CENTRE FOR ECONOMIC POLICY RESEARCH

Australian National University

DISCUSSION PAPERS

Econometrics for Summative Evaluations: An Introduction to Recent Developments

Deborah A. Cobb-Clark[1] and Thomas Crossley[2]

DISCUSSION PAPER NO. 454

December 2002

ISSN: 1442-8636
ISBN: 0 7315 3524 3

1 Deborah Cobb-Clark, Director, Social Policy Evaluation, Analysis and Research Centre, , Research School of Social Sciences, Australian National University, Canberra ACT 0200, Australia.

2 Thomas Crossley, Department of Economics, McMaster University, Canada and Centre for Economic Policy Research, Research School of Social Sciences, Australian National University, Canberra ACT 0200, Australia.

We gratefully acknowledge financial support form the Commonwealth Department of Family and Community Services, Australia.

CONTENTS

ABSTRACT

There has recently been a rapid expansion of interest in the econometrics of summative program evaluation, both within Australia and around the world. We provide a review of the key issue and recent developments in this field. A central feature of recent developments is the attempt to allow for program impacts that vary across individuals. This contrasts with earlier econometric approaches which implicitly assumed a homogenous treatment effect. We survey alternative non- experimental estimation strategies, and note that they can be characterised by (1) an assumption about how untreated outcomes vary across individuals: this assumption in turn suggests how the counter-factual untreated outcomes of program participants should be estimated, and (2) the way in which the estimator aggregates or weights the program impacts of different individuals in the treatment group.

JEL Classifications: J68, C1

Keywords program, evaluation, econometrics

1 Introduction

In Australia, as in many other countries, there is an increasing interest in the methods that can be used to evaluate the effects of social programs and public policies. While academic researchers are increasingly focused on assessing the strengths and weaknesses of the evaluation methodology itself, policy makers are turning to the results to provide the foundations for evidence-based policy.

In Australia, these trends are perhaps most evident in recent government initiatives to expand the capacity for social policy evaluation generally, but particularly in regards to income-support policies. Considerable resources are being devoted to developing additional data sources - for example, the Household Income Longitudinal Data for Australia (HILDA) and the Longitudinal Data Set (LDS) from the Department of Family and Community Services (FaCS) - and in some cases to making existing data more available to researchers.[1] Social policy evaluation centres have been established at several universities and there is increased interest in the use of a range of evaluation techniques including social experiments.[2] Finally, the government's 2001 budget included a substantial budget for the evaluation of the proposed welfare reform package-the Australians Working Together Act.

It is often useful when thinking about evaluation to make two distinctions. First we distinguish between research describing the outcomes of different groups (which we might refer to as "monitoring") and evaluation, which focuses on particular policies or interventions. Second, in the evaluation context, it is useful to distinguish between descriptive research that outlines the intended design and actual implementation of a particular policy, and evaluative research that assesses how well the policy works. Human Resources Development Canada (1998), for example, classifies evaluation methods into first, methods that aim to determine if a program has been implemented as planned ("process evaluation") and second, methods that measure the program's success in meeting its objectives ("summative evaluations").

[1] In particular, the Department of Immigration and Multicultural Affairs has recently adopted a policy of making unit-record data from the Longitudinal Survey of Immigrants to Australia available to researchers without charge.

[2] Since 1999, for example, FaCS has conducted at least four randomized trials involving samples of Parenting Payment, Mature Age Allowance, and NewStart Allowance clients.

Monitoring, process evaluation and summative evaluation can all provide vital information to policy makers. It is important to understand the differences between them, however. Essentially, summative evaluation involves working out how individuals' outcomes were *altered* by a program or policy. To do this we need to know what individuals' outcomes would have been had they not participated in the program (or been affected by the policy) and how this differs from what we in fact observe. The difference in these two outcomes provides a measure of program impact. Unfortunately, this counter-factual outcome is not observed and a way of estimating it must be devised. The key difficulty in summative evaluation lies in identifying a suitable counter-factual and both experimental and non-experimental methods can be used to solve this problem. In contrast, monitoring and process evaluation involve a description of observed outcomes and program activities, but neither involves the construction of a "counter-factual". Monitoring and process evaluation tell us about what happened - in terms of individual outcomes and program implementation respectively - but neither tells us how observed outcomes differed from those that would have occured in the absence of the program. Only summative evaluation is informative about program *impacts*.

We provide an overview of the summative evaluation problem and of various methods available for constructing a sensible counter-factual. While experimental approaches will be discussed, the primary focus will be on non-experimental or econometric solutions. Process evaluation will not be discussed at all and we will use 'evaluation' to mean summative evaluation only.[3] We will pay careful attention to the nature of the evaluation problem and will emphasize that the key issue is how to best model individual heterogeneity in both outcomes and impacts. Our review of methods is not meant to be comprehensive but rather is designed to build intuition about the nature and merits of various approaches. More comprehensive and detailed surveys can be found in Blundell and Costa Dias (2000), Heckman, Lalonde and Smith (1999) and Angrist and Krueger (1999).

In the next section we introduce the 'potential outcomes framework' which has become the standard way of thinking about the evaluation problem. To help connect the evaluation problem to "textbook" econometrics, we illustrate that the potential outcomes

[3] Like most of the literature, we will focus on evaluation in a partial (rather than general) equilibrium framework. See Heckman, LaLonde, and Smith (1999) for discussion of general equilibrium effects.

framework is in fact a random coefficients model. In Section 3 we discuss the relative merits of econometrics and social experiments as solutions to the evaluation problem. The fourth and largest section of the paper surveys alternative econometric methods for estimating program impacts. Section 5 offers some concluding remarks.

2 The Potential Outcomes Framework

Let us suppose that we wish to evaluate the impact of some program - say a training program- on some outcome of interest - for example, earnings.[4] We shall refer to the program as the 'treatment' or 'intervention'. Let Y_1 and Y_0 be random variables capturing the outcome for an individual if she does and does not receive training respectively. The realizations of these random variables for individual i are given by Y_{1i} and Y_{0i} and the impact of training for this individual is given by $\Delta_i = Y_{1i} - Y_{0i}$.

For those who receive training we observe only the training outcome (Y_{1i}), while for those who do not train we observe only the non-training outcome (Y_{0i}). That is, for each individual, the only observed outcome is:

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i} \tag{1}$$

where D_i is a dummy variable indicating the incidence of training. For no individual do we observe more than one of (Y_{1i}, Y_{0i}).[5] This is often referred to as the "potential outcomes framework" which was first developed by Rubin (1974) and is closely related to the Roy (1951) model.

There are several things to note about this framework. First, it emphasizes the distinction between outcomes (Y_{1i} and Y_{0i}) and impacts (Δ_i), thus formalizing the difference between outcome monitoring and (summative) evaluation discussed above. Monitoring is about outcomes (Y_{1i} and Y_{0i}), whereas evaluation is about impacts (Δ_i). The importance of this distinction is emphasized by the research on various labour market programs which shows that outcomes for program participants are often poorly correlated with program

[4] Thinking about training and earnings may help our intuition, however the ideas in this section are quite general. Outcomes could be wages or employment, but also health, family formation or dissolution, or any number of other things. Similarly the type of conceivable interventions is almost without limit.

[5] The distinction between realized and observed outcomes is a bit subtle. The idea is that each individual does have realizations of Y_{1i} and Y_{0i} - i.e., earnings if trained and not trained respectively. However, for each individual only one of these two realizations is ever observed.

impacts (Heckman, Heinrich and Smith, 1999). Measuring only the former often tells us little about the latter.

Second, the potential outcomes framework explicitly allows for heterogeneity in the program impacts as well as in untreated outcomes. Indeed, we can make this explicit by rearranging equation 1 to yield

$$Y_i = Y_{0i} + \Delta_i D_i \tag{2}$$

The provision for heterogeneity in treatment effects is an important distinction between this framework and models which are more familiar to applied econometricians. We develop this further below.

Third, note that while the potential outcomes framework is quite rich in allowing for heterogeneity in both outcomes and impacts, it is also in some ways restrictive. In particular, this model makes a stable-unit-treatment-value assumption. In other words, the impact of an intervention may vary across individuals but is assumed to be constant for a particular individual. This means, for example, that the impact of a program on an individual is assumed to be independent of whether other individuals are also participating in the program so that this framework is not well suited to thinking about interaction effects or general equilibrium effects.

2.1 Equity and Efficiency

Although the potential outcomes framework is a statistical model, it is nevertheless very useful for thinking about economic issues. For example, we can use it to highlight both efficiency and equity objectives. Economic efficiency require maximizing $\sum \Delta_i$ net of costs. Assuming constant program costs across individuals, those individuals with the largest Δ_i should be selected for treatment in order to maximize the per dollar impact of program expenditures (and we certainly would not want to treat any individual for whom Δ_i was less than the program costs). It is efficient to treat those who have the most to gain from treatment. However, if the program is beneficial (i.e, Δ_i is positive), equity considerations might dictate that we treat individuals with the worst outcomes in the absence of treatment. Thus, efficiency suggests treating individuals with high Δ_i, while equity may require treating individuals with low Y_{0i}. Efficiency and equity goals will be

at odds if $cov(Y_{0i}, \Delta_i) > 0$, that is, if those who would benefit most from the intervention are also those who do best in the absence of the intervention.

