

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Recent Developments in the Econometrics of Program Evaluation

Citation

Imbens, Guido W. and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. Journal of Economic Literature 47, no. 1: 5-86.

Published version

https://doi.org/10.1257/jel.47.1.5

Link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:3043416

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

NBER WORKING PAPER SERIES

RECENT DEVELOPMENTS IN THE ECONOMETRICS OF PROGRAM EVALUATION

Guido M. Imbens
Jeffrey M. Wooldridge

Working Paper 14251
http://www.nber.org/papers/w14251

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2008

Financial support for this research was generously provided through NSF grants SES 0136789 and 0452590. We are grateful for comments by Caroline Hoxby, Roger Gordon, Larry Katz, The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

© 2008 by Guido M. Imbens and Jeffrey M. Wooldridge. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Recent Developments in the Econometrics of Program Evaluation
Guido M. Imbens and Jeffrey M. Wooldridge
NBER Working Paper No. 14251
August 2008
JEL No. C01

ABSTRACT

Many empirical questions in economics and other social sciences depend on causal effects of programs or policiy interventions. In the last two decades much research has been done on the econometric and statistical analysis of the effects of such programs or treatments. This recent theoretical literature has built on, and combined features of, earlier work in both the statistics and econometrics literatures. It has by now reached a level of maturity that makes it an important tool in many areas of empirical research in economics, including labor economics, public finance, development economics, industrial organization and other areas of empirical micro-economics. In this review we discuss some of the recent developments. We focus primarily on practical issues for empirical researchers, as well as provide a historical overview of the area and give references to more technical research.

Guido M. Imbens
Department of Economics
Littauer Center
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
and NBER
imbens@fas.harvard.edu

Jeffrey M. Wooldridge
Department of Economics, Michigan State
University.
wooldri1@msu.edu

1 Introduction

Many empirical questions in economics and other social sciences depend on causal effects of programs or policies. In the last two decades, much research has been done on the econometric and statistical analysis of the effects of such programs or treatments. This recent theoretical literature has built on, and combined features of, earlier work in both the statistics and econometrics literatures. It has by now reached a level of maturity that makes it an important tool in many areas of empirical research in economics and suitable for a review. In this article we attempt to present such a review. We will focus on practical issues for empirical researchers, as well as provide an historical overview of the area and give references to more technical research. This review complements and extends other reviews and discussions by Blundell and Costa-Dias (2002), Imbens (2004), Angrist and Krueger (2000), and the books by Rosenbaum (1995), Pearl (2000), Lee (2005a), Rubin (2006), Caliendo (2006), Angrist and Pischke (2008), and Morgan and Winship (2007). In addition the reviews in Heckman, Lalonde and Smith (2000), Heckman and Vytlacil (2007a, 2007b), and Abbring and Heckman (2007) provide an excellent overview of the important theoretical work by Heckman and his coauthors in this area.

The central problem studied in this literature is that of evaluating the effect of the exposure of a set of units to a program or treatment on some outcome. In economic studies, the units are typically economic agents such as individuals, households, markets, firms, counties, states or countries, but in other disciplines where evaluation methods are used the units can be animals, plots of land, or pieces of material. The treatments can be job search assistance programs, educational programs, vouchers, laws or regulations, medical drugs, environmental exposure, or technologies. A critical feature is that, in principle, each unit can be exposed to one or more different levels of the treatment. An individual may enroll or not in a training program, or he or she may receive or not receive a voucher, or be subject to a particular regulation or not. The object of interest is a comparison of the two outcomes for the same unit when exposed, and when not exposed, to the treatment. The problem is that we can at most observe one of these outcomes because the unit can be exposed to only one level of the treatment. Holland (1986) refers to this as the "fundamental problem of causal inference." In order to evaluate the effect of the treatment we therefore always need to compare distinct units receiving the different levels of the treatment. Such a comparison can involve different physical units, or the same physical unit at different times.

The problem of evaluating the effect of a binary treatment or program is a well studied problem with a long history in both econometrics and statistics. This is true both in the theoretical literature as well as in the more applied literature. The econometric literature goes back to early work by Ashenfelter (1978) and subsequent work by Ashenfelter and Card (1985), Heckman and Robb (1985), Lalonde (1986), Fraker and Maynard (1987), Card and Sullivan (1988), and Manski (1990). Motivated primarily by applications to the evaluation of labor market programs in observational settings, the focus in the econometric literature is traditionally on endogeneity, or self-selection, issues. Individuals who choose to enroll in a training program are by definition different from those who choose not to enroll. These differences, if they influence the response, may invalidate causal comparisons of outcomes by treatment status, possibly even after ad-

justing for observed covariates. Consequently, many of the initial theoretical studies focused on the use of traditional methods for dealing with endogeneity, such as fixed effect methods from panel data analyses and instrumental variables methods. Subsequently, the econometrics literature has combined insights from the semiparametric literature to develop new estimators for a variety of settings, requiring fewer functional form and homogeneity assumptions.

The statistics literature starts from a different perspective. This literature originates in the analysis of randomized experiments by Fisher (1925) and Neyman (1923). From the early seventies, Rubin (1973a,b, 1974, 1977, 1978), in a series of papers, formulated the now dominant approach to the analysis of causal effects in observational studies. Rubin proposed the interpretation of causal statements as comparisons of so-called potential outcomes: pairs of outcomes defined for the same unit given different levels of exposure to the treatment. Models are developed for the pair of potential outcomes rather than solely for the observed outcome. Rubin's formulation of the evaluation problem, or the problem of causal inference, labeled the Rubin Causal Model (RCM) by Holland (1986), is by now standard in both the statistics and econometrics literature. One of the attractions of the potential outcomes setup is that from the outset it allows for general heterogeneity in the effects of the treatment. Such heterogeneity is important in practice, and it is important theoretically as it is often the motivation for the endogeneity problems that concern economists. One additional advantage of the potential outcome set up is that the parameters of interest can be defined, and the assumptions stated, without reference to particular parametric models.

Of particular importance in Rubin's approach is the relationship between treatment assignment and the potential outcomes. The simplest case for analysis is when assignment to treatment is randomized, and thus independent of covariates as well as the potential outcomes. In such classical randomized experiments, it is straightforward to obtain attractive estimators for the average effect of the treatment, e.g. the difference in means by treatment status. Randomized experiments have been used in some areas in economics. In the seventies, negative income tax experiments received widespread attention. In the late eighties, following an influential paper by Lalonde (1986) that concluded econometric methods were unable to replicate experimental results, more emphasis was put on experimental evaluations of labor market programs, although more recently this emphasis seems to have weakened a bit. In the last couple of years, some of the most interesting experiments have been conducted in development economics (e.g., Miguel and Kremer, 2004; Duflo, 2001; Angrist, Bettinger and Kremer, 2005; Banerjee, Duflo, Cole and Linden, 2007) and behavioral economics (e.g., Bertrand and Mullainathan, 2004). Nevertheless, experimental evaluations remain relatively rare in economics. More common is the case where economists analyze data from observational studies. Observational data generally create challenges in estimating causal effects, but in one important special case, variously referred to as unconfoundedness, exogeneity, ignorability, or selection on observables, questions regarding identification and estimation of the policy effects are fairly well understood. All these labels refer to some form of the assumption that adjusting treatment and control groups for differences in observed covariates, or pretreatment variables, remove all biases in comparisons between treated and control units. This case is of great practical relevance, with many studies relying on some form of this assumption. The semiparametric efficiency bound has been cal-

culated for this case (Hahn, 1998) and various semi-parametric estimators have been proposed (Hahn, 1998; Heckman, Ichimura, and Todd, 1998; Hirano, Imbens and Ridder, 2003; Chen, Hong, and Tarozzi, 2005; Imbens, Newey and Ridder, 2005; Abadie and Imbens, 2006). We discuss the current state of this literature, and the practical recommendations coming out of it, in detail in this review.

Without unconfoundedness there is no general approach to estimating treatment effects. Various methods have been proposed for special cases, and in this review we will discuss several of them. One approach (Rosenbaum and Rubin, 1983; Rosenbaum, 1995) consists of sensitivity analyses, where robustness of estimates to specific limited departures from unconfoundedness are investigated. A second approach, developed by Manski (1990, 2003, 2007), consists of bounds analyses, where ranges of estimands consistent with the data and the limited assumptions the researcher is willing to make, are derived and estimated. A third approach, instrumental variables, relies on the presence of additional treatments, the so-called instruments, that satisfy specific exogeneity and exclusion restrictions. The formulation of this method in the context of the potential outcomes framework is presented in Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996). A fourth approach applies to settings where, in its pure form, overlap is completely absent because the assignment is a deterministic function of covariates, but comparisons can be made exploiting contintuity of average outcomes as a function of covariates. This setting, known as the regression discontinuity design, has a long tradition in statistics (see Shadish, Campbell, and Cook, (2002), Cook (2007) for a historical perspective), but has recently been revived in the economics literature through work by VanderKlaauw (2002), Hahn, Todd, and VanderKlaauw (2000), Lee (2001), and Porter (2003). Finally, a fifth approach, referred to as difference–in–differences, relies on the presence of additional data in the form of samples of treated and control units before and after the treatment. An early application is Ashenfelter and Card (1985). Recent theoretical work includes Abadie (2005), Bertrand, Duflo and Mullainathan (2004), Donald and Lang (2008), and Athey and Imbens (2006).

In this review we will discuss in detail some of the new methods that have been developed in this literature. We will pay particular attention to the practical issues raised by the implementation of these methods. At this stage, the literature has matured to the extent that it has much to offer the empirical researcher. Although the evaluation problem is one where identification problems are important, there is currently a much better understanding of which assumptions are most useful, as well as a better set of methods for inference given different sets of assumptions.

Most of this review will be limited to settings with binary treatments. This is in keeping with the literature, which has largely focused on binary treatment case. There are some extensions of these methods to multivalued, and even continuous, treatments (e.g., Imbens, 2000; Lechner, 2001; Lechner and Miquel, 2005; Gill and Robins, 2001; Hirano and Imbens, 2004), and some of these extensions will be discussed in the current review. But the work in this area is ongoing, and much remains to be done here.

The running example we will use throughout the paper is that of a job market training program. Such programs have been among the leading applications in the economics litera-

ture, starting with Ashenfelter (1978) and including Lalonde (1986) as a particularly influential study. In such settings, a number of individuals enroll or not in a training program, with labor market outcomes, such as yearly earnings or employment status, as the main outcome of interest. An individual not participating in the program may have choosen not to do so, or may have been ineligible for various reasons. Understanding the choices made and constraints faced by the potential participants is a crucial component of any analysis. In addition to observing participation status and outcome measures, we typically observe individual background characteristics, such as education levels and age, as well as information regarding prior labor market histories, such as earnings at various levels of aggregation (e.g., yearly, quarterly or monthly). In addition, we may observe some of the constraints faced by the individuals, including measures used to determine eligibility, as well as measures of general labor market conditions in the local labor markets faced by potential participants.

2 The Rubin Causal Model: Potential Outcomes, the Assignment Mechanism, and Interactions

In this section we describe the essential elements of the modern approach to program evaluation, based on the work by Rubin. Suppose we wish to analyze a job training program using observations on N individuals, indexed by $i = 1, \ldots, N$. Some of these individuals were enrolled in the training program. Others were not enrolled, either because they were ineligible or chose not to enroll. We use the indicator W_i to indicate whether individual i enrolled in the training program, with $W_i = 0$ if individual i did not, and $W_i = 1$ if individual i did, enroll in the program. We use $\mathbf{W}$ to denote the N-vector with i-th element equal to W_i, and N_0 and N_1 to denote the number of control and treated units, respectively. For each unit we also observe a K-dimensional column vector of covariates or pretreatment variables, X_i, with $\mathbf{X}$ denoting the $N \times K$ matrix with i-th row equal to X_i'.

2.1 Potential Outcomes

The first element of the RCM is the notion of potential outcomes. For individual i, for $i = 1, \ldots, N$, we postulate the existence of two potential outcomes, denoted by $Y_i(0)$ and $Y_i(1)$. The first, $Y_i(0)$, denotes the outcome that would be realized by individual i if he or she did not participate in the program. Similarly, $Y_i(1)$ denotes the outcome that would be realized by individual i if he or she did participate in the program. Individual i can either participate or not participate in the program, but not both, and thus only one of these two potential outcomes can be realized. Prior to the assignment being determined, both are potentially observable, hence the label potential outcomes. If individual i participates in the program $Y_i(1)$ will be realized and $Y_i(0)$ will *ex post* be a counterfactual outcome. If, on the other hand individual i does not participate in the program, $Y_i(0)$ will be realized and $Y_i(1)$ will be the *ex post* counterfactual. We will denote the realized outcome by Y_i, with $\mathbf{Y}$ the N-vector with i-th element equal to Y_i.

The preceeding discussion implies that

$$Y_i = Y_i(W_i) = Y_i(0) \cdot (1 - W_i) + Y_i(1) \cdot W_i = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

The potential outcomes are tied to the specific manipulation that would have made one of them the realized outcome. The more precise the specification of the manipulation, the more well-defined the potential outcomes are.

This distinction between the pair of potential outcomes $(Y_i(0), Y_i(1))$ and the realized outcome Y_i is the hallmark of modern statistical and econometric analyses of treatment effects. We offer some comments on it. The potential outcomes framework has important precursors in a variety of other settings. Most directly, in the context of randomized experiments, the potential outcome framework was introduced by Neyman (1923) to derive the properties of estimators and confidence intervals under repeated sampling.

The potential outcomes framework also has important antecedents in econometrics. Specifically, it is interesting to compare the distinction between potential outcomes $Y_i(0)$ and $Y_i(1)$ and the realized outcome Y_i in Rubin's approach to Haavelmo's (1943) work on simultaneous equations models (SEMs). Haavelmo discusses identification of supply and demand models. He makes a distinction between "any imaginable price π"as the argument in the demand and supply functions, $q^d(\pi)$ and $q^s(\pi)$, and the "actual price p", which is the observed equilibrium price satisfying $q^s(p) = q^d(p)$. The supply and demand functions play the same role as the potential outcomes in Rubin's approach, with the equilibrium price similar to the realized outcome. Curiously, Haavelmo's notational distinction between equilibrium and potential prices has gotten blurred in many textbook discussions of simultaneous equations. In such discussions, the starting point is often the general formulation $\mathbf{Y\Gamma} + \mathbf{XB} = \mathbf{U}$ for $N \times M$ vectors of realized outcomes $\mathbf{Y}$, $N \times L$ matrices of exogenous covariates $\mathbf{X}$, and an $N \times M$ matrix of unobserved components $\mathbf{U}$. A nontrivial byproduct of the potential outcomes approach (POA) is that it forces users of SEMs to articulate what the potential outcomes are, thereby leading to better applications of SEMs.

Another area where potential outcomes are used explicitly is in the econometric analyses of production functions. Like the potential outcomes framework, a production function $g(x, \varepsilon)$ describes production levels that would be achieved for each value of a vector of inputs, some observed (x) and some unobserved (ε). Observed inputs may be choosen partly as a function of (expected) values of unobserved inputs. Potential outcomes are also used explicitly in labor market settings by Roy (1951). Roy models individuals choosing from a set of occupations. Individuals know what their earnings would be in each of these occupations and choose the occupation (treatment) that maximizes their earnings. Here we see the explicit use of the potential outcomes, combined with a specific selection/assignment mechanism, namely, choosing the treatment with the highest potential outcome.

The potential outcomes framework has a number of advantages over a framework based directly on realized outcomes. The first advantage of the potential outcome framework is that it allows us to define causal effects before specifying the assignment mechanism, and without making functional form or distributional assumptions. The most common definition of the

causal effect at the unit level is as the difference $Y_i(1) - Y_i(0)$, but we may wish to look at ratios $Y_i(1)/Y_i(0)$, or other functions. Such definitions do not require us to take a stand on whether the effect is constant or varies accross the population. Further, defining individual-specific treatment effects using potential outcomes does not require us to assume endogeneity or exogeneity of the assignment mechanism. By contrast, the causal effects are more difficult to define in terms of the realized outcomes. Often, researchers write down a regression function $Y_i = \alpha + \tau \cdot W_i + \varepsilon_i$. This regression function is then interpreted as a structural equation, with τ as the causal effect. Left unclear is whether the causal effect is constant or not, and what the properties of the unobserved component, ε_i, are. The potential outcomes approach separates these issues, and allows the researcher to first define the causal effect of interest without considering probabilistic properties of the outcomes or assignment.

The second advantage of the POA is that it links the analysis of causal effects to explicit manipulations. Considering the two potential outcomes forces the researcher to think about scenarios under which each outcome could be observed, that is, to consider the kinds of experiments that could reveal the causal effects. Doing so clarifies the interpretation of causal effects. For illustration, consider a couple of recent examples from the economics literature. First, consider the causal effects of gender or ethnicity on outcomes of job applications. Simple comparisons of economic outcomes by ethnicity are difficult to interpret. Are they the result of discrimination by employers, or are they the result of differences between applicants, possibly arising from discrimination at an earlier stage of life? Now, one can obtain unambiguous causal interpretations by linking comparisons to specific manipulations. A recent example is the study by Bertrand and Mullainathan (2004), who compare call-back rates for job applications submitted with names that suggest African-American or Caucasian ethnicity. Their study has a clear manipulation – a name change – and therefore a clear causal effect. As a second example, consider some recent economic studies that have focused on causal effects of individual characteristics such as beauty (Hamermesh and Biddle, 1994), or height. Do the differences in earnings by ratings on a beauty scale represent causal effects? One possible interpretation is that they represent causal effects of plastic surgery. Such a manipulation would make differences causal, but it appears unclear whether cross-sectional correlations between beauty and earnings in a survey from the general population represent causal effects of plastic surgery.

A third advantage of the POA is that it separates the modelling of the potential outcomes from that of the assignment mechanism. Modelling the realized outcome is complicated by the fact that it combines the potential outcomes and the assignment mechanism. The researcher may have very different sources of information to bear on each. For example, in the labor market program example we can consider the outcome, say, earnings, in the absence of the program: $Y_i(0)$. We can model this in terms of individual characteristics and labor market histories. Similarly, we can model the outcome given enrollment in the program, again conditional on individual characteristics and labor market histories. Then finally we can model the probability of enrolling in the program given the earnings in both treatment arms conditional on individual characteristics. This sequential modelling will lead to a model for the realized outcome, but it may be easier than directly specifying a model for the realized outcome.

A fourth advantage of the potential outcomes approach is that it allows us to formulate

probabilistic assumptions in terms of potentially observable variables, rather than in terms of unobserved components. In this approach, many of the critical assumptions will be formulated as (conditional) independence assumptions involving the potential outcomes. Assessing their validity requires the researcher to consider the dependence structure if all potential outcomes where observed. By contrast, models in terms of realized outcomes often formulate the critical assumptions in terms of errors in regression functions. To be specific, consider again the regression function $Y_i = \alpha + \tau \cdot W_i + \varepsilon_i$. Typically (conditional independence) assumptions are made on the relationship between ε_i and W_i. Such assumptions implicitly bundle a number of assumptions, including functional-form assumptions and substantive exogeneity assumptions. This bundling makes the plausibility of these assumptions more difficult to assess.

A fifth advantage of the POA is that it clarifies where the uncertainty in the estimators comes from. Even if we observe the entire (finite) population (as is increasingly common with the growing availability of administrative data sets) – so we can estimate population averages with no uncertainty – causal effects will be uncertain because for each unit at most one of the two potential outcomes is observed. One may still use super population arguments to justify approximations to the finite sample distributions, but such arguments are not required to motivate the existence of uncertainty about the causal effect.

