WIDER Working Paper No. 2013/011

The porous dialectic

Experimental and non-experimental methods in development economics

Rajeev Dehejia*

February 2013

Abstract

This paper provides a survey of six widely used non-experimental methods for estimating the impact of programmes in the context of developing economies (instrumental variables, regression discontinuity, direct matching, propensity score matching, linear regression and non-parametric methods, and difference-in-differences), and assesses their internal and external validity relative both to each other and to randomized controlled trials. While randomized controlled trials can achieve the highest degree of internal validity when cleanly implemented in the field, the availability of large, nationally representative datasets offers the opportunity for a high degree of external validity using non-experimental methods. Whereas these methods are often presented as competing alternatives, we argue that each method has merits in some context and that experimental and non-experimental methods are complements rather than substitutes.

Keywords: randomized controlled trials, observational studies, programme evaluation, external validity

JEL classification: C18, C52, O12

*Robert F. Wagner Graduate School of Public Service, New York University, New York
Email: rajeev@dehejia.net

This study has been prepared within the UNU-WIDER project 'ReCom—Research and Communication on Foreign Aid', directed by Tony Addison and Finn Tarp.

UNU-WIDER gratefully acknowledges specific programme contributions from the governments of Denmark (Ministry of Foreign Affairs, Danida) and Sweden (Swedish International Development Cooperation Agency—Sida) for ReCom. UNU-WIDER also gratefully acknowledges core financial support to its work programme from the governments of Denmark, Finland, Sweden, and the United Kingdom.

ISSN 1798-7237 ISBN 978-92-9230-588-8

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by Liisa Roponen at UNU-WIDER.

The views expressed in this publication are those of the author(s). Publication does not imply endorsement by the Institute or the United Nations University, nor by the programme/project sponsors, of any of the views expressed.

1 Introduction

Development economics has become a battleground for a methodological debate that has raged in economics since the 1980s (Heckman and Robb 1985; Lalonde 1986; and Fraker and Maynard 1987) and more broadly in the social and medical sciences for almost half a century (e.g., Chalmers, Block and Lee 1970; Meier 1972). The contention concerns the validity of experimental versus non-experimental (or observational or econometric) methods in estimating the impact of interventions. Randomized experiments offer a research design which, if implemented cleanly, yields unbiased estimates of the impact of interventions on a wide range of outcomes of interest, and from this perspective non-experimental methods are questioned as unreliable in the sense of requiring strong, untestable, and often implausible assumptions. Observational studies instead use econometric methods (including simple regression analysis, instrumental variables, diff-in-diffs, regression discontinuity, and matching), each with its own assumptions, strengths, and weaknesses to estimate the impact of programmes by analysing data ranging from small-scale surveys to nationally representative censuses. From this perspective, observational studies have the potential that experiments lack: of tackling big questions and estimating impacts for entire populations.

This paper provides a survey of six widely used non-experimental methods for estimating the impact of programmes in the context of developing economies, and assesses their strengths and weaknesses relative both to each other and to randomized controlled trials. While this paper will not delve into the nuances of experiments, it is worthwhile at the outset considering their evident advantages, in order to set up a benchmark of comparison for non-experimental methods.

The distinction between internal and external validity will be important to our discussion. Internal validity refers to the biasedness of an estimator within a given research sample or experiment. External validity refers to the ability to forecast (predict, extrapolate) the estimate from the research sample to a population of policy interest. Randomized experiments' strength is their internal validity: with a well-implemented experiment, the experimental design guarantees an unbiased estimate of the treatment impact of interest. At the same time, the challenge of implementing a randomized experiment in the real world implies that experimental populations are rarely representative of the population of policy interest and indeed are often small, non-representative groups chosen for their convenience and willingness to participate in an experiment. While development economists have been at the forefront of moving experiments out of labs in elite colleges and into the real world, the basic issue remains that despite the impressive powers of persuasion of field researchers, it is difficult to implement experiments over the whole population of policy interest. As we discuss below, although there is no simple ranking of non-experimental methods vis-à-vis experiments or among themselves, internal and external validity proves to be useful criteria for a comparison.

A third dimension along which estimators are usually compared in evaluation settings is efficiency; while the bias-efficiency tradeoff is conceptually important, it is not typically useful in comparing experimental and non-experimental methods, as both sets of methods can be efficient under their respective assumptions.

While the comparison between experiments and non-experimental methods is often seen as a choice between mutually exclusive alternatives or even a contest of competing methods, we argue below that both of these sets of methods face similar conceptual challenges, although they tackle them quite differently, and furthermore can sometimes be fruitfully combined.

2 LATE versus ATE and ideal experiments versus feasible experiments versus ideal non-experiments

It will be useful to define some simple terminology for thinking through the relative merits of different evaluation methods. Suppose that we are trying to evaluate the effectiveness of a treatment (e.g., an educational innovation in primary schools) for a population of interest (e.g., the universe of elementary-aged school children). Let i index individuals in the population, $T_i = 1$ denote individuals exposed to the new educational programme, and $T_i = 0$ denote individuals exposed to the default (control) programme. For each individual i and outcome of interest Y (e.g., grades on a standardized test), Y_{1i} denotes the potential grade outcome if individual i is exposed to the treatment, and Y_{0i} denotes the potential grade outcome if individual i is exposed to the control.