We can also use this framework to think about the relationship between the impact of a particular program and the incentives of managers administrating that program. Suppose, for example, that an agency monitors outcomes rather than evaluating a program and that program administrators are rewarded on the basis of the outcomes of program participants (Y_{1i}) not on program impacts, Δ_i. Many economists have noted that this is likely to induce 'cream-skimming'. Because

$$Y_{1i} = Y_{0i} + \Delta_i \tag{3}$$

the program administrator may be able to maximize realizations of Y_{1i} (and hence her rewards) by choosing program participants with high Y_{0i}. In other words, the program administrator may choose to treat those individuals who would do well even without the intervention. Cream-skimming is normally thought of as a perverse outcome of bureaucratic incentives and it is clearly iniquitous. However, cream-skimming is inefficient only if $cov(Y_{0i}, \Delta_i) < 0$.[6]

Some empirical evidence on the correlation between Y_{0i} and Δ_i is beginning to accumulate. Heckman, Lalonde and Smith (1999), for example, review several training programs for the economically disadvantaged in which there is evidence of a negative correlation between Y_{0i} and Δ_i.This suggests a modest equity-efficiency trade-off for these types of programs and target populations. At the same time, the post-program outcomes – on which program administrators are often rewarded – are generally only weakly related to estimated program impacts (Heckman, Heinrich, and Smith, 1999; Barnow, 2000).

2.2 The Parameters of Interest

Individual heterogeneity in the impact of particular programs implies that we need to take some care in defining the parameter that we want to estimate. Two parameters are most frequently estimated in the literature. The first, is the population average treatment

[6] In this case, cream skimming is likely to be less effective in insuring good outcomes (Y_{1i}) and hence less tempting for program admistrators. The reason is that in the face of this negative correlation, chosing individuals with large non-treated outcomes also on average means choosing individuals who gain little from the program itself. Since $Y_{1i} = Y_{0i} + \Delta_i$, a small Δ_i tends to offset the advantage of choosing a large Y_{0i}.

effect, i.e. $E[\Delta_i]$. This is simply a weighted average of the average treatment effect for the treated and non-treated populations. Using the law of iterated expectations:

$$\begin{aligned} E[\Delta_i] &= E[Y_{1i} - Y_{0i}] \\ &= E[Y_{1i} - Y_{0i}|D_i = 1]P(D_i = 1) + E[Y_{1i} - Y_{0i}|D_i = 0]P(D_i = 0) \end{aligned} \tag{4}$$

This parameter tells us what the expected effect of the intervention would be on average for the entire population. At the same time, it is not clear why we would be interested in the effect of the treatment on those who did not - and might never - receive the treatment. For this reason, researchers are often more interested in the average effect of the treatment on the treated:

$$\begin{aligned} E[\Delta_i|D_i = 1] &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]. \end{aligned} \tag{5}$$

The most common evaluation context is one of *ex post* evaluation, where we wish to know what change in outcomes an intervention delivered for those who were subject to the intervention. The average effect of the treatment on the treated answers this question and also tells us what the effect of the treatment is likely to be if similar groups of individuals were to receive the same treatment.

One can also imagine situations where one would be interested in estimating the impact of the treatment on *some* of the untreated - such as a situation where the expansion of a program is under consideration. Although less frequently estimated, the marginal treatment effect – i.e., the mean effect of the program for those individuals at the margin of program participation – focuses on the effect of the treatment for those additional individuals who would be treated if the existing program were expanded.[7]

In the face of homogenous treatment effects, the effect of the program would be constant for all individuals and the average population treatment effect, the average effect of the treatment on the treated, and the marginal treatment effect would all be equivalent. However, a large empirical literature demonstrates that individuals generally respond

[7] This parameter was first introduced into the evaluation literature by Björklund and Moffit (1987) and extended by Heckman and Vytlacil in a series of articles. (See Heckman, 2001 for a review.)

quite differently to the same policy or program leading the marginal entrant into a social program to be quite different from the average participant (Heckman, 2001). The very real possibility of heterogeneity in the impact of a particular program implies that we need to think carefully about our research questions and the population for whom we would like to estimate an effect.

2.3 Linking the Evaluation Problem to Linear Regression

Recall the evaluation problem discussed above. While $E[Y_{1i}|D_i = 1]$ is easily estimated using data on outcomes for program participants, Y_{0i} is not observed for those individuals participating in the treatment. Thus, statisticians often think of the evaluation problem as a 'missing data problem'. How can $E[Y_{0i}|D_i = 1]$ can be estimated when Y_{0i} is not observed for those individuals for whom $D_i = 1$? One possibility is to use the outcomes of non-treated individuals as a measure of what treated individuals would have received had they not received treatment. In this case, we have

$$\begin{aligned} &E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &\qquad = E[Y_{1i} - Y_{0i}|D_i = 1] + (E[Y_{0i}|D_i{=}\ 1] - E[Y_{0i}|D_i{=}\ 0]) \\ &\qquad = E[\Delta_i|D_i = 1] + (E[Y_{0i}|D_i{=}\ 1] - E[Y_{0i}|D_i{=}\ 0]) \end{aligned} \tag{6}$$

Simply comparing the observed outcomes of those who do and do not get training yields a biased estimate of the average effect of treatment on the treated if trainees would have had different earnings than non-trainees in the absence of training. That is, the second right-hand-side term in equation (6) captures the bias due to selection effects: i.e., that those who get training have different untreated outcomes than do non-trainees. This is why economists often think of the evaluation problem as a 'selection problem'. Later, we will sometimes find it useful to distinguish between selection on observables - where the trainees and non-trainees differ in ways that are observable to the econometrician and selection on unobservables which occurs if trainees and non-trainees differ in ways that are not observable to the econometrician.

It is interesting to contrast the way the potential outcomes framework allows for heterogeneity with the standard linear regression model that is much more familiar to

applied econometricians. Consider again equation (2):

$$Y_i = Y_{0i} + \Delta_i D_i$$

Since Δ_i and Y_{0i} are random variables, they can be written as the sum of a mean and a deviation from that mean:

$$\begin{aligned} Y_i &= \{E[Y_0] + (Y_{0i} - E[Y_0])\} \\ &\quad + \{E[\Delta] + (\Delta_i - E[\Delta])\} D_i \\ Y_i &= (\alpha + \varepsilon_i) + (\beta + \nu_i) D_i \end{aligned} \tag{7}$$

with

$$\begin{aligned} \alpha &= E[Y_0] \\ \beta &= E[\Delta] \\ \varepsilon &= (Y_{0i} - E[Y_0]) \\ \upsilon &= (\Delta_i - E[\Delta]) \end{aligned}$$

Thus, the potential outcomes framework leads to a random coefficients model (see Greene, 1997) in which there is heterogeneity in both the slope and intercepts. Note that the missing data or selection problem outlined above - i.e., that $E[Y_{0i}|D_i = 1] \neq E[Y_{0i}|D_i = 0]$ - can now be seen to be a violation of one of the usual regression assumptions because it implies that $E[D_i \varepsilon_i] \neq 0$. In addition, the usual, constant-slope, linear regression model assumes that:

$$\nu_i = 0$$

$$(Y_{1i} - E[Y_1]) - (Y_{0i} - E[Y_0]) = 0$$

$$(Y_{1i} - Y_{0i}) = E[Y_1 - Y_0] = \Delta$$

That is, the standard, constant-slope linear regression model implies complete homogeneity across the population in the treatment effect ($\Delta = Y_{1i} - Y_{0i}$) - though not in the specific outcomes Y_{1i} and Y_{0i}. In this case, the distinction between the average effect of the treatment on the treated and the population average treatment effect is moot because

the effect of the treatment is identical for every one in the population. In many cases, this is an untenable assumption. Its hard to imagine, for example, that all potential trainees would benefit equally from training. Indeed, it seems likely that some individuals select into training precisely because either they, or program administrators, believe that they would benefit substantially relative to the rest of the eligible population. In most cases, a sensible model of public programs will lead to heterogeneity in program impacts, and any reasonable econometric estimator of those impacts must deal with this heterogeneity.

3 Social Experiments versus Econometrics

Experimental data provide one solution to the evaluation problem. If the receipt of the treatment (say training) is randomly assigned, then D_i is independent of other variables – including non-treated outcomes (Y_{0i}). This implies that:

$$E[Y_{0i}|D_i{=}\ 1] = E[Y_{0i}|D_i{=}\ 0] \tag{8}$$

Given this, the second term in equation (6) vanishes leaving:

$$E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = E[Y_{1i} - Y_{0i}|D_{1i} = 1]$$

and the average effect of the treatment on the treated (AETT) can be estimated by a simple comparison of the outcomes of treated and untreated individuals:

$$\widehat{AETT} = \frac{\sum Y_i D_i}{\sum D_i} - \frac{\sum Y_i(1-D_i)}{\sum(1-D_i)} = \overline{Y}_{Treated} - \overline{Y}_{Untreated} \tag{9}$$

This is why social experiments - situations where a policy or intervention is randomly assigned by design - are such attractive options for evaluating the effects of interventions. Econometricians would refer to equation (8) as a situation with 'no selection bias'. An alternative terminology, which is sometimes favoured by statisticians is that equation (8) corresponds to 'ignorable' treatment assignment (Rubin, 1978).