2.2 The Assignment Mechanism

The second ingredient of the RCM is the assignment mechanism. This is defined as the conditional probability of receiving the treatment, as a function of potential outcomes and observed covariates. We distinguish three classes of assignment mechanisms, in order of increasing complexity of the required analysis.

The first class of assignment mechanisms is that of randomized experiments. In randomized experiments, the probability of assignment to treatment does not vary with potential outcomes, and is a known function of covariates. The leading case is that of a completely randomized experiment where, in a population of N units, $N_1 < N$ randomly chosen units are assigned to the treatment and the remaining $N_0 = N - N_1$ units are in the control group. There are important variations on this example, such as pairwise randomization, where initially units are matched in pairs, and in a second stage one unit in each pair is randomly assigned to the treatment. Another variant is a general stratified experiment, where randomization takes place within a finite number of strata. In any case, there are in practice few experiments in economics, and most of those are of the completely randomized experiment variety, so we shall limit our discussion to this type of experiment. It should be noted though that if one has the opportunity to design a randomized experiment, and if pretreatment variables are available, stratified experiments are at least as good as completely randomized experiments, and typically better, in terms of expected mean squared error, even in finite samples. See Imbens, King, McKenzie and Ridder (2008) for more details. The use of formal randomization has become more widespread in the social sciences in recent years, sometimes as a formal design for an evaluation and sometimes as an acceptable way of allocating scarce resources. The analysis of such experiments is often straightforward. In practice, however, researchers have typically limited themselves to simple mean differences by assignment. Such analyses are valid, but often

they are not the most powerful tools available to exploit the randomization. We discuss the analysis of randomized experiments, including more powerful randomization-based methods for inference, in Section 4.

The second class of assignment mechanisms maintains the restriction that the assignment probabilities do not depend on the potential outcomes, or

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i,$$

where $A \perp\!\!\!\perp B|C$ denotes conditional independence of A and B given C. However, in contrast to randomized experiments, the assignment probabilities are no longer assumed to be a known function of the covariates. The precise form of this critical assumption, not tied to functional form or distributional assumptions, was first presented in Rosenbaum and Rubin (1983a). Following Rubin (1990) we refer to this assignment mechanism as unconfounded assignment. Somewhat confusingly, this assumption, or variations on it, are in the literature also referred to by various other labels. These include selection on observables[1], exogeneity[2], and conditional independence[3]. Although the analysis of data with such assignment mechanisms is not as straightforward as that of randomized experiments, there are now many practical methods available for this case. We review them in Section 5.

The third class of assignment mechanisms contains all remaining assignment mechanisms with some dependence on potential outcomes.[4] Many of these create substantive problems for the analysis, for which there is no general solution. There are a number of special cases that are by now relatively well understood, and we discuss these in Section 6. The most prominent of these cases are instrumental variables, regression discontinuity, and differences-in-differences. In addition, we discuss two general methods that also relax the unconfoundedness assumption but do not replace it with additional assumptions. The first relaxes the unconfoundedness assumption in a limited way and investigates the sensitivity of the estimates to such violations. The second drops the unconfoundedness assumption entirely and establishes bounds on estimands of interest. The latter is associated with the work by Manski (1990, 1995, 2008).

[1] Although Heckman, Ichimura and Todd (1997, page 611) write that "In the language of Heckman and Robb (1985) matching assumes that selection is *on observables*" (their italics), the original definition in Heckman and Robb (1985, page 163) is not equivalent to unconfoundedness. In the context of a single cross-section version of their two equation selection model, $Y = X_i'\beta + W_i\alpha + \varepsilon_i$ and $W_i = 1\{Z_i'\gamma + \nu_i > 0\}$, they define selection bias to refer to the case where $\mathbb{E}[\varepsilon_i W_i] \neq 0$, and selection-on-observables to the case where selection bias is present and caused by correlation between ε_i and Z_i, rather than by correlation between ε_i and ν_i.

[2] Although X_i is not exogenous for $\mathbb{E}[Y_i(1) - Y_i(0)]$, according to the definitions in Engle, Hendry and Richard (1983), because knowledge of its marginal distribution contains information about $\mathbb{E}[Y_i(1) - Y_i(0)]$, standard usage of the term "exogenous" does appear to capture the notion of unconfoundedness, e.g., Manski, Sandefur, McLanahan, and Powers (1992), and Imbens (2004).

[3] E.g., Lechner, 2001; Cameron and Trivedi, 2005.

[4] This includes some mechanisms where the dependence on potential outcomes does not create any problems in the analyses. Most prominent in this category are sequential assignment mechanisms. For example, one could randomly assign the first ten units to the treatment or control group with probability 1/2. From then on one could skew the assignment probability to the treatment with the most favorable outcomes so far. For example, if the active treatment looks better than the control treatment based on the first N units, then the $(N+1)$th unit is assigned to the active treatment with probability 0.8 and vice versa. Such assignment mechanisms are not very common in economics settings, and we ignore them in this discussion.

2.3 Interactions and General Equilibrium Effects

In most of the literature it is assumed that treatments received by one unit do not affect outcomes for another unit. Only the level of the treatment applied to the specific individual is assumed to potentially affect outcomes for that particular individual. In the statistics literature this assumption is referred to as the Stable-Unit-Treatment-Value-Assumption (SUTVA, Rubin, 1978). In this paper we mainly focus on settings where this assumption is maintained. In the current section we discuss some of the literature motivated by concerns about this assumption.

This lack-of-interaction assumption is very plausible in many biomedical applications. Whether one individual receives or does not receive a new treatment for a stroke or not is unlikely to have a substantial impact on health outcomes for any other individual. However, there are also many cases in which such interactions are a major concern and the assumption is not plausible. Even in the early experimental literature, with applications to the effect of various fertilizers on crop yields, researchers were cognizant of potential problems with this assumption. In order to minimize leaking of fertilizer applied to one plot into an adjacent plot experimenters used guard rows to physically separate the plots that were assigned different fertilizers. A different concern arises in epidimiological applications when the focus is on treatments such as vaccines for contagious diseases. In that case, it is clear that the vaccination of one unit can affect the outcomes of others in their proximity, and such effects are a large part of the focus of the evaluation.

In economic applications, interactions between individual are also a serious concern. It is clear that a labor market program that affects the labor market outcomes for one individual potentially has an effect on the labor market outcomes for others. In a world with a fixed number of jobs, a training program could only redistribute the jobs, and ignoring this constraint on the number of jobs by using a partial, instead of a general, equilibrium analysis could lead one to erroneously conclude that extending the program to the entire population would raise aggregate employment. Such concerns have rarely been addressed in the recent program evaluation literature. Exceptions include Heckman, Lochner, and Taber (1999) who provide some simulation evidence for the potential biases that may result from ignoring these issues.

In practice these general equilbrium effects may, or may not, be a serious problem. The indirect effect on one individual of exposure to the treatment of a few other units is likely to be much smaller than the direct effect of the exposure of the first unit itself. Hence, with most labor market programs both small in scope and with limited effects on the individual outcomes, it appears unlikely that general equilibrium effects are substantial and they can probably be ignored for most purposes.

One general solution to these problems is to redefine the unit of interest. If the interactions between individuals are at an intermediate level, say a local labor market, or a classroom, rather than global, one can analyze the data using the local labor market or classroom as the unit and changing the no-interaction assumption to require the absence of interactions among local labor markets or classrooms. Such aggregation is likely to make the no-interaction assumption more plausible, albeit at the expense of reduced precision.

An alternative solution is to directly model the interactions. This involves specifying which

individuals interact with each other, and possibly relative magnitudes of these interactions. In some cases it may be plausible to assume that interactions are limited to individuals within well-defined, possibly overlapping, groups, with the intensity of the interactions equal within this group. This would be the case in a world with a fixed number of jobs in a local labor market. Alternatively, it may be that interactions occur in broader groups but decline in importance depending on some distance metric, either geographical distance or proximity in some economic metric.

The most interesting literature in this area views the interactions not as a nuisance but as the primary object of interest. This literature, which includes models of social interactions and peer effects, has been growing rapidly in the last decade, following the early work by Manski (1993). See Manski (2000) and Brock and Durlauf (2000) for recent surveys. Empirical work includes Kling, Liebman and Katz (2007), who look at the effect of households moving to neighborhoods with higher average socio-economic status; Sacerdote (2001), who studies the effect of college roommate behavior on a student's grades; Glaeser, Sacerdote and Scheinkman (1996), who study social interactions in criminal behavior; Case and Katz (1991), who look at neighbourhood effects on disadvantaged youths, Graham (2006), who infer interactions from the effect of class size on the variation in grades; and Angrist and Lang (2004), who study the effect of desegregation programs on students' grades. Many identification and inferential questions remain unanswered in this literature.

3 What are We Interested In? Estimands and Hypotheses

In this section we discuss some of the questions that researchers have asked in this literature. A key feature of the current literature, and one that makes it more important to be precise about the questions of interest, is the accommodation of general heterogeneity in treatment effects. In contrast, in many early studies it was assumed that the effect of a treatment was constant, implying that the effect of various policies could be captured by a single parameter. The essentially unlimited heterogeneity in the effects of the treatment allowed for in the current literature implies that it is generally not possible to capture the effects of all policies of interest in terms of a few summary statistics. In practice researchers have reported estimates of the effects of a few focal policies. In this section we describe some of these estimands. Most of these estimands are average treatment effects, either for the entire population or for some subpopulation, although some correspond to other features of the joint distribution of potential outcomes.

Most of the empirical literature has focused on estimation. Much less attention has been devoted to testing hypotheses regarding the properties or presence of treatment effects. Here we discuss null and alternative hypotheses that may be of interest in settings with heterogeneous effects. Finally, we discuss some of the recent literature on decision-theoretic approaches to program evaluation that ties estimands more closely to optimal policies.

3.1 Average Treatment Effects

The econometric literature has largely focused on average effects of the treatment. The two most prominent average effects both rely on a superpopulation perspective. The sample of size N is viewed as a random sample from a large (super-)population, and interest is in the average effect in the superpopulation. The most popular one is the Population Average Treatment Effect (PATE), the population expectation of the unit-level causal effect, $Y_i(1) - Y_i(0)$:

$$\tau_{\text{pate}} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

If the policy under consideration would expose all units to the treatment or none at all, this is the most relevant quantity. Another popular estimand is the Population Average Treatment effect on the Treated (PATT), the average over the subpopulation of treated units:

$$\tau_{\text{patt}} = \mathbb{E}\left[Y_i(1) - Y_i(0)|W_i = 1\right].$$

In many observational studies τ_{patt} is a more interesting estimand than the overall average effect. As an example, consider the case where a well defined population was exposed to a treatment, say a job training program. There may be various possibilities for a comparison group, including subjects drawn from public use data sets. In that case it is generally not interesting to consider the effect of the program for the comparison group: for many members of the comparison group (e.g., individuals with stable, high-wage jobs) it is difficult and uninteresting to imagine their being enrolled in the labor market program. (Of course, the problem of averaging across units that are unlikely to receive future treatments can be mitigated by more carefully constructing the comparison group to be more like the treatment group, making τ_{pate} a more meaningful parameter. See the discussion below.) A second case where τ_{patt} is the estimand of most interest is in the setting of a voluntary program where those not enrolled will never be required to participate in the program. A specific example is the effect of serving in the military where an interesting question concerns the foregone earnings for those who served (Angrist, 1998).

In practice, there is typically little motivation presented for the focus on the overall average effect or the average effect for the treated. Take a job training program. The overall average effect would be the parameter of interest if the policy under consideration is a mandatory exposure to the treatment versus complete elimination. It is rare that these are the alternatives, with more typically exemptions granted to various subpopulations. Similarly the average effect for the treated would be informative about the effect of entirely eliminating the current program. More plausible regime changes would correspond to a modest extension of the program to other jurisdictions, or a contraction to a more narrow population.

A somewhat subtle issue is that we may wish to separate the extrapolation from the sample to the superpopulation fromthe problem of inference for the sample at hand. This suggests that, rather than focusing on PATE or PATT, we might first focus on the average causal effect conditional on the covariates in the sample,

$$\tau_{\text{cate}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[Y_i(1) - Y_i(0)|\, X_i\right],$$

and, similarly, the average over the subsample of treated units:

$$\tau_{\text{catt}} = \frac{1}{N_1} \sum_{i|W_i=1} \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i \right].$$

If the effect of the treatment or intervention is constant ($Y_i(1) - Y_i(0) = \tau$ for some constant τ), all four estimands, τ_{pate}, τ_{patt}, τ_{cate}, and τ_{catt}, are obviously identical. However, if there is heterogeneity in the effect of the treatment, the estimands may all be different. The difference between τ_{pate} and τ_{cate} (and between τ_{patt} and τ_{catt}) is relatively subtle. Most estimators that are attractive for the population treatment effect are also attractive for the corresponding conditional average treatment effect, and vice versa. Therefore, we do not have to be particularly concerned with the distinction between the two estimands at the estimation stage. However, there is an important difference between the population and conditional estimands at the inference stage. If there is heterogeneity in the effect of the treatment, we can estimate the sample average treatment effect τ_{cate} more precisely than the population average treatment effect τ_{pate}. When one estimates the variance of an estimator $\hat{\tau}$ – which can serve as an estimate for τ_{pate} or τ_{cate} – one therefore needs to be explicit about whether one is interested in the variance relative to the population or to the conditional average treatment effect. We will return to this issue in Section 5.

A more general class of estimands includes average causal effects for subpopulations and weighted average causal effects. Let $\mathbb{A}$ be a subset of the covariate space $\mathbb{X}$, and let $\tau_{\text{cate},\mathbb{A}}$ denote the conditional average causal effect for the subpopulation with $X_i \in \mathbb{A}$:

$$\tau_{\text{cate},\mathbb{A}} = \frac{1}{N_{\mathbb{A}}} \sum_{i:X_i \in \mathbb{A}} \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i \right],$$

where $N_{\mathbb{A}}$ is the number of units with $X_i \in \mathbb{A}$. Crump, Hotz, Imbens and Mitnik (2008a) argue for considering such estimands. Their argument is not based on the intrinsic interest of these subpopulations. Rather, they show that such estimands may be much easier to estimate than τ_{cate} (or τ_{catt}). Instead of solely reporting an imprecisely estimated average effect for the overall population, they suggest it may be informative to also report a precise estimate for the average effect of some subpopulation. They then propose a particular set $\mathbb{A}$ for which the average effect is most easily estimable. See Section 5.10.2 for more details. The Crump et al estimates would not necessarily have as much external validity as estimates for the overall population, but they may be much more informative for the sample at hand. In any case, in many instances the larger policy questions concern extensions of the interventions or treatments to other populations, so that external validity may be elusive irrespective of the estimand.

In settings with selection on unobservables the enumeration of the estimands of interest becomes more complicated. A leading case is instrumental variables. In the presence of heterogeneity in the effect of the treatment one can typically not identify the average effect of the treatment even in the presence of valid instruments. There are two new approaches in the recent literature. One is to focus on bounds for well-defined estimands such as the average effect τ_{pate} or τ_{cate}. Manski (1990, 2003) developed this approach in a series of papers. An alternative is to focus on estimands that can be identified under weaker conditions than those

requird for the average treatment effect. Imbens and Angrist (1994) show that one can, under much weaker conditions than required for identification of τ_{pate}, identify the average effect for the subpopulation of units whose treatment status is affected by the instrument. They refer to this subpopulation as the compliers. This does not directly fit into the classification above since the subpopulation is not defined solely in terms of covariates. We discuss this estimand in more detail in Section 6.3.

3.2 Quantile and distributional Treatment Effects and other estimands

An alternative class of estimands concerns quantile treatment effects. These have only recently been studied and applied in the economics literature, although they were introduced in the statistics literature in the seventies. Doksum (1974) and Lehman (1974) define

$$\tau_q = F^{-1}_{Y(1)}(q) - F^{-1}_{Y(0)}(q), \tag{1}$$

as the q-th quantile treatment effect. There are some important issues in interpreting these quantile treatment effects. First, note that these quantiles effects are defined as differences between quantiles of the two marginal potential outcome distributions, rather than as quantiles of the unit level effect,

$$\tilde{\tau}_q = F^{-1}_{Y(1)-Y(0)}(q). \tag{2}$$

In general the quantile of the difference, $\tilde{\tau}_q$, differs from the difference in the quantiles, τ_q, unless there is perfect rank correlation between the potential outcomes $Y_i(0)$ and $Y_i(1)$ (the leading case of this is the constant additive treatment effect). The quantiles of the treatment effect, $\tilde{\tau}_q$, have received much less attention than the quantile treatment effects, τ_q. The main reason is that the $\tilde{\tau}_q$ are generally not identified without assumptions on the rank correlation between the potential outcomes, even with data from a randomized experiment. Note that this issue does not arise if we look at average effects because the mean of the difference is equal to the difference of the means: $\mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right]$.

A complication facing researchers interested in quantile treatment effects is that the difference in a marginal quantile, τ_q, is in general not equal to the average difference in the conditional quantiles, where the latter are defined as

$$\tau_q(x) = F^{-1}_{Y(1)|X}(q|x) - F^{-1}_{Y(0)|X}(q|x).$$

In other words, even if we succeed in estimating $\tau_q(x)$, we cannot simply average $\tau_q(X_i)$ across i to consistently estimate τ_q. Bitler, Gelbach, and Hoynes (2002) estimate quantile treatment effects in a randomized evaluation of a job training program. Firpo (2006) develops methods for estimating τ_q in observational studies given unconfoundedness. Abadie, Angrist and Imbens (2002) and Chernozhukov and Hansen (2005) study quantile treatment effects in instrumental variables settings.

3.3 Testing

The literature on hypothesis testing in program evaluation is relatively limited. Most of the testing in applied work has focused on the null hypothesis that the average effect of interest is zero. Because many of the commonly used estimators for average treatment effects are asymptotically normally distributed with zero asymptotic bias, it follows that standard confidence intervals (the point estimate plus or minus a constant times the standard error) can be used for testing such hypotheses. However, there are other interesting hypotheses to consider.

One question of interest is whether there is any effect of the program, that is whether the distribution of $Y_i(1)$ differs from that of $Y_i(0)$. This is equivalent to the hypothesis that not just the mean, but all moments, are identical in the two treatment groups. Abadie (2002) studies such tests in the settings with randomized experiments as well as settings with instrumental variables using Kolmogorov-Smirnov type testing procedures.

A second set of questions concerns treatment effect heterogeneity. Even if the average effect is zero, it may be important to establish whether a targeted implementation of the intervention, with only those who can expect to benefit from the intervention assigned to it, could improve average outcomes. In addition, in cases where there is not sufficient information to obtain precise inferences for the average causal effect τ_{pate}, it may still be possible to establish whether there are any subpopulations with an average effect positive or different from zero, or whether there are subpopulations with an average effect exceeding some threshold. It may also be interesting to test whether there is any evidence of heterogeneity in the treatment effect by observable characteristics. This bears heavily on the question whether the estimands are useful for extrapolation to other populations which may differ in terms of some observable characteristics. Crump, Hotz, Imbens and Mitnik (2008b) study these questions in settings with unconfounded treatment assignment.