The average treatment effect of interest in an ideal experiment is defined as:

$$ATE = E\left(Y_{1i} - Y_{0i}\right).$$

However, in practice, experimenters are rarely able to randomize the treatment over the full population, P, and instead assign the treatment within an experimental population $E \subset P$. For example, Muralidharan and Sundararaman (2011) are interested in the effect of performance pay on teacher and student performance. Given that the policy experiment was conducted within the state of Karnataka, one would image the population of interest being all schools or students within the state. As a matter of practice they are able to randomize within 5 of 30 districts in the state. In this case, define:

$$ATEX = E\left(Y_{1i} \middle| i \in E\right) - E\left(Y_{0i} \middle| i \in E\right).$$

ATEX differs from *ATE* to the extent that the experimental population is systematically different from the overall population of interest. In the Muralidharan and Sundararaman example, to the extent they select districts at random they might be representative of the overall population; however, even with an experimental selection protocol, by random chance it might be that the sampled states differ systematically from the state as a whole.

Assuming the experimental protocol is reliably implemented in the field, T_i is randomly assigned which implies that:

$$E\left(Y_{ji}\right) = E\left(Y_{ji} \middle| T_i = 0\right) = E\left(Y_{ji} \middle| T_i = 1\right), \text{ for } j = 0, 1,$$

and that *ATE* and *ATEX* can be estimated by their sample analogues. Any failure of the experimental protocol would imply that the sample analogues of these two would be (further) biased estimators.

In a non-experimental setting, there are also three potential steps that can lead away from the *ATE*. First, if the data that are being used to estimate the treatment effect are not representative of the full population of interest (e.g., a non-representative survey), then again we would find ourselves focusing on *ATEX*. Second, in a non-experimental setting some assumptions are required to estimate $E(Y_{1i})$ and $E(Y_{0i})$ from sample data. We discuss specific assumptions in sections 3 and 4 below, but examples include selection on observables or linearity. Third, violations of the assumptions (or relatedly approximation error, e.g., assuming linearity when the true functional form is non-linear) would bias estimates of *ATE* and *ATEX*.

Setting up experimental and non-experimental estimators in parallel is not meant to imply that each of these steps plays an equally important role in randomized trials and observational studies, although in some instances it does. The first step, from *ATE* to *ATEX*, is the result of different constraints in experimental and non-experimental settings. In experiments the constraint is the ability to run an experiment on the full population of interest, which in turn is because of the cost and challenges of maintaining an experimental protocol for a very large population and the unwillingness of most policymakers to experiment on a large scale. While there are instances of representative, large-scale experiments (e.g., Muralidharan and Sundararaman 2011), these have become more challenging over time, as developing countries are increasingly adopting explicit and rigorous human subject standards for development research (e.g., India, Brazil and China; see VandenBosch 2011). In non-experiments, the challenge concerns the data available for the evaluation exercise, which are not always representative of the overall population of interest. However, the fact that representative datasets exist and can readily be gathered using well-developed sampling techniques suggests that the *ATE* is a less challenging target in a non-experiment than an experiment.

A feature which both experiments and non-experiments share in common is the challenge of extrapolating from a given time-period, location, or range of values to another setting. Although in non-experimental settings this is sometimes more explicit (e.g., through the use of time effects in a regression framework) the same problem is implicit in experiments as well; as we discuss below, a similar set of assumptions, with their concomitant strengths and weaknesses, can be used to extrapolate experimental results, for example through meta-analysis.

The second step, from either *ATE* or *ATEX* to an estimator, marks the sharpest difference between experiments and non-experiments. In experiments, randomized experimental design guarantees that sample analogues (typically means) are unbiased estimators of their population equivalents, whereas in non-experiments, this step involves assumptions that are both stringent and usually untestable (although depending on the setting, perhaps plausible; further discussion below).

The third step, violation of assumptions, is common across both types of settings, although qualitatively different. In an experiment, the violation of assumptions is usually a breakdown of the experimental protocol or design—situations in which the experiment was not perfectly implemented—that could be observed and possibly prevented through a careful process of monitoring and supervision accompanying the experiment. In non-experimental settings, the assumptions that fail are restrictions on the form of the expectation or selection process; these are difficult to verify, although there are specific situations in which they can be rejected based on testable implications in the data. In the next two sections we discuss the relative strengths and weaknesses of specific non-experimental methods.

3 Non-experimental alternatives

3.1 Instrumental variables estimators

Instrumental variables (IV) estimators are a natural extension of experiments with two intermediate steps (provided by Imbens and Angrist 1994; Angrist, Imbens and Rubin 1996). First, in addition to random assignment to treatment, consider natural experiments or other plausibly exogenous ('instrumental') variables that influence assignment to treatment. (Instruments must be independent of potential outcomes and not influence the outcome directly, only indirectly through assignment to treatment). Second, allow that the instrument only imperfectly predicts assignment to treatment. The simplest example is an encouragement design or a natural experiment in which: units are randomly assigned an encouragement to take the treatment; encouragement increases the probability of being treated; and non-compliance is possible (opting into or out of the treatment despite encouragement status). (Note that if the instrument perfectly predicted assignment to treatment, then conceptually one would be back in the world of randomized controlled trials).

For example, Angrist et al. (2002) use a lottery that assigned school-choice vouchers in Colombia as an instrumental variable for using a school-choice voucher. Even though the lottery assigned the vouchers randomly, not all winners used them. While exogeneity is satisfied by the randomization of the lottery, the exclusion restriction—that winning a school voucher lottery only affects school outcomes through use of the voucher—is a necessary, albeit plausible, assumption. Of course, the same framework can be used when exogeneity is assumed rather than implemented through randomization of the treatment. For example, Duflo (2001) uses a school construction programme in Indonesia as an instrument for the level of education. The exogeneity assumption cannot be, and is not, taken for granted, since incremental schools were not randomly located, but using both knowledge of the programme and additional evidence she argues for its plausibility.

