time effect. Any transformation of the data that eliminates the unobserved c_i can be used to estimate the above regression model.

Nevertheless, the two most popular ways of estimating panel data models are a first-differencing estimation (FD) or a fixed-effects estimation (FE). In the case of T=2, the estimates from the FD and FE are identical but they will differ for T=3. Although the FE and FD should asymptotically be the same when there are more than two time periods, the most popular approach in applied work is the FE transformation. This is partly due to the fact that the FE estimator is more efficient if the errors are homoscedastic and serially uncorrelated but also due to the fact that the FE is less sensitive to violation of the strict exogeneity assumption, especially with large T, than an FD estimator.[13]

To see how the FE transformation works in practice, suppose that we have the following population model

(4.20) $$Y_{it} = ßX_{it} + c_i + v_{it}$$

where, for expositional ease, we have excluded the time effects. The FE transformation is obtained by first averaging the above equation over, t=1, 2,…,T, in order to get the cross-section equation

(4.21) $$\overline{Y}_i = b\overline{X}_i + c_i + \overline{v}_i$$

where $\overline{Y}_i = T^{-1}\sum_{t=1}^{T} Y_{it}$ and so on. Subtracting equation (4.20) from (4.22) for each t gives the FE transformed equation

(4.22) $$Y_{it} - \overline{Y}_i = b(X_{it} - \overline{X}_i) + v_{it} - \overline{v}_i .$$

The time demeaning of the original equation has removed the unit fixed effect c_i. The FE is also called the within estimator, because it uses the time variation in Y and within each cross-sectional observation, i. We can estimate equation (4.22) by pooled OLS but the standard errors and test statistics would have to be correct since pooled OLS do not use the correct degrees of freedom. However, we can use another procedure called the dummy variable regression that leads to the same estimate but also produces the correct standard errors and test statistics if the errors v_{it} are homoscedastic and serially uncorrelated. The dummy variable regression amounts to include a separate dummy variable for each of the N units in (4.20).

[13] Both the FE and the FD estimators assume that causing variable X_{it} is strictly exogenous conditional on the unobserved effect, namely for each t, the expected value of the idiosyncratic error u given X in all time periods and the unobserved effect is zero.

An illustration of this method is Reinikka and Svensson (2007), which exploits a four-year panel of school data to assess the impact of reducing corruption (due to a transparency campaign) on schooling.

There are some potential problems with a fixed-effect approach. The first is that it can only be used when there is time series variation in the regressor. Another problem is that the bias due to any measurement error in the regressor is usually aggravated. For example, many economic variables tend to be persistent (health status) while measurement errors often change from year-to-year (health status may be misreported or miscoded this year but not the next year). Thus, the observed year-to-year changes in the economic variables may mostly be noise and therefore, there is likely to be more measurement error in fixed effect regressions than in cross-sectional regressions. The lagged dependent variable also creates a bias unless the time period is large and the FE estimator is used. Finally, there is typically little theoretical support that suggests unobserved characteristics to be constant over time. See Wooldridge (2008) for a discussion of the fixed-effect approach.

4.3.5 Instrumental variables

The instrumental variable (IV) approach is a general method for obtaining a consistent estimator of the unknown parameter of the population model when the regressor is correlated with the error term. Thus, IV works when there are omitted variables, measurement errors, simultaneity and selection bias while the regression-control framework, DD or fixed effect approaches will not solve any measurement errors or simultaneity issues. The IV approach can be described in a number of ways. We will start by describing it in terms of the traditional regressions language and the potential outcomes framework.

Before we formally describe the IV, it is useful to give an intuitive understanding of the workings of the IV method. One can think of the variation of a regressor as having two parts: one part that is correlated with the error term and another part that is not correlated with the error term. The IV method only uses the uncorrelated part to identify the parameter of interest. To exemplify this idea, suppose that we are interested in the following population model

(4.25) $$Y = b_0 + bX + u ,$$

but the regressor is correlated with the error term, i.e., $E(u \mid X) ? 0$, thereby implying that the OLS estimator is not consistent. Assume that we have another variable Z, which is called an

instrumental variable. The instrument must be uncorrelated with the error term u, Cov(Z, u)=0 but must be correlated with the endogenous variable X, Cov(Z, X) ? 0. The IV is considered as a two-step procedure (which is why it is often labeled two-stage least squares, 2SLS). The first stage decomposes X into two components: a problematic component that may be correlated with the error term and a problem-free component that is uncorrelated with the error. The second stage uses the problem-free component to estimate parameter b. The first step begins with a population regression linking X to Z, i.e.,

(4.26) $$X = a_0 + a_1 Z + v .$$

This regression provides the required decomposition of X. The first stage decomposes X into two components, one problematic component v that is related to the error term u and another component, $a_0 + a_1 Z$, that is unrelated to u since Z is exogenous. The idea of the IV method is to only use the problem free component of X to estimate parameter b. The only complication is that the values of a_0 and a_1 are unknown so that $a_0 + a_1 X$ cannot be calculated. Nevertheless, we can use OLS to estimate equation (4.26) and calculate the predicted value $\hat{X} = \hat{a}_o + \hat{a}_2 Z$ where $\hat{a}_o$ and $\hat{a}_1$ are the OLS estimates. The second stage is to regress (using OLS) Y on $\hat{X}$. The resulting estimator from the second stage is the IV estimator $\hat{b}^{IV}$. This two-step procedure illustrates the basic idea behind the IV approach.[14]

We are now ready to more formally describe the IV method. The population model is $Y = b_0 + b X + u$, where X is endogenous. We have a valid instrumental variable Z, i.e., the instrument is exogenous and relevant. The first condition states that the instrument must be uncorrelated with the regression error while the second condition states that the instrument must correlated with the endogenous variable, i.e., Cov(Z, X) ? 0. With these two conditions, parameter b can be identified, that is, we can write β in terms of population moments that can be estimated using a sample of data. The population model together with the exogeneity of the instrument imply that we can write parameter b as

(4.27) $$b = \text{Cov}(Z, Y)/\text{Cov}(Z, X) .$$[15]

[14] Although the two-step procedure produces the IV estimate, it should not be used since the standard errors will not be correct as they do not take into account the sampling uncertainty of estimates of the first-stage parameters.