3.4 Decision-theoretic Questions

Recently, a small but very innovative literature has started to move away from the focus on summary statistics of the distribution of treatment effects or potential outcomes to directly address policies of interest. This is very much a literature in progress. Manski (2000, 2002, 2004), Dehejia (2005), and Hirano and Porter (2005) study the problem faced by program administrators who can assign individuals to the active treatment or to the control group. These administrators have available two pieces of information. First, covariate information for these individuals, and second, information about the efficacy of the treatment based on a finite sample of other individuals for whom both outcome and covariate information is available. The administrator may care about the entire distribution of outcomes, or solely about average outcomes, and may also take into account costs associated with participation. If the administrator knew exactly the conditional distribution of the potential outcomes given the covariate information this would be a simple problem: the administrator would simply compare the expected welfare for different rules and choose the one with the highest value. However, the administrator does not have this knowledge and needs to make a decision given uncertainty about these distributions. In these settings, it is clearly important that the statistical model allows for heterogeneity in the

treatment effects.

Graham, Imbens, and Ridder (2006) extend the type of problems studied in this literature by incorporating resource constraints. They focus on problems that include as a special case the problem of allocating a fixed number of slots in a program to a set of individuals on the basis of observable characteristics of these individuals given a random sample of individuals for whom outcome and covariate information is available.

4 Randomized Experiments

Experimental evaluations have traditionally been rare in economics. The few experiments that have been conducted have generally been influential, including some of the labor market training programs. More recently, many small scale experiments have been conducted in development economics. We discuss some of these in Section 4.1.

With experimental data the statistical analysis is generally straightforward. Differencing the means by treatment status or, equivalently, regressing the outcome on an intercept and an indicator for the treatment, leads to an unbiased estimator for the average effect of the treatment. Adding covariates to the regression function typically improves precision without jeopardizing consistency because the randomization implies that in large samples the treatment indicator and the covariates are independent. In practice, researchers have rarely gone beyond basic regression methods. In principle, however, there are additional methods that can be useful in these settings. In Section 4.2 we review one important experimental technique, Fisher's method for calculating exact p-values, that deserves wider usage.

4.1 Randomized Experiments in Economics

Randomized experiments have a long tradition in biostatistics. In this literature they are often viewed as the only credible approach to establishing causality. For example, the US Food and Drug Administration requires evidence from randomized experiments in order to approve new drugs and medical procedures. A first comment concerns the fact that even randomized experiments rely to some extent on substantive knowledge. It is only once the researcher is willing to limit interactions between units that randomization can establish causal effects. In settings with potentially general interactions between units, randomization by itself cannot solve the identification problems required for establishing causality. In biomedical settings, where such interaction effects are often arguably absent, randomized experiments are therefore attractive. Moreoever, it is often possible to keep the units ignorant of their treatment status, further enhancing the interpretation of the estimated effects as causal effects of the treatment, and thus improving the external validity.

In the economics literature randomization has played a much less prominent role. At various times social experiments have been conducted, but they have never been viewed as the sole method for establishing causality, and in fact they have often been regarded with some suspicion concerning the relevance of the results for policy purposes. Part of this may be due to the fact that for the treatments of interest to economists, e.g., education and labor market programs,

it is generally impossible to do blind or double-blind experiments, creating the possibility of placebo effects that compromise the external validity of the estimates.

Among the early social experiments in economics were the negative income tax experiments in Seattle and Denver in the early seventies, formallly referred to as the Seattle and Denver Income Maintenance Experiments (SIME and DIME). In the eighties, a number of papers called into question the reliability of econometric and statistical methods for estimating causal effects in observational studies. In particular, Lalonde (1986) and Fraker and Maynard (1987), using data from the National Supported Work (NSW) programs, suggested that widely used econometric methods were unable to replicate the results from experimental evaluations. These conclusions encouraged government agencies to include experimental evaluation components in job training programs. Examples of such programs include the Greater Avenues to INdependence (GAIN) programs (e.g., Riccio and Friedlander, 1992, the WIN programs (e.g., Gueron and Pauly, 1991; Friedlander and Gueron, 1992), the Self Sufficiency Project (SPP) in Canada (Card and Hyslop, 2005), and the Statistical Assistance for Programme Selection (SAPS) in Switzerland (Behncke, Froelich, and Lechner, 2006). Like the NSW evaluation, these experiments have been useful not merely in establishing the effects of particular programs, but also in providing fertile testing grounds for new evaluations methods.

Recently there has been a large number of experiments in development economics. These range from the large scale Progressa school subsidy program in Mexico (see Schulz, 2001; Attansio, Meghir and Santiago, 2001) to much smaller experiments, such as the teacher incentive study by Duflo and Hanna (2006) and the corruption study by Olken (2007). Many of these address questions in education (e.g., Banerjee, Duflo, Cole, and Linden, 2007; Miguel and Kremer, 2003). In a number of these experiments, economists have been involved from the beginning in the design of the evaluations, raising questions of optimal design. These issues are discussed in Duflo, Glennester and Kremer (2007), Bruhn and McKenzie (2007), and Imbens, King, McKenzie and Ridder (2008).

4.2 Fisher's Exact P-values

Fisher (1925) was interested in calculating p-values for hypotheses regarding the effect of treatments. The aim is to provide exact inferences for a finite population of size N. This finite population may be a random sample from a large superpopulation, but that is not exploited in the analysis. The inference is nonparametric in that it does not make functional form assumptions regarding the effects; it is exact in that it does not rely on large sample approximations. In other words, the p-values coming out this analysis are exact and valid irrespective of the sample size.

The most common null hypothesis in Fisher's framework is that of no effect of the treatment for any unit in this population, against the alternative that, at least for some units, there is a non-zero effect:

$$H_0: \ Y_i(0) = Y_i(1), \ \forall i = 1, \ldots, N, \qquad \text{against} \quad H_a: \ \exists i \text{ such that } Y_i(0) \neq Y_i(1).$$

It is not important that the null hypothesis is that the effects are all zero. What is essential is that the null hypothesis is sharp, that is, the null hypothesis specifies the value of all unobserved

potential outcomes for each unit. A more general null hypothesis could be that $Y_i(0) = Y_i(1)+c$ for some pre-specified c, or that $Y_i(0) = Y_i(1) + c_i$ for some set of pre-specified c_i. Importantly, this framework cannot accommodate null hypotheses such as the average effect of the treatment is zero, against the alternative hypothesis of a non-zero average effect, or

$$H_0' : \frac{1}{N}\sum_i \Big(Y_i(1) - Y_i(0)\Big) = 0, \qquad \text{against} \quad H_a' : \frac{1}{N}\sum_i \Big(Y_i(1) - Y_i(0)\Big) \neq 0.$$

Whether the null of no effect for any unit versus the null of no effect on average is more interesting was the subject of a testy exchange between Fisher (who focused on the first) and Neyman (who thought the latter was the interesting hypothesis, and who stated that the first was only of "academic" interest) in Neyman (1923). Putting the argument about its ultimate relevance aside, Fisher's test is a powerful tool for establishing whether a treatment has any effect. It is not essential in this framework that the probabilities of assignment to the treatment group are equal for all units. It is crucial, however, that the probability of any particular assignment vector is known. These probabilities may differ by unit provided the probabilities are known.

The implication of Fisher's framework is that, under the null hypothesis, we know the exact value of all the missing potential outcomes. Thus there are no nuisance parameters under the null hypothesis. As a result, we can deduce the distribution of any statistic, that is, any function of the realized values of $(Y_i, W_i)_{i=1}^N$, generated by the randomization. For example, suppose the statistic is the average difference between treated and control outcomes, $T(\mathbf{W}, \mathbf{Y}) = \overline{Y}_1 - \overline{Y}_0$, where $\overline{Y}_w = \sum_{i:W_i=w} Y_i/N_w$, for $w = 0, 1$. Now suppose we had assigned a different set of units to the treatment. Denote the vector of alternative treatment assignments by $\tilde{\mathbf{W}}$. Under the null hypothesis we know all the potential outcomes and thus we can deduce what the value of the statistic would have been under that alternative assignment, namely $T(\tilde{\mathbf{W}}, \mathbf{Y})$. We can infer the value of the statistic for all possible values of the assignment vector $\mathbf{W}$, and since we know the distribution of $\mathbf{W}$ we can deduce the distribution of $T(\mathbf{W}, \mathbf{Y})$. The distribution generated by the randomization of the treatment assignment is referred to as the randomization distribution. The p-value of the statistic is then calculated as the probability of a value for the statistic that is at least as large, in absolute value, as that of the observed statistic, $T(\mathbf{W}, \mathbf{Y})$.

In moderately large samples it is typically not feasible to calculate the exact p-values for these tests. In that case one can approximate the p-value by basing it on a large number of draws from the randomization distribution. Here the approximation error is of a very different nature than that in typical large sample approximations: it is controlled by the researcher, and if more precision is desired one can simply increase the number of draws from the randomization distribution.

In the form described above, with the statistic equal to the difference in averages by treatment status, the results are typically not that different from those using Wald tests based on large sample normal approximations to the sampling distribution to the difference in means $\overline{Y}_1 - \overline{Y}_0$, as long as the sample size is moderately large. The Fisher approach to calculating p-values is much more interesting with other choices for the statistic. For example, as advocated by Rosenbaum in a series of papers (Rosenbaum 1984a, 1995), a generally attractive

choice is the difference in average ranks by treatment status. First the outcome is converted into ranks (typically with, in case of ties, all possible rank orderings averaged), and then the test is applied using the average difference in ranks by treatment status as the statistic. The test is still exact, with its exact distribution under the null hypothesis known as the Wilcoxon distribution. Naturally, the test based on ranks is less sensitive to outliers than the test based on the difference in means.

If the focus is on establishing whether the treatment has some effect on the outcomes, rather than on estimating the average size of the effect, such rank tests are much more likely to provide informative conclusions than standard Wald tests based differences in averages by treatment status. To illustrate this point, we took data from eight randomized evaluations of labor market programs. Four of the programs are from the WIN demonstration programs. The four evaluations took place in Arkansas, Baltimore, San Diego, and Virginia. See Gueron and Pauly (1991), Friedlander and Gueron (1992), Greenberg and Wiseman (1992), and Friedlander and Robins (1995) for more detailed discussions of each of these evaluations. The second set of four programs is from the GAIN programs in California. The four locations are Alameda, Los Angeles, Riverside, and San Diego. See Riccio and Friedlander (1992), Riccio, Friedlander, and Freeman (1994) for more details on these programs and their evaluations. In each location we take as the outcome total earnings for the first (GAIN) or second (WIN) year following the program, and we focus on the subsample of individuals who had positive earnings at some point prior to the program. For each location we calculate three p-values. The first p-value is based on the normal approximation to the t-tstatistic calculated as the difference in average outcomes for treated and control individuals divided by the estimated standard error. The second p-value is based on randomization inference using the difference in average outcomes by treatment status. And the third p-value is based on the randomization distribution using the difference in average ranks by treatment status as the statistic. The results are in Table 1.

In all eight cases the p-values based on the t-test are very similar to those based on randomization inference. This outcome is not surprising given the reasonably large sample sizes, ranging from 71 (Arkansas, WIN) to 4,779 (San Diego, GAIN). However, in a number of cases the p-value for the rank test is fairly different from that based on the level difference. In both sets of four locations there is one location where the rank test suggests a clear rejection at the 5% level whereas the level-based test would suggest that the null hypothesis of no effect should not be rejected at the 5% level. In the WIN (San Diego) evaluation, the p-value goes from 0.068 (levels) to 0.024 (ranks), and in the GAIN (San Diego) evaluation, the p-value goes from 0.136 (levels) to 0.018 (ranks). It is not surprising that the tests give different results. Earnings data are very skewed. A large proportion of the populations participating in these programs have zero earnings even after conditioning on positive past earnings, and the earnings distribution for those with positive earnings is skewed. In those cases a rank-based test is likely to have more power against alternatives that shift the distribution towards higher earnings than tests based on the difference in means.

As a general matter it would be useful in randomized experiments to include such results for rank-based p-values, as a generally applicable way of establishing whether the treatment has any effect. As with all omnibus tests, one should use caution in interpreting a rejection, as

the test can pick up interesting changes in the distribution (such as a mean or median effect) but also less interesting changes (such as higher moments about the mean).

5 Estimation and Inference under Unconfoundedness

Methods for estimation of average treatment effects under unconfoundedness are the most widely used in this literature. Often this assumption, which requires that conditional on observed covariates there are no unobserved factors that are associated both with the assignment and with the potential outcomes, is controversial. Nevertheless, in practice, where often data have been collected in order to make this assumption more plausible, there are many cases where there is no clearly superior alternative, and the only alternative is to abandon the attempt to get precise inferences. In this section we discuss some of these methods and the issues related to them. A general theme of this literature is that the concern is more with biases than with efficiency.

This setting is closely related to that underlying standard multiple regression analysis with a rich set of controls. Unconfoundedness implies that we have a sufficiently rich set of predictors for the treatment indicator, contained in the vector of covariates X_i, such that adjusting for differences in these covariates leads to valid estimates of causal effects. Combined with linearity assumptions of the conditional expectations of the potential outcomes given covariates, the unconfoundedness assumption justifies linear regression. But in the last fifteen years the literature has moved away from the earlier emphasis on regression methods. The main reason is that, although locally linearity of the regression functions may be a reasonable approximation, in many cases the estimated average treatment effects based on regression methods can be severely biased if the linear approximation is not accurate globally. To assess the potential problems with (global) regression methods, it is useful to report summary statistics of the covariates by treatment status. In particular, one may wish to report, for each covariate, the difference in averages by treatment status, scaled by the square root of the sum of the variances, as a scale-free measure of the difference in distributions. To be specific, one may wish to report the normalized difference

$$\Delta_X = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{S_0^2 + S_1^2}}, \tag{3}$$

where for $w = 0, 1$, $S_w^2 = \sum_{i:W_i=w}(X_i - \overline{X}_w)^2/(N_w - 1)$, the sample variance of X_i in the subsample with treatment $W_i = w$. Imbens and Rubin (2007) suggest as a rule of thumb that with a normalized difference exceeding one quarter, linear regression methods tend to be sensitive to the specification. Note the difference with the often reported t-statistic for the null hypothesis of equal means,

$$T = \frac{\overline{X}_1 - \overline{X}_0}{\sqrt{S_0^2/N_0 + S_1^2/N_1}}. \tag{4}$$

The reason for focusing on the normalized difference, (3), rather than on the t-statistic, (4), as a measure of the degree of difficulty in the statistical problem of adjusting for differences in

covariates, comes from their relation to the sample size. Clearly, simply increasing the sample size does not make the problem of inference for the average treatment effect inherently more difficult. However, quadrupling the sample size leads, in expectation, to a doubling of the t-statistic. In contrast, increasing the sample size does not systematically affect the normalized difference. In the landmark Lalonde (1986) paper the normalized difference in mean exceeds unity for many of the covariates, immediately showing that standard regression methods are unlikely to lead to credible results for those data, even if one views unconfoundedness as a reasonable assumption.

As a result of the concerns with the sensitivity of results based on linear regression methods to seemingly minor changes in specification, the literature has moved to more sophisticated methods for adjusting for differences in covariates. Some of these more sophisticated methods use the propensity score – the conditional probability of receiving the treatment – in various ways. Others rely on pairwise matching of treated units to control units, using values of the covariates to match. Although these estimators appear at first sight to be quite different, many (including nonparametric versions of the regression estimators) in fact achieve the semiparametric efficiency bound; thus, they would tend to be similar in large samples. Choices among them typically rely on small sample arguments, which are rarely formalized, and which do not uniformly favor one estimator over another.

Most estimators currently in use can be written as the difference of a weighted average of the treated and control outcomes, with the weights in both groups adding up to one:

$$\hat{\tau} = \sum_{i=1}^{N} \lambda_i \cdot Y_i, \qquad \text{with} \sum_{i:W_i=1} \lambda_i = 1, \qquad \sum_{i:W_i=0} \lambda_i = -1.$$

The estimators differ in the way the weights λ_i depend on the full vector of assignments and matrix of covariates (including those of other units). For example, some estimators implicitly allow the weights to be negative for the treated units and positive for controls units, whereas others do not. In addition, some depend on essentially all other units whereas others depend only on units with similar covariate values. Nevertheless, despite the commonalities of the estimators and large sample equivalence results, in practice the performance of the estimators can be quite different, particularly in terms of robustness and bias. On a more positive note, some understanding has been reached regarding the sensitivity of specific estimators to particular configurations of the data, such as limited overlap in covariate distributions. Currently, the best practice is to combine linear regression with either propensity score or matching methods in ways that explicitly rely on local rather than global linear approximations to the regression functions.

In this section, we first discuss the key assumptions underlying an anlysis based on unconfoundedness. We then review some of the efficiency bound results for average treatment effects. Next, in Sections 5.3 to 5.5 we briefly review the basic methods relying on regression, propensity score methods, and matching. Although still fairly widely used, we do not recommend these methods in practice. In Sections 5.6 to 5.8 we discuss three of the combination methods that we view as more attractive and recommend in practice. In Section 5.9 we discuss estimating variances. Next we discuss implications of lack of overlap in the covariate distributions. In

particular, we discuss two general methods for constructing samples with improved covariate balance. In Section 5.11 we describe methods that can be used to assess the plausibility of the unconfoundedness assumption, even though this assumption is not directly testable. In Section 5.12 we discuss methods for testing for the presence of average treatment effects and for the presence of treatment effect heterogeneity under unconfoundedness.

5.1 Identification

The key assumption is unconfoundedness, introduced by Rosenbaum and Rubin (1983),

Assumption 1 (Unconfoundedness)

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i.$$

The unconfoundedness assumption is often controversial, as it assumes that beyond the observed covariates X_i there are no (unobserved) characteristics of the individual associated both with the potential outcomes and the treatment. Nevertheless, this kind of assumption is used routinely in multiple regression analysis. In fact, suppose we assume that the treatment effect, τ, is constant, so that, for each random draw i, $\tau = Y_i(1) - Y_i(0)$. Further, assume that $Y_i(0) = \alpha + \beta' X_i + \varepsilon_i$, where $\varepsilon_i = Y_i(0) - \mathbb{E}[Y_i(0)|X_i]$ is the residual capturing the unobservables affecting the response in the absense of treatment. Then, with the observed outcome defined as $Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1)$, we can write

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \varepsilon_i,$$

and unconfoundedness is equivalent to independence of ε_i and of W_i, conditional on X_i. Imbens (2004) discusses some economic models that imply unconfoundedness. These models assume agents choose to participate in a program if the benefits, equal to the difference in potential outcomes, exceed the costs associated with participation. Unconfoundedness is implied by independence of the costs and benefits, conditional on observed covariates.

The second assumption used to identify treatment effects is that for all possible values of the covariates, there are both treated and control units.

Assumption 2 (Overlap)

$$0 < \text{pr}(W_i = 1|X_i = x) < 1, \qquad \text{for all } x.$$

We call this the overlap assumption as it implies that the support of the conditional distribution of X_i given $W_i = 0$ overlaps completely with that of the concditional distribution of X_i given $W_i = 1$.

With a random sample $(W_i, X_i)_{i=1}^N$ we can estimate the propensity score $e(x) = \text{pr}(W_i = 1|X_i = x)$, and this can provide some guidance for determining whether the overlap assumption holds. Of course common parametric models, such as probit and logit, ensure that all estimated probabilities are strictly between zero and one, and so examining the fitted probabilities from such models can be misleading. We discuss approaches for improving overlap in 5.10.