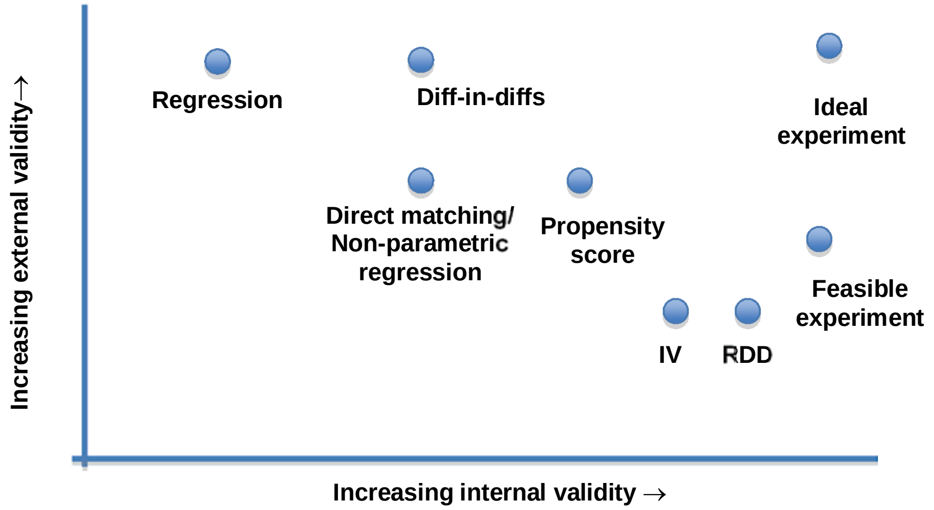
Because assignment to treatment is potentially self-selected, ordinary least squares or a difference in means based on the treatment is a potentially biased estimator of the treatment impact. Likewise the same estimators based on the instrument or encouragement variable are causal, but are causal effects of the instrument or encouragement and not of the underlying treatment (known as the intention-to-treat effect or ITT). Imbens and Angrist (1994) show that the IV estimator provides a consistent estimate of the impact of the treatment on the outcome for the subset of individuals known as compliers (the local average treatment effect, where E = compliers); these are individuals who will take the treatment if encouraged, but who would not take the treatment if not encouraged. In particular, they show that this is the only subset of the population for whom there is observable variation from the instrument, because those who don't comply are either always in the treatment or always in the control.[1]

The internal validity of the IV estimator depends on the source of the instrument; if it is from a randomized trial or plausible natural experiment, its internal validity rivals that of an experiment (subject to the plausibility of the exclusion restriction). However, because it identifies a local average treatment effect, the IV estimator has a lower degree of external

[1] The fourth possibility—individuals who take the treatment if not encouraged, and do not take it if they are encouraged—is assumed away.

validity than an ideal experiment (although not necessarily a feasible experiment in the sense defined in section 2). This is depicted in Figure 1.

Figure 1



Regression
Diff-in-diffs
Ideal experiment
Direct matching/ Non-parametric regression
Propensity score
Feasible experiment
IV
RDD
Increasing external validity→
Increasing internal validity →

Source: See text.

3.2 Regression discontinuity

Regression discontinuity (RD) methods have gained widespread acceptance in programme evaluation in both developed and developing country settings (see *inter alia* Thistlewaite and Campbell 1960; Van Der Klaauw 2002, 2008; Imbens and Lemieux 2008; Buddlemeyer and Skoufias 2004). While we will not review the technical details of RD methods here, it is worthwhile sketching the intuition of this approach. Imagine a situation in which assignment to treatment depends on a single 'forcing' variable, and in particular being above some cutoff of that variable. For example, in Ozier (2011) scores on a standardized national exam are used to determine high school admission and the probability of admission rises sharply at the national mean.

Assuming that the only discontinuous change at the cutoff is the probability of exposure to the treatment (i.e., that either you are assigned to the treatment on one side of the cutoff and not on the other or that the probability of being assigned to the treatment jumps at the cutoff), then looking in a neighbourhood of the cutoff an individual just above and just below the cutoff should be essentially identical except for exposure to the treatment. In particular, the assumptions are that the distribution of observed and unobserved covariates is continuous at the cutoff, that no other treatments are assigned using the same cutoff, and that agents are unable to manipulate their value of the forcing variable. The essence of the assumption in RD is that the cutoff for assignment to treatment is at least locally arbitrary with respect to other economic variables, i.e., that there is no other economically relevant economic variable that changes at the specific cutoff in the forcing variable other than the treatment.

In the context of our discussion in section 2, RD is very much a method that is estimating *ATEX* rather than *ATE*, where $i \in E$ is defined by individuals in a neighbourhood of the cutoff. Although one might assume a constant treatment effect (*ATE* = *ATEX*), RD identifies the treatment effect only for those individuals who find themselves just above or just below the cutoff in the forcing variable.

In the terms of the second step (the assumptions needed to identify *ATEX* in the sample), RD requires a little more than a true experiment. Rather than ensuring random assignment through experimental design, in RD we must make key identifying assumptions. While the plausibility of these assumptions will vary depending on the context, one of the virtues of the RD approach is that these assumptions can be corroborated or tested, at least in part.

The assumption that individuals are not manipulating their value of the forcing variable in order to opt into or out of treatment can be in part tested by looking for evidence of manipulation in the density of the forcing variable. For example, if a poverty score cutoff is used to determine eligibility for many social assistance programmes, then it is likely that individuals will find a way to manipulate their poverty scores and that we will see a bunching up of the poverty distribution just below the poverty threshold. Indeed this has been shown for Colombia's SISBEN (Camacho and Conover 2011; see also Miller, Pinto and Vera-Hernández 2009, and for a formalization of this test McCrary 2008). In contrast, in Ozier (2011), the threshold of the test score at which students' chances of high school admission jump discontinuously is based on the mean of the test score; if this is stable over time it might be known quite precisely and possibly manipulated. If instead the mean is sufficiently variable, it would be difficult for students to fine tune their test scores in order to manipulate their chances of admission.