[15] This is derived using $\text{Cov}(Z, Y) = \text{Cov}(Z, b_0 + bX + u) = \beta\text{Cov}(Z, X) + \text{Cov}(Z, u)$.

This equation shows that b is the population covariance between Z and Y, divided by the population covariance between Z and X.

Given a random sample, we can estimate the population covariance by the sample analog

(4.28) $$\hat{b}^{IV} = S_{zy}/S_{zx},$$

where S_{zy} is the sample covariance between Z and Y and S_{zx} is the sample covariance between Z and X. Since the sample covariance is a consistent estimator of the population covariance, the IV estimator will be consistent, i.e., plim($\hat{b}^{IV}$) $=b$. In large samples, the IV estimator will also be normally distributed. Intuitively, this is because the IV estimator is an average of random variables and when the sample size is large, the central limit theorem tells us that averages of random samples are normally distributed. This means that we can perform a hypothesis test about β by computing t-statistics and a 95-percent large-sample confidence interval by $\hat{b}^{IV} \pm$ 1.96SE($\hat{b}^{IV}$).

So far, we have only used one regressor but the IV approach can, of course, be used with more than one explanatory variable and more than one instrument. In this case, the general IV population model (the structural equation) is the following

(4.29) $$Y = b_0 + b X_1 + b X_2 + \ldots + b X_r + ?_1 W_1 + ?_2 W_2 + \ldots + ?_k W_k + u,$$

where X_1, X_2, + X_r are r endogenous regressors (those that need to be instrumented), W_1, W_2,…W_k are k additional exogenous regressors (i.e., uncorrelated with u), and Z_1, Z_2, … Z_m are m instrumental variables. The regression coefficients are said to be exactly identified if the number of instruments (m) equals the number of endogenous variables (r). The coefficients are overidentified if the number of instruments exceeds the number of endogenous regressors, that is, $m > r$. The coefficients are underidentified if the number of instruments is less than the number of endogenous regressors, that is, $m < r$. The parameters must either be exactly identified or overidentified if they are to be estimated by IV regressions. The IV assumptions in the general case: exogeneity of the instruments Z_1, Z_2, … Z_m, i.e., E(u | Z_1, Z_2, … Z_m) = 0, and the relevance of the instrument; i.e., the instrument must be partially correlated with the endogenous regressors once all other exogenous variables W_1, W_2,…W_k have been netted out. In the case of one endogenous variable, say X_1, and multiple instruments, this condition can be formally expressed as

(4.30) $$X_1 = ?_0 + ?_1 Z_1 + ?_2 Z_2 + \ldots + ?_K Z_K + p_1 W_1 + p_2 W_2 + \ldots + p_K W_K + v,$$

where $?_1 = ?_2 = ... = ?_K ? 0$.[16]

An important cost of performing IV estimates when the regressor is not endogenous is that the asymptotic variance of the IV estimator is always larger, and sometimes much larger, than the asymptotic variance of the OLS estimator. In other words, the estimates from IV will always have larger standard errors than OLS, i.e., the estimates will be less precisely measured.

Another potential concern with the IV method is the problem of weak instruments i.e., when the instrument is only weakly correlated with the endogenous regressor. One way of thinking of instrument relevance is that it plays a role akin to sample size. The more relevant the instrument, that is, the more variation in the endogenous variable that is explained by the instrument, the more information is available for use in the IV regression. Thus, a more relevant instrument produces a more accurate estimator, just as a larger sample size produces a more accurate estimator. Statistical inference using the IV estimator is predicated on the IV estimator having a normal sample distribution, but according to the central limit theorem, the normal distribution is a good approximation for large – but not necessarily small – samples. If having a more relevant instrument is like having a larger sample size, this suggests that for the normal distribution to provide a good approximation of the sampling distribution of the IV estimator, the instruments should not just be relevant, but highly relevant. Instruments that explain little of the variation in the endogenous variable are called weak instruments. The 2SLS estimator is most biased when the instruments are weak and there are many instruments. In that case, the 2SLS estimator will be biased towards the probability limit of the corresponding OLS estimate.[17] The intuition for this is that in a small sample, even a valid instrument will pick up some small amounts of endogenous variation in X, and if one starts adding more and more irrelevant instruments, then the amount of random and hence, endogenous, variation in X will become increasingly important.

More generally, if the instruments are weak, the IV estimator can be badly biased and the normal distribution provides a poor approximation of the sampling distribution of the IV, even if the sample is large. The question is now how relevant must the instruments be for the IV

[16] In order to assess whether the instruments are weak or strong when there is more than one endogenous variable, it is necessary to look at a matrix version of the F-statistic, which assesses all first-stage equations at once. This is called the Cragg-Donald or minimum eigenvalue statistic. References can be found in Stock, Wright, and Yogo (2002),

[17] If the instrument is totally irrelevant, i.e., $\text{Cov}(Z, X)=0$, then the population parameter b is not even defined since $b = \text{Cov}(Z, Y)/\text{Cov}(Z, X)$.

estimator to be reliable? There is a simple rule of thumb in the case of a single endogenous regressor: compute the F-statistic testing the hypothesis that the coefficients on the instruments are zero in the first-stage regression of IV. This first-stage F-statistic provides a measure of the information content contained in the instrument: the more information content, the larger is the expected value of the F-statistic. If the F-statistic is larger than 10, there is no need to worry about weak instruments.