The combination of unconfoundedness and overlap was referred to by Rosenbaum and Rubin's (1983) as strong ignorability. There are various ways to establish identification of various average treatment effects under strong ignorability. Perhaps the easiest is to note that $\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ is identified for x in the support of the covariates:

$$\begin{aligned}\tau(x) &= \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x] \\ &= \mathbb{E}[Y_i(1)|W_i = 1, X_i = x] - \mathbb{E}[Y_i(0)|W_i = 0, X_i = x] \\ &= \mathbb{E}[Y_i|W_i = 1, X_i = x] - \mathbb{E}[Y_i|W_i = 0, X_i = x], \end{aligned} \tag{5}$$

where the second equality follows by unconfoundedess: $\mathbb{E}[Y_i(w)|W_i = w, X_i]$ does not depend on w. By the overlap assumption, we can estimate both terms in the last line, and therefore we can identify $\tau(x)$. Given that we can identify $\tau(x)$ for all x, we can identify the expected value across the population distribution of the covariates,

$$\tau_{\text{pate}} = \mathbb{E}[\tau(X_i)], \tag{6}$$

as well as τ_{patt} and other estimands.

5.2 Efficiency Bounds

Before discussing specific estimation methods, it is useful to see what we can learn about the parameters of interest, given just the strong ignorability of treatment assignment assumption, without functional form or distributional assumptions. In order to do so, we need some additional notation. Let $\sigma_0^2(x) = \mathbb{V}(Y_i(0)|X_i = x)$ and $\sigma_1^2(x) = \mathbb{V}(Y_i(1)|X_i = x)$ denote the conditional variances of the potential outcomes given the covariates. Hahn (1998) derives the lower bounds for asymptotic variances of $\sqrt{N}$-consistent estimators for τ_{pate} as

$$\mathbb{V}_{\text{pate}} = \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1 - e(X_i)} + (\tau(X_i) - \tau)^2\right], \tag{7}$$

where $p = \mathbb{E}[e(X_i)]$ is the unconditional treatment probability. Interestingly, this lower bound holds irrespective of whether the propensity score is known or not. The form of this variance bound is informative. It is no surprise that τ_{pate} is more difficult to estimate the larger are the variances $\sigma_0^2(x)$ and $\sigma_1^2(x)$. However, as shown by the presence of the third term, it is also more difficult to estimate τ_{pate}, the more variation there is in the average treatment effect conditional on the covariates. If we focus instead on estimating τ_{cate}, the conditional average treatment effect, the third term drops out, and the variance bound for τ_{cate} is

$$\mathbb{V}_{\text{cate}} = \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1 - e(X_i)}\right]. \tag{8}$$

Still, the role of heterogeneity in the treatment effect is potentially important. Suppose we actually had prior knowledge that the average treatment effect conditional on the covariates is constant, or $\tau(x) = \tau_{\text{pate}}$ for all x. Given this assumption the model is closely related to the

partial linear model (Robinson, 1988; Stock, 1989). Given this prior knowledge, the variance bound is

$$\mathbb{V}_{\text{const}} = \left(\mathbb{E}\left[\left(\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} \right)^{-1} \right] \right)^{-1}. \tag{9}$$

This variance bound can be much lower than (8) if there is variation in the propensity score. Knowledge of lack of variation in the treatment effect can be very valuable, or, conversely, allowing for general heterogeneity in the treament effect can be expensive in terms of precision.

In addition to the conditional variances of the counterfactual outcomes, a third important determinant of the efficiency bound is the propensity score. Because it enters in (7) in the denominator, the presence of units with the propensity score close to zero or one will make it difficult to obtain precise estimates of the average effect of the treatment. One approach to address this problem, developed by Crump, Hotz, Imbens and Mitnik (2008a) and discussed in more detail in Section 5.10, is to drop observations with the propensity score close to zero and one, and focus on the average effect of the treatment in the subpopulation with propensity scores away from zero. Suppose we focus on $\tau_{\text{cate},\mathbb{A}}$, the average of $\tau(X_i)$ for $X_i \in \mathbb{A}$. Then the variance bound is

$$\mathbb{V}_{\mathbb{A}} = \frac{1}{\text{pr}(X_i \in \mathbb{A})} \cdot \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} \middle| X_i \in \mathbb{A} \right], \tag{10}$$

By excluding from the set $\mathbb{A}$ subsets of the covariate space where the propensity score is close to zero or one, we may be able to estimate $\tau_{\text{cate},\mathbb{A}}$ more precisely than τ_{cate}. (If we are instead interested in τ_{catt}, we only need to worry about covariate values where $e(x)$ is close to one.)

Having displayed these lower bounds on variances for the average treatment effects, a natural question is: Are there estimators that achieve these lower bounds that do not require parametric models or functional form restrictions on either the conditional means or the propensity score? The answer in general is yes, and we now consider different classes of estimators in turn.

5.3 Regression Methods

To describe the general approach to regression methods for estimating average treatment effects, define $\mu_0(x)$ and $\mu_1(x)$ to be the two regression functions for the potential outcomes.

$$\mu_0(x) = \mathbb{E}[Y_i(0)|X_i = x] \text{ and } \mu_1(x) = \mathbb{E}[Y_i(1)|X_i = x].$$

By definition, the average treatment effect conditional on $X = x$ is $\tau(x) = \mu_1(x) - \mu_0(x)$. As we discussed in the identification subsection, under the unconfoundedness assumption, $\mu_0(x) = \mathbb{E}[Y_i|W_i = 0, X_i = x]$ and $\mu_1(x) = \mathbb{E}[Y_i|W_i = 1, X_i = x]$, which means we can estimate $\mu_0(\cdot)$ using regression methods for the untreated subsample and $\mu_1(\cdot)$ using the treated subsample. Given consistent estimators $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$, a consistent estimator for either τ_{pate} or τ_{cate} is

$$\hat{\tau}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \Big). \tag{11}$$

Given parametric models for $\mu_0(\cdot)$ and $\mu_1(\cdot)$, estimation and inference are straightforward. In the simplest case, we assume each conditional mean can be expressed as functions linear in parameters, say

$$\mu_0(x) = \alpha_0 + \beta_0'(x - \psi_X), \qquad \mu_1(x) = \alpha_1 + \beta_1'(x - \psi_X), \tag{12}$$

where we take deviations from the overall population covariate mean ψ_X so that the treatment effect is the difference in intercepts. (Naturally, as in any regression context, we can replace x with general functions of x.) Of course, we rarely know the population mean of the covariates is rarely known, so in estimation we replace ψ_X with the sample average across all units, $\overline{X}$. Then $\hat{\tau}_{\rm reg}$ is simply

$$\hat{\tau}_{\rm reg} = \hat{\alpha}_1 - \hat{\alpha}_0. \tag{13}$$

This estimator is also obtained from the coefficient on the treatment indicator W_i in the regression Y_i on $1, W_i, X_i, W_i \cdot (X_i - \overline{X})$. Standard errors can be obtained from standard least square regression output. (As we show below, in the case of estimating $\tau_{\rm pate}$, the usual standard errors, whether or not they are made robust to heteroskedasticity, ignore the estimation error in $\overline{X}$ as an estimator of ψ_X, and so formally the conventional standard errors are only valid for $\tau_{\rm cate}$ and not for $\tau_{\rm pate}$.)

A different representation of $\hat{\tau}_{\rm reg}$ is useful in order to illustrate some of the concerns with regression estimators in this setting. Suppose we do use the linear model in (12). It can be shown that

$$\hat{\tau}_{\rm reg} = \overline{Y}_1 - \overline{Y}_0 - \left(\frac{N_0}{N_0 + N_1} \cdot \hat{\beta}_1 + \frac{N_1}{N_0 + N_1} \cdot \hat{\beta}_0 \right)' \left(\overline{X}_1 - \overline{X}_0 \right). \tag{14}$$

To adjust for differences in covariates between treated and control units, the simple difference in average outcomes, $\overline{Y}_1 - \overline{Y}_0$, is adjusted by the difference in average covariates, $\overline{X}_1 - \overline{X}_0$, multiplied by the weighted average of the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ in the two treatment regimes. This is a useful representation. It shows that if the averages of the covariates in the two treatment arms are very different, then the adjustment to the simple mean difference can be large. We can see that even more clearly by inspecting the predicted outcome for the treated units had they been subject to the control treatments:

$$\hat{\mathbb{E}}\left[Y_i(1) \middle| W_i = 0 \right] = \overline{Y}_0 + \hat{\beta}_0'(\overline{X}_1 - \overline{X}_0).$$

The regression parameter $\hat{\beta}_0$ is estimated on the control sample, where the average of the covariates is equal to $\overline{X}_0$. It therefore likely provides a good approximation to the conditional mean function around that value. However, this estimated regression function is then used to predict outcomes in the treated sample, where the average of the covariates is equal to $\overline{X}_1$. If these covariate averages are very different, and thus the regression model is used to predict outcomes far away from where the parameters were estimated, the results can be sensitive to minor changes in the specification. Unless the linear approximation to the regression function is globally accurate, regression may lead to severe biases. Another way of interpreting this problem

is as a multicollinearity problem. If the averages of the covariates in the two treatment arms are very different, the correlation between the covariates and the treatment indicator is relatively high. Although conventional least squares standard errors take the degree of multicollinearity into account, they do so conditional on the specification of the regression function. Here the concern is that any misspecification may be exacerbated by the collinearity problem. As noted in the introduction to Section 5, an easy way to establish the severity of this problem is to inspect the normalized differences $(\overline{X}_1 - \overline{X}_0)/\sqrt{S_0^2 + S_1^2)}$.

In the case of the standard regression estimator it is straightforward to derive and to estimate the variance when we view the estimator as an estimator of τ_{cate}. Assuming the linear regression model is correctly specified, we have

$$\sqrt{N}(\hat{\tau}_{\text{reg}} - \tau_{\text{cate}}) \xrightarrow{d} \mathcal{N}(0, V_0 + V_1), \qquad \text{where} \quad V_w = N \cdot \mathbb{E}\left[(\hat{\alpha}_w - \alpha_w)^2\right], \tag{15}$$

which can be obtained directly from standard regression output. Estimating the variance when we view the estimator as an estimator of τ_{pate} requires adding a term capturing the variation in the treatment effect conditional the covariates. The form is then

$$\sqrt{N}(\hat{\tau}_{\text{reg}} - \tau_{\text{pate}}) \xrightarrow{d} \mathcal{N}(0, V_0 + V_1 + V_\tau),$$

where the third term in the normalized variance is

$$V_\tau = (\beta_1 - \beta_0)' \, \mathbb{E}\left[(X_i - \mathbb{E}[X_i])\,(X_i - \mathbb{E}[X_i])'\right](\beta_1 - \beta_0),$$

which can be estimated as

$$\hat{V}_\tau = \left(\hat{\beta}_1 - \hat{\beta}_0\right)' \frac{1}{N} \sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)' \left(\hat{\beta}_1 - \hat{\beta}_0\right).$$

In practice this additional term is rarely incorporated, and researcher instead report the variance corresponding to τ_{cate}. In cases where the slope coefficients do not differ substantially across the two regimes – equivalently, the coefficients on the interaction terms $W_i \cdot (X_i - \overline{X})$ are "small" – this last term is likely to be swamped by the variances in (15).

In many cases researchers have sought to go beyond simple parametric models for the regression functions. Two general directions have been explored. The first relies on local smoothing, and the second on increasingly flexible global approximations. We discuss both in turn.

Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) consider local smoothing methods for estimate the two regression functions. The first method they consider is kernel regression. Given a kernel $K(\cdot)$, and a bandwith h, the kernel estimator for $\mu_w(x)$ is

$$\hat{\mu}_w(x) = \sum_{i:W_i=w} Y_i \cdot \lambda_i \qquad \text{with weight} \quad \lambda_i = K\left(\frac{X_i - x}{h}\right) \Big/ \sum_{i:W_i=w} K\left(\frac{X_i - x}{h}\right).$$

Although the rate of convergence of the kernel estimator to the regression function is slower than the conventional parametric rate $N^{-1/2}$, the rate of convergence of the implied estimator for the

average treatment effect, $\hat{\tau}_{\rm reg}$ in (11), is the regular parametric rate under regularity conditions. These conditions include smoothness of the regression functions and require the use of higher order kernels (with the order of the kernel depending on the dimension of the covariates). In practice researchers have not used higher order kernels, and with positive kernels the bias for kernel estimators is a more severe problem than for the matching estimators discussed in Section 5.5.

Kernel regression of this type can be interpreted as locally fitting a constant regression function. A general alternative is to fit locally a polynomial regression function. The leading case of this is local linear regression (Fan and Gijbels, 1996), applied to estimation of average treatment effects by Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998). Define $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ as the local least squares estimates, based on locally fitting a linear regression function:

$$\Big(\hat{\alpha}(x),\hat{\beta}(x)\Big) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \lambda_i \cdot \big(Y_i - \alpha - \beta'(X_i - x)\big)^2,$$

with the same weights λ_i as in the standard kernel estimator. The regression function at x is then estimated as $\hat{\mu}(x) = \hat{\alpha}(x)$. In order to achieve convergence at the best possible rate for $\hat{\tau}_{\rm reg}$, one needs to use higher order kernels, although the order required is less than that for the standard kernel estimator.

For both the standard kernel estimator and the local linear estimator an important choice is that of the bandwidth h. In practice, researchers have used *ad hoc* methods for bandwidth selection. Formal results on bandwidth selection from the literature on nonparametric regression are not directly applicable. Those results are based on minimizing a global criterion such as the expected value of the squared difference between the estimated and true regression function, with the expectation taken with respect to the marginal distribution of the covariates. Thus, they focus on estimating the regression function well everywhere. Here the focus is on a particular scalar functional of the regression function, and it is not clear whether the conventional methods for bandwidth choices have good properties.

Although formal results are given for the case with continuous regressors, modifications have been developed that allows for both continuous and discrete covariates (Qi and Racine, 2004). All such methods require choosing the degree of smoothing (often known as bandwidths), and there has not been much work on choosing bandwidths for the particular problem of estimating average treatment effects where the parameter of interest is effectively the average of a regression function, and not the entire function. See Imbens (2004) for more discussion. Although the estimators based on local smoothing have not been shown to attain the variance efficiency bound, it is likely that they can be constructed to do so under sufficient smoothness conditions.

An alternative to local smoothing methods are global smoothing methods, such as series or sieve estimators. Such estimators are parametric for a given sample size, with the number of parameters and the flexibility of the model increasing with the sample size. One attraction of such methods is that often estimation and inference can proceed as if the model is completely parametric. The amount of smoothing is determined by the number of terms in the series, and the large-sample analysis is carried out with the number of terms growing as a function of the

sample size. Again, little is known about how to choose the number of terms when interest lies in average treatment effects. For the average treatment case Hahn (1998), Imbens, Newey, and Ridder (2004), and Chen, Hong, and Tarozzi (2007) have developed estimators of this type. Hahn shows that estimators in this class can achieve the variance lower bounds for estimating τ_{pate}. For a simple version of such an estimator, suppose that X_i is a scalar. Then we can approximate $\mu_w(x)$ by a K-th order polynomial

$$\mu_{w,K}(x) = \sum_{k=0}^{K} \beta_{w,k} \cdot x^k.$$

We then estimate $\beta_{w,k}$ by least squares regression, and estimate the average treatment effect using (11). This is a special case of the estimator discussed in Imbens, Newey and Ridder (2004) and Chen, Hong and Tarozzi (2007), with formal results presented for the case with general X_i. Imbens, Newey and Ridder (2004) also discuss methods for choosing the number of terms in the series based on expected squared error for the average treatment effect.

If the outcome is binary, or more generally of a limited dependent variable form, a linear series approximation to the regression function is not necessarily attractive. It is likely that one can use increasingly flexible approximations based on models that exploit the structure of the outcome data. For the case with binary outcomes, Hirano, Imbens and Ridder (2003) show how using a polynomial approximation to the log odds ratio leads to an attractive estimator for the conditional mean. See Chen (2007) for general discussion of such models. One can imagine that in cases with nonnegative response variables, exponential regression functions, or those derived from specific models, such as Tobit (when the response can pile up at zero), combined with polynomial approximations in the linear index function, might be useful.

Generally, methods based on global approximations suffer from the same drawbacks as linear regression. If the covariate distributions are substantially different in both treatment groups, estimates based on such methods rely, perhaps more than is desired, on extrapolation. Using these methods in cases with substantial differences in covariate distributions is therefore not recommended (except possibly in cases where the sample has been trimmed so that the covariates across the two treatment regimes have sufficient considerable overlap).

Before we turn to propensity score methods, we should comment on estimating the average treatment effects on the treated, τ_{patt} and τ_{catt}. In this case, $\hat{\tau}(X_i)$ gets averaged across observations with $W_i = 1$, rather than across the entire sample as in (11) Because $\hat{\mu}_1(x)$ is estimated on the treated subsample, in estimating ATT or ATT there is no problem if $\mu_1(x)$ is poorly estimated at covariate values that are common in the control group but scarce in the treatment group. But we must have a good estimate of $\mu_0(x)$ at covariate values common in the treatment group, and this is not ensured because we can only use the control group to obtain $\hat{\mu}_0(x)$. Nevertheless, in many settings $\mu_0(x)$ can be estimated well over the entire range of covariates because the control group often includes units that are similar to those in the treatment group. By contrast, often many units in the control group are quite different from any units in the control group, and this is what makes the ATE parameters considerably more difficult to estimate than ATT parameters in many applications.

5.4 Methods Based on the Propensity Score

The first set of alternatives to regression estimators relies on estimates of the propensity score. An important result here is due to Rosenbaum and Rubin (1983a). They show that under unconfoundedness, independence of potential outcomes and treatment indicators also holds after conditioning solely on the propensity score, $e(x) = \text{pr}(W_i = 1|X_i = x)$:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i \quad \Longrightarrow \quad W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; e(X_i).$$

The basic insight is that for any binary variable W_i, and any random vector X_i, it is true (without assuming unconfoundedness) that

$$W_i \perp\!\!\!\perp X_i \mid e(X_i).$$

Hence, within subpopulations with the same value for the propensity score, covariates are independent of the treatment indicator and thus cannot lead to biases (the same way in a regression framework omitted variables that are uncorrelated with included covariates do not introduce bias). Since under unconfoundedness all biases can be removed by adjusting for differences in covariates, this means that within subpopulations homogenous in the propensity score there are no biases in comparisons between treated and control units.

Given the Rosenbaum-Rubin result, it is sufficient, under the maintained assumption of unconfoundedness, to adjust solely for differences in the propensity score between treated and control units. This result can be exploited in a number of ways. Here we discuss three of these that have been used in practice. The first two of these methods exploit the fact that the propensity score can be viewed as a covariate that is sufficient to remove biases in estimation of average treatment effects. For this purpose, any one-to-one function of the propensity score could also be used. The third method further uses the fact that the propensity score is the conditional probability of receiving the treatment.

The first method simply uses the propensity score in place of the covariates in regression analysis. Define $\nu_w(e) = \mathbb{E}[Y_i|W_i = w, e(X_i) = e)$. Unconfoundedness in combination with the Rosenbaum-Rubin result implies that $\nu_w(e) = \mathbb{E}[Y_i(w)|e(X_i) = e]$. Then we can estimate $\nu_w(e)$ very generally using kernel or series estimation on the propensity score, something which is greatly simplified by the fact that the propensity score is a scalar. Heckman, Ichimura, and Todd (1998) consider local smoothers and Hahn (1998) considers a series estimator. In either case we have the consistent estimator

$$\hat{\tau}_{\text{regprop}} = \frac{1}{N} \cdot \sum_{i=1}^{N} \Big(\hat{\nu}_1\left(e(X_i)\right) - \hat{\nu}_0\left(e(X_i)\right)\Big),$$

which is simply the average of the differences in predicted values for the treated and untreated outcomes. Interestingly, Hahn shows that, unlike when we use regression to adjust for the full set of covariates, the series regression estimator based on adjusting for the known propensity score does not achieve the efficiency bound.