The assumption that covariates do not jump discontinuously at the same threshold used to assign the treatment can be tested for observed covariates by looking at the distribution of the covariates in the neighbourhood of the threshold. For example Ozier (2011) shows that mothers' and fathers' educations do not jump at the threshold in test scores where admission to high school jumps. Of course, for unobserved covariates this assumption cannot be tested and remains just that, an assumption.

Figure 1 summarizes the relationship between RD, randomized trials, and IV with respect to internal and external validity. Experiments through design and control in implementation typically offer the highest degree of internal validity. At the same time, in principle an experiment can be conducted on many populations, not just the individuals who happen to be at the cutoff of the forcing variable, and so even feasible experiments potentially offer greater external validity than RD. RD in turn has greater internal validity than IV, in the sense that many of the key identifying assumptions can be tested or corroborated in the data. Its external validity is more difficult to compare to IV. While RD and IV both identify 'local' treatment effects, for RD the local sub-population can be identified (i.e., those individuals just above and below the threshold in the forcing variable) while the population of compliers for whom IV identifies the treatment effect cannot be identified.[2]

3.3 Direct matching methods

The idea behind direct matching methods is closely related to regression discontinuity. Beginning with an RD-type assignment rule, which depends on a single forcing variable,

[2] Although the population of compliers cannot be directly identified in the data, it is possible to characterize this population to some extent; e.g., their proportion of the total population.

consider a more general assignment mechanism that is a potentially non-linear and stochastic function of the forcing variable *X*, $\Pr\left(T_i = 1 \middle| X_i\right) = f(X_i)$; for example the probability of treatment could decrease in the forcing variable. The key insight of matching is that conditional on *X* (whether univarate or multidimensional) individuals with $X_i = X_0$ have the same probability of assignment to treatment (i.e., $f(X_0)$) and whether they are treated or untreated is a matter of random chance, hence essentially an experiment.

The key difference between experiments and non-experiments is that in the former, individuals' probability of assignment to treatment is known (specified in the experimental design) whereas in a non-experiment it is *assumed* to be a function of a vector of pre-treatment covariates, also known as the selection on observables assumption (Heckman and Robb 1985). The plausibility of the selection on observables assumption depends on the context. When the process of selection into treatment is well understood and the relevant variables observed, the assumption is plausible and matching has the same internal validity as an experiment. For example Diaz and Handa (2006) show that, since PROGRESSA is targeted using community and individual poverty measures, matching on a broad set of community and household poverty and demographic measures replicates the experimentally estimated treatment impact of the programme.

Instead, in situations in which unobservables (i.e., variables that are relevant for selection into treatment and the outcome of interest and that are inherently difficult to observe or simply not available in the researcher's dataset) are believed to be important, matching is not likely to yield internally valid estimates. Again, Diaz and Handa (2006) show that a comparison group drawn from survey similar to the one used in PROGRESSA is better able to replicate the experimentally estimated treatment effect than a comparison group drawn from surveys using a different survey instrument. In this case, the unobservable factor is the difference in measurement of the relevant variables. In other instances, the unobservable could simply be a key missing covariate (see also the discussion in Heckman, LaLonde and Smith 1999; Dehejia and Wahba 1999, 2002).

The best matching applications use knowledge of the specific context of the application to make the case for selection on observables, and a variety of studies have tried to demonstrate which observables are key determinants of selection into treatment in different contexts. While the internal validity of matching rests on an assumption, its strength is that the method can be applied to any dataset; hence it has a high degree of external validity.

As a matter of practice, matching on a single variable is straightforward. Some standard methods include: one to one matching (match a given treated unit to the comparison unit with the closest value of *X*, with or without replacement); one to *n* matching (match each treated unit to the *n* closest comparison units, with or without replacement); radius matching (match each treated unit to the comparison units within a radius defined on *X*); and kernel matching (match each treated unit to nearby comparison units with weights determined by a kernel).

The greater challenge is when matching is carried out with a multivalued *X*. The most intuitive method is to use a distance metric to map *X* into a single Euclidian distance, and then matching on that distance. But as shown by Abadie and Imbens (2006), standard matching methods are inconsistent when matching is carried out on more than two continuous covariates. Fortunately, Abadie and Imbens also propose a bias-corrected matching estimator.

In Figure 1, we compare direct matching to RD, IV, and RCT's. Since matching can be carried out on large-scale representative datasets, it offers a greater degree of external validity than RD, IV, and feasible experiments (of course an ideal, but rarely feasible, experiment on a representative population could have the same degree of external validity as matching), but its internal validity is potentially lower since it rests on the assumption of selection on observables, which is both stronger and less amenable to testing than the underlying assumptions of the other three methods.

3.4 Propensity score matching and extensions

Propensity score matching is closely related to direct matching on covariates. Indeed, one can think of it as a method for converting direct matching on a vector of covariates into direct matching on a scalar. The key insight, due to Rosenbaum and Rubin (1993), is that rather than conditioning on the entire vector of covariates that determines (the probability of) assignment to treatment, one can directly condition on the probability of assignment to treatment. For individuals with the same probability of assignment to treatment which individuals are treated and which are not is a matter of random chance.