Turning to instrument exogeneity, if the instruments are not exogenous, then IV is inconsistent. Can we test whether the instruments are exogenous? The answer is basically no, since the assumption about instrument exogeneity involves the covariance between the instruments and the unobservable error term, u. Assessing whether the instrument is exogenous necessarily requires making an expert judgement based on personal knowledge of the empirical problem at hand. However, if there are more instruments than endogenous variables, it is possible to statistically test whether the other instruments are exogenous. This test is known as a test of overidentifying restrictions. Suppose that you have one endogenous variable but two instruments. Then two different IV estimates are computed. The two estimates will not be the same because of sampling variation, but if both instruments are exogenous, the estimates will be close to each other. However, if the two instruments produce very different estimates, there is something wrong with one of the instruments – or both. The test of overidentifying restrictions implicitly makes this comparison. The idea of the test is that the instruments should be uncorrelated with the error term, which suggests that the instruments should be approximately uncorrelated with the residual from the IV regression

$$\hat{u}_i^{IV} = Y_i - (\hat{b}_0^{IV} + \hat{b}_1^{IV} X_{1i} + \hat{b}_2^{IV} X_{2i} + ... \hat{g}_{Ki}^{IV} W_{Ki}). \quad (4.31)$$

One way of testing for overidentifying restrictions is to run the OLS regression $\hat{u}_i^{IV}$ on Z_1, Z_2,…Z_m, W_1, W_2,…, and W_K and compute the F-statistic from testing that all instruments are jointly zero.[18] The overidentifying restriction test statistic is J=mF. Under the null hypothesis that all instruments are exogenous, J is distributed as $?^2_{m-r}$, where m-r is the "degree of overidentification" i.e., the number of instruments minus the number of endogenous variables.

A recent example of an impact evaluation using an IV-strategy in a Sub-Saharan Africa context is Dinkelman (2008). She evaluates the employment effects of a mass roll-out of

[18] This test is known as a J-test, which was first developed by Hansen (1980).

household electrification in rural South Africa, using a land gradient that directly affects the cost of grid expansion as an instrument for project placement.

Sometimes it can be better to use OLS than IV if the instruments are very weak and there is some correlation between the instrument and the error term. (Here it is important to point out that IV is a consistent but not unbiased estimator). IV is consistent only if Cov(Z, u) = 0, that is

(4.32) $$\text{plim}\ \hat{b}_{IV} = b + \frac{Cov(Z,u)}{Var(X,z)} = b + \frac{Corr(Z,u)s_u}{Corr(Z,X)s_x}$$

while OLS is consistent only if Cov(X, u)=0

(4.33) $$\text{plim}\ \hat{b}_{OLS} = b + \frac{Cov(X,u)}{Var(X)} = b + \frac{Corr(X,u)s_u}{s_x}.$$

OLS can be preferred to IV on asymptotic bias grounds if the following inequality is valid

(4.34) $$\text{Corr}(X, \text{u}) > \text{Corr}(Z, u)/\text{Corr}(Z, X).$$

In practice, the most difficult aspect of an IV estimation is finding instruments that are both exogenous and relevant. There are two main approaches, which reflect two different perspectives on econometrics and statistical modeling. One way of generating instruments is to write down theoretical models. This strategy, which is known as structural model estimation, produces a framework that is complete (theoretical model and data application) and estimates that are fully meaningful in the context of the model. The other approach is the "experimentalist" approach. In the experimentalist approach, there was a search for some exogenous source of variation in the endogenous regressor of interest. The variation may come from a true randomized experiment but usually comes from some sort of quasi or natural experiment, that is, situations where human institutions or the forces of nature provide something close to a random assignment.

4.3.6 Regression-discontinuity designs

Regression-discontinuity (RD) methods exploit detailed knowledge of the rules determining the regressor of interest. The RD comes in two flavors: sharp and fuzzy designs. The sharp RD design is based on the selection on observable assumption (i.e. a regression-control framework) while the fuzzy design can be considered as an instrumental variable set up.

In the sharp design, the treatment rule is perfectly known since the treatment-determining rule is deterministic, i.e.,

(4.35) $$T = T(X) = 1 \text{ if } X = X_0,$$

where T is a treatment indicator, X is the treatment-determining or assignment variable, and X_0 is a known treatment threshold. This assignment rule means that treatment is a deterministic function of the treatment-determining variable, that is, $T = 1$ if $X = X_0$ or $T = 0$ if $X < X_0$. The basic idea in an RD design is to compare outcomes for subjects whose values of X are "just below" and "just above" the discontinuity X_0 since, on average, they will have similar characteristics except for the treatment. In other words, those subjects slightly below the threshold will provide the counterfactual outcome for those subjects slightly above, since the treatment status will be as good as random in a neighborhood of X_0. Thus, the causal inference from a regression discontinuity analysis can be as internally valid as those drawn from a randomized experiment.

There are a number of different ways of estimating the treatment effect in a sharp RD design. One approach would be to restrict the estimation close to the discontinuity X_0, which is basically the idea behind the non-parametric estimation approaches as discussed by Hahn el al (2001) and Porter (2003). A drawback of this method is that it requires large sample sizes close to the treatment threshold while in the typical application of the RD design, there are rather few observations around the discontinuity. As a result, the most common approach in applied work is to use a larger sample and try to model the relationship between the treatment determining variable and the outcome of interest. Since the assignment variable X is the only systematic determinant of treatment status T, this means that the conditional mean assumption CI will hold, i.e., $E[u|T, X] = E[u|X]$. We can then estimate the following regression using the entire sample of data

(4.36) $$Y = a + \beta T + f(X) + u.$$

The OLS estimate of β will be unbiased and consistent since $f(X)$ will capture any dependence between T and u, that is, $E(u \mid T)=0$. The problem is that we do not know the functional form of $f(.)$. In practice, f(.) therefore needs to be approximated and one popular approach is to approximate it with a low-order polynomial.