Furthermore, experiments are often credited with other virtues, such as producing results which are easily understood by policy makers (Burtless, 1995). Nevertheless, experiments are not a solution to every evaluation problem. First, social experiments are quite costly and practically limits them to a small number - much smaller than the number of policy interventions we might like to know about. Second, some interventions, for

either ethical or practical reasons, are simply not amenable to experimental manipulation. Third, experiments do not always work. They can be undermined by noncompliance and other problems. Fourth, experiments do not – in and of themselves – allow us to estimate all the potential parameters of interest. In particular, if (8) holds we can estimate an average treatment effect by comparing average outcomes for participants and nonparticipants. However, its important to note that because this procedure does not involve estimating a counter-factual untreated outcome for each member of the treatment group, we are unable to calculate other parameters of interest (Heckman and Smith, 1995). For example, we cannot estimate, $cov(Y_{0i}, \Delta_i))$, a parameter that, as noted in the previous section, determines whether the targeting of the program faces an equity-efficiency tradeoff. Nor can we estimate the median impact since the difference in the medians is, in general, not the median difference.[8]

3.1 The Internal and External Validity of Social Experiments

In discussing the relative merits of experimental and non-experimental methods of program evaluation, it is sometimes useful to distinguish between 'threats to internal validity' and 'threats to external validity'(Campbell and Stanley, 1966, Reicken and Boruch, 1978, Burtless, 1995). Internal validity refers to the ability to get an unbiased estimate of the parameter of interest. The missing data or selection problem characterized above (Equation (6)) is one potential threat to internal validity. If $E[Y_{0i}|D_i = 1] \neq E[Y_{0i}|D_i = 0]$ then a simple comparison of treated and untreated subjects does not estimate the average effect of the treatment on the treated. External validity relates to the applicability of the results of an evaluation to other situations. Would the same intervention generate the same outcomes if it were repeated, expanded or extended to another population?

Social experiments are subject to threats to both internal and external validity. The former typically arise when, in real world situations, there is noncompliance with the experimental design – that is, the random assignment of the treatment breaks down.

[8] Other strategies for estimating the distributional aspects of social policies do exist, however. Carniero, Hansen, and Heckman (2002) discuss a method for identifying the joint distribution of potential outcomes (Y_0, Y_1) when potential outcomes are generated by a small set of factors. In this case, it is possible to use a social experiment to estimate the distribution of those factors and to consequently generate the joint distribution of potential outcomes. Note that while it is possible to estimate differences in mean outcomes without reference to a structural model, the Carniero, Hansen, and Heckman approach requires specifying the process through which potential outcomes are generated.

Noncompliance occurs, for example, if the individuals in the control group obtain the treatment (or a close substitute) outside the scope of the study.[9] In the literature, this is referred to as substitution bias (Heckman and Smith, 1995). Noncompliance can also occur if some individuals in the treatment group do not undertake or complete the treatment. While techniques are available for dealing with dropouts (see for example, Heckman, Smith and Taber, 1998), they rely on standard econometrics moving the researcher from experimental to non-experimental methods. Finally, it is important to note that it is not just program clients who generate these types of noncompliance. Operationalizing a randomized trial can be difficult if program staff are unfamiliar with or poorly trained in evaluation methods or if they have ethical concerns about denying treatment to the control group (Riecken and Boruch, 1978).

The costs of social experiments often lead them to being of rather limited scale and this can generate threats to external validity. An intervention might have quite a different effect if, for example, a large number of individuals in a local labour market experience the intervention. Furthermore, if participation in a social experiment is not universal, it may well be that the group of individuals who are willing to participate in the trial (and hence be randomly assigned to a treatment or control group) is not representative of those individuals who would apply to participate in the subsequent program (when they could expect treatment with certainty). This effect of the randomization on enrollment in the program is referred to as randomization bias (Heckman and Smith, 1995, Blundell and Costa Dias, 2000), and it is an often cited example of a threat to external validity.

Even if the subjects in a trial are representative of participants in a potential program, it may be that these subjects' behaviour in the trial is different from what their behaviour in the actual program would be simply because they are part of an experiment. This might occur, for example, because they are aware that their outcomes are being observed and recorded. It may also be that their awareness of the experiment changes the social dynamics between participants and program staff or case workers or between the participants themselves. This possibility is often referred to as the Hawthorne Effect.[10]

[9] For example, some individuals not assigned to a training course might take some private training on their own.

[10] A series of experimental studies of workplace behaviour were conducted at the Hawthorne Plant of the Western Electric Company in the 1920s and 1930s. It is commonly reported that *all* of the experimental treatments (including reverting to the pre-experiment work arrangements) increased productivity, and

Finally it could be that subjects in an experiment behave differently because they are better informed about the nature of the intervention than would be the case in a real policy setting.

For further discussion of the merits and limitations of social experiments, see the exchange between Heckman and Smith (1995) and Burtless (1995) or Ricken and Boruch (1978). For our purposes, it is sufficient to conclude that for reasons of cost, ethics, logistics, validity, and the desire to estimate parameters other than average outcomes, we would also like to have non-experimental alternatives to program evaluation.

3.2 The Ability of Non-Experimental Methods to Solve the Evaluation Problem

In the absence of random assignment, we need an econometric method for dealing with heterogeneity in the untreated outcome (Y_{0i}). That is we need to find a solution to the selection problem (See equation (6)). In an influential paper in the mid 1980s (Lalonde, 1986), Robert Lalonde compared experimental estimates of the impact of a training program with a range of non-experimental estimates of the impact of the same program. The non-experimental estimates were generated by comparing the outcomes of treated individuals with control individuals drawn from large population surveys, and correcting for selection with a variety of popular econometric procedures. The results were rather dismal. The non-experimental estimators generated a huge range of estimates, none of which was particularly close to the experimental estimate.

However, in the intervening years the profession has learned a lot. We now have better econometrics.[11] Perhaps even more importantly, we now understand a lot more about why Lalonde's non-experimental estimators performed poorly, and under what circumstances a non-experimental evaluation might be expected to produce more satisfactory results. The potential outcomes framework has helped in this regard, as has accumulating experience with non-experimental estimators, especially in situations where an experimental benchmark is available. Examples of the latter include re-analyses of data from

this is attributed to changing social relations among the workers and between the workers and the managers/experimenters. This potential threat to the external validity of experiments has thus come to be called the Hawthorne Effect (see, for example, French (1953) and Bracht and Glass (1968)). Interestingly, more recent research has called into question whether there actually was a Hawthorne Effect in the Hawthorne Experiments (Jones, 1992).

[11] For example, the development of nonparametric econometrics has been very significant.

the same experiment as analyzed by Lalonde (see Heckman and Hotz, 1989, Dehejia and Wahba, 1999 and Smith and Todd, 2000) as well as the analysis by Heckman, Ichamura and Todd (1997) and Heckman, Ichamura, Smith and Todd (1998) of experimental data from the Job Training Partnership Act (JTPA) in the United States.

In particular, Heckman and Hotz (1989) show that of the non-experimental estimators that Lalonde employed, those that performed poorly are typically ruled out by specification tests which can be performed with the data at hand. At the same time, Freedman (1991) argues that while Heckman and Hotz were able to ex post rule out those estimators that performed poorly using the experimental data as a benchmark, the real question is whether those estimators would have been ruled out ex ante had the experimental data not been available. Dehejia and Wahba (1999) report a successful non-experimental analysis of Lalonde's data (in the sense of replicating the experimental impact estimates) using an econometric estimator based propensity score matching (to be discussed below) though the robustness of their success has been questioned by Smith and Todd (2000). Thus, the development of econometric tools for program evaluation is promising, but ongoing, and sometimes controversial. The next section of this paper provides an introductory review of some of the econometric tools that are currently applied to the evaluation problem.

4 Alternative Non-Experimental Estimators of Program Impacts

This section surveys a number of econometric or statistical approaches to estimating program impacts. These include regression-based methods, matching estimators, the method of instrumental variables, and quasi-experimental comparisons (also called "natural experiments"). These methods can be characterized in two ways. First, each makes an assumption about how untreated outcomes vary across individuals and this assumption in turn suggests how the counter-factual untreated outcome of program participants should be estimated. Second, a distinguishing feature of all of these econometric methods is the way in which they, either explicitly or implicitly, aggregate or weight the estimated treatment effects of different individuals.

4.1 Regression

As discussed in Section 2, the evaluation problem hinges on dealing appropriately with both heterogeneity in treatment effects (Δ_i) and heterogeneity in untreated outcomes (Y_{0i}). Equation (7) demonstrated that this could be equivalently thought of as heterogeneity in an intercept term and in the coefficient on a treatment status dummy.

4.1.1 Selection Effects

In order to understand selection (heterogeneity in untreated outcomes Y_{0i}) in a regression context, assume for the moment that treatment effects are homogeneous, so that the effect of treatment is to shift the intercept for those individuals receiving treatment. Specifically,

$$Y_i = Y_{0i} + (Y_1 - Y_0)D_i = \alpha + \beta D_i + \varepsilon_i \tag{10}$$

This is a simple linear regression model and Ordinary Least Squares (OLS) regression of Y_i on D_i provides an unbiased estimate of β if:

$$E[\varepsilon_i|D_i{=}\ 1] = E[\varepsilon_i|D_i{=}\ 0] = E[\varepsilon_i] = 0$$

where $E[\varepsilon_i] = 0$ by construction (see Equation (7)). As previously noted, the equality of $E[\varepsilon_i|D_i{=}\ 1]$ and $E[\varepsilon_i|D_i{=}\ 0]$ is equivalent to:

$$E[Y_{0i}|D_i{=}\ 1] = E[Y_{0i}|D_i{=}\ 0]$$

implying that we have no selection.