Two sets of issues arise in implementing propensity score matching. The first is how to estimate the propensity score. A variety of methods have been proposed, although the most common are probit or logit functional forms in conjunction with specification tests (see Dehejia and Wahba 1999; Shaikh et al. 2009; Huber 2011). The second is how to use the propensity score in an estimator. As suggested above, the intuitive choice is to use the estimated propensity score in one of the univariate matching methods listed in section 3.3. An alternative choice is to use the propensity score as a weight in a linear regression (see Dehejia and Wahba 1997; Wooldridge 2001; Hirano, Imbens and Ridder 2003; Busso, DiNardo, and McCrary 2009). The advantage of weighting is that it is an efficient estimator and that it shoehorns the propensity score into the easily-implemented framework of regression analysis. The disadvantage is the potential of significant misspecification and small sample bias (see Frölich 2004; Busso, DiNardo, and McCrary 2009).

The key advantage of propensity matching with respect to direct matching is that it eschews the bias of direct matching estimators, and hence achieves a great degree of internal validity. This is depicted in Figure 1.

Propensity score matching methods have found wide application within development economics. For example Jalan and Ravallion (2003) use propensity score matching to estimate the impact of piped water on children's health in India. They argue that it is important to match on both village and individual-level characteristic because selection into piped water occurs at both levels: first a village has to be connected (e.g., predicted by size of the village or having access to a paved road) and then an individual has to opt in (e.g., predicted by living in an electrified home or a female-headed household). See Behrman, Cheng and Todd (2004) and Godtland et al. (2004) for other examples.

4 Regression-based methods

In this section we discuss methods that are implemented using linear regressions, albeit with different identifying assumptions. Although the methods discussed in the previous section can be reformulated in a regression framework, conceptually their identifying assumptions are more closely related to randomized experiments. Indeed, regression discontinuity, instrumental variables, and matching explicitly try to recreate or exploit naturally occurring circumstances that create the advantages of randomized trials. In this section we discuss instead a set of methods that are more naturally understood directly within the regression framework.

4.1 Linear regression

The most natural evaluation estimator, at least for economists, is a linear regression of the form:

$$Y_i = \alpha + \beta X_i + \delta T_i + \varepsilon_i .$$

The most widely discussed challenge with this estimator is the problem of sample selection bias, namely that assignment to treatment, T_i, might dependent on variables not captured in X, hence creating a correlation between T and ε and a bias in the ordinary least squares estimate of the regression. In this sense, sample selection bias can be reformulated as a selection on unobservables or omitted variables problem.

As discussed in section 3.3, the severity of selection on unobservables, or the plausibility of selection on observables, is very much context dependent. For example in labour training programmes in the US, it has been shown that selection on observable factors such as labour market conditions and work history can be more important than unobservables (Heckman, Lalonde and Smith 1999; Heckman et al. 1998); while this is a context-specific conclusion it does illustrate the plausibility of the selection on observables assumption in some applications.

Even if we accept the selection on observables assumption, then a second assumption embedded in the regression approach is linearity. While this refers specifically to the linearity of the regression function, in the evaluation context the key question is the extent to which the pre-treatment covariates, X, differ between the treatment group and comparison groups. For example, Lalonde (1986) finds that linear regressions are unable to control for differences between the treatment and comparison groups in a setting where these groups differ by more than five standard deviations for some covariates. A lack of overlap in the covariates implies that the treatment effect is being estimated by an extrapolation based on the linear functional form. In contrast, the matching methods discussed in section 3 make no such assumption on linearity.[3]

As a result, in Figure 1, we denote regression methods as having lower internal validity than matching methods. At the same time since regressions can be run on the full sample of a dataset (rather than simply the set of matched treatment and comparison units in matching), we denote it as having potentially greater external validity than matching. There are three responses to the assumption of linearity required for regression methods. First, one can test

[3] As discussed above for experiments, a special but important case of extrapolation is extending the estimated treatment impact from the research sample to another geographical setting or longer time horizon. While this can be naturally incorporated within a regression framework though an appropriate choice of functional form, the validity of the resulting inference is entirely dependent on untestable functional form assumptions.

for overlap in the covariates between the treatment group and comparison groups; if there is considerable overlap then there is a lesser degree of extrapolation and the estimated treatment impact should be less sensitive to the choice of functional form. Several tests are possible, ranging from a simple comparison of means for each covariate to comparing the estimated propensity score across treatment and comparison groups (see Dehejia and Wahba 1999, 2002). Second, one can use matching methods, which explicitly do not assume a linear relationship between the covariates and the outcome. Third, one can use non-parametric regression methods, which we discuss in the next section.

4.2 The non-parametric approach

In some ways the most natural response to the problem of overlap or non-linearity in the covariates, even if one is willing to assume selection on observables, is to use non-parametric regression estimators, such as the kernel or series estimator. See for example the pioneering work of Deaton (e.g., 1989). The evident advantage of this is that it allows the data to choose the most appropriate non-linear functional form.

A classic example of this approach is Subramanian and Deaton (1996). Subramanian and Deaton examine the relationship between nutrition, as measured by calorie consumption, and a household's economic status, measured by household expenditure. While this question can be formulated in the framework of section 3, this would not be the most natural approach. The data do not offer a natural source of exogenous variation in food prices to use an experimental or experimentally inspired framework. At the same time, the goal is specifically to understand the relationship between expenditure and calorie consumption for a range of values, not just a local effect at whichever point one might have exogenous variation. Their analysis illustrates the rich rewards of this approach, since the relationship turns out to be non-linear and varies meaningfully at different levels of economic status. For example, they find that the elasticity of calories declines gradually with increasing economic status, but that substitution between food categories is less important for the poor than the better off.

The advantage of this approach is that, setting aside problems of self-selection or endogeneity, the lack of functional form assumptions yields conclusions with a high degree of internal validity. At same time, in settings where endogeneity or selection into treatment is a paramount concern, this method does not in itself offer a solution. In this sense, the natural comparison is to matching methods. If one is comfortable with the selection on observables assumption, then both methods can yield internally valid estimates of treatment impacts, with two key differences. First, matching methods are naturally set up to estimate treatment impacts, so from binary or more generally discrete treatments, whereas non-parametric regression methods are more useful for continuous treatments or variables of interest. Second, as with direct matching (although unlike propensity score matching), non-parametric methods are subject to the curse of dimensionality; that is, these methods are not easy to implement and lack appealing properties when there are a large number of covariates.