There is a number of specification checks that are often used in the RD approach. Many papers make a visual plot of the data to show the presence of a discontinuity in the outcome at

the treatment threshold (see, for example, Lee 2008 for an illustration). They also typically check whether the treatment estimate is sensitive to the specification of f(.), and examine the robustness of the results by restricting the sample to a subsample of observations more closely clustered around the cut-off. Another important check is to test whether individuals on either side of the cut-off are observably similar since if individuals can exercise control over their values of the assignment variable, then individuals just below and above the threshold will not be similar and will thus invalidate the RD design.

In the "fuzzy" RD design, the probability of treatment is no longer zero or one, as in a sharp RD design. However, there is still a jump in the probability of treatment at the treatment threshold. One approach in the fuzzy RD is to use the method of instrumental variables (e.g., Angrist and Lavy 1999, and Hahn et al 2001) where the instrumental variable Z is defined as Z=1 if $X = X_0$, and as zero otherwise. In other words, in the fuzzy RD design, the discontinuity becomes an instrumental variable for the treatment status instead of deterministically switching treatment on or off.

An illustration of the regression-discontinuity method is (Barrera-Osorio, Linden and Urquiola, 2007). They evaluate the impact of a fee reduction program launched by the city of Bogota in 2004. The program was targeted using a proxy-mean index, implying that the probability that households benefit from the fee reduction was a discontinuous function of their proxy-mean score, allowing the authors to implement a regression discontinuity design.

See Hahn et al (2001), Imbens and Lemiux (2008), and Angrist and Pischke (2008) for more discussions of the RD approach.

5. Other issues in causal modeling

In this section, we discuss the implications of population heterogeneity, i.e., the causal effect varies across the population and across issues regarding statistical inference.

5.1 Heterogeneous effects

Previously, we assumed a constant causal effect, that is, $Y_i = b_0 + b_i X_i + u_i$, with $b_i = b$ for all i. In reality, the causal effect can vary from one subject i to another, based on the subject's circumstances, background and other characteristics.

In the case of population heterogeneity, we can consider b_i as a random variable which, just like u_i, reflects unobserved variation across units. If there is population heterogeneity, the OLS estimator of b is still a consistent estimator of an average treatment effect. Specifically, if X is uncorrelated with the error term u, then the treatment effect is the causal effect among those who receive the treatment (TOT), while if X is randomly assigned the treatment effect is the average causal effect in the population (ATE).

For the IV estimator, the situation is more complicated when the population is heterogeneous. To illustrate IV with heterogeneous causal effects, suppose Z_i to be a valid instrument and related to X_i by the linear model

(5.1) $$X_i = p_0 + p_i Z_i + v_i ,$$

where the coefficient p_i varies from one subject to the next. This equation is the first stage in a 2SLS with the modification that the effect on X_i of a change in Z_i is allowed to vary from one unit to the next. The 2SLS estimator is $\hat{b}^{IV} = S_{zy}/S_{zx}$ where S_{zy} is the sample covariance between Z and Y and S_{zx} is the sample covariance between Z and X. Suppose that (i) p_i and $ß_i$ are distributed independently of u_i, v_i and z_i, (ii) $E(u_i \mid Z_i) = E(v_i \mid Z_i) = 0$ and $E(p_i) \neq 0$. Under these assumptions,

(5.2) $$\text{plim}\,(\hat{b}^{IV}) = E(ß_i p_i)/E(p_i) .$$

The ratio in (5.2) can be interpreted as a weighted average of the individual causal effects, $ß_i$. The weights are p_i which measure the degree to which the instrument influences whether the i^{th} subject receives treatment. Thus, the 2SLS estimator is a consistent estimator of a weighted average of the subjects' causal effects, where the individuals who receive most weight are those for whom the instrument is most influential. To see this, consider two cases where the 2SLS estimator is a consistent estimator of the average causal effect and one case where it is not: (i) $b_i = b$ for all i (constant causal effect) (ii) $p_i = p$ for all i (the instrument affects each unit equally) and (iii) suppose that Z_i has no influence on the treatment decision for half the population, i.e., $p_i = 0$, and that Z_i has a constant influence for the other part. In the last case, 2SLS is a consistent estimator of the average treatment effect in the half of the population for which the instrument influences the treatment decision. To sum up, the 2SLS estimates a weighted

average of the causal effects, where the causal effects of the units that are most influenced by the instrument receive the greatest weight. This causal effect is also known as LATE (Local Average Treatment Effect), originally derived by Imbens and Angrist (1994).

5.2 Inference issues

The statistical analysis of cross-sectional data is based on the assumption that the data is independent, i.e., each observation is treated as a random draw from the same population, unrelated to the observation before or after. This sampling model is often unrealistic and analysts must also worry about the correlation between observations in cross-sectional and panel data. The most important form of dependencies is data with a group structure. This may give rise to a clustering problem (or the Moulton problem after Moulton 1986) if there is a correlation within groups, or it may give rise to a problem of serial correlation if the data is repeated cross-section or panel data.

However, before we start discussing the inference problems caused by clustering or serial correlation, we briefly discuss the implications of when the error terms are heteroskedastic rather than homoskedastic in independent samples. The error term is said to be homoskedastic if the variance in the conditional distribution of the error term given the regressor of interest is constant for all observations; otherwise the error term is said to be heteroskedastic. The OLS estimator remains unbiased and consistent, even if the errors are heteroskedastic. However, if the errors are heteroskedastic then the homoskedaticity-only standard errors are inappropriate. For example, the t-static computed using the homoskedaticity-only standard errors does not have a standard normal distribution, even in large samples. In such a case, one can compute heteroskedasticity-robust standard errors that would lead to valid statistical inference notwithstanding if the errors are heteroskedastic or homoskedastic.[19]

The main issue of practical relevance is whether one should use heteroskedasticity-robust standard errors or homoskedaticity-only standard errors. Most econometric textbooks suggest that one should use heteroskedasticity-robust standard errors since they are more reliable. However, the robust standard errors may be more biased in small samples than homoskedaticity-only standard errors when heteroskedasticity is modest. Thus, a simple rule of thumb is to

[19] These are known as Eicker-Huber-White standard errors.

compute both standard errors and use whichever is largest so as to avoid any gross misjudgements of statistical precision due to small sample problems.