Although methods of this type have been used in practice, probably because of their simplicity, regression on simple functions of the propensity score is not recommended. Because the

propensity score does not have a substantive meaning, it is difficult to motivate a low order polynomial as a good approximation to the conditional expectation. For example, a linear model in the propensity score is unlikely to provide a good approximation to the conditional expectation: individuals with propensity scores of 0.45 and 0.50 are likely to be much more similar than individuals with propensity scores equal to 0.01 and 0.06. Moreover, no formal asymptotic properties have been derived for the case with the propensity score unknown.

The second method, variously referred to as blocking, subclassification, or stratification, also adjusts for differences in the propensity score in a way that can be interpreted as regression, but in a more flexible manner. Originally suggested by Rosenbaum and Rubin (1983), the idea is to partition the sample into strata by (discretized) values of the propensity score, and then analyse the data within each stratum as if the propensity score were constant and the data could be interpreted as coming from a completely randomized experiment. This can be interpreted as approximating the conditional mean of the potential outcomes by a step function. To be more precise, let $0 = c_0 < c_1 < c_2 < \ldots < c_J = 1$ be boundary values. Then define B_{ij}, for $i = 1, \ldots, N$, and $j = 1, \ldots, J-1$, as the indicators

$$B_{ij} = \begin{cases} 1 & \text{if } c_{j-1} \leq e(X_i) < c_j \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \quad B_{iJ} = 1 - \sum_{j=1}^{J-1} B_{ij}.$$

Now estimate within stratum j the average treatment effect $\tau_j = \mathbb{E}[Y_i(1) - Y_i(0) | B_{ij} = 1]$ as

$$\hat{\tau}_j = \overline{Y}_{j1} - \overline{Y}_{j0} \qquad \text{where} \quad \overline{Y}_{jw} = \frac{1}{N_{jw}} \sum_{i:W_i=w}^{N} B_{ij} \cdot Y_i, \quad \text{and } N_{jw} = \sum_{i:W_i=w}^{N} B_{ij}.$$

If J is sufficiently large, and the differences $c_j - c_{j-1}$ small, there is little variation in the propensity score within a stratum or block, and one can analyze the data as if the propensity score is constant, and thus as if the data within a block were generated by a completely randomized experiment (with the assignment probabilities constant within a stratum, but varying between strata). The average treatment effect is then estimated as the weighted average of the within-stratum estimates:

$$\hat{\tau}_{\text{block}} = \sum_{j=1}^{J} \hat{\tau}_j \cdot \left(\frac{N_{j0} + N_{j1}}{N} \right).$$

With J large, the implicit step function approximation to the regression functions $\nu_w(e)$ will be accurate. Cochran (1969) shows in a Gaussian example that with five equal-sized blocks the remaining bias is less than 5% of the bias in the simple difference between average outcomes among treated and controls. Motivated by Cochran's calculations, researchers have often used five strata, although depending on the sample size and the joint distribution of the data, fewer or more blocks will generally lead to a lower expected mean squared error.

The variance for this estimator is typically calculated conditional on the strata indicators, and assuming random assignment within the strata. That is, for stratum j, the estimator is $\hat{\tau}_j$, and its variance is estimated as $\hat{V}_j = \hat{V}_{j0} + \hat{V}_{j1}$, where

$$\hat{V}_{jw} = \frac{S_{jw}^2}{N_{jw}}, \qquad \text{where} \quad S_{jw}^2 = \frac{1}{N_{jw}} \sum_{i:B_{ij}=1, W_i=w} \left(Y_i - \overline{Y}_{jw} \right)^2.$$

The overall variance is then estimated as

$$\hat{\mathbb{V}}(\hat{\tau}_{\text{block}}) = \sum_{j=1}^{J} \left(\hat{V}_{0j} + \hat{V}_{1j} \right) \cdot \left(\frac{N_{j0} + N_{j1}}{N} \right)^2 .$$

This variance estimator is appropriate for τ_{cate}, although it ignores biases arising from variation in the propensity score within strata.

The third method exploiting the propensity score is based on weighting. Recall that $\tau_{\text{pate}} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$. We consider the two terms separately. Because $W_i \cdot Y_i = W_i \cdot Y_i(1)$, we have

$$\begin{aligned} \mathbb{E}\left[\frac{W_i \cdot Y_i}{e(X_i)} \right] &= \mathbb{E}\left[\frac{W_i \cdot Y_i(1)}{e(X_i)} \right] = \mathbb{E}\left[\mathbb{E}\left[\frac{W_i \cdot Y_i(1)}{e(X_i)} \middle| X_i \right] \right] = \mathbb{E}\left[\frac{\mathbb{E}(W_i|X_i) \cdot \mathbb{E}(Y_i(1)|X)}{e(X_i)} \right] \\ &= \mathbb{E}\left[\frac{e(X_i) \cdot \mathbb{E}(Y_i(1)|X_i)}{e(X_i)} \right] = \mathbb{E}[\mathbb{E}(Y_i(1)|X_i] = \mathbb{E}[Y_i(1)], \end{aligned}$$

where the second and final inequalities follow by iterated expectations and the third equality holds by unconfoundedness. The implication is that weighting the treated population by the inverse of the propensity score recovers the expectation of the unconditional response under treatment. A similar calculation shows $\mathbb{E}[((1 - W_i)Y_i)/(1 - e(X_i))] = \mathbb{E}[Y_i(0)]$, and together these imply

$$\tau_{\text{pate}} = \mathbb{E}\left[\frac{W_i \cdot Y_i}{e(X_i)} - \frac{(1 - W_i) \cdot Y_i}{1 - e(X_i)} \right]. \tag{16}$$

Equation (16) suggests an obvious estimator of τ_{pate}:

$$\hat{\tau}_{\text{weight}} = \frac{1}{N} \cdot \sum_{i=1}^{N} \left[\frac{W_i \cdot Y_i}{e(X_i)} - \frac{(1 - W_i) \cdot Y_i}{1 - e(X_i)} \right], \tag{17}$$

which, as a sample average from a random sample, is consistent for τ_{pate} and $\sqrt{N}$ asymptotically normally distributed. The estimator in (17) is essentially due to Horvitz and Thompson (1952).

In practice (17) is not a feasible estimator because it depends on the propensity score function $e(\cdot)$, which is rarely known. A surprising result is that, even if we know the propensity score, $\hat{\tau}_{\text{weight}}$ does not achieve the efficiency bound given in (7). It turns out to be better, in terms of large sample efficiency, to weight using the estimated rather than the true propensity score. Hirano, Imbens, and Ridder (2003) establish conditions under which replacing $e(\cdot)$ with a logistic sieve estimator results in a weighted propensity score estimator that achieves the variance bound. The estimator is practically simple to compute, as estimation of the propensity score involes a straightforward logit estimation involving flexible functions of the covariates. Theoretically, the number of terms in the approximation should increase with the sample size. In the second step, given the estimated propensity score $\hat{e}(x)$, one estimates

$$\hat{\tau}_{\text{ipw}} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{\hat{e}(X_i)} - \sum_{i=1}^{N} \frac{(1 - W_i) \cdot Y_i}{1 - \hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{1 - \hat{e}(X_i)}. \tag{18}$$

We refer to this as the inverse probability weighting (IPW) estimator. See Hirano, Imbens, and Ridder (2003) for intuition as to why estimating the propensity score leads to a more efficient estimator, asymptotically, than knowing the propensity score.

Ichimura and Linton (2001) studied $\hat{\tau}_{\text{ipw}}$ when $\hat{e}(\cdot)$ is obtained via kernel regression, and they consider the problem of optimal bandwidth choice when the object of interest is τ_{pate}. More recently, Li, Racine, and Wooldridge (2008) consider kernel estimation for discrete as well as continuous covariates. The estimator proposed by Li, Racine and Wooldridge achieves the variance lower bound. See Hirano, Imbens and Ridder (2006) and Wooldridge (2007) for methods for estimating the variance for these estimators.

Note that the blocking estimator can also be interpreted as a weighting estimator. Consider observations in block j. Within the block the N_{j1} treated observations all get equal weight $1/N_{j1}$. In the estimator for the overall average treatment effect this block gets weight $(N_{j0} + N_{j1})/N$, so we can write $\hat{\tau} = \sum_{i=1}^{N} \lambda_i \cdot Y_i$, where for treated observations in block j the weight normalized by N is $N \cdot \lambda_i = (N_{j0} + N_{j1})/N_{j1})$, and for control observations it is $N \cdot \lambda_i = (N_{j0} + N_{j1})/N_{j0})$. Implicitly this estimator is based on an estimate of the propensity score in block j equal to $N_{j1}/(N_{j0} + N_{j1})$. Compared to the ipw estimator the propensity score is smoothed within the block. This has the advantage of avoiding particularly large weights, but comes at the expense of introducing bias if the propensity score is correctly specified.

A particular concern with IPW estimators arises again when the covariate distributions are substantially different for the two treatment groups. That implies that the propensity score gets close to zero or one for some values of the covariates. Small or large values of the propensity score raises a number of issues. One concern is that alternative parametric models for the binary data, such as probit and logit models that can provide similar approximations in terms of estimated probabilities over the middle ranges of their arguments, tend to be more different when the probabilities are close to zero or one. Thus the choice of model and specification becomes more important, and it is often difficult to make well motivated choices in treatment effect settings. A second concern is that for units with propensity scores close to zero or one, the weights can be be large, making those units particularly influential in the estimates of the average treatment effects, and thus making the estimator imprecise. These concerns are less serious than those regarding regression estimators because at least the IPW estimates will accurately reflect uncertainty. Still, these concerns make the simple ipw estimators less attractive. (As for regression cases, the problem can be less severe for the ATT parameters because propensity score values close to zero play no role. Problems for estimating ATT arise when some units, as described by their observed covariates, are almost certain to receive treatment.)

5.5 Matching

Matching estimators impute the missing potential outcomes using only the outcomes of a few nearest neighbors of the opposite treatment group. In that sense, matching is similar to non-parametric kernel regression, with the number of neighbors playing the role of the bandwidth in the kernel regression. A formal difference with kernel methods is that the asymptotic distribution for matching estimators is derived conditional on the implicit bandwidth, that is, the number of neighbors, often fixed at a small number, e.g., one. Using such asymptotics, the

implicit estimate $\hat{\mu}_w(x)$ is (close to) unbiased, but not consistent, for $\mu_w(x)$. In contrast, the kernel regression estimators discussed in the previous section implied consistency of $\hat{\mu}_w(x)$.

Matching estimators have the attractive feature that the smoothing parameters are easily interpretable. Given the matching metric, the researcher only has to choose the number of matches. Using only a single match leads to the most credible inference with the least bias, at the cost of sacrificing some precision. This sits well with the focus in the literature on reducing bias rather than variance. It also can make the matching estimator easier to use than those estimators that require more complex choices of smoothing parameters, and this may be another explaination for its popularity.

Matching estimators have been widely studied in practice and theory (e.g., Gu and Rosenbaum, 1993; Rosenbaum, 1989, 1995, 2002; Rubin, 1973b, 1979; Rubin and Thomas, 1992ab, 1996, 2000; Heckman, Ichimura and Todd, 1998; Dehejia and Wahba, 1999; Abadie and Imbens, 2002). Most often they have been applied in settings where, (i) the interest is in the average treatment effect for the treated, and (ii) there is a large reservoir of potential controls, although recent work (Abadie and Imbens, 2006) shows that matching estimators can be modified to estimate the overall average effect. The setting with many potential controls allows the researcher to match each treated unit to one or more distinct controls, hence the label "matching without replacement." Given the matched pairs, the treatment effect within a pair is estimated as the difference in outcomes, and the overall average as the average of the within-pair difference. Exploiting the representation of the estimator as a difference in two sample means, inference is based on standard methods for differences in means or methods for paired randomized experiments, ignoring any remaining bias. Fully efficient matching algorithms that take into account the effect of a particular choice of match for treated unit i on the pool of potential matches for unit j are computationally cumbersome. In practice, researchers use greedy algorithms that sequentially match units. Most commonly the units are ordered by the value of the propensity score with the highest propensity score units matched first. See Gu and Rosenbaum (1993) and Rosenbaum (1995) for discussions.

Abadie and Imbens (2006) study formal asymptotic properties of matching estimators in a different setting, where both treated and control units are (potentially) matched and matching is done with replacement. Code for the Abadie-Imbens estimator is available in Matlab and Stata (see Abadie, Drukker, Herr, and Imbens, 2003).[5] Formally, given a sample, $\{(Y_i, X_i, W_i)\}_{i=1}^N$, let $\ell_1(i)$ be the nearest neighbor to i, that is, $\ell_1(i)$ is equal to the nonnegative integer j, for $j \in \{1, \ldots, N\}$, if $W_j \neq W_i$, and

$$\|X_j - X_i\| = \min_{k:W_k \neq W_i} \|X_k - X_i\| .$$

More generally, let $\ell_m(i)$ be the index that satisfies $W_{\ell_m(i)} \neq W_i$ and that is the m-th closest to unit i:

$$\sum_{l:W_l \neq W_i} 1\Big\{\|X_l - X_i\| \leq \|X_{\ell_m(i)} - X_i\|\Big\} = m,$$

[5] See Becker and Ichino (2002) and Sianesi (2001) for alternative Stata implementations of matching estimators.

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is the m-th closest to unit i in terms of the distance measure based on the norm $\|\cdot\|$. Let $\mathcal{J}_M(i) \subset \{1,\ldots,N\}$ denote the set of indices for the first M matches for unit i: $\mathcal{J}_M(i) = \{\ell_1(i),\ldots,\ell_M(i)\}$. Now impute the missing potential outcomes as the average of the outcomes for the matches, by defining $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ as

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } W_i = 1, \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

The simple matching estimator discussed in Abadie and Imbens is then

$$\hat{\tau}_{\text{match}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{Y}_i(1) - \hat{Y}_i(0)\Big). \tag{19}$$

Abadie and Imbens show that the bias of this estimator is of order $O(N^{-1/K})$, where K is the dimension of the covariates. Hence, if one studies the asymptotic distribution of the estimator by normalizing by $\sqrt{N}$ (as can be justified by the fact that the variance of the estimator is of order $O(1/N)$), the bias does not disappear if the dimension of the covariates is equal to two, and will dominate the large sample variance if K is at least three. To put this result in perspective, it is useful to relate it to bias properties of estimators based on kernel regression. Kernel estimators can be viewed as matching estimators where all observations within some bandwidth h_N receive some weight. As the sample size N increases, the bandwidth h_N shrinks, but sufficiently slow in order to ensure that the number of units receiving non-zero weights diverges. If all the weights are positive, the bias for kernel estimators would generally be worse. In order to achieve root$-N$ consistency it is therefore critical that some weights are negative through the device of higher order kernels, with the exact order required dependent on the dimension of the covariates (see, e.g., Heckman, Ichimura and Todd, 1998). In practice, however, researchers have not used higher order kernels, and so bias concerns for nearest-neighbor matching estimators are even more relevant for kernel matching methods.

There are three caveats to the Abadie-Imbens bias result. First, it is only the continuous covariates that should be counted in the dimension of the covariates. With discrete covariates the matching will be exact in large samples, and as a result such covariates do not contribute to the order of the bias. Second, if one matches only the treated, and the number of potential controls is much larger than the number of treated units, one can justify ignoring the bias by appealling to an asymptotic sequence where the number of potential controls increases faster with the sample size than the number of treated units. Specifically, if the number of controls, N_0, and the number of treated, N_1, satisfy $N_1/N_0^{4/K} \to 0$, then the bias disappears in large samples after normalization by $\sqrt{N}_1$. Third, even though the order of the bias may be high, the actual bias may still be small if the coefficients in the leading term are small. This is possible if the biases for different units are at least partially offsetting. For example, the leading term in the bias relies on the regression function being nonlinear, and the density of the covariates having a nonzero slope. If either the regression function is well approximated by a linear function, or the density is approximately flat, the bias may be fairly limited.

Abadie and Imbens (2006) also show that matching estimators are generally not efficient. Even in the case where the bias is of low enough order to be dominated by the variance, the estimators do not reach the efficiency bound given a fixed number of matches. To reach the bound the number of matches would need to increase with the sample size. If $M \to \infty$, with $M/N \to 0$, then the matching estimator is essentially like a nonparametric regression estimator. However, it is not clear that using an approximation based on a sequence with an increasing number of matches improves the accuracy of the approximation. Given that in an actual data set one uses a specific number of matches, M, it would appear appropriate to calculate the asymptotic variance conditional on that number, rather than approximate the distribution as if this number is large. Calculations in Abadie and Imbens show that the efficiency loss from even a very small number of matches is quite modest, and so the concerns about the inefficiency of matching estimators may not be very relevant in practice. Little is known about the optimal number of matches, or about data-dependent ways of choosing it.

All of the distance metrics used in practice standardize the covariates in some manner. Abadie and Imbens use a diagonal matrix with each diagional element equal to the inverse of the corresponding covariate variance. The most common metric is the Mahalanobis metric, which is based on the inverse of the full covariance matrix. Zhao (2004), in an interesting discussion of the choice of metrics, suggests some alternatives that depend on the correlation between covariates, treatment assignment, and outcomes. So far there is little experience with any metrics beyond inverse-of-the-variances and the Mahalanobis metrics. Zhao (2004) reports the results of some simulations using his proposed metrics, finding no clear winner given his specific design.

5.6 Combining Regression and Propensity Score Weighting

In the Sections 5.3 and 5.4, we described methods for estimating average causal effects based on two strategies: the first is based on estimating $\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x]$ for $w = 0, 1$ and averaging the difference as in (11), and the second is based on estimating the propensity score $e(x) = \text{pr}(W_i = 1|X_i = x)$ and using that to weight the outcomes as in (18). For each approach we have discussed estimators that achieve the asymptotic efficiency bound. If we have large sample sizes, relative to the dimension of X_i, we might think our nonparametric estimators of the conditional means or propensity score are sufficiently accurate to invoke the asymptotic efficiency results described above.

In other cases, however, we might choose flexible parametric models without being confident that they necessarily approximate the means or propensity score well. As we discussed earlier, one reason for viewing estimators of conditional means or propensity scores as flexible parametric models is that it greatly simplifies standard error calculations for treatment effect estimates. In such cases, one might want to adopt a strategy that combines regression and propensity score methods in order to achieve some robustness to misspecification of the parametric models. It may be helpful to think about the analogy to omitted variable bias. Suppose we are interested in the coefficient on W_i in the (long) linear regression of Y_i on a constant, W_i and X_i. Suppose we omit X_i from the long regression, and just run the short regression of Y_i on a constant and W_i. The bias in the estimate from the short regression is equal to the product of

the coefficient on X_i in the long regression, and the coefficient on X_i in a regression of W_i on a constant and X_i. Weighting can be interpreted as removing the correlation between W_i and X_i, and regression as removing the direct effect of X_i. Weighting therefore removes the bias from omitting X_i from the regression. As a result, combining regression and weighting can lead to additional robustness by both removing the correlation between the omitted covariates, and by reducing the correlation between the omitted and included variables. This is the idea behind the doubly-robust estimators developed in Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995), and Van Der Laan and Robins (2003).