Hence, in Figure 1, we depict non-parametric methods as having the same relative merits as direct matching methods.

4.3 Difference-in-differences estimators

One of the most widely used regression-based estimators is the difference-in-differences (diff-in-diffs) estimator (see Duflo 2001 for an example). The idea behind this method is to use both cross-sectional and time-series variation to relax the assumptions needed for plausible identification of the treatment effect. In particular, rather than trying to find the ideal cross-sectional comparison group, a diff-in-diffs estimator uses a comparison group that can differ in levels from the treatment group but has the same time trend. The within-treatment group, pre-post treatment comparison is used to estimate the impact of the treatment, which is disentangled from the underlying time trend by using the within-comparison group, pre-post comparison.

Returning to Duflo (2001), the intensity of school building in Indonesia varied by region and over time. This allows Duflo to use the low intensity school-building regions as comparisons for regions where more schools were built. Diff-in-diffs doesn't require the assumption that the treatment and comparison regions are similar in the pre-period; instead it allows them to differ up to an additive fixed effect in the pre-period but assumes that absent treatment two sets of regions would have had the same time trend in the outcome.

In this sense, diff-in-diffs has a greater degree of internal validity than a simple regression approach. Its external validity is similar, in the sense that both methods use widely available (and often representative) datasets. To the extent that diff-in-diffs is often used in situations with large-scale policy variation, external validity relates to the comparability of treated and untreated jurisdictions, e.g., New Jersey and Pennsylvania in Card and Krueger's classic minimum wage paper (Card and Krueger 1994) or the regions with different patterns of school building in Duflo (2001).

In Figure 1, diff-in-diffs is depicted as having the same degree of external validity as, but greater internal validity than, an OLS linear-regression approach since it relaxes key assumptions of the latter approach. It is more difficult to rank the internal validity of diff-in-diffs and matching methods. In many applications, the diff-in-diffs common time-trend assumption is testable (e.g., if there is more than one year of pre-intervention data on the treatment and comparison groups), but at the same time diff-in-diffs still relies on the linearity assumption (by assuming that group and time differences drop out with first differencing). Direct matching does not rely on linearity, but does rely on the selection on observables assumption. Thus in Figure 1 we depict them as having the same degree of internal validity, although this will of course depend on the application and which of these assumptions is more plausible. It is also worth noting that matching methods can be combined with differencing (see, for example, Smith and Todd 2005).

5 Meta-issues

In this section we address a set of issues that are faced both by experimental and non-experimental methods and by each of the non-experimental approaches addressed above.

5.1 Replication: all knowledge is local until it is not

While we have argued that the range of experimental and non-experimental methods discussed above are likely to have differing degrees of external validity, it is worth considering the implications of the obvious fact that every feasible analysis is externally valid only to some extent. In the context of development economics, randomized controlled trials are rarely (and reasonably) implemented on the full population of policy interest. While non-experimental methods are more likely to use data representative at the national level, issues of comparability over time and to other countries remain a concern.

Since no one feasible study is likely to be able to provide both externally and internally valid results for an intervention of interest for a population of interest, I would argue for the central importance of ‘external’ replication (see also Hammermesh 2007). Whereas replication usually refers to rerunning experiments in settings identical to the original experiment (which we could call ‘internal’ replication), external replication takes on two forms. The first is applying the same research design (either an experimental protocol or non-experimental specification) to a different population. The second is applying a different research design to address the same question.

External replication can be used both to test formally and to substantiate informally the external validity of a finding. In the context of experiments, the most natural question is whether the treatment effect in the population (individuals, organization, and jurisdiction) that was willing to participate in an experiment differs from the population of interest. While experiments in development economics use an impressive array of subjects (e.g., compared to the standard lab rats in biomedical research; see Martin et al. 2010), they still cannot span the full population of interest. Likewise in non-experimental studies, not all datasets are representative of the population of interest, and even when they are, they will rarely be from the same point in time. For example, in the context of labour training programmes in the US, Heckman, Lalonde and Smith (1999) have argued that local labour market conditions are very important for external validity in evaluating labour training programmes.

While it is natural for researchers to focus on producing the ‘perfect’ study, this line of reasoning suggests a complementary view: that it is important to produce many good pieces of evidence from a wide range of contexts.

5.2 Meta-analysis

One response to the proposal in the previous section is to aggregate studies on the same question in a formal meta-analysis. As with replication, meta-analysis is more common in the medical field where studies are frequently repeated in different populations. The advantage of meta-analysis is that it tackles head on the question of how the treatment effect in a given study is affected by various features of the analysis and research sample. The two key challenges are that it requires a sufficient number of studies, and that studies within a meta-analysis have to be broadly comparable to a degree that is not very common within micro-empirical economics.

The most common use of meta-analysis in the development economics literature has been to summarize large and contradictory macro-empirical literatures. For example, there are a plethora of results on relationship between development assistance and economic growth, ranging from negative to neutral to positive. Doucouliagos and Paldam (2008) review sixty-eight comparable papers that find no systematic evidence of positive link. Just as in primary research, meta-analysis is not without its own methodological and interpretative debates.

Mekasha and Tarp (2011) revisit this question and reach a more positive conclusion from the literature.