Turning to the inference problem with grouped data, the clustering problem can be illustrated using a bivariate regression estimated in data with a group structure. Suppose that we are interested in the following relationship

(5.3) $$Y_{ig} = ß_0 + ßX_g + u_{ig} ,$$

where Y_{ig} is the outcome variable for individual i in cluster or group g, with G groups. Importantly, the explanatory variable, X_g, only varies at the group level. We model the correlation within groups as an additive random effect, i.e., $u_{ig} = v_g + e_{ig}$, where v_g is the random component specific to group g, and e_{ig} is the usual error term. When the explanatory variable only varies at the group level and there is a group random component, standard errors can increase dramatically.

In the case where the regressor is fixed at the group level, and the groups are of equal size n, it can be shown that the relationship between the OLS variance formula $V(\hat{b})^{OLS}$ and the corrected sampling variance formula $V(\hat{b})$ is $V(\hat{b})^{OLS} = V(\hat{b})/[1+ (n-1)r]$ where $r = s^2_v/(s^2_v + s^2_e)$ and s^2_v is the variance of v_g and s^2_e is the variance of e_{ig}. Parameter r is called the intra-class correlation coefficient. This equation tells us how much we overestimate precision by ignoring intra class correlation. To make a stark example, suppose that one makes n identical copies of a data set in order to increase its sample size. This is the same as assuming $r = 1$ and the OLS variance $V(\hat{b})^{OLS}$ should therefore be scaled up by a factor of n, since copying a data set does not generate any new information. There is a number of solutions to the clustering problem. Perhaps the most common solution is to compute cluster standard errors (e.g., using Stata cluster), but this method is only appropriate when there is a reasonably large number of clusters or groups. For example, Angrist and Piscke (2008) suggest that 50 clusters are typically sufficiently large for the statistical inference based on the clusters' standard errors to be reliable. Another popular solution is to use group averages instead of micro data. The standard errors that are based on group averages are more reliable than clustered standard errors in samples with few groups.

The inference problem caused by serial correlation can be illustrated in a panel data setting. Suppose that we have the following panel data regression

(5.4) $$Y_{it} = c_i + l_t + bX_{it} + u_{it},$$

where Y_{itt} is the outcome for individual i in year t, c_i is an individual fixed-effect and l_t is a time-fixed effect. Typically, observations for an individual tend to be correlated over time. This means that the error term u_{it} will be serially correlated and therefore, the standard errors would need to be corrected. If the number of individuals (groups) is large and the number of time periods is small (observations within groups), one can once more compute cluster standard errors (e.g., using Stata cluster).

Sometimes there are both clustering and serial correlation problems. This often occurs in difference-in-difference (DD) settings. Suppose that we have the following DD set up

(5.5) $$Y_{igt} = g_g + l_t + bX_{gt} + u_{igt},$$

where Y_{ig_t} is the outcome for individual i in group g in year t, γ_g is a group fixed-effect and λ_t is a time-fixed effect. We can consider the error term u_{igt} as the sum of a group-year shock, v_{gt}, and an idiosyncratic individual component, e_{igt}. Since there is a group-year shock in the error term and the regressor of interest X_{gt} also only varies at the group×year level, there will be a clustering problem. With only two groups and two time periods, as in many DD applications, there is no solution to the clustering problem. Even worse, if there are only two groups and two time periods, the DD estimator will not even be inconsistent if there are random group-year shocks. Intuitively, adding more individual observations to the four different groups (treatment group before, treatment group after, control group before, and control group after) does not help distinguish the causal effect from the random shock. The solution to the clustering problem is to have multiple time periods or many groups (or both). However, when there are more than two time periods there will typically also be a serial correlation problem in addition to the Moulton problem, since observation within groups tends to be correlated across time. In this case, the most important inference issue is the behavior of the common shock, v_{gt}. If the group-year shocks are serially uncorrelated, the standard errors on the by group×time can be clustered (e.g., using Stata cluster) to take into account any correlation within clusters (group×time). This takes into account the Moulton problem if there is a reasonably large amount of clusters. However, if the group-year shocks are serially correlated, the standard errors for the serial correlation in the v_{gt} themselves must be adjusted. There is a number of ways of doing this, not all equally effective in all situations. The simplest and most common approach is to cluster the standard at a higher level, i.e., the group level instead of by group×time. This means that we need to have a

large number of groups in this case since few clusters mean biased standard errors and misleading inference. The question of how to solve the serial correlation problem when the number of clusters is few is currently under study and a consensus has not yet emerged.

For papers discussing the Moulton and serial correlation problems see, for example, Donald and Lang (2007) and Bertrand et al (2004). For a textbook treatment, see Angrist and Pischke (2008).

6. Data and Power Issues

6.1. Data

There has been a spectacular increase in the availability and quality of data from developing countries in recent years. Many of these datasets are either in the public domain or can be obtained at a modest cost from the data collection agency.[20] While randomized evaluations rely on collecting original data through fieldwork, a lion's share of the evaluations based on quasi-experimental methods typically exploit already existing data sources. This second option has become more fruitful, given that over the last 10-15 years, high-quality, large scale, multipurpose data sets have become readily available. The World Bank's Living Standard Measurement Surveys (LSMS) and the Rand Corporation Family Life Surveys are two prime examples of this. Demographic and Health Surveys (DHS) are another source of fertility, mortality and health data. To date, DHS have been implemented in 40 countries in sub-Saharan Africa and, in several cases, with more than one round per country. Many developing countries also collect their own data, including large household survey data and the quality of these data has been steadily improving.