If this is not the case – and selection effects are present – we can always write:

$$Y_i = \alpha + \beta D_i + E[\varepsilon_i|D_i]{+}\varepsilon_i^* \tag{11}$$

where $\varepsilon_i^* = \varepsilon_i - E[\varepsilon_i|D_i]$ and:

$$E[\varepsilon_i^*|D_i{=}\ 1] = E[\varepsilon_i^*|D_i{=}\ 0] = E[\varepsilon_i^*] = 0$$

by construction.[12] The above highlights the fact that the selection problem results from

[12] To see this note that applying the law of iterated expectations yields

$$\begin{aligned} E[\varepsilon_i^*|D_{1i} = 1] &= E[(\varepsilon_i - E[\varepsilon_i|D_{1i}])|D_{1i} = 1] \\ &= E[\varepsilon_i|D_{1i} = 1] - E[E[\varepsilon_i|D_{1i}])|D_{1i} = 1] \\ &= E[\varepsilon_i|D_{1i} = 1]{-}E[\varepsilon_i|D_{1i} = 1] = 0 \end{aligned}$$

Similarly for $D_{1i} = 0$.

the fact that there is an omitted or confounding variable, $E[\varepsilon_i|D_i]$,(see Heckman, 1979). Regression-based approaches to the evaluation problem in the face of heterogenous untreated outcomes rely on the notion that variation in untreated outcomes across individuals can be captured by observable characteristics (such as age, education, gender or labour market experience).

To see this more clearly, suppose that the linear projection of Y_{0i} on x is given by:

$$E[Y_{0i}|x_i] = \alpha + \gamma x_i$$

where x_i is an observable individual characteristic which we express in lower case to indicate that $x_i = X_i - E[X_i]$ so that $E[x_i] = 0$.[13] This implies[14]:

$$\begin{aligned} Y_{0i} &= \alpha + \gamma x_i + u_i \\ E[u_i|x] &= E[u_i] = 0 \end{aligned}$$

If u_i are 'random', so that $E[u_i|D_i{=}\,1] = E[u_i|D_i{=}\,0] = E[u_i] = 0$, then

$$Y_i = \alpha + \beta D_i + \gamma x_i + u_i \tag{12}$$

and equation (12) could be estimated by simple linear regression to give an unbiased estimate of β. The key here is that the homogeneous treatment effect β is identified by a conditional mean independence assumption which (if it holds) implies that (potential) untreated outcomes do not vary systematically between treated and untreated groups once we control for differences in x. In other words,

$$E[Y_{0i}|D_i{=}\,1, x_i] = E[Y_{0i}|D_i{=}\,0, x_i] = E[Y_{0i}|x_i] = \alpha + \gamma x_i. \tag{13}$$

4.1.2 Regression When Treatment Effects are Homogeneous

Recall that the linear regression estimator is simply a weighted average of the observed outcomes. That is, if we regress Y_i on D_i the OLS estimate of β is:

$$\begin{aligned} \widehat{\beta} &= \frac{\sum(D_i - \overline{D})(Y_i - \overline{Y})}{\sum(D_i - \overline{D})^2} = \sum \frac{(D_i - \overline{D})}{\sum(D_i - \overline{D})^2} Y_i \\ &= \sum W_i Y_i \text{ with } W_i = \frac{(D_i - \overline{D})}{\sum(D_i - \overline{D})^2} \end{aligned} \tag{14}$$

[13] $\gamma E[X]$ can be subsumed in α without loss of generality.

[14] An alternative way of writing this is: $E[\varepsilon_i] = \gamma x_i$, $\varepsilon_i = \gamma x_i + u_i$, $E[u_i|x] = E[u_i] = 0$.

which identifies the homogeneous treatment effect if there is no selection. In equation (14), $\overline{D}$ is the sample average of the treatment dummy or equivalently the sample probability of treatment which we shall denote as p for convenience. Since $D_i = 1$ for treated individuals and 0 for untreated individuals, the weights (W_i) take a constant value $1-p$ for all treated individuals and a constant value $-p$ for all untreated individuals. With some additional algebraic manipulation its easy to show that:

$$\widehat{\beta} = \frac{1}{np} \sum^{treated} Y_i - \frac{1}{n(1-p)} \cdot \sum^{untreated} Y_i \tag{15}$$

where n is the number of individuals in the sample, np is the number of treated individuals, and $n(1-p)$ is the number of untreated individuals. Thus, when treatment effects are homogenous a regression of Y_i on D_i simply returns the difference in the sample average outcomes of treated and untreated individuals.

A property of linear regression is that the regression of Y_i on D_i and x_i is exactly equivalent to a two step 'residual regression' procedure which involves first, regressing D_i on x_i to get predicted values D_x and then second, regressing Y_i on $(D_i - D_x)$ – the residuals from the first stage regression.[15] Using this residual regression procedure highlights the fact that to use linear regression to get an unbiased estimate of a homogenous treatment effect we do not require strict independence (equation (8)) but only conditional independence (equation (13)). At the same time, it can also be used to show that if we regress Y_i on D_i and x_i the OLS estimate of the treatment effect – though unbiased – is a different weighted average of the outcomes than in equation (14). In particular:

$$\begin{aligned} \widehat{\beta} &= \frac{\sum(D_i - D_x)(Y_i - \overline{Y})}{\sum(D_i - D_x)^2} = \sum \frac{(D_i - D_x)}{\sum(D_i - D_x)^2} Y_i \\ &= \sum W_i^* Y_i \text{ with } W_i^* = \frac{(D_i - D_x)}{\sum(D_i - D_x)^2} \end{aligned} \tag{16}$$

where the weight W_i^* is constant for a given value of D_i and x_i so that we can write:

$$\widehat{\beta} = \frac{1}{\sum(D_i - D_x)^2} \sum_x \left\{ (1 - D_x) \cdot \sum_{treated} Y_i - D_x \cdot \sum_{untreated} Y_i \right\} \tag{17}$$

To gain intuition into the different OLS estimates generated by equations (15) and (17) recall that given homogenous treatment effects Δ, $Y_i = Y_{1i} = Y_{0i} + \Delta$ for treated

[15] See Johnson and DiNardo, pp. 101 - 103.

individuals and $Y_i = Y_{0i}$ for untreated individuals. Define $f(x)$ as the sample probability of participating in the treatment given a particular realization of x (ie., given a certain set of observable characteristics). After some further manipulations we can show that:

$$\begin{aligned} \widehat{\beta} &= \left(\frac{n}{\sum (D_i - D_x)^2}\right) \bullet \\ &\qquad \sum_x f(x)(1-D_x)D_x \left\{\Delta + \frac{1}{D_x}\sum_{treated} Y_{0i} - \frac{1}{(1-D_x)}\sum_{untreat} Y_{i0}\right\} \\ &= \frac{1}{K}\sum_x \omega_x \{\Delta + v_i\} \end{aligned} \tag{18}$$

where

$$\omega_x = f(x)(1-D_x)D_x \tag{19}$$

The term $\frac{1}{K}$ is simply a scaling factor so that equation (18) is effectively a weighted sum (over each of the realizations of x) of the term in curled brackets $\{\}$. If the conditional mean independence assumption holds then:

$$E[v_i] = E\left[\frac{1}{D_x}\cdot\sum_{treated} Y_{0i} - \frac{1}{(1-D_x)}\cdot\sum_{untreated} Y_{i0}\right] = 0 \tag{20}$$

Thus, the term in curled brackets in equation (18) is an unbiased estimator of the homogenous treatment effect Δ. Intuitively, OLS regression calculates an unbiased estimate of Δ at each value of x and then generates a weighted average of these estimates. The weights, ω_x, are proportional to $f(x)$ so that realizations of x that that correspond to a higher probability of receiving treatment receive more weight. The weights are also proportional to $(1-D_x)D_x$ which is the (binomial) variance of D_i at that value of x. This makes intuitive sense because the more variable D_i is at a given value of x, the more precisely Δ can be estimated at that value of x. Thus, OLS regression is averaging different estimates of the homogenous treatment effect Δ and placing more weight on the more precise estimates. In fact, any basic econometrics textbook will tell us that under standard assumptions least squares is the best linear unbiased estimator, so this weighting is not just sensible but optimal.[16]

[16] This is true under the standard assumptions, including homoskedasticity of the error terms. See, for example, Greene (1997) Chapter 6.

4.1.3 Regression When Treatment Effects are Heterogenous

What does least squares regression estimate in a heterogeneous impact context? For ease of exposition, let us assume that individuals with the same value of x have the same treatment effect, but that the impact of treatment can differ across individuals with different observable characteristics. In this case, we can write:

$$\Delta_i = \Delta_x$$

and we have:

$$\widehat{\beta} = \frac{1}{K} \sum_x \omega_x \{\Delta_x + v_i\} \tag{21}$$

The terms K, ω_x and v_i are defined exactly as above. In the face of heterogenous treatment effects, least squares regression still calculates a weighted average of a series of unbiased estimates of the treatment effect conditional on each value of x. The difference between this case and the one discussed in Section 4.1.2 is that the true treatment effect is *different* at each x, so that rather than taking a weighted average of multiple estimates of a single treatment effect Δ, we are taking a weighted average of unbiased estimates of different treatment effects. This gives an unbiased estimate of a weighted average of treatment effects. However, as noted by Angrist (1998), it is not clear that this particular weighted average is something that a researcher would ever be interested in. In particular, the weights, ω_x, given in equation (19) put greater weight on the estimates of the treatment effect at those values of x where D_i has greater (conditional) variance. When treatment effects are heterogeneous, regression does not estimate the population average treatment effect or the average effect of the treatment on the treated.