5.3 Publication bias

One challenge of replication and meta-analysis is that it relies on the dissemination of both positive and negative findings. The wider dissemination of working papers through the internet has alleviated to some extent the problem of publication bottlenecks and lags, but does not directly address the problem that positive findings are more likely to be written, circulated, and eventually published (e.g., Delong and Lang 1992). One could argue that publication bias is more severe in non-experimental studies because of the emerging tradition for journals to publish even negative experimental results (e.g., going back to Glewwe, Kremer and Moulin's [2009] famous finding of the ineffectiveness of textbooks in classrooms in Kenya). On the other hand, one could argue that experiments are more subject to meta-selection bias: selection not of assignment into treatment, but of the settings in which experiments are implemented.

6 Conclusion: the porous dialectic

There has been much debate within the field of development and in economics more broadly on the relative merits of experimental and non-experimental methods. These alternatives are sometimes perceived as mutually exclusive choices. But I would argue that there are three senses in which it is more productive to consider the complementarities between these approaches.

First, while an ideal experiment has the virtue that data analysis can be confined to differences in means, experiments are in reality subject to flaws in execution and design. Furthermore, even when well executed there can be legitimate random variation in exposure to and implementation of the treatment. Non-experimental adjustment can help to mitigate bias and improve efficiency in experiments (e.g., Rosenbaum 1995; Barnard et al. 2003; Gelman, Meng and Sacerdote 2005). For example, regression or matching methods can be used to control for random or implementation-failure-related differences in covariates between the treatment and control groups or simply to improve ex-post efficiency. Likewise random effects methods, along with regression analysis, can be used to assess the significance of variation in treatment implementation (e.g., Dehejia 2003).

Second, non-experimental studies can be informed by a design philosophy in the spirit of experimental design (e.g., Rubin 2005, 2008). In particular, just as an experimental researcher is forced to specify the experimental protocol in advance of the execution of a randomized trial, a non-experimental researcher can also follow a set of transparent protocols in designing an observational study. Among these are selecting the specification or comparison units based on pre-treatment covariates and a transparent statement of the identifying assumptions along with a reasoned defence of their plausibility within the context of the specific application.

Third, in thinking of evaluation as an accretion of knowledge rather than the design of a single study, experiments and non-experiments offer an array of methods—each with strengths and weaknesses—which represent different opportunities to learn about the impact of a programme.

References

Abadie, A., and G. Imbens (2006). ‘Large Sample Properties of Matching Estimators for Average Treatment Effects’. *Econometrica*, 74(1): 235–67.

Angrist, J., G. Imbens, and D. Rubin (1996). ‘Identification of Causal Effects Using Instrumental Variables’. *Journal of the American Statistical Association*, 91(434): 444–55.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002). ‘Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment’. *American Economic Review,* 92(5): 1535–58.

Barnard, J., C. E. Frangakis, J. L. Hill, and D. B. Rubin (2003). ‘Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City’. *Journal of the American Statistical Association*, 98(462): 299–323.

Behrman, J., Y. Cheng, and P. Todd (2004). ‘Evaluating Preschool Programmes When Length of Exposure to the Programme Varies: A Non-Parametric Approach’. *Review of Economics and Statistics*, 86(1): 108–32.

Buddlemeyer, H., and E. Skoufias (2004). ‘An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA’. Working Paper 3386; IZA Discussion Paper 827. Washington, DC: World Bank.

Busso, M., J. DiNardo, and J. McCrary (2009). ‘New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators’. IZA Working Paper 3998. Bonn: IZA.

Camacho, A., and E. Conover (2011). ‘Manipulation of Social Programme Eligibility’. *American Economic Journal: Economic Policy*, 3(2): 41–65.

Card, D., and A. Krueger (1994). ‘Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania’. *American Economic Review*, 84(4): 772–93.

Chalmers, T., J. Block, and S. Lee (1970). ‘Controlled Studies in Clinical Cancer Research’. *New England Journal of Medicine*, 287: 75–8.

Diaz, J. J., and S. Handa (2006). ‘An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico’s PROGRESA Programme’. *Journal of Human Resources*, 41(2).

Deaton, A. (1989). ‘Rice Prices and Income Distribution in Thailand: A Non-Parametric Analysis’. *Economic Journal*, 99: 1–7.

Dehejia, R. (2003). ‘Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programmes with Grouped Data’. *Journal of Business and Economic Statistics*, 21(1): 1–11.

Dehejia, R., and S. Wahba (1997). ‘Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programmes’. In R. Dehejia, *Econometric Methods for Programme Evaluation*. Cambridge: Harvard University. Ph.D. Dissertation.

Dehejia, R., and S. Wahba (1999). ‘Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programmes’. *Journal of the American Statistical Association*, 94(448): 1053–62.

Dehejia, R., and S. Wahba (2002). ‘Propensity Score Matching Methods for Non-Experimental Causal Studies’. *Review of Economics and Statistics*, 84: 151–61.

Delong, B., and K. Lang (1992). ‘Are All Economic Hypotheses Wrong?’. *Journal of Political Economy*, 100(6): 1257–72.

Doucouliagos, H., and M. Paldam (2008). ‘Aid Effectiveness on Growth: A Meta Study’. *European Journal of Political Economy*, 24: 1-24.

Duflo, E. (2001). ‘Schooling and Labour Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment’. *American Economic Review*, 91(4): 795–813.

Fraker, T., and R. Maynard (1987). ‘The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programmes’. *Journal of Human Resources*, 22: 194–227.

Frölich, M. (2004). ‘Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators’. *Review of Economics and Statistics*, 86(1): 77–90.

Gelman, A., X. Meng, and B. Sacerdote (2005). ‘Fixing Broken Experiments Using the Propensity Score’. In A. Gelman and X. Meng (eds), *Applied Bayesian Modelling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin’s Statistical Family*. New York: Wiley.