Census data from developing countries is available from the IPUMS-International web site (https://international.ipums.org/international/), although only a handful of countries from sub-Saharan Africa are included in their sample.[21]

6.2. Power issues

[20] The Bureau for Research and Economic Analysis of Development (BREAD) provides a useful link to data from developing countries at http://chd.ucla.edu/dev_data/index.htmllink.

[21] Survey design issues (as well as methodological issues regarding the analysis of household survey data) are discussed in great detail in Deaton (1997).

In case the impact evaluation requires the collection of new data, as is the case when conducting randomized experiments, power calculations are of great importance. In principle, power calculations should be conduced ex-ante to determine the necessary sample to obtain a given power. In practice, however, sample size is often largely determined by budget or implementation constraints. This raises the risks that the evaluators will make "type II errors"; i.e. not detect a significant difference of an intervention that would have been found to have a significant impact had the sample size been large enough.

The basic principle of power calculations can be illustrated in a simple regression framework.[22] Consider the regression model in (6.1)

(6.1) $$Y_i = a + bT_i + e_i,$$

where Y is a continuous outcome variable. The variance in the OLS estimator, $\hat{b}$, is

(6.2) $$\text{Var}(\hat{b}) = \frac{s^2}{\sum_i^n (T_i - \overline{T})^2},$$

where $s^2 = \text{Var}(Y)$ and $\overline{T}$ is the mean of T_i. If a fraction f of the sample belongs to the treatment group, (6.2) simplifies to

(6.3) $$\text{Var}(\hat{b}) = \frac{s^2}{nf(1-f)}.$$

Consider the decision rule to be used to determine whether an experiment has an effect of size ß. First, note that we would reject the null hypothesis of zero impact ($ß = 0$) if

(6.4) $$\frac{|\hat{b}|}{SE_{\hat{b}}} > t_a,$$

for a one-sided test (for a two-sided test t_a is replaced by $t_{a/2}$) where $SE_{\hat{b}}$ is the standard error of the OLS estimator. That is, if the impact estimate ß is greater (in absolute terms) that the critical value $t_a * SE_{\hat{b}}$, the null hypothesis of zero impact would be rejected. Now, consider what will happen if the true impact equals ß. Then, with ?% power, ?% of the sampling distribution of the OLS estimator of ß must lie above the threshold value, t_a,[23] that is

[22] For an introduction to power calculations, see Duflo et al (2006). A more in-depth treatment is provided in Bloom (2004) and Donner and Klar (2000). The exposition here partly follows from Blom (2004).

[23] A standard protocol in both social and medical sciences requires 80% power of detecting a significant difference at the 0.05 significance level for a given effect size, ß.

(6.5) $$|b| > (t_a + t_{1-k}) SE_{\hat{b}}.$$

Using (6.3), we can rewrite (6.5) to get the minimum required sample size *n* as,

(6.6) $$n = (t_a + t_{1-k})^2 \left(\frac{s^2}{b^2} \right) \left(\frac{1}{f(1-f)} \right).$$

When the treatment and control groups are of the same size, (6.6) is reduced to

(6.7) $$n = 4(t_a + t_{1-k})^2 \left(\frac{s^2}{b^2} \right).$$

In many cases, the outcome variable is not continuous (examples include child mortality, whether or not a student drops out of school or whether the person is infected by a STD). In this case, the formula for the required sample size must be slightly modified. Specifically, with a binary outcome measure, the disturbance term can only take two values. With probability P_l: $Y = 1$ and with probability $1\text{-}P_l$: $Y = 0$, where subscript *l* denotes treatment or control group. Using the fact that $E(e_i) = 0$, it follows from (6.1) that

(6.8) $$p(1 - a - bT_i) + (1 - p)(-a - bT_i) = 0,$$

implying that $P = a + ßT_i$. Therefore, an estimate of s^2, for simplicity assuming that $f = 0.5$, is

(6.9) $$\hat{s}^2 = \frac{1}{2}\left[(P_T(1 - P_T) + P_C(1 - P_C) \right],$$

implying a required sample size of

(6.10) $$n = 2(t_a + t_{1-k})^2 \left(\frac{P_T(1 - P_T) + P_C(1 - P_C)}{(P_T - P_C)^2} \right),$$

where $ß = (P_T - P_C)$.

As discussed in section 4, many randomized impact evaluations are not randomized to individuals but to groups or clusters. However, the evaluator may still have access to individual data. As discussed in section 5, when analyzing individual data from programs randomized at the group level, it is important to take into account that the error term will most likely be correlated within clusters. This has implications for the sample size requirements. Specifically, Bloom (2004) shows the required sample size to be

(6.11) $$n = (t_a + t_{1-k})^2 \left(\frac{s^2}{b^2} \right) \left(\frac{1}{f(1-f)} \right) \left(1 + (n_j - 1) r \right),$$

where n_j is the number of individuals per cluster and *?* is the intracluster correlation, i.e. the proportion of the overall variance explained by the within-group variance.[24]

For sample size determination, comparing expressions (6.6) and (6.11) implies that the usual estimate of the required number of individuals should be multiplied by the "inflation factor" *(1 + (*n_j *- 1)?)* when randomizing across groups instead of across individuals. The difference in the required number of individuals can be substantial if *?* is large. We can also note from (6.11) that the required number of individuals is minimized when *f = 0.5*, i.e., when the treatment and the control group are of the same size. Finally, dividing by n_j, and rearranging, we have

$$\frac{b}{s} = (t_a + t_{1-k})\sqrt{r + \frac{1-r}{n_j}}\sqrt{\frac{1}{f(1-f)}}\sqrt{\frac{1}{J}} \quad , \tag{6.12}$$

where *ß/s* is the minimum detectable effect size, i.e. the smallest effect that, if true, has a *?*% chance (or power) of producing an impact estimate that is statistically significant at the *a* level, and $J = n/n_j$ is the number of clusters. Equation (6.12) shows the trade-off between power and size. Ignoring the effects through the critical values of the *t* distributions, an increase in the number of clusters (*J*) or the number of individuals per cluster (n_j) reduces the minimum detectable effect size. Note, though, that while the minimum detectable effect size declines in roughly inverse proportions to the square root of the number of randomized groups (*J*), the size of the randomized groups often makes far less difference to the precision of the estimate. As noted in Bloom (2004), if *? = 0.05*, the values of $\sqrt{r + (1-r)/n_j}$ for randomized groups of 50, 100, 200, and 500 individuals would, correspondingly, be approximately 0.26, 0.24, 0.23, and 0.23. Thus, even a tenfold increase in the size of the randomized groups makes little difference to the precision of the impact estimator.