Suppose we model the two regression functions as $\mu_w(x) = \alpha_w + \beta_w'(x - \overline{X})$, for $w = 0, 1$ (where we abuse notation a bit and insert the sample averages of the covariates for their population means). More generally, we may use a nonlinear model for the conditional expectation, or just a more flexible linear approximation. Suppose we model the propensity score as $e(x) = p(x;\gamma)$, for example as $p(x;\gamma) = \exp(\gamma_0 + x'\gamma_1)/(1 + \exp(\gamma_0 + x'\gamma_1))$. In the first step we estimate γ by maximum likelihood and obtain the estimated propensity scores as $\hat{e}(X_i) = p(x;\hat{\gamma})$. In the second step, we use linear regression, where we weight the objective function by the inverse probability of treatment or non-treatment. Specifically, to estimate (α_0, β_0) and (α_1, β_1), we would solve the weighted least squares problems

$$\min_{\alpha_0,\beta_0} \sum_{i:W_i=0} \frac{\big(Y_i - \alpha_0 - \beta_0'(X_i - \overline{X})\big)^2}{p(X_i;\hat{\gamma}))}, \qquad \text{and} \quad \min_{\alpha_1,\beta_1} \sum_{i:W_i=1} \frac{\big(Y_i - \alpha_1 - \beta_1'(X_i - \overline{X})\big)^2}{1 - p(X_i;\hat{\gamma})}. \tag{20}$$

Given the estimated conditional mean functions, we estimate τ_{pate}, using the expression for $\hat{\tau}_{\text{reg}} = \hat{\alpha}_1 - \hat{\alpha}_0$ as in equation (13). But what is the motivation for weighting by the inverse propensity score when we did not use such weighting in Section 5.3? The motivation is the double robustness result due to Robins and Rotznitzky (1995); see also Scharfstein, Rotnitzky, and Robins (1999).

First, suppose that the conditional expectation is indeed linear, or $\mathbb{E}[Y_i(w)|X_i = x] = \alpha_w + \beta_w'(x - \overline{X})$. Then, as discussed in the treatment effect context by Wooldridge (2007), weighting the objective function by any nonnegative function of X_i does not affect consistency of least squares.[6] As a result, even if the logit model for the propensity score is misspecified, the binary response MLE $\hat{\gamma}$ still has a well-defined probability limit, say γ^*, and the IPW estimator that uses weights $1/p(X_i;\hat{\gamma})$ for treated observations and $1/(1-p(X_i;\hat{\gamma}))$ for control observations is asymptotically equivalent to the estimator that uses weights based on γ^*.[7] It does not matter that for some x, $e(x) \neq p(x;\gamma^*)$. This is the first part of the double robustness result: if the parametric conditional means for $\mathbb{E}[Y(w)|X = x]$ are correctly specified, the model for the propensity score can be arbitrarily misspecified for the true propensity score. Equation (20) still leads to a consistent estimator for τ_{pate}.

When the conditional means are correctly specified, weighting will generally hurt in terms of asymptotic efficiency. The optimal weight is the inverse of the variance, and in general there is

[6] More generally, it does not affect the consistency of any quasi-likelihood method that is robust for estimating the parameters of the conditional mean. These are likelihoods in the linear exponential family, as described in Gourieroux, Monfort and Trognon (1984).

[7] See Wooldridge (2007).

no reason to expect that weighting the inverse of (one minus) the propensity score gives a good approximation to that. Specifically, under homoskedasticity of $Y_i(w)$ so that $\sigma_w^2 = \sigma_w^2(x)$, in the context of least squares – the IPW estimator of (α_w, β_w) is less efficient than the unweighted estimator; see Wooldridge (2007). The motivation for propensity score weighting is different: it offers a robustness advantage for estimating τ_{pate}.

The second part of the double robustness result assumes that the logit model (or an alternative binary response model) is correctly specified for the propensity score, so that $e(x) = p(x;\gamma^*)$, but allows the conditional mean functions to be misspecified. The result is that in that case $\hat{\alpha}_w \to \mathbb{E}[Y_i(w)]$, and thus $\hat{\tau} = \hat{\alpha}_1 - \hat{\alpha}_0 \to \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \tau_{\text{pate}}$ and the estimator is still consistent. Let the weight for control observations be $\lambda_i = (1 - p(X_i;\gamma^*))^{-1} / \sum_{j:W_j=0}(1 - p(X_i;\gamma^*))^{-1}$. Then the least squares estimator for $\hat{\alpha}_0$ is

$$\hat{\alpha}_0 = \sum_{i=1}^{N}(1 - W_i) \cdot \lambda_i \cdot \left(Y_i - \hat{\beta}_0'(X_i - \overline{X})\right). \tag{21}$$

The weights imply that $\mathbb{E}\left[(1 - W_i)\lambda_i Y_i\right] = \mathbb{E}[Y_i(0)]$ and $\mathbb{E}\left[(1 - W_i)\lambda_i(X_i - \overline{X})\right] = \mathbb{E}[X_i - \overline{X}] = 0$, and as a result $\hat{\alpha}_0 \to \mathbb{E}[Y_i(0)]$. Similarly, the average of the predicted values for $Y_i(1)$ converges to $\mathbb{E}[Y_i(1)]$, and so the resulting estimator $\hat{\tau}_{\text{ipw}} = \hat{\alpha}_1 - \hat{\alpha}_0$ is consistent for τ_{pate} and τ_{cate} irrespective of the shape of the regression functions. This is the second part of the double robustness part, at least for linear regression.

For certain kinds of responses, including binary responses, fractional responses, and count responses, linearity of $\mathbb{E}[Y_i(w)|X_i = x]$ is a poor assumption. Using linear conditional expectations for limited dependent variables effectively abdicates the first part of the double robustness result. Instead, we should use coherent models of the conditional means, as well as a sensible model for the propensity score, with the hope that the mean functions, propensity score, or both are correctly specified. Beyond specifying logically coherent for $\mathbb{E}[Y_i(w)|X_i = x]$ so that the first part of double robustness has a chance, for the second part we need to choose functional forms and estimators with the following property: even when the mean functions are misspecified, $\mathbb{E}[Y_i(w)] = \mathbb{E}[\mu(X_i, \delta_w^*)]$, where δ_w^* is the probability limit of $\hat{\delta}_w$. Fortunately, for the common kinds of limited dependent variables used in applications, such functional forms and estimators exist; see Wooldridge (2007) for further discussion.

Once we estimate τ based on (20), how should we obtain a standard error? The normalized variance still has the form $V_0 + V_1$, where $V_w = \mathbb{E}[(\hat{\alpha}_w - \mu_w)^2]$. One option is to exploit the representation of $\hat{\alpha}_0$ as a weighted average of $Y_i + \hat{\beta}_0(X_i - \overline{X})$, and use the naive variance estimator based on weighted least squares with known weights:

$$\hat{V}_0 = \sum_{i:W_i=0} \lambda_i^2 \cdot \left(Y_i + \hat{\beta}_0'(X_i - \overline{X}) - \hat{\alpha}_0\right)^2, \tag{22}$$

and similar for V_1. In generally we may again want to adjust for the estimation of the parameters in γ. See Wooldridge (2007) for details.

Although combining weighting and regression is more attractive then either weighting or regression on their own, it still requires at least one of the two specifications to be accurate

globally. It has been used regularly in the epidemiology literature, partly through the efforts of Robins and his coauthors, but has not been widely used in the economics literature.

5.7 Subclassification and Regression

We can also combine subclassification with regression. The advantage relative to weighting and regression is that we do not use global approximations to the regression function. The idea is that within stratum j, we estimate the average treatment effect by regressing the outcome on a constant, an indicator for the treatment, and the covariates, instead of simply taking the difference in averages by treatment status as in Section 5.4. The latter can be viewed as a regression estimate based on a regression with only an intercept and the treatment indicator. The further regression adjustment simply adds (some of) the covariates to that regression. The key difference with using regression in the full sample is that, within a stratum, the propensity score varies relatively little. As a result, the covariate distributions are similar, and the regression function is not used to extrapolate far out of sample.

To be precise, we estimate on the observations with $B_{ij} = 1$, the regression function

$$Y_i = \alpha_j + \tau_j \cdot W_i + \beta_j' X_i + \varepsilon_i,$$

by least squares, obtaining the estimates $\hat{\tau}_j$ and estimated variances $\hat{V}_j$. Dropping X_i from this regression leads to $\hat{\tau}_j = \overline{Y}_{j1} - \overline{Y}_{j0}$, which is the blocking estimator we discussed in Section 5.4 We average the estimated stratum-specific average treatment effects, weighted by the relative stratum size:

$$\hat{\tau} = \sum_{j=1}^{J} \left(\frac{N_{j0} + N_{j1}}{N} \right) \cdot \hat{\tau}_j, \qquad \text{with estimated variance} \quad \hat{V} = \sum_{j=1}^{J} \left(\frac{N_{j0} + N_{j1}}{N} \right)^2 \cdot \hat{V}_j.$$

With a modest number of strata, this already leads to an estimator that is considerably more flexible and robust than either subclassification alone, or regression alone. It is probably one of the more attractive estimators in practice. Imbens and Rubin (2008) suggest data-dependent methods for choosing the number of strata.

5.8 Matching and Regression

Once we have the N pairs $(\hat{Y}_i(0), \hat{Y}_i(1))$, the simple matching estimator given in (19) averages the difference. This estimator may still be biased due to discrepancies between the covariates of the matched observations and their matches. One can attempt to reduce this bias by using regression methods. This use of regression is very different from using regression methods on the full sample. Here the covariate distributions are likely to be similar in the matched sample, and so regression is not used to extrapolate far out of sample.

The idea behind the regression adjustment is to replace $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ by

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if} \quad W_i = 0, \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} \left(Y_j + \beta_0'(X_i - X_j) \right) & \text{if} \quad W_i = 1, \end{cases}$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\left(Y_j + \beta_1(X_i - X_j)\right) & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases}$$

where the average of the matched outcomes is adjusted by the difference in covariates relative to the matched observation. The only question left is how to estimate the regression coefficients β_0 and β_1. For various methods see Quade (1982), Rubin (1979), and Abadie and Imbens (2006). The methods differ in whether the difference in outcomes is modeled as linear in the difference in covariates, or the original conditional outcome distributions are approximated by linear regression functions, and on what sample the regression functions are estimated.

Here is one simple regression adjustment. To be clear, it is useful to introduce some additional notation. Given the set of matching indices $\mathcal{J}_M(i)$, define

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} X_j & \text{if } W_i = 1, \end{cases} \qquad \hat{X}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} X_j & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1, \end{cases}$$

and let $\hat{\beta}_w$ be based on a regression of $\hat{Y}_i(w)$ on a constant and $\hat{X}_i(w)$:

$$\begin{pmatrix} \hat{\alpha}_w \\ \hat{\beta}_w \end{pmatrix} = \left(\sum_{i=1}^{N} \begin{pmatrix} 1 & \hat{X}_i(w)' \\ \hat{X}_i(w) & \hat{X}_i(w)\hat{X}_i(w)' \end{pmatrix}\right)^{-1} \begin{pmatrix} \hat{Y}_i(w) \\ \hat{X}_i(w)\hat{Y}_i(w) \end{pmatrix}.$$

Like the combination of subclassification and regression, this leads to relatively robust estimators. Abadie and Imbens (2008) find that the method works well in simulations based on the Lalonde data.

5.9 A General Method for Estimating Variances

For some of the estimators discussed in the previous sections, particular variance estimators have been used. Assuming that a particular parametric model is valid, one can typically use standard methods based on likelihood theory or generalized method of moments theory. Often, these methods rely on consistent estimation of components of the variance. Here we discuss two general methods for estimating variances that apply to all estimators.

The first approach is to use bootstrapping (Efron and Tibshirani, 1993; Davison and Hinkley, 1997; Horowitz, 2002). Bootstrapping has been widely used in the treatment effects literature, as it is straightforward to implement. It has rarely been formally justified, although in many cases it is likely to be valid given that many of the estimators are asymptotically linear. However, in some cases it is known that bootstrapping is not valid. Abadie and Imbens (2008) show that, for a fixed number of matches, bootstrapping is not valid for matching estimators. It is likely that the problems that invalidate the bootstrap disappear if the number of matches increases with the sample size (thus, the bootstrap might be valid for kernel estimators). Nevertheless, because in practice researchers often use a small number of matches, or nonnegative kernels, it is not clear whether the bootstrap is an effective method for obtaining standard errors and constructing confidence intervals. In cases where bootstrapping is not valid, often subsampling (Politis and Romano, 1999) remains valid, but this has not been applied in practice.

There is an alternative, general, method for estimating variances of treatment effect estimators, developed by Abadie and Imbens (2006), that does not require additional nonparametric

estimation. First, recall that most estimators are of the form

$$\hat{\tau} = \sum_{i=1}^{N} \lambda_i \cdot Y_i, \qquad \text{with} \sum_{i:W_i=1} \lambda_i = 1, \qquad \sum_{i:W_i=0} \lambda_i = -1,$$

with the weights λ_i generally functions of all covariates and all treatment indicators. Conditional on the covariates and the treatment indicators (and thus relative to τ_{cate}), the variance of such an estimator is

$$\mathbb{V}(\hat{\tau}|X_1, \ldots, X_N, W_1, \ldots, W_N) = \sum_{i=1}^{N} \lambda_i^2 \cdot \sigma^2_{W_i}(X_i).$$

In order to use this representation, we need estimates of $\sigma^2_{W_i}(X_i)$, for all i. Fortunately, these need not be consistent estimates, as long as the estimation errors are not too highly correlated so that the weighted average of the estimates is consistent for the weighted average of the variances. This is similar in the way robust (Huber-Eicker-White) standard errors allow for general forms of heteroskedasticity without having to consistently estimate the conditional variance function.

Abadie and Imbens (2006) suggested using a matching estimator for $\sigma^2_{W_i}(X_i)$. The idea behind this matching variance estimator is that if we can find two treated units with $X_i = x$, we can estimate $\sigma_1^2(x)$ as $\hat{\sigma}_1^2(x) = (Y_i - Y_j)^2/2$. In general, it is difficult to find exact matches, but, again, this is not necessary. Instead, one uses the closest match within the set of units with the same treatment status. Let $\nu(i)$ be the unit closest to i, with the same treatment indicator ($W_{\nu(i)} = W_i$), so that

$$\left\|X_{\nu(i)} - X_i\right\| = \min_{j:W_j=i} \|X_j - X_i\|.$$

Then we can estimate $\sigma^2_{W_i}(X_i)$ as

$$\hat{\sigma}^2_{W_i}(X_i) = \left(Y_i - Y_{\nu(i)}\right)^2/2.$$

This way we can estimate $\sigma^2_{W_i}(X_i)$ for all units. Note that these are not consistent estimators of the conditional variances. As the sample size increases, the bias of these estimators will disappear, just as we saw that the bias of the matching estimator for the average treatment effect disappears under similar conditions.

We then use these estimates of the conditional variance to estimate the variance of the estimator:

$$\hat{V}(\hat{\tau}) = \sum_{i=1}^{N} \lambda_i^2 \cdot \hat{\sigma}^2_{W_i}(X_i).$$

5.10 Overlap in Covariate Distributions

In practice a major concern in applying methods under the assumption of unconfoundedness is lack of overlap in the covariate distributions. In fact, once one is committed to the unconfoundedness assumption, this may well be the main problem facing the analyst. The overlap issue was

highlighted in papers by Dehejia and Wahba (1999) and Heckman, Ichimura, and Todd (1998). Dehejia and Wahba re-analyzed data on a job training program originally analyzed by Lalonde (1986). Lalonde (1986) had attempted to replicate results from an experimental evaluation of a job training program, the National Supported Work (NSW) program, using a comparison group constructed from two public use data sets, the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS). The NSW program targeted individuals who were disadvantaged with very poor labor market histories. As a result, they were very different from the raw comparison groups constructed by Lalonde from the CPS and PSID. Lalonde partially addressed this problem by limiting his raw comparison samples based on single covariate criteria (e.g., limiting it to individuals with zero earnings in the year prior to the program). Dehejia and Wahba looked at this problem more systematically and find that a major concern is the lack of overlap in the covariate distributions.

Traditionally, overlap in the covariate distributions was assessed by looking at summary statistics of the covariate distributions by treatment status. As discussed before in the introduction to Section 5, it is particularly useful to report differences in average covariates normalized by the square root of the sum of the within-treatment group variances. In Table 2 we report, for the Lalonde data, averages and standard deviations of the basic covariates, and the normalized difference. For four out of the ten covariates the means are more than a standard deviation apart. This immediately suggests that the technical task of adjusting for differences in the covariates is a challenging one. Although reporting normalized differences in covariates by treatment status is a sensible starting point, inspecting differences one covariate at a time is not generally sufficient. Even if all these differences are small, there may still be areas with limited overlap. Formally, we are concerned with regions in the covariate space where the density of covariates in one treatment group is zero and the density in the other treatment group is not. This corresponds to the propensity score being equal to zero or one. Therefore, a more direct way of assessing the overlap in covariate distributions is to inspect histograms of the estimated propensity score by treatment status.

Once it has been established that overlap is a concern, several strategies can be used. We briefly discuss two of the earlier specific suggestions, and then describe in more detail two general methods. In practice, researchers have often simply dropped observations with propensity score close to zero or one, with the actual cutoff value chosen in an *ad hoc* fashion. Dehejia and Wahba (1999) focus on the average effect for the treated. After estimating the propensity score, they find the smallest value of the estimated propensity score among the treated units, $\underline{e}_1 = \min_{i:W_i=1} \hat{e}(X_i)$. They then drop all control units with an estimated propensity score lower than this threshold $\underline{e}_1$. The idea behind this suggestion is that control units with very low values for the propensity score may be so different from treated units that including them in the analysis is likely to be counterproductive. (In effect, the population over which the treatment effects are calculated is redefined.) A concern is that the results may be sensitive to the choice of specific threshold $\underline{e}_1$. If, for example, one used as the threshold the K-th order statistic of the estimated propensity score among the treated (Lechner, 2002ab), the results might change considerably. In the sixth column of Table 2 we report the normalized difference (normalized using the same denominator equal to the square root of the sum of the within treatment

group sample variances) after removing 9891 (out of a total 16177) control observations whose estimated propensity score was smaller than the smallest value of the estimated propensity score among the treated, $\underline{e}_1 = 0.00051$. This improves the covariate balance, but many of the normalized differences are still substantial.

Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998) develop a different method. They focus on estimation of the set where the density of the propensity score conditional on the treatment is bounded away from zero for both treatment regimes. Specifically, they first estimate the density functions $f(e|W = w)$, for $w = 0, 1$, nonparametrically. They then evaluate the estimated density $\hat{f}(\hat{e}(X_i)|W_i = 0)$ for all N values X_i, and the same for the estimated density $\hat{f}(\hat{e}(X_i)|W_i = 1)$ for all N values X_i. Given these $2N$ values they calculate the $2N \cdot q$ order statistic of these $2N$ estimated densities. Denote this order statistic by $\hat{f}_q$. Then, for each unit i, they compare the estimated density $\hat{f}(\hat{e}(X_i)|W_i = 0)$ to $\hat{f}_q$, and $\hat{f}(\hat{e}(X_i)|W_i = 1)$ to $\hat{f}_q$. If either of those estimated densities is below the order statistic, the observation gets dropped from the analysis. Smith and Todd (2005) implement this method with $q = 0.02$, but provide no motivation for the choice of the threshold.

5.10.1 Matching to Improve Overlap in Covariate Distributions

A systematic method for dropping control units who are different from the treated units is to construct a matched sample. This approach has been pushed by Rubin in a series of studies, see Rubin (2006). It is designed for settings where the interest is in the average effect for the treated (e.g., as in the Lalonde application). It relies on the control sample being larger than the treated sample, and works especially well when the control sample is much larger.