Godtland, E., E. Sadoulet, A. De Janvry, R. Murgai, and O. Ortiz (2004). ‘The Impact of Famer Field Schools on Knowledge and Productivity: A Study of Potato Farmers in the Peruvian Andes’. *Economic Development and Cultural Change*, 53(1): 63–93.

Hammermesh, D. (2007). ‘Replication in Economics’. *Canadian Journal of Economics*, 40(5): 715–33.

Heckman, J. J., H. Ichimura, J. A. Smith, and P. E. Todd (1998). ‘Characterizing Selection Bias Using Experimental Data’. *Econometrica*, 66: 1017–98.

Heckman, J., and R. Robb (1985). ‘Alternative Methods for Evaluating the Impact of Interventions’. In J. Heckman and B. Singer (eds), *Longitudinal Analysis of Labour Market Data*. Econometric Society Monograph 10. Cambridge: Cambridge University Press.

Heckman, J., R. LaLonde, and J. Smith (1999). ‘The Economics of Active Labour Market Programmes’. In O. Ashenfelter and D. Card (eds), *Handbook of Labour Economics*, vol. 3. Amsterdam: Elsevier Science.

Glewwe, P., M. Kremer, and S. Moulin (2009). ‘Many Children Left Behind: Textbooks and Test Scores in Kenya’. *American Economic Journal: Applied Economics*, 1(1): 112–35.

Hirano, K., G. Imbens, and G. Ridder (2003). ‘Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score’. *Econometrica*, 71: 1161–89.

Huber, M. (2011). ‘Testing for Covariate Balance Using Quantile Regression and Resampling Methods’. *Journal of Applied Statistics*, 38(12): 2881–99.

Imbens, G., and J. Angrist (1994). ‘Identification and Estimation of Local Average Treatment Effects’. *Econometrica*, 62(2): 467–75.

Imbens, G., and T. Lemieux (2008). ‘Regression Discontinuity: A Guide to Practice’. *Journal of Econometrics*, 142(2): 615–35.

Jalan, J., and M. Ravallion (2003). ‘Estimating the Benefit Incidence for an Anti-Poverty Programme Using Propensity Score Matching’. *Journal of Business and Economic Statistics*, 21(1): 19–30.

Lalonde, R. (1986). ‘Evaluating the Econometric Evaluation of Training Programmes with Experimental Data’. *American Economic Review*, 76: 604–20.

Martin, B,. S. Ji, S. Maudsley, and M. Mattson (2010). ‘ “Control” Laboratory Rodents Are Metabolically Morbid: What It Matters’. *Proceedings of the National Academy of Sciences*, 107(4): 6127-–33.

McCrary, J. (2008). ‘Testing for Manipulation of the Running Variable in Regression Discontinuity Design’. *Journal of Econometrics*, 142(2).

Meier, P. (1972). ‘The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine’. In J. Tanur (ed.), *Statistics: A Guide to the Unknown*. San Francisco: Holden Day, 2-13.

Mekasha, T., and F. Tarp (2011). ‘Aid and Growth: What Meta-Analysis Reveals’. WIDER Working Paper 2011/22. Helsinki: UNU-WIDER.

Miller, G., D. Pinto, and M. Vera-Hernández (2009). ‘High-Powered Incentives in Developing Country Health Insurance’. NBER Working Paper 15456. Cambridge, MA: National Bureau of Economic Research.

Muralidharan, K., and V. Sundararaman (2011). ‘Teacher Performance Pay: Experimental Evidence from India’. *Journal of Political Economy*, 119(1): 39–77.

Ozier, O. (2011). ‘The Impact of Secondary School in Kenya: A Regression Discontinuity Analysis’. Berkeley: University of California at Berkeley. Manuscript.

Ravallion, M., E. Galasso, T. Lazo, and E. Philipp (2005). ‘What Can Ex-Participants Reveal about a Programme’s Impact?’. *Journal of Human Resources*, 40(1): 208–30.

Rosenbaum, P. (1995). *Observational Studies*. New York: Springer-Verlag.

Rosenbaum, P., and D. Rubin (1993). ‘The Central Role of the Propensity Score in Observational Studies for Causal Effects’. *Biometrika*, 40: 41–55.

Rubin, D. (2005). ‘The Design *versus* the Analysis of Observational Studies for Causal Effects: Parallels with the Designs of Randomized Trials’. *Statistics in Medicine*, 26(1): 20–36.

Rubin, D. (2008). ‘For Objective Causal Inference, Design Trumps Analysis’. *The Annals of Applied Statistics*, 2(4): 808–40.

Subramanian, S., and A. Deaton (1996). ‘The Demand for Food and Calories’. *Journal of Political Economy*, 104(1): 133–62.

Shaikh, A. M., M. Simonsen, E. Vytlacil, and N. Yildiz (2009). ‘A Specification Test for the Propensity Score Using its Distribution Conditional on Participation’. *Journal of Econometrics*, 151(1): 33–46.

Smith, J., and P. Todd (2005). ‘Does Matching Overcome Lalonde’s Critique of Nonexperimental Estimators?’. *Journal of Econometrics*, 125(1-2): 305–53.

Thistlewaite, D., and D. Campbell (1960). ‘Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment’. *Journal of Educational Psychology*, 51: 309–17.

VandenBosch, T. (2011). ‘International Human Subjects Research Risks’. Office of Human Research Compliance Review. Ann Arbour, MI: University of Michigan.

van der Klaauw, W. (2002). ‘Estimating the Effect of Financial Aid Offers on College Enrolment: A Regression Discontinuity Approach’. *International Economic Review*, 43: 1249–87.

van der Klaauw, W. (2008). ‘Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics’. *Labor*, 22(2): 219–45.

Wooldridge, J. (2001). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.