6.3. Measurement

There is a fairly large literature on measurement in education and health, but a much more limited literature on measurement in water and sanitation.[25] However, apart from direct measures of water quality (see, for instance, Kremer et al (2007)), and sanitation infrastructure

[24] See section 5.
[25] On education, see Glewwe and Kremer (2008) for details and references. On health, see Strauss and Thomas (1998).

and quality, or measures of connections to public (or private) water connection systems, such as the fraction of homes with latrines or access to communal standpipes or protected springs, evaluations in water and sanitation sectors typically have health, and sometimes education, outcomes as their prime target – thus, the focus here is on education and health. When studying the impact of water and sanitation projects on health outcomes, water-related diseases, such as diarrhea, respiratory, eyes and skin infections, are prime target for measurement.[26]

There is somewhat of a consensus in the literature that the number of years of schooling is a reasonably good indicator of education attainments, or the quantity of schooling. However, as an individual's completed years of schooling are only known several years after he or she first enrolled in school, measures of current schooling are often used in practice when researchers evaluate the impact of an intervention in education. This raises a couple of measurement issues.

The first is primarily conceptual. While increasing the probability of current schooling, for instance the probability of completing a given grade, may increase the number of years of schooling eventually completed, it could primarily "just" create intertemporal substitution in the timing of education. The second concern refers to how to measure current schooling. In developing countries, it is not uncommon for students to attend school erratically and the difference between frequently absent students and dropouts may be unclear. Thus, by looking at measures such as the completion of a given grade, or the decision to drop out, a large variation in the quantity of schooling would be overlooked. A way of partly dealing with the latter concern is to focus on participation, measured as the proportion of days that the students are present at school for a given number of days that the school is open. As classroom attendance registers are often inaccurate in developing countries, participation data typically requires independent data collection. Miguel and Kremer (2004), for instance, used unannounced visits by enumerators during a handful days over the school year to record which children were actually in class.

Education quality is sometimes measured by input proxies (such as student-teacher ratios, or the share of qualified teachers) but these measures are obviously imperfect measures of learning outcomes. Thus, too a large extent, education quality is directly measured by looking at student performance on academic tests. In many cases, these tests are organized by the evaluators. This has the obvious advantage that the tests could be designed in such a way as to get sufficient variation in and accurate measures of learning outcomes. Many countries also

[26] Kremer et al (2007) measure water contamination by the fecal indicator of bacteria *E. coli*.

record results from standardized tests that could be exploited. Reinikka and Svensson (2007), for instance, use test scores from Primary Leaving Exam records in Uganda. One advantage with this type of data is that students have incentives to do their very best on the test (since passing the test is a requirement for acceptance into secondary school).

There is less consensus on the measurement of health, partly because health is fundamentally multidimensional.[27] Moreover, while it is typically assumed that measurement error in schooling is random (Griliches, 1977), many health indicators are measured with errors that are systematically related to demand for health and other behavior (Strauss and Thomas, 1998).

The simplest form of health measurement is self-evaluation, most often self-reported general health status. As discussed in Strauss and Thomas (1998), while popular, these measures are fraught with problems.

Self-reported health problems, i.e. illnesses or death of family members, are also common in household surveys. This raises two problems, i.e. recall biases (see Deaton, 1997) and biases due to difficulties in interpreting what is deemed as illness or symptom that may systematically vary across individuals. Recall biases may be less problematic when probing information about major events in people's lives (like the birth or death of a child).

Anthropometric data, such data on height, weight, or a combination of the two, is increasingly becoming a standard module in many large-scale surveys. For instance, the latest round of DHS data includes data on height and weight for women and children. Child height has proven to be an informative longer-run indicator of nutritional status, as well as a cumulative measure of health investment for adults. Weight, on the other hand, varies more in the short run and thus provides a more current indicator of nutritional status. Since a light person may also be small, it is common to analyze weight given height. There are many potential ways of expressing this ratio, the most common being the body mass index – the ratio of weight (in kilograms) to height (in meters) squared.

7. Concluding remarks

[27] See Strauss and Thomas (1998) for a more detailed discussion and further references.

Impact evaluations ought to be an integral part of the policy formation process. The benefits of knowing which programs work and which do not extend far beyond any program or agency. A credible impact evaluation is also a global public good in the sense that it can offer reliable guidance to international organizations, governments, donors, and nongovernmental organizations in their ongoing search for effective ideas.

In this paper, we provide an introduction, and to some extent a practical guide, for researchers and practitioners interested in an impact evaluation in education, health, water, and sanitation. We refer the reader to the references given herein for a more in-depth treatment of the methods, concepts, and data issues we have discussed.

References

Altonji, J., Elder, T., and C. Taber, 2005, "Selection on Observed and Unobserved: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151-184

Angrist, J., 1998, "Estimating the Labor Market Impact of Voluntary Military Service Using Social SecurityData on Military Applicants," Econometrica, March.

Angrist J. and A. Kreuger, 1999, "Empirical Strategies in Labor Economics," *Handbook of Labor Economics*, Volume 3, Ashenfelter, A. and D. Card, eds., Amsterdam: Elsevier Science.

Angrist, J. and S. Pischke, 2008, "Mostly Harmless Econometrics: An Empiricist's Companion" forthcoming at Princeton University Press.

Ashenfelter and D. Card, 1985, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs on Earnings," *The Review of Economics and Statistics* 67:648-66.