First, the treated observations are ordered, typically by decreasing values of the estimated propensity score, since treated observations with high values of the propensity score are generally more difficult to match. Then the first treated unit (e.g., the one with the highest value for the estimated propensity score) is matched to the nearest control unit. Next, the second treated unit is matched to the nearest control unit, excluding the control unit that was used as a match for the first treated unit. Matching without replacement all treated units in this manner leads to a sample of $2 \cdot N_1$ units, (where N_1 is the size of the original treated subsample), half of them treated and half of them control units. Note that the matching is not necessarily used here as the final analysis. We do not propose to estimate the average treatment effect for the treated by averaging the differences within the pairs. Instead, this is intended as a preliminary analysis, with the goal being the construction of a sample with more overlap. Given a more balanced sample, one can use any of the previously discussed methods for estimating the average effect of the treatment, including regression, propensity score methods, or matching. Using those methods on the balanced sample is likely to reduce bias relative to using the simple difference in averages by treatment status.

The last two columns in Table 2 report the balance in the ten covariates after constructing a matched sample in this fashion. In both cases the treated units were matched in reverse order of the estimated propensity score. The seventh column is based on matching on the estimated propensity score, and the last column is based on matching on all the covariates, using the Mahalanobis metric (the inverse of the covariance matrix of the covariates). Matching, either

on the estimated propensity score or on the full set of covariates dramatically improves the balance. Whereas before some of the covariates differed by as much as 1.7 times a standard deviation, now the normalized differences are all less than one tenth of a standard deviation. The remaining differences are not negligible, however. For example, average differences in 1974 earnings are still on the order of $700, which, given the experimental estimate from the Lalonde (1986) paper of about $2000 is substantial. As a result, simple estimators such as the average of the within-matched-pair differences are not likely to lead to credible estimates. Nevertheless, maintaining unconfoundedness, this matched sample is sufficiently well balanced that one may be able to obtain credible and robust estimates from it in way that the original sample would not allow.

5.10.2 Trimming to Improve Overlap in Covariate Distributions

Matching with replacement does not work if the estimand of interest is the overall average treatment effect. For that case Crump, Hotz, Imbens and Mitnik (2008a) suggest an easily implementable way of selecting the subpopulation with overlap, consistent with the current practice of dropping observations with propensity score values close to zero or one. Their method is generally applicable and in particular does not require that the control sample is larger than the treated sample. They consider estimation of the average treatment effect for the subpopulation with $X_i \in \mathbb{A}$. They suggest choosing the set $\mathbb{A}$ from the set of all subsets of the covariate space to minimize the asymptotic variance of the efficient estimator of the average treatment effect for that set. Under some conditions (in particular homoskedasticity), they show that the optimal set $\mathbb{A}^*$ depends only on the value of the propensity score. This method suggests discarding observations with a propensity score less than α away from the two extremes, zero and one:

$$\mathbb{A}^* = \Big\{ x \in \mathbb{X} | \alpha \leq e(x) \leq 1-\alpha \Big\},$$

where α satisifies a condition based on the marginal distribution of the propensity score:

$$\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{1}{e(X)\cdot(1-e(X))} \middle| \frac{1}{e(X)\cdot(1-e(X))} < \frac{1}{\alpha\cdot(1-\alpha)} \right].$$

Based on empirical examples and numerical calculations with Beta distributions for the propensity score, Crump, Hotz, Imbens and Mitnik (2008a) suggest that the rule-of-thumb fixing α at 0.10 gives good results.

To illustrate this method, Table 3 presents summary statistics for data from Imbens, Rubin and Sacerdote (2001) on lottery players, including "winners" who won big prizes, and "losers" who did not. Even though winning the lottery is obviously random, variation in the number of tickets bought, and nonresponse, creates imbalances in the covariate distributions. In the full sample (sample size $N = 496$), some of the covariates differ by as much as 0.64 standard deviations. Following the Crump-Hotz-Imbens-Mitnik calculations leads to a bound of 0.0914. Discarding the observations with an estimated propensity score outside the interval $[0.0914, 0.9086]$ leads to a sample size 388. In this subsample the largest normalized difference

is 0.35, about half of what it is in the full sample, with this improvement obtained by dropping approximately 20% of the original sample.

A potentially controversial feature of all these methods is that they change what is being estimated. Instead of estimating τ_{pate}, the Crump, Hotz, Imbens and Mitnik (2008a) approach estimates $\tau_{\text{cate},\mathbb{A}}$. This results in reduced external validity, but it is likely to improve internal validity.

5.11 Assessing the Unconfoundedness Assumption

The unconfoundedness assumption used in Section 5 is not testable. It states that the conditional distribution of the outcome under the control given receipt of the active treatment and covariates, is identical to the distribution of the control outcome conditional on being in the control and covariates. A similar assumption is made for the distribution of the treatment outcome. Yet since the data are completely uninformative about the distribution of $Y_i(0)$ for those who received the active treatment and of $Y_i(1)$ for those receiving the control, the data can never reject the unconfoundedness assumption. Nevertheless, there are often indirect ways of assessing this assumption. The most important of these were developed in Rosenbaum (1987) and Heckman and Hotz (1989). Both methods rely on testing the null hypothesis that an average causal effect is zero, where the particular average causal effect is known to equal zero. If the testing procedure rejects the null hypothesis, this is interpreted as weakening the support for the unconfoundedness assumption. These tests can be divided into two groups.

The first set of tests focuses on estimating the causal effect of a treatment that is known not to have an effect. It relies on the presence of two or more control groups (Rosenbaum, 1987). Suppose one has two potential control groups, for example eligible nonparticipants and ineligibles, as in Heckman, Ichimura and Todd (1997). One can estimate a "pseudo" average treatment effect by analyzing the data from these two control groups as if one of them is the treatment group. In that case the treatment effect is known to be zero, and statistical evidence of a non-zero effect implies that at least one of the control groups is invalid. Again, not rejecting the test does not imply the unconfoundedness assumption is valid (as both control groups could suffer the same bias), but non-rejection in the case where the two control groups could potentially have different biases makes it more plausible that the unconfoundness assumption holds. The key for the power of this test is to have available control groups that are likely to have different biases, if they have any at all. Comparing ineligibles and eligible nonparticipants as in Heckman, Ichimura and Todd (1997) is a particularly attractive comparison. Alternatively one may use geographically distinct comparison groups, for example from areas bordering on different sides of the treatment group.

To be more specific, let G_i be an indicator variable denoting the membership of the group, taking on three values, $G_i \in \{-1, 0, 1\}$. For units with $G_i = -1$ or 0, the treatment indicator W_i is equal to 0:

$$W_i = \begin{cases} 0 & \text{if } G_i = -1, 0, \\ 1 & \text{if } G_i = 1. \end{cases}$$

Unconfoundedness only requires that

$$Y_i(0), Y_i(1) \;\perp\!\!\!\perp\; W_i \;\Big|\; X_i, \tag{23}$$

and this is not testable. Instead we focus on testing an implication of the stronger conditional independence relation

$$Y_i(0), Y_i(1) \;\perp\!\!\!\perp\; G_i \;\Big|\; X_i. \tag{24}$$

This independence condition implies (23), but in contrast to that assumption, it also implies testable restrictions. In particular, we focus on the implication that

$$Y_i(0) \;\perp\!\!\!\perp\; G_i \;\Big|\; X_i, G_i \in \{-1, 0\} \quad\Longleftrightarrow\quad Y_i \;\perp\!\!\!\perp\; G_i \;\Big|\; X_i, G_i \in \{-1, 0\}, \tag{25}$$

because $G_i \in \{-1, 0\}$ implies that $Y_i = Y_i(0)$.

Because condition (24) is slightly stronger than unconfoundedness, the question is whether there are interesting settings where the weaker condition of unconfoundedness holds, but not the stronger condition. To discuss this question, it is useful to consider two alternative conditional independence conditions, both of which are implied by (24):

$$\Big(Y_i(0), Y_i(1)\Big) \;\perp\!\!\!\perp\; W_i \;\Big|\; X_i, G_i \in \{-1, 1\}, \tag{26}$$

and

$$\Big(Y_i(0), Y_i(1)\Big) \;\perp\!\!\!\perp\; W_i \;\Big|\; X_i, G_i \in \{0, 1\}. \tag{27}$$

If (26) holds, then we can estimate the average causal effect by invoking the unconfoundedness assumption using only the first control group. Similarly, if (27) holds, then we can estimate the average causal effect by invoking the unconfoundedness assumption using only the second control group. The point is that it is difficult to envision a situation where unconfoundedness based on the two comparison groups holds – that is, (23) holds – but it does not hold using only one of the two comparison groups at the time. In practice, it seems likely that if unconfoundedness holds then so would the stronger condition (24), and we have the testable implication (25).

Next, we turn to implementation of the tests. We can simply test whether there is a difference in average values of Y_i between the two control groups, after adjusting for differences in X_i. That is, we effectively test whether

$$\mathbb{E}\Big[\mathbb{E}\left[Y_i \middle| G_i = -1, X_i\right] - \mathbb{E}\left[Y_i \middle| G_i = 0, X_i\right]\Big] = 0.$$

More generally we may wish to test

$$\mathbb{E}\Big[\mathbb{E}\left[Y_i \middle| G_i = -1, X_i = x\right] - \mathbb{E}\left[Y_i \middle| G_i = 0, X_i = x\right]\Big] = 0$$

for all x in the support of X_i, using the methods discussed in Crump, Hotz, Imbens adn Mitnik (2008b). We can also include transformations of the basic outcomes in the procedure to test for difference in other aspects of the conditional distributions.

A second set of tests of unconfoundedness focuses on estimating the causal effect of the treatment on a variable known to be unaffected by it, typically because its value is determined prior to the treatment itself. Such a variable can be time-invariant, but the most interesting case is in considering the treatment effect on a lagged outcome. If it is not zero, this implies that the treated observations are distinct from the controls; namely that the distribution of $Y_i(0)$ for the treated units is not comparable to the distribution of $Y_i(0)$ for the controls. If the treatment is instead zero, it is more plausible that the unconfoundedness assumption holds. Of course this does not directly test the unconfoundedness assumption; in this setting, being able to reject the null of no effect does not directly reflect on the hypothesis of interest, unconfoundedness. Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the test arguably has more power. For these tests it is clearly helpful to have a number of lagged outcomes.

First partition the vector of covariates X_i into two parts, a (scalar) pseudo outcome, denoted by $X_i^{\rm p}$, and the remainder, denoted by $X_i^{\rm r}$, so that $X_i = (X_i^{\rm p}, X_i^{\rm r\prime})'$. Now we will assess whether the following conditional independence relation holds:

$$X_i^{\rm p} \quad \perp\!\!\!\perp \quad W_i \quad \Big| \quad X_i^{\rm r}. \tag{28}$$

The two issues are, first, the interpretation of this condition and its relationship to the unconfoundedness assumption, and second, the implementation of the test.

The first issue concerns the link between the conditional independence relation in (28) and original unconfoundedness. This link, by necessity, is indirect, as unconfoundedness cannot be tested directly. Here we lay out the arguments for the connection. First consider a related condition:

$$Y_i(0), Y_i(1) \quad \perp\!\!\!\perp \quad W_i \quad \Big| \quad X_i^{\rm r}. \tag{29}$$

If this modified unconfoundedness condition were to hold, one could use the adjustment methods using only the subset of covariates $X_i^{\rm r}$. In practice, though not necesarily, this is a stronger condition than the original unconfoundedness condition which requires conditioning on the full vector X_i. One has to be careful here, because it is theoretically possible that conditional on a subset of the covariates unconfoundedness holds, but at the same time unconfoundedness does not hold conditional on the full set of covariates. In practice this situation is rare though. For example, it is difficult to imagine in an evaluation of a labor market program that unconfoundedness would hold given age and the level of education, but not if one additionally conditions on gender. Generally making subpopulations more homogenous in pretreatment variables tends to improve the plausibility of unconfoundedness.

The modified unconfoundedness condition (29) is not testable for the same reasons the original unconfoundedness assumption is not testable. Nevertheless, if one has a proxy for either of the potential outcomes, and in particular a proxy that is observed irrespective of

the treatment status, one can test independence for that proxy variable. We use the pseudo outcome $X_i^{\rm p}$ as such a proxy variable. That is, we view $X_i^{\rm p}$ as a proxy for, say, $Y_i(0)$, and assess (29) by testing (28).

The most convincing applications of these assessments are settings where the two links are plausible. One of the leading examples is where X_i contains multiple lagged measures of the outcome. For example, in the evaluation of the effect of a labor market program on annual earnings, one might have observations on earnings for, say, six years prior to the program. Denote these lagged outcomes by $Y_{i,-1},\ldots,Y_{i,-6}$, where $Y_{i,-1}$ is the most recent and $Y_{i,-6}$ is the most distant pre-program earnings measure. One could implement the above ideas using earnings for the most recent pre-program year $Y_{i,-1}$ as the pseudo outcome $X_i^{\rm p}$, so that the vector of remaining pretreatment variables $X_i^{\rm r}$ would still include the five prior years of pre-program earnings $Y_{i,-2},\ldots,Y_{i,-6}$ (ignoring additional pre-treatment variables). In that case one might reasonably argue that if unconfoundedness holds given six years of pre-program earnings, it is plausible that it would also hold given only five years of pre-program earnings. Moreoever, under unconfoundedness $Y_i(c)$ is independent of W_i given $Y_{i,-1},\ldots,Y_{i,-6}$, which would suggest that it is plausible that $Y_{i,-1}$ is independent of W_i given $Y_{i,-2},\ldots,Y_{i,-6}$. Given those arguments, one can plausibly assess unconfoundedness by testing whether

$$Y_{i,-1} \perp\!\!\!\perp W_i \,\Big|\, Y_{i,-2},\ldots,Y_{i,-6}.$$

The implementation of the tests is the same as in the first set of tests for assessing unconfoundedness. We can simply test whether estimates of the average difference between the groups adjusted for differences in $X_i^{\rm r}$ are zero, or test whether the average difference is zero for all values of the covariates (e.g., Crump, Hotz, Imbens and Mitnik, 2008b).

5.12 Testing

Most of the focus in the evaluation literature has been on estimating average treatment effects. Testing has largely been limited to the null hypothesis that the average effect is zero. In that case testing is straightforward since many estimators exist for the average treatment effect that are approximately normally distributed in large samples with zero asymptotic bias. In addition there is some testing based on the Fisher approach using the randomization distribution. In many cases, however, there are other null hypotheses of interest. Crump, Hotz, Imbens and Mitnik (2008b) develop tests of the null hypotheses of zero average effects conditional on the covariates, and of a constant average effect conditional on the covariates. Formally, in the first case the null hypothesis

$$H_0:\ \tau(x)=0,\ \forall\ x, \tag{30}$$

against the alternative hypothesis

$$H_a:\ \tau(x)\neq 0,\ \text{for some } x.$$

Recall that $\tau(x)=\mathbb{E}[Y_i(1)-Y_i(0)|X_i=x]$ is the average effect for the subpopulation with covariate value x. The second hypothesis studied by Crump, Hotz, Imbens and Mitnik is

$$H_0:\ \tau(x)=\tau_{\rm pate},\ \forall\ x, \tag{31}$$

against the alternative hypothesis

$$H_a: \ \tau(x) \neq \tau_{\text{pate}}, \ \text{for some } x.$$

Part of their motivation is that in many cases there is substantive interest in whether the program is beneficial for some groups, even if on average it does not affect outcomes.[8] They show that in some data sets they reject the null hypothesis (30) even though they cannot reject the null hypothesis of a zero average effect.

Taking the motivation in Crump, Hotz, Imbens and Mitnik (2008b) one step further, one may also be interested in testing the null hypothesis that the conditional distribution of $Y_i(0)$ given $X_i = x$ is the same as the conditional distribution of $Y_i(1)$ given $X_i = x$. Under the maintained hypothesis of unconfoundedness, this is equivalent to testing the null hypothesis that

$$H_0: \ Y_i \perp\!\!\!\perp W_i \ \Big| \ X_i,$$

against the alternative hypothesis that Y_i is not independent of W_i given X_i. Tests of this type can be implemented using the methods of Gozalo and Linton (2003). There have been no applications of these tests in the program evaluation literature.

5.13 Selection of Covariates

A very important set of decisions in implementing all of the methods described in this section involves the choice of covariates to be included in the regression functions or the propensity score. Here the literature has not been very helpful, and as a result researchers have just included all covariates linearly, without much systematic effort to find more compelling specifications. Most of the technical results using nonparametric methods include rates at which the smoothing parameters should change with the sample size. For example, using regression estimators, one would have to choose the bandwidth if using kernel estimators, or the number of terms in the series if using series estimators. The program evaluation literature does not provide much guidance as to how to choose these smoothing parameters in practice. More generally, the nonparametric estimation literature has little to offer in this regard. Most of the results in this literature offer optimal choices for smoothing parameters if the criterion is integrated squared error. In the current setting the interest is in a scalar parameter, and the choice of smoothing parameter that is optimal for the regression function itself need not be close to optimal for the average treatment effect.

Hirano and Imbens (2001) consider an estimator that combines weighting with the propensity score and regression. In their application they have a large number of covariates, and they suggest deciding which ones to include on the basis of t-statistics. They find that the results are fairly insensitive to the actual cutoff point if they use the weight/regression estimator, but find more sensitivity if they only use weighting or regression. They do not provide formal properties for these choices.

[8] A second motivation is that it may be impossible to obtain precise estimates for τ_{pate} even in cases where one can convincingly reject some of the hypotheses regarding $\tau(x)$.

Ichimura and Linton (2005) consider inverse probability weighting estimators and analyze the formal problem of bandwidth selection with the focus on the average treatment effect. Imbens, Newey and Ridder (2006) look at series regression estimators and analyze the choice of the number of terms to be included, again with the objective being the average treatment effect. Imbens and Rubin (2008) discuss some stepwise covariate selection methods for finding a specification for the propensity score.

It is clear that more work needs to be done in this area, both for the case where the choice is which covariates to include from a large set of potential covariates, and in the case where the choice concerns functional form and functions of the a small set of covariates.

6 Selection on Unobservables

In this section we discuss a number of methods that relax the pair of assumptions made in Section 5. Unlike in the setting under unconfoundedness, there is not a unified set of methods for this case. In a number of special cases there are well-understood methods, but there are many cases without clear recommendations. We will highlight some of the controversies and different approaches. First we discuss some methods that simply drop the unconfoundedness assumption. Next, in Section 6.2, we discuss sensivitity analyses that relax the unconfoundedness assumption in a more limited manner. In Section 6.3 we discuss instrumental variables methods. Then, in Section 6.4 we discuss regression discontinuity designs, and in Section 6.5 we discuss difference-in-differences methods.

6.1 Bounds

In a series of papers and books, Manski (1990, 1995, 2003, 2005, 2007) has developed a general framework for inference in settings where the parameters of interest are not identified. Manski's key insight is that even if in large samples one cannot infer the exact value of the parameter, one may be able to rule out some values that one could not rule out *a priori*. Prior to Manski's work, researchers had typically dismissed models that are not point-identified as not useful in practice. This framework is not restricted to causal settings, and the reader is referred to Manski (2007) for a general discussion of the approach. Here we limit the discussion to program evaluation settings.