4.1.4 Summary

To summarize, regression approaches to the evaluation problem do not suffer from selection bias if a conditional mean independence assumption holds. In fact, regression is often characterized as a method for dealing with selection on observables. On the other hand, if (potential) untreated outcomes vary across treatment status in a way that is not completely captured by observable characteristics, then the conditional mean independence assumption does not hold, and regression estimates will suffer from selection bias.

If treatment effects are homogenous (and the conditional mean independence assumption holds) regression returns an (statistically) efficient estimate of that treatment effect. At the same time, if treatment effects vary across individuals, then regression returns a weighted average of the treatment effects of different individuals where the weight for each individual is determined by her observable characteristics (in particular those characteristics which are used as control variables in the regressions). These weights are designed (as OLS is) to return an a efficient estimate when treatment effects are homogenous. There is no reason for this weighted average to correspond to any parameter of interest in a heterogeneous effect context.

4.2 Matching

Regression is not the only way to deal with selection on observables. Matching estimators are another class of estimators which rely on the same conditional mean independence assumption as regression. However, the weighting of estimated treatment effects across different individuals remains under the explicit control of the researcher rather than being implicit in the estimator, as in OLS. Thus matching methods are likely to be more amenable to heterogeneous treatment effect contexts. We turn to these methods here.

4.2.1 The Basics

A matching estimator is based on the simple idea that the best estimate of the (unobserved) counterfactual untreated outcome for an individual in the treatment group, is the outcome of the individual or individuals most like them (in terms of observable characteristics) in the control group. Specifically, matching uses (an) observation(s) ($Y_{0i,}D_{i}= 0, X_i$ close to x) drawn from the control group ($D_i= 0$) to generate an estimate of the counterfactual ($\widehat{Y_{0i}}, D_i= 1, X_i = x$) for each observation ($Y_{1i,}D_{i}= 1, X_i = x$) in the treatment group ($D_i= 1$).

Because we explicitly calculate a counterfactual untreated outcome ($\widehat{Y_{0i}}$) and hence impact ($\widehat{\Delta_i} = Y_{1i} - \widehat{Y_{0i}}$) for each individual in the treatment group, we can aggregate these estimated impacts in any way we chose, depending on our parameter of interest. For example, the estimate of the average effect of the treatment on the treated is:

$$\frac{1}{\sum D_i}\sum(Y_{1i} - \widehat{Y_{0i}})D_i = \frac{1}{\sum D_i}\sum\widehat{\Delta_i}D_i$$

This ability to explicitly control the weighting is an advantage of matching techniques over regression-based approaches in a heterogeneous treatment effect context.

Matching methods rely on a conditional independence assumption**.** Observations on Y_{0i} from (matched) individuals in the control group can act as a counterfactual for the unobserved, untreated outcomes of their intervention group counterparts if the distribution of untreated outcomes is independent of assignment (D_i) conditional on x. In fact, if the aim is just to estimate various average treatment effects (and not, for example, medians), then this is unnecessarily strong. All that is required in that case is conditional mean independence - the same assumption that underlies a regression-based approach.

Rosenbaum and Rubin (1983) noted that in addition to the conditional independence assumption, matching estimators require that a common support condition holds. Defining $P(x)$ as the probability of treatment for an individual with observable characteristics x, the common support condition holds if $0 < P(x) < 1$ for all values of x observed in the treatment group. The logic behind this observation is straightforward. If $P(x) = 1$ for some x, then the conditional independence assumption does not help us because we have no observations on $(Y_{0i,}D_i{=}\, 0, X_i = x)$ from which to construct a counterfactual for the treated individuals with $X_i = x$.

4.2.2 Propensity Score Methods

One potential problem with matching methods has to do with the 'curse of dimensionality'. Since the method corrects for selection by controlling for observable characteristics, and because individuals may potentially differ in many dimensions, the matching problem can quickly become intractable with finite data. However, in a somewhat surprising result, Rosenbaum and Rubin (1983) showed that if outcomes (Y_{0i}) are independent of D_i conditional on $X_i = x$ then they are also independent of D_i conditional on $P(X_i) = P(x)$. Often $P(x) = \Pr(D_i{=}\, 1|X_i = x)$ is referred to as the propensity score. The practical importance of this result is that it is only necessary to match on $P(X_i)$ – or a consistent estimate $\widehat{P(X_i)}$ – which significantly reduces the dimension of the problem and makes it feasible to match on a large number of covariates.[17]

Note that propensity score matching requires conditional independence (or conditional

[17] This property of the propensity score is sometimes called the 'balancing property'.

mean independence) and common support conditions just as other matching approaches. In fact, one additional virtue of propensity score methods is that it becomes easy to check common support conditions by simply examining the support of the propensity score distributions in the treated and untreated samples.

4.2.3 Summary

The principal difference between regression-based and matching techniques is the flexibility the latter gives the researcher in choosing how to aggregate heterogeneous impacts. In an environment where program impacts vary across individuals, regression imposes a particular, and perhaps uninteresting, weighting when calculating an average treatment effect. In contrast, in a matching estimator the weighting is easily manipulated so that interesting parameters (that is, interesting averages) like the average effect of the treatment on the treated, can be estimated. On the other hand, if treatment effects are homogeneous, then the regression-based estimator is more efficient.

An important implication of this is that different econometric methods may yield different point estimates all of which could be correct in the sense of not suffering from selection bias. It might simply be the case that they are estimating the impact of the program on different groups of people, or implicitly weighting the groups differently in the generation of an average impact estimate. Thus, considerable care must be taken in the interpretation and comparison of the results of different econometric methods.

4.3 Instrumental Variables

Both regression-based and matching methods rely on conditional comparisons – i.e., a conditional independence assumption – in producing unbiased estimates. Consider the following simple example. Suppose we believe that untreated outcomes vary only by the region in which an individual resides. In this case, the untreated outcomes of program participants should be estimated using the untreated outcomes of nonparticipants from same region. Conversely, if untreated outcomes are constant across regions while the probability of program participation varies across regions, we can get unbiased estimates of the program's impact by comparing outcomes across regions. Simply put, if outcomes do not vary across regions *except* as a consequence of differences in program participation,

then we can attribute outcome differences across regions to regional differences in participation. Furthermore, if we know how much participation varies across regions then we can quantify program effects by relating regional differences in outcomes to regional differences in the intensity of participation. This is the basic intuition behind instrumental variable (IV) methods and related techniques.

Instrumental variable methods – like regression-based techniques – have been around along time and are discussed in depth standard econometric textbooks. In this survey we are interested in instrumental variables estimates of treatment effects and we focus on three issues. First, we want to contrast the assumptions required for an instrumental variables approach to yield unbiased estimates of a program's impact with the conditional independence assumption that justifies a regression-based or matching approach. Second, we discuss 'natural experiments' as a source of instruments. Third, we examine instrumental variable estimates in the presence of heterogeneity in treatment effects.

4.3.1 The Basics

In order to contrast instrumental variable with regression-based and matching approaches, we begin by considering the case of homogenous treatment effects. Substituting homogenous treatment effects into equation (2) we have:

$$\begin{aligned} Y_i &= Y_{0i} + \Delta D_i && (22) \\ &= \overline{Y_0} + \Delta D_i + \varepsilon_i && (23) \end{aligned}$$

where $\varepsilon_i = Y_{0i} - \overline{Y_0}$, (so that $\overline{\varepsilon_i} = 0$ by construction). For convenience, rewrite the above taking deviations from means so that

$$y_i = \Delta d_i + \varepsilon_i \tag{24}$$

where $y_i = Y_i - \overline{Y}$ and $d_i = D_i - \overline{D}$. Then note that:

$$\begin{aligned} cov(d_i y_i) &= \Delta var(d_i) + cov(d_i \varepsilon_i) && (25) \\ \frac{cov(d_i y_i)}{var(d_i)} &= \Delta + \frac{cov(d_i \varepsilon_i)}{var(d_i)} && (26) \end{aligned}$$

If $cov(d_i\varepsilon_i) = 0$ then a sensible estimator of Δ is simply the ratio of the sample analogues of $cov(d_i y_i)$ and $var(d_i)$, that is $\widehat{\Delta} = \frac{d'y}{d'd}$. Of course, this is just the OLS estimator, and the requirement that $cov(d_i\varepsilon_i) = 0$ is equivalent to the conditional mean independence assumption discussed in Section 4.1. If this assumption does not hold and $cov(d_i\varepsilon_i) \neq 0$ then this approach does not provide an attractive estimation strategy.