Banerjee, A. , Banerji, R., Duflo, E., Glennerster, R., Khemani, S., 2008, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India", CEPR Discussion Paper Series no. 6781, Centre for Economic Policy Research, London.

Banerjee, A, S. Cole, E. Duflo, L. Linden, 2007, "Remedying Education: Evidence from Two Randomized Experiments in India", *Quarterly Journal of Economics*, 122(3): 1235-1264.

Barrera-Osorio, F., L.L. Linden, M. Urquiola, 2007, "The Effects of User Fee Reductions on Enrollment: Evidence from a quasi-experiment, Working Paper, Columbia University.

Bertrand, M., Duflo, E., and S. Mullainathan, 2004, "How Much Should We Trust Difference-in-Differences Estimates," *Quarterly Journal of Economics*, 119, 249-275.

Bjorkman, M. and J. Svensson, 2007, "Power to the People: Evidence from a Randomized Field Experiment of a Community-Based Monitoring Project in Uganda", CEPR Working Paper Series no. 6344, Centre for Economic Policy Research, London.

Black, R.E., S.S. Morris, J. Bryce, 2003 "Where and why are 10 million children dying every year?", *Lancet* 361: 2226–34.

Bloom, H. S., 2004, "Randomizing Groups to Evaluate Place-based Programs, Manuscript.

Cox, D., and N. Reid, 2000, *Theory of the Design of Experiments*, London: Chapman and Hall.

Deaton, A., 1997, *The Analysis of Household Surveys*, World Bank, International Bank for Reconstruction and Development.

Dinkelman, T., 2008, "The Effects of Rural Electrification on Employment: New Evidence from South Africa, Working Paper University of Michigan.

Donald, S., and K. Lang, 2007, "Inference with Difference in Differences and Other Panel Data," *The Review of Economics and Statistics*, 89, 221-233.

Donner A., and N. Klar, *Design and Analysis of Cluster Randomization Trials in Health Research*, London: Arnold.

Duflo, E., 2001, "Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment," *American Economic Review*, 91, 795-813.

Duflo, E., 2005, Field Experiments in Development Economics, BREAD Policy Paper No. 012, MIT.

Duflo, E., Glennerster R., and M. Kremer, 2007, "Using Randomization in Development Economics Research: A Tool Kit," CEPR Working Papers Series no. No 6059, Centre for Economic Policy Research, London.

Duflo, E., M. Kremer, and J. Robinson, 2006, "Understanding Technology Adoption: Fertilizer in Western Kenya", Working Paper, MIT.

Glewwe, P., 1999, *The Economics of School Quality Investments in Developing Countries*, St. Martin's Press, New York.

Glewwe, P., 2005, "The Impact of Child Health and Nutrition on Education in Developing Countries: Theory, Econometric Issues and Recent Empirical Evidence", *Food and Nutritional Bulletin* (forthcoming).

Glewwe, P. and M. Kremer, 2008, "Schools, Teachers, and Education Outcomes in Developing Countries", in E. Hanushek and F. Welch (eds.) *Handbook of the Economics of Education, 2*, Elsevier B.V.

Griliches, Z., 1977, "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica* 45(1): 1-22.

Holland, P., 1986, "Statistics and Causal Inference," *Journal of the American Statistical Association*, Vol 81, No. 396, 945-960.

Hahn, J., Todd, P., and W., Van der Klaauw, 2001, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica, 69, 201-9.

Imbens, G., and T. Lemieux, 2008, "Regression Discontinuity Designs: A Guide to Practice," forthcoming in Journal of Econometrics.

Imbens G. 2000, "The Role of the Propensity Score in Estimating Dose–Response Functions", Biometrika, Vol. 87, No. 3, 706-710.

Imbens G. 2004, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, Vol 86, 4-30.

Imbens, G., G. King, and G. Ridder, 2006, "On the Benefits of Stratification in Randomized Experiments, Working Paper, Harvard.

Jones, G., R.W Steketee, R.E. Black, Z.A. Bhutta, S.S. Morris, and the Bellagio Child Survival Study Group, 2003, "How many child deaths can we prevent this year?", *Lancet* 362: 65-71.

Karlan, D. and J. Zinman, 2008, "Expanding Credit Access: Using Randomized Supply Decisions To Estimate the Impacts", Working Paper, Yale University.

Kremer, M., J. Leino, E. Miguel, A. Peterson Zwane, 2006, "Spring Cleaning: A Randomized Evaluation of Source Water Improvement", Working Paper, Harvard University.

Levensohn, J., Z. McLaren, and K. Zuma, 2008, "HIV Status and Labor Market Participation in South Africa", Working Paper, University of Michigan.

Meyer, B., 1995, "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics*, 13, 151-161.

Miguel, E. and M. Kremer, 2004, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities, *Econometrica* 72(1): 159-217.

Kremer, M., J. Leino, E. Miguel, and A. P. Zwane, 2007, Spring Cleaning: A Randomized Evaluation of Source Water Quality Improvement, Working Paper, Berkeley.

Moulton, B., 1986, Random group effects and the precision of regression estimates," *Journal of Econometrics*, Elsevier, vol. 32(3): 385-397.

Olken, B., 2007, "Monitoring Corruption: Evidence from a Field Experiment in Indonesia", *Journal of Political Economy* 115 (2): 200-249.

Reinikka, R. and J. Svensson, 2007, "The Returns from Reducing Corruption: Evidence from Education in Uganda", CEPR Working Paper Series no. 6363, Centre for Economic Policy Research, London.

Stock, J., M. Yogo and J. Wright, 2002, "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20, 518-529.

Strauss, J. and D. Thomas, 1998, "Health, Nutrition, and Economic Development", *Journal of Economic Literature* 36(2): 766-817.

Wooldridge, 2008, Econometric Analysis of Cross Section and Panel Data, 2nd Edition, MIT Press.