Now consider another random variable, Z_i and again use the lower case to denote deviations from mean. Then

$$cov(z_i y_i) = \Delta cov(z_i d_i) + cov(z_i \varepsilon_i) \quad (27)$$

$$\frac{cov(z_i y_i)}{cov(z_i d_i)} = \Delta + \frac{cov(z_i \epsilon_i)}{cov(z_i d_i)} \quad (28)$$

Equations (27) and (28) point to another possible estimator of Δ that could be constructed, that is the ratio of the sample analogues of $cov(z_i y_i)$ and $cov(z_i d_i)$. This is the instrumental variables estimator,

$$\Delta^{IV} = \frac{z'y}{z'd}$$

and Z is the instrumental variable or "instrument". Obviously, this makes sense if two conditions are met. First, we require that $cov(z_i d_i) \neq 0$ - i.e., the instrument must be correlated with the receipt of treatment. Second, we need that $cov(z_i\varepsilon_i) = 0$ - the instrument must be uncorrelated with the heterogeneity in untreated outcomes. If these conditions are satisfied then the IV estimator is consistent. It is not, however, unbiased which implies that we must rely on large-sample properties to justify the use of IV estimation.[18]

The practical difficulty is in finding an instrument which meets the above conditions. Instruments that are only weakly correlated with the receipt of treatment can result in biased and incorrect inferences even in very large samples (see for example, Angrist and Krueger, 2002; Bound, Jaeger and Baker 1995; Wooldridge, 2002) and good practice requires subsidiary analysis to check the strength of the instruments. Furthermore, if $cov(z_i\varepsilon_i) \neq 0$ then the instrument is not exogenous and is therefore not valid. In this case, IV estimation will not be consistent and indeed may give answers that are less useful than simple comparisons of the treated and controls or regression estimates.

[18] See Wooldridge (2002), Chapter 5. It is also important to note that when the classical assumptions hold – and OLS is justified – IV estimation is less efficient than OLS estimation.

With a single instrument, $cov(z_i\varepsilon_i) = 0$ is an identifying assumption that cannot be tested. If one has more than one instrument, they can be combined to give a more efficient estimator. The general formula is

$$\Delta^{IV} = (d'P_Z d)^{-1} d'P_Z y$$

where now Z is a matrix of instruments including a constant and $P_z = Z'(Z'Z)^{-1}Z$ is the projection matrix that projects d onto Z. Moreover, an overidentification test can be used to test the hypothesis that all the instruments are exogenous.

4.3.2 Some Connections to Common Estimators

In the literature, there are several different estimators which are at their roots IV estimators. They vary in mostly in the way they are motivated. We began section 4.3 with the notion of comparing treatment intensities and average outcomes across groups, in the case where we believe that the mean average outcome does not differ across the groups. This quite natural idea is sometimes referred to (especially where there are two groups) as a 'Wald Estimator' (see Angrist, 1991). Durbin (1954) first noted that the Wald estimator is in fact an IV estimator. It is easy to see this by noting that if Z is a dummy variable indicating membership in one of two groups, then $\frac{1}{n}z\prime y$ captures the difference in the average outcomes of the two groups and $\frac{1}{n}z\prime d$ captures the difference in the treatment intensity across the two groups. Thus the IV estimator simply scales the difference in average outcomes by the difference in treatment intensities, as our intuition suggests.

Another approach to the IV estimator is as follows. Suppose that treatment is correlated with heterogeneity in untreated outcomes, i.e., $cov(d_i\varepsilon_i) \neq 0$. Then a natural estimation strategy is to replace d with a prediction $\hat{d}$ that is not correlated with ε. If Z is uncorrelated with ε then the predicted values from a regression of d on Z would have this property. This is of course the 'two stage least squares' (TSLS) estimator. The predictions from such a regression are $P_Z d$ and because P_Z is idempotent,

$$\begin{aligned}
\Delta^{TSLS} &= (\widehat{d}'\widehat{d})^{-1}\widehat{d}'y \\
&= (d'P_Z'P_Z d)^{-1} d'P_Z y \\
&= \Delta^{IV}
\end{aligned}$$

To connect two stage least squares to the Wald estimator, note that the group averages of the treatment intensity are just the predicted value from a first stage regression of treatment status on the group dummy.

Finally, consider the following procedure. First, do the first stage regression of treatment status on the instrument. Then regress observed outcomes on the treatment adding the residual from the first stage regression as an additional control variable. That is, estimate the equation:

$$y = \Delta d + M_Z d + \varepsilon \tag{29}$$

where $M_Z = (I - Z'(Z'Z)^{-1}Z)$ is the projection matrix that projects d onto the orthogonal complement of Z. By the projection property of regression – sometimes called the Frisch-Waugh-Lovell theorem (see, Johnston and DiNardo, 1997) – this procedure gives:

$$\widehat{\Delta} = (d' M_{M_Z} d)^{-1} d' M_{M_Z} y$$

where M_{M_Z} is projection of d onto $M_z d$. Note, however that $M_{M_Z} = P_z$. That is, the part of d which is orthogonal to Z is $M_z d$ and the part of d that is orthogonal to $M_z d$ is $P_Z d$. This is because $d = P_Z d + M_Z d$ and $P_Z d$ is orthogonal to $M_Z d$. Thus,

$$\widehat{\Delta} = (d' P_Z d)^{-1} d' P_Z y = \Delta^{IV}$$

The IV estimator can be implemented in a regression frame work by using the residuals from a first stage regression (that is $M_Z d$) as an additional control variable in a regression of y on d as in equation (29) (Dhrymes, 1970).

This observation is important for several reasons. First, this 'residual stuffing' or 'control function' approach allows for a convenient way to test whether $cov(d_i \varepsilon_i) = 0$. If the estimated residuals from first stage regression ($M_Z d$) are not significant in the second-stage regression then it is reasonable to accept the exogeneity of the treatment indicator (d) and assume that $cov(d_i \varepsilon_i) = 0$. This test is equivalent to the more common Hausman (1978) which compares the coefficients of the IV and OLS estimators (Holly, 1982). The control function idea also is the basis for handling endogenous variables in nonlinear contexts (Blundell and Powell, 2001).

Perhaps even more importantly, these results illustrate why finding valid instruments is difficult, and that, at one level, IV estimation and strategies that 'control' for differences

between treatments and controls (such as regression) are not fundamentally different. If Z is a valid instrument, then $M_Z d$ is an adequate set of controls. If Z is not a valid instrument, then $M_Z d$ is not an adequate set of controls and may or may not be better than some other set of controls, X. IV strategies are not inherently better nor more plausible than regression strategies since finding a valid instrument is equivalent to – and hence in general no more plausible than – finding an adequate set of controls.

In some settings, however, it is easier to think about finding instruments than to think about finding control variables. We turn now to one such popular source of instruments.

4.3.3 Natural Experiments as a Basis for IV Estimation

Over the past decade, a growing body of applied research in economics has attempted to exploit 'natural experiments' to estimate causal effects. Natural experiments are situations in which, although the social scientist cannot manipulate the treatment, nature or policy provides something resembling an experiment in which there are natural 'treatment' and 'control' groups.[19] The key assumption that is that the process that led individuals to be in the two groups is unrelated to their untreated outcomes so that the required conditional mean independence (8) holds and a comparison of the 'treatment' and 'control' groups gives an unbiased estimate of the average effect of the treatment on the treated.

David Freedman (1991,1999) has written about what may be the original natural experiment, John Snow's research dating from the 1950s into the causes of cholera. Snow hypothesized that cholera was waterborne. He observed that there were several companies providing water service in London at the time. They varied in the location of their water intake - upstream or downstream of the sewage outflow on the Thames river. He also noted that the pattern of supply by the different companies to households in London appeared to be largely random. Even on the same street, households might get their water from different companies. This provided the natural experiment. The 'treatment' (water taken downstream of the sewage outflow) was allocated by a process - perhaps historical accident - which was, for Snow's purposes, plausibly as good as experimental randomization. Under the assumption that the specific water company serving a household was unrelated to other potential determinants of cholera, a comparison of cholera rates across customers of

[19] Natural experiments are sometimes referred to as "quasi-experiements", particularly in the education and psychology literature (see for example Cook and Campbell, 1979).

different water companies identified the effect of sewage contaminated water on cholera incidence. This assumption is both plausible, and equally important, transparent. The comparison did not require extensive controls or a sophisticated statistical procedure. Indeed, Snow showed that households served by companies taking water downstream of the sewage outflow had much higher cholera rates, and this was convincing evidence that cholera was a water borne disease.

In economics, a variety of types of natural experiments resulting from policy changes or the forces of nature have been considered.[20] Natural experiments – can generally be evaluated using simple difference or difference-in-difference estimators. Difference estimators use comparisons of a single group across time (that is pre- and post-treatment) or two groups (treated and control) at a point in time as a estimator for the effect of treatment on the treated. Difference-in-difference estimators exploit variation across two dimensions – most commonly policy jurisdiction or eligibility and time – to estimate the effect of the treatment.[21] Recent examples of natural experiments in the economics literature include Card and Krueger's use of variation in minimum wages across U.S. states to estimate the effect of minimum wages on employment (Card and Krueger, 1994), Angrist's use of the Vietnam draft lottery to estimate the effects of veteran status on subsequent earnings (Angrist, 1990) and Paxson's use of weather to investigate the ability of rural households in Thailand to smooth transitory income fluctuations Paxson (1993).

Difference estimators - whether through time or across groups – and difference-in-difference estimators are in fact IV estimators. This is perhaps most easily seen by thinking about these estimators in the regression context. Difference and difference-in-difference estimators involve a comparison of mean outcomes for different groups or time periods and these comparisons can easily be made by regressing outcomes on a group or time dummy (difference estimator) or on a group dummy, time dummy, and a group/time interaction (difference-in-difference estimator) (see Meyer, 1995). In other words, the means can be derived from the projection of outcomes on to variables (group, time, and/or interaction dummies) that are believed be unrelated to untreated outcomes. Thus, these variables are effectively instruments.[22]

[20] See Rosenzweig and Wolpin (2000) who review the use of 'natural' natural experiments in economics.
[21] See Meyer (1995) for a discussion of the internal and external validity of these two strategies.
[22] Note that Heckman (1996) makes a similar point by noting that random assignment in social exper-

The natural experiment approach has been strongly advocated by statisticians, particularly, Freedman (1991,1999) and Rosenbaum (1999), as well as economists (Angrist and Krueger, 2002). The heart of their argument is that the identifying assumption inherent in this approach is both more plausible and more transparent than in alternative methods. However, the natural experiment literature - which has grown rapidly in recent years - is not without its critics. One concern is that the policy changes that are often used as an instrument may not be exogenous if politicians are responding to economic circumstances. That is, policy variation in treatment may be driven by politicians' or administrators' perceptions of untreated outcomes, thus invalidating the identifying assumption. For example Besley and Case (2000) analyse the determinants of workers' compensation benefits across time and state in the United States. They conclude that evolving state economic and demographic conditions are important determinants of benefits. Because these conditions also affect outcomes of interest (such as employment) directly, this relationship invalidates a standard "difference-in-difference" natural experiment design.

Governments do not chose policies randomly, and the Besley and Case analysis is an important warning about the interpretation of policy variation as a natural experiment. However, it is sometimes possible to identify policy reforms that are more plausibly exogenous. A nice example of policy variation not subject to the Besley and Case critique is due to Green and Riddell (1997) who study a how unemployment insurance qualification rules affect employment durations. In Canada, workers must have a minimum number of weeks of covered employment within the last year in order to qualify for unemployment insurance benefits. Green and Riddell study an increase in the qualification requirement in certain regions in Canada and show that job durations lengthened correspondingly to ensure that many affected workers continued to qualify. A key feature of the study is that Green and Riddell are able to plausibly argue that the policy change was unrelated to economic conditions. It resulted because a bill with a "sunset clause" (expiry date) needed to be renewed in order for the shorter qualification periods to continue. However, the passage of all legislation through the upper house of the Canadian Parliament was blocked by an unrelated dispute between the government and an opposition party which controlled the senate (upper house). The bill expired resulting in a period of more strin-

iments also can be thought of as an instrument.

gent qualification requirements which was followed – after the dispute was resolved – by a return to shorter qualification periods.

The bottom line is that while natural experiments are an attractive source of instruments not everything that is proposed as a natural experiment will be convincing. At the same time, one of the virtues of natural experiments is that – as with social experiments – the identifying assumption becomes transparent making it relatively easy to make sensible judgements about what is and is not being identified.

4.3.4 IV Estimation when Treatment Effects are Heterogeneous

The IV estimator, like the OLS estimator, is a weighted average of outcomes and so, unsurprisingly, it must be interpreted with care when treatment effects are heterogeneous. However,with a dummy, endogenous variable (as in the case of a variable measuring 'treatment' versus not) the weighted average which is the instrumental variables estimator has a useful interpretation. Imbens and Angrist (1994) show that under a reasonable set of assumptions the IV estimator of a treatment effect captured by a dummy endogenous generates estimates of the average effect of the treatment for those whose treatment status would be changed by being allocated different values of the instrument. They call this the Local Average Treatment Effect (LATE).

Note this has several important implications. First, with heterogeneous treatment effects, the "overidentification" tests available in a world of homogeneous treatment effects are no longer sensible. The basis of these "overidentification" tests is the idea that if two instruments are valid they both yield consistent estimates of the (homogeneous) treatment effect and thus the two estimates should be "close" (in a way that can be quantified statistically). If the two estimates are not close, then it is likely that (at least) one of the instruments is not valid. However, when treatment effects are heterogenous, each instrument estimates the average treatment effect for the group whose treatment status is switched by that particular instrument. If the two estimates are not close, it could indicate that one of the instruments is not valid. Or it might simply imply that the two instruments result in different groups of 'switchers' who have very dissimilar average treatment effects.

Second, and more broadly, this emphasizes yet again the limits to the external validity

of estimates generated in a heterogeneous treatment world. Even if IV estimates are internal valid (in the sense of not suffering from selection bias so that the instrument is uncorrelated with untreated outcomes) it is only an estimate of the treatment effect for a particular group, and may not be informative about treatment effects for other groups, or about the effect of increases in the incidence of treatment brought about by another instrument. Thus, while it may tell you about the average effect of treatment for individuals (or a subset of individuals) receiving treatment, it may say little about what the effect of extending treatment to other groups might be.

4.3.5 Selection Models

The well-known "Heckit" correction (Heckman, 1979) - is another approach to dealing with the sample selection problem which is closely related to IV estimators. It is now widely recognized that without an instrument for selection into the treatment group (in other words, a variable that has explanatory power in a selection equation but does not affect outcomes except through selection) these models are identified only by assumptions about functional form and error distributions. Identification through functional form alone has been shown to be quite tenuous resulting in standard errors that are often very large, and results which are very sensitive to the particular distributional assumptions invoked. In fact, in a recent paper, Vytlacil (2000) has shown that the general assumptions underlying the selection model approach (that it is not any particular distributional assumptions) are equivalent to the assumptions that Imbens and Angrist (1994) show identify the Local Average Treatment Effect in the linear instrumental variables context. Thus, there is a close equivalence between IV and selection model approaches.

4.3.6 Summary

IV methods provide an approach that uses alternative identifying assumptions to those adopted in regression-based and matching techniques. However, IV methods are not a magic bullet. Finding good instruments is difficult. Indeed, as discussed above, finding a valid instrument is akin to finding a perfect control variable. IV methods using a poor instruments (ones that are either correlated with heterogeneity in untreated outcomes or 'weak' in the sense of being weakly correlated with treatment) can lead to poor results - in fact, results that may be worse than simple regression strategies.

Natural experiments have been a fruitful source of instruments, and are likely to remain so. They nevertheless have their detractors. It seems to us that the key to sensible IV estimation lies in establishing a firm understanding of the source of the variation used to estimate the treatment effect. We must understand what is being identified – and for whom – and be very clear about the exact question we are answering.

4.4 Combining Estimation Strategies

In the preceding subsections, we have presented regression-based, matching, and instrumental variable techniques as alternative nonexperimental strategies for estimating program impacts. Perhaps not surprisingly, some very good empirical strategies involve combining these approaches, or incorporating elements of one into another.

For example, difference-in-difference estimators - which as discussed above are a type of IV estimator - can be 'regression adjusted' by including a set of control variables in the model. A 'matched' or 'conditional' difference-in-difference estimator first differences the data to eliminate the effect of time-invariant unobservables and then uses matching techniques to construct a comparison group for each treated individual. These estimators can be quite powerful. For example, in a comparison of nonexperimental and experimental results in the case of the JTPA program, Heckman et al. (1997) find support for the conditional difference-in-difference estimator, though not for traditional difference-in-difference techniques or for straight matching estimators.

It is also possible for one to apply lessons from matching techniques to improve regression-based analysis. For example, one concern about linear regression and related non-linear models (such as logit or probit models) is that the assumption of linearity (either in the model itself or in the underlying index functions) may allow inappropriate comparisons between treatment and control groups that are in fact very different - that is, it is possible (though very undesirable) to proceed with estimation even though the common support condition does not hold. In fact, common support considerations are important in regression-based approaches as well. To the extent that regression-based approaches circumvent common support problems they do so only by the dubious application of functional form assumptions. This amounts to using regression to make out of sample predictions (ie., predicting Y_0 outside the range of x for which observations on

Y_0 are available).

Regression-based techniques can be strengthened by adopting procedures developed in the context of matching estimators for checking common support conditions. Specifically, Rubin (1979, 1973) suggests several (necessary but not sufficient) criteria for checking the adequacy of regression controls in an observational study. Rubin suggests that the average propensity score (for treatment, conditional on the regression covariates) in the two groups (treatment and control) should not be more than 1 standard deviation apart and that the ratio of the variance of the propensity scores in the two groups should be close to 1 and not more than 2. More generally, one can simply inspect the distribution of propensity scores and impose the common support condition (by trimming the data) before proceeding with subsequent analysis.[23]

5 Concluding Remarks

The economics profession has made significant progress in developing non-experimental approaches to program evaluation in the last 15 years. We have both better econometrics and, perhaps more importantly, a much better understanding of when econometrics is likely to work. While there remain numerous unresolved issues and outstanding difficulties, the payoff to good non-experimental evaluations is much too great for us to stop doing these kind of evaluations or to stop working on ways to improve them. We conclude by revisiting two of the most important themes that this review has touched upon: dealing with heterogeneity and the need for good data.

A fundamental challenge in evaluation - as in all applications of microeconometrics - is finding appropriate but tractable ways of handling the heterogeneity that is such an important feature of the real world. Indeed, much of the new estimation technology – and accompanying terminology – reflects the attempt to accommodate heterogeneity in treatment effects. Both common sense and a large empirical literature suggest that individuals often respond quite differently to the same program or intervention. Thus, evaluations methods that do not consider the possibility of heterogeneity in treatment effects are likely to be off limited value.

[23] See Chapman, Crossley and Kim (2002) for an example with discrete choice models.

As many authors have noted, good data helps a lot. All econometric approaches to the evaluation problem, including those discussed above, are only as good as the data to which they are applied. For example, Heckman et al. (1998) suggest some necessary (but not sufficient) conditions for matching estimators to work well. First, treated and non-treated observations should be sampled in the same way. Second treated and nontreated observations should respond to the survey instrument, so that estimates of treatment effects are not confounded by survey instrument effects. Third, and very important, treated and untreated individuals should be drawn from the same local labour markets. Note this means that good geographic information on survey respondents is required.

Panel data are particularly helpful. Panel data provide good predictors of program participation, allow time invariant bias due to unobservables to be 'differenced' out, and provide good controls for differences in (potential) untreated outcomes. The intuition for this is quite simple. Suppose we are interested in estimating the impact of a training program on earnings. Our concern is that the counterfactual (unobserved) earnings that trained individuals would have earned in the absence of training are different to the observed earnings of those who did not get training. One obvious way to check this is to examine the earnings of these groups in the past. To do so, we need panel data on both groups

Given this, policy makers interested in knowing about the impact of the policies and programs they implement must also be interested in the collection of good data.

6 References

Angrist, Joshua D., 1900. Lifetime Earnings and the Vietnam-era Draft Lottery: Evidence from Social Security Administration Records, *American Economic Review*, June 1990, 80(3), pp. 313-336.

Angrist, Joshua D., 1991. Grouped-data estimation and testing in simple labor-suppy models, *Journal of Econometrics*, 47 , pp. 243-266.

Angrist, Joshua D., 1998. Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants, *Econometrica*, Vol. 66, No.2 (March, 1998), 249-288.

Angrist, Joshu D. and Alan B. Krueger, 1999. Empirical Strategies in Labor Economics, in A. Ashenfelter and D. Card eds. *Handbook of Labor Economics* Vol. 3.

Angist, J.D. and AB Krueger, 2002, Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives, 15(4)* pp. 69-87.

Barnow, Burt S. 2000. Exploring the Relationship between Performance Management and Program Impact: A Case Study of the Job Training Partnership Act, *Journal of Policy Analysis and Management*, 19(1), pp. 118 - 141.

Besley, Timothy and Anne Case, 2000. Unnatural Experiments? Estimating the Incidence of Endogenous Policies. *Economic Journal*, 110, F672-F694.

Björklund, Anders and Robert Moffit, 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models" *Review of Economics and Statistics*, February, pp. 42 - 49.

Blundell, Richard and Monica Costa Dias, 2000. Evaluation Methods for Non-Experimental Data. *Fiscal Studies*, 21(4):47-468.

Blundell, Richard and James L. Powell, 2001. "Endogeniety in Nonparametric and Semiparametric Regression Models". CEMMAP working paper CWP0901. Institue for Fiscals Studies/Department of Economics, University College London.

Bound, John, David A. Jaeger, and Regina M. Baker, 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak", *Journal of the American Statistical Association*, 90(430), pp. 443 - 450.

Bracht, G.H. and G.V: Glass, 1968. The External Validity of Experiments. *American Educational Research Journal* 5: 437-74.

Burtless, Gary. 1995. The Case for Randomized Field Trials in Economic and Policy Research. *Journal of Economic Perspectives*, 9(2):63-84.

Campbell, Donald T. and Julian Stanley, 1966. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Card, David and Alan Krueger, 1994. "Minimum Wages and Employment: A Case Study of the Fast Food Industry", American Economic Review, 84, pp. 772 - 793.

Carneiro, Pedro, Karsten T. Hansen, and James J. Heckman, 2002, Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies, NBER Working Paper 8840, March 2002.

Chapman, Bruce, Thomas F. Crossley and Taejong Kim, 2001. Training after Job-Loss. Working paper, McMaster University.

Cook , T.D. and Campbell, D.T. 1979. Quasiexperimentation: Design and Analysis Issues for Field Settings. Rand-McNally, Chicago.

Dehejia, Rajeev and Sadek Wahba, 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Ealuation of Training Programs. *Journal of the American Statistical Association*, 94(448), 1053-1062.

Dhrymes, 1970. *Econometrics, Statistical Foundations and Applications*, New York: Springer Verlag.

Durbin, J. 1954 "Errors in Variables", *Review of the International Statistical Institute*, 22, pp. 23-32.

Freedman, David, 1991, Statistical Models and Shoe Leather. Sociological Methodology, 21, pp. 291-313

Feedman, David, 1999, From Association to Causation: Some Remarks on the History of Statistics. *Statistical Science*. 14(3),pp. 243-258.

French, J.R. P. 1953. Expreiments in Field Settings. pgs. 98-135 in *Research Methods in the Behavioural Sciences*, edited by L. Festinger and D. Katz. New York : Holt, Rinehart and Winston.

Green, David, A. and W. Craig Ridell, 1997, Qualifying for Unemployment Insurance: An Empirical Analysis. *Economic Journal*, 107(440), 67-84.

Greene, William H., 1997. *Econometric Analysis*. 3rd Edition. Upper Saddle River, New Jersey: Prentice Hall.

Hausman, Jerry, 1978. Specification Tests in Econometrics, *Econometrica*, 46, pp. 1251-1271.

Heckman, James, J., 1979. Sample Selection Bias as a Specification Error. *Econometrica*, 47:153-161.

Heckman, James J., 1996. "Randomization as an Instrumental Variable", Review of Economics and Statistics, 78(20), pp. 336-341.

Heckman, James J. 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture" *Journal of Political Economy*, 109(4), pp. 673 - 748.

Heckman, James J., Carolyn Heinrich and Jeffrey Smith, 1999. Understanding Incentives in Public Organizations. Working paper, University of Maryland.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, 1997. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme, *Review of Economic Studies*, 64, pp. 605-654.

Heckman, James J., Hidehiko Ichimura, Jefffrey Smith and Petra Todd, 1998, September. Characterizing Selection Bias Using Experimental Data. *Econometrica* 66(5):1017-1098.

Heckman, James J., Robert J. LaLonde, and Jeffrey A. Smith, 1999. The Economics and Econometrics of Active Labor Market Programs, in *Handbook of Econometrics*, Volume 3, A. Ashenfelter and D. Card eds.

Heckman, James J. and V. Joseph Hotz, 1989. Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of The American Statistical Association*, 84(408):862-874.

Heckman, James J. and Jeffrey A. Smith, 1995. Assessing the Case for Social Experiments. *Journal of Economic Perspectives, 9(2):85-110.*

Heckman, James J., Jeffrey A. Smith and Chris Taber, 1998. Accounting for Dropouts in Evaluations of Social Programs. *Review of Economics and Statistics, 80(1):1-14.*

Holly, A. 1982, A remark on Hausmans Specification test. *Econometrica,* 50, pp. 749-759.

Human Resources Development Canada, Evaluation and Data Development Branch, 1998. Quasi-Experimental Evaluation, SP-AH053E-01-98.

Imbens, G. and J. Angrist, 1994. Identification and Estimation of Local Average Treatement Effects. *Econometrica*, 62:467-476.

Johnston, Jack and John DiNardo, 1997. *Econometric Methods, Fourth Edition*, New York: McGraw-Hill Companies, Inc.

Jones, Stephen R., 1992. Was There a Hawthorne Effect? *American Journal of Sociology* 98(3):451-468.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *American Economic Review*, 76, pp. 604-620.

LaLonde, Robert J., 1995. "The Promise of Public Sector-Sponsored Training Programs", *Journal of Economic Perspectives*, 9(2):149-168.

Meyer, Bruce D. 1995. "Natural and Quasi-Experimental Experiments in Economics", *Journal of Business and Economic Statistics*, 13(2), April, pp. 151-161.

Paxson, Christina H., 1993. "Consumption and Income Seasonality in Thailand", *The Journal of Political Economy*, 101(1), Feburary, pp. 39-72.

Ricken, Henry W. and Robert F. Boruch, 1978. "Social Experiments", *Annual Review of Sociology*, Vol 4, pp. 511- 532.

Rosenbaum, Paul R. and Donald B. Rubin, 1983. The Central Role of the Propensity Score in Obervational Studies for Causal Effects, *Biometrika*, 70pp, 41 - 55.

Rosenbaum, Paul, R. 1999. Choice as an Alternative to Control in Observation Studies. Statistical Science. 14(3)259-304.

Rosenbaum, P.R. and Donald .B.Rubin, 1983. The Central Role of The Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70:41-55.

Rosenzweig, Mark R. and Kenneth I. Wolpin, 2000. Natural 'Natural Experiements' in Economics. *Journal of Economic Literature* 38, 827-874.

Roy, A.D., 1951. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, 3:135-146.

Rubin, Donald B., 1974. Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies. *Journal of Educational Psychology*, 66:688-701.

Rubin, Donald B., 1978. Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*, 6, 34-58.

Rubin, Donald, B. 1973, Matching to Remove Bias in Observational Studies. Biometrics, 29(1):159-183.

Rubin, Donald, 1979, Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. Journal of the American Statistical Association, 74(366):318-328.

Smith, Jeffrey and Petra Todd, 2000. Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators? Working paper, University of Maryland.

Vytlacil, Edward, 2000. Independence, Monotonicity and Latent Index Models: An Equivalence Result. Working Paper, Stanford University.

Wooldridge, Jeffrey M, 2002. *Econometric Analysis of Cross Section and Panel Data,* Cambridge, Massachusetts: MIT Press.