We start by discussing Manksi's perspective in a very simple case. Suppose we have no covariates and a binary outcome $Y_i \in \{0, 1\}$. Let the goal be inference for the average effect in the population, τ_{pate}. We can decompose the population average treatment effect as

$$\tau_{\text{pate}} = \mathbb{E}[Y_i(1)|W_i = 1] \cdot \text{pr}(W_i = 1) + \mathbb{E}[Y_i(1)|W_i = 0] \cdot \text{pr}(W_i = 0)$$

$$-[\mathbb{E}[Y_i(0)|W_i = 1] \cdot \text{pr}(W_i = 1) + \mathbb{E}[Y_i(0)|W_i = 0] \cdot \text{pr}(W_i = 0)].$$

Of the eight components of this expression, we can estimate six. The data contain no information about the remaining two, $\mathbb{E}[Y_i(1)|W_i = 0]$ and $\mathbb{E}[Y_i(0)|W_i = 1]$. Because the outcome is binary, and before seeing any data, we can deduce that these two conditional expectations

must lie inside the interval $[0, 1]$, but we cannot say any more without additional assumptions. This implies that without additional assumptions we can be sure that

$$\tau_{\text{pate}} \in [\tau_l, \tau_u] \,,$$

where we can express the lower and upper bound in terms of estimable quantities,

$$\tau_l = \mathbb{E}[Y_i(1)|W_i = 1] \cdot \text{pr}(W_i = 1) - \text{pr}(W_i = 1) - \mathbb{E}[Y_i(0)|W_i = 0] \cdot \text{pr}(W_i = 0),$$

and

$$\tau_u = \mathbb{E}[Y_i(1)|W_i = 1] \cdot \text{pr}(W_i = 1) + \text{pr}W_i = 0) - \mathbb{E}[Y_i(0)|W_i = 0] \cdot \text{pr}(W_i = 0).$$

In other words, we can bound the average treatment effect. In this example the bounds are tight, meaning that without additional assumptions we cannot rule out any value inside the bounds. For an empirical example of these particular bounds, see Manski, Sandefur, McLanahan, and Powell (1992).

In this specific case the bounds are not particularly informative. The width of the bounds, the difference in $\tau_u - \tau_l$, with τ_l and τ_u given above, is always equal to one, implying we can never rule out a zero average treatment effect. (In some sense this is obvious: if we refrain from making any assumptions regarding the treatment effects we cannot rule out that the treatment effect is zero for any unit.) In general, however, we can add some assumptions, short of making the type of assumption as strong as unconfoundedness that gets us back to the point-identified case. With such weaker assumptions we may be able to tighten the bounds and obtain informative results, without making the strong assumptions that strain credibility. The presence of covariates increases the scope for additional assumptions that may tighten the bounds. Examples of such assumptions include those in the spirit of instrumental variables, where some covariates are known not to affect the potential outcomes (e.g., Manski, 2007), or monotonicity assumptions where expected outcomes are monotonically related to covariates or treatments (e.g., Manski and Pepper, 2000). For an application of these methods, see Hotz, Mullin, and Sanders (1997). We return to some of these settings in Section 6.3.

This discussion has focused on identification and demonstrated what can be learned in large samples. In practice these bounds need to be estimated which leads to additional uncertainty regarding the estimands. A fast developing literature (e.g., Horowitz and Manski, 2000;Imbens and Manski, 2004; Chernozhukov, Hong and Tamer, 2007; Beresteanu, and Molinari, 2006; Romano and Shaikh, 2006ab; Pakes, Porter, Ho and Ishii, 2006; Rosen, 2007; Andrews and Soares, 2007; Canay, 2007; and Stoye, 2007) discusses construction of confidence intervals in general settings with partial identification. One point of contention in this literature has been whether the focus should be on confidence intervals for the parameter of interest (τ_{pate} in this case), or for the identified set. Imbens and Manski (2004) develop confidence sets for the parameter. In large samples, and at a 95% confidence level the Imbens-Manski confidence intervals amount to taking the lower bound minus 1.645 times the standard error of the lower bound and the upper bound plus 1.645 times its standard error. The reason for using 1.645 rather than 1.96 is to take account of the fact that even in the limit the width of the confidence

set will not shrink to zero, and therefore one only needs to be concerned with one-sided errors. Chernozhukov, Hong, and Tamer (2007) focus on confidence sets that include the entire partially identified set itself with fixed probability. For a given confidence level the latter approach generally leads to larger confidence sets than the Imbens-Manski approach. See also Romano and Shaikh (2006ab) for subsampling approaches to inference in these settings.

6.2 Sensitivity Analysis

Unconfoundedness has traditionally been seen as an all or nothing assumption: either it is satisfied and one proceeds accordingly using the methods appropriate under unconfoundedness, such as matching, or the assumption is deemed implausible and one considers alternative methods. The latter include the bounds approach discussed in Section 6.1, as well as approaches relying on alternative assumptions, such as instrumental variables, which will be discussed in Section 6.3. However, there is an important alternative that has received much less attention in the economics literature. Instead of completely relaxing the unconfoundedness assumption, the idea is to relax it slightly. More specifically, violations of unconfoundedness are interpreted as evidence of the presence of unobserved covariates that are correlated, both with the potential outcomes and with the treatment indicator. The size of bias these violations of unconfoundedness can induce depends on the strength of these correlations. Sensitivity analyses investigate whether results obtained under the maintained assumption of unconfoundedness can be changed substantially, or even overturned entirely, by modest violations of the unconfoundedness assumption.

To be specific, consider a job training program with voluntary enrollment. Suppose that we have monthly labor market histories for a two year period prior to the program. We may be concerned that individuals choosing to enroll in the program are more motivated to find a job than those that choose not to enroll in the program. This unobserved motivation may be related to subsequent earnings both in the presence and in the absence of training. Conditioning on the recent labor market histories of individuals may limit the bias associated with this unobserved motivation, but it need not eliminate it entirely. However, we may be willing to limit how highly correlated unobserved motivation is with the enrollment decision and the earnings outcomes in the two regimes, conditional on the labor market histories. For example, if we compare two individuals with the same labor market history for the last two years, e.g., not employed the last six months and working the eighteen months before, and both with one two-year old child, it may be reasonable to assume that these cannot differ radically in their unobserved motivation given that their recent labor market outcomes have been so similar. The sensitivity analyses developed by Rosenbaum and Rubin (1983) formalize this idea and provides a tool for making such assessments. Imbens (2003) applies this sensitivity analysis to data from labor market training programs. The second approach is associated with work by Rosenbaum (1995). Similar to the Rosenbaum-Rubin approach Roseenbaum's method relies on an unobserved covariate that generates the deviations from unconfoundedness. The analysis differs in that sensitivity is measured using only the relation between the unobserved covariate and the treatment assignment, with the focus on the correlation required to overturn, or change substantially, p-values of statistical tests of no effect of the treatment.

6.2.1 The Rosenbaum-Rubin Approach to Sensitivity Analysis

The starting point is that unconfoundedness is satisfied only conditional on the observed covariates X_i and an unobserved scalar covariate U_i:

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \,\Big|\, X_i, U_i.$$

This set up in itself is not restrictive, although once parametric assumptions are made the assumption of a scalar unobserved covariate U_i is resctritive.

Now consider both the conditional distribution of the potential outcomes given observed and unobserved covariates and the conditional probability of assignment given observed and unobserved covariates. Rather than attempting to estimate both these conditional distributions, the idea behind the sensitivity analysis is to specify the form and the amount of dependence of these conditional distributions on the unobserved covariate, and estimate only the dependence on the observed covariate. Conditional on the specification of the first part estimation of the latter is typically straightforward. The idea is then to vary the amount of dependence of the conditional distributions on the unobserved covariate and assess how much this changes the point estimate of the average treatment effect.

Typically the sensitivity analysis is done in fully parametric settings, although since the models can be arbitrarily flexible, this is not particularly restrictive. Following Rosenbaum and Rubin (1984) we illustrate this approach in a setting with binary outcomes. See Imbens (2003) for an example with continuous outcomes. Rosenbaum and Rubin (1984) fix the marginal distribution of the unobserved covariate to be binomial with $p = \text{pr}(U_i = 1)$, and assume independence of U_i and X_i. They specify a logistic distribution for the treatment assignment:

$$\text{pr}(W_i = 1|X_i = x, U_i = u) = \frac{\exp(\alpha_0 + \alpha_1' x + \alpha_2 \cdot u)}{1 + \exp(\alpha_0 + \alpha_1' x + \alpha_2 \cdot u)}.$$

They also specify logistic regression functions for the two potential outcomes:

$$\text{pr}(Y_i(w) = 1|X_i = x, U_i = u) = \frac{\exp(\beta_{w0} + \beta_{w1}' x + \beta_{w2} \cdot u)}{1 + \exp(\beta_{w0} + \beta_{w1}' x + \beta_{w2} \cdot u)}.$$

For the subpopulation with $X_i = x$ and $U_i = u$, the average treatment effect is

$$\mathbb{E}[Y_i(1) - Y_i(0|X_i = x, U_i = u] = \frac{\exp(\beta_{10} + \beta_{11}' x + \beta_{12} \cdot u)}{1 + \exp(\beta_{10} + \beta_{11}' x + \beta_{12} \cdot u)} - \frac{\exp(\beta_{00} + \beta_{01}' x + \beta_{02} \cdot u)}{1 + \exp(\beta_{00} + \beta_{01}' x + \beta_{02} \cdot u)}.$$

The average treatment effect τ_{cate} can be expressed in terms of the parameters of this model and the distribution of the observable covariates by averaging over X_i, and integrating out the unobserved covariate U:

$$\begin{aligned}\tau &\equiv \tau(p, \alpha_2, \beta_{02}, \beta_{12}, \alpha_0, \alpha_1, \beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}) \\ &= \frac{1}{N}\left\{\sum_{i=1}^{N} p\left(\frac{\exp(\beta_{10} + \beta_{11}' X_i + \beta_{12})}{1 + \exp(\beta_{10} + \beta_{11}' X_i + \beta_{12})} - \frac{\exp(\beta_{00} + \beta_{01}' X_i + \beta_{02})}{1 + \exp(\beta_{00} + \beta_{01}' X_i + \beta_{02})}\right)\right.\end{aligned}$$

$$+(1-p)\left(\frac{\exp(\beta_{10}+\beta_{11}'X_i)}{1+\exp(\beta_{10}+\beta_{11}'X_i)}-\frac{\exp(\beta_{00}+\beta_{01}'X_i)}{1+\exp(\beta_{00}+\beta_{01}'X_i)}\right)\Bigg\}.$$

We do not know the values of the parameters (p, α, β), but the data are somewhat informative about them. One conventional approach would be to attempt to estimate all parameters, and then use those estimates to obtain an estimate for the average treatment effect. Given the specific parametric model this may be possible, although in general this would be difficult given the inclusion of unobserved covariates in the basic model. A second approach, as discussed in Section 6.1, is to derive bounds on τ given the model and the data. A sensivitity analysis offers a third approach.

The Rosenbaum-Rubin sensitivity analysis proceeds by dividing the parameters into two sets. The first set includes the parameters that would be set to boundary values under unconfoundedness, $(\alpha_2, \beta_{02}, \beta_{12})$, plus the parameter p capturing the marginal distribution of the unobserved covariate U_i. Together we refer to these as the sensitivity parameters, $\theta_{\text{sens}} = (p, \alpha_2, \beta_{02}, \beta_{12})$. The second set consists of the remaining parameters, $\theta_{\text{other}} = (\alpha_0, \alpha_1, \beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})$. The idea is that θ_{sens} is difficult to estimate. Estimates of the other parameters under unconfoundedness could be obtained by fixing $\alpha_2 = \beta_{02} = \beta_{12} = 0$ and p at an arbitrary value. The data are not directly informative about the effect of an unobserved covariate in the absence of functional form assumptions, and so attempts to estimate θ_{sens} are therefore unlikely to be effective. Given θ_{sens}, however, estimating the remaining parameters is considerably easier. In the second step the plan is therefore to fix the first set of parameters and estimate the others by maximum likelihood, and then translate this into an estimate for τ. Thus, for fixed θ_{sens}, we first estimate the remaining parameters through maximum likelihood:

$$\hat{\theta}_{\text{other}}(\theta_{\text{sens}}) = \arg\max_{\theta_{\text{other}}} L(\theta_{\text{other}}|\theta_{\text{sens}}),$$

where $L(\cdot)$ is the logarithm of the likelihood function. Then we consider the function

$$\tau(\theta_{\text{sens}}) = \tau(\theta_{\text{sens}}, \hat{\theta}_{\text{other}}(\theta_{\text{sens}})),$$

Finally, in the third step, we consider the range of values of the function $\tau(\theta_{\text{sens}})$ for a reasonable set of values for the sensitivity parameters (θ_{sens}), and obtain a set of values for τ_{cate}.

The key question is how to choose the set of reasonable values for the sensitivity parameters. If we do not wish to restrict this set at all, we end up with unrestricted bounds along the lines of Section 6.1. The power from the sensitivity approach comes from the researcher's willingness to put real limits on the values of the sensitivity parameters $(p, \alpha_2, \beta_{02}, \beta_{12})$. Among these parameters it is difficult to put real limits on p, and typically it is fixed at $1/2$, with little sensitivity to its choice. The more interesting parameters are $(\alpha_2, \beta_{02}, \beta_{12})$. Let us assume that the effect of the unobserved covariate is the same in both treatment arms, $\beta_2 \equiv \beta_{02} = \beta_{21}$, so that there are only two parameters left to fix, α_2 and β_2. Imbens (2003) suggests linking the parameters to the effects of the observed covariates on assignment and potential outcomes. Specifically he suggests to calculate the partial correlations between observed covariates and the treatment and potential outcomes, and then as a benchmark look at the sensitivity to an unobserved covariate that has partial correlations with treatment and potential outcomes as high as any of

the observed covariates. For example, Imbens considers, in the labor market training example, what the effect would be of omitting unobserved motivation, if in fact motivation had as much explanatory power for future earnings and for treatment choice as did earnings in the year prior to the training program. A bounds analysis, in contrast, would implicitly allow unobserved motivation to completely determine both selection into the program and future earnings. Even though putting hard limits on the effect of motivation on earnings and treatment choice may be difficult, it may be reasonable to put some limits on it, and the Rosenbaum-Rubin sensivity analysis provides a useful framework for doing so.

6.2.2 Rosenbaum's Method for Sensitivity Analysis

Rosenbaum (1995) developed a slightly different approach. The advantage of his approach is that it requires fewer tuning parameters than the Rosenbaum-Rubin approach. Specifically, it only requires the researcher to consider the effect unobserved confounders may have on the probability of treatment assignment. Rosenbaum's focus is on the effect the presence of unobserved covariates could have on the p-value for the test of no effect of the treatment based on the unconfoundedness assumption, in contrast to the Rosenbaum-Rubin focus on point estimates for average treatment effects. Consider two units i and j with the same value for the covariates, $x_i = x_j$. If the unconfoundedness assumption conditional on X_i holds, both units must have the same probability of assignment to the treatment, $e(x_i) = e(x_j)$. Now suppose unconfoundedness only holds conditional on both X_i and a binary unobserved covariate U_i. In that case the assignment probabilities for these two units may differ. Rosenbaum suggest bounding the ratio of the odds ratios $e(x_i)/(1-e(x_i))$ and $e(x_j)/(1-e(x_j))$:

$$\frac{1}{\Gamma} \leq \frac{e(x_i)\cdot(1-e(x_j))}{(1-e(x_i))\cdot e(x_j)} \leq \Gamma.$$

If $\Gamma = 1$ we are back in the setting with unconfoundedness. If we allow $\Gamma = \infty$ we are not restricting the association between the treatment indicator and the potential outcomes. Rosenbaum investigates how much the odds would have to be different in order to substantially change the p-value. Or, starting from the other side, he investigates for fixed values of Γ what the implication is on the p-value.

For example, suppose that a test of the null hypothesis of no effect has a p-value of 0.0001 under the assumption of unconfoundedness. If the data suggest it would take the presence of an unobserved covariate that changes the odds of participation by a factor ten in order to increase that p-value to 0.05, then one would likely consider the result to be very robust. If instead a small change in the odds of participation, say with a value of $\Gamma = 1.5$, would be sufficient for a change of the p-value to 0.05, the study would be much less robust.

6.3 Instrumental Variables

In this section we review the recent literature on instrumental variables. We focus on the part of the literature concerned with heterogenous effects. In the current section we limit the discussion to the case with a binary endogenous variable. The early literature focused on

identification of the population average treatment effect and the average effect on the treated. Identification of these estimands ran into serious problems once researchers wished to allow for unrestricted heterogeneity in the effect of the treatment. In an important early result, Bloom (1984) showed that if eligibility for the program is used as an instrument, then one can identify the average effect of the treatment for those who received the treatment. Key for the Bloom result is that the instrument changes the probability of receiving the treatment to zero. In order to identify the average effect on the overall population, the instrument would also need to shift the probability of receiving the treatment to one. This type of identification is sometimes referred to as identification at infinity (Chamberlain, 1986; Heckman, 1990) in settings with a continuous instrument. The practical usefulness of such identification results is fairly limited outside of cases where eligibility is randomized. Finding a credible instrument is typically difficult enough, without also requiring that the instrument shifts the probability of the treatment close to zero and one. In fact, the focus of the current literature on instruments that can credibly be expected to satisfy exclusion restrictions makes it even more difficult to find instruments that even approximately satisfy these support conditions. Imbens and Angrist (1994) got around this problem by changing the focus to average effects for the subpopulation that is affected by the instrument.

Initially we focus on the case with a binary instrument. This case provides some of the clearest insight into the identification problems. In that case the identification at infinity arguments are obviously not satisfied and so one cannot (point-)identify the population average treatment effect.

6.3.1 A Binary Instrument

Imbens and Angrist adopt a potential outcome notation for the receipt of the treatment, as well as for the outcome itself. Let Z_i denote the value of the instrument for individual i. Let $W_i(0)$ and $W_i(1)$ denote the level of the treatment received if the instrument takes on the values 0 and 1 respectively. As before, let $Y_i(0)$ and $Y_i(1)$ denote the potential values for the outcome of interest. The observed treatment is, analogously to the relation between the observed outcome Y_i and the potential outcomes $Y_i(0)$ and $Y_i(1)$, is

$$W_i = W_i(0) \cdot (1 - Z_i) + W_i(1) \cdot Z_i = \begin{cases} W_i(0) & \text{if } Z_i = 0, \\ W_i(1) & \text{if } Z_i = 1. \end{cases}$$

Exogeneity of the instrument is captured by the assumption that all potential outcomes are independent of the instrument, or

$$(Y_i(0), Y_i(1), W_i(0), W_i(1)) \perp Z_i$$

Formulating exogeneity in this way is attractive compared to conventional residual-based definitions, as it does not require the researcher to specify a regression function in order to define the residuals. This assumption captures two properties of the instrument. First, it captures random assignment of the instrument so that causal effects of the instrument on the outcome and treatment received can be estimated consistently. This part of the assumption, which is implied by explicitly randomization of the instrument, as for example in the seminal draft lottery

study by Angrist (1990), is not sufficient for causal interpretations of instrumental variables methods. The second part of the assumption captures an exclusion restriction that there is no direct effect of the instrument on the outcome. This second part is captured by the absence of z in the definition of the potential outcome $Y_i(w)$. This part of the assumption is not implied by randomization of the instrument and it has to be argued on a case by case basis. See Angrist, Imbens and Rubin (1996) for more discussion on the distinction between these two assumptions, and for a formulation that separates them.

Imbens and Angrist introduce a new concept, the compliance type of an individual. The type of an individual describes the level of the treatment that an individual would receive given each value of the instrument. In other words, it is captured by the pair of values $(W_i(0), W_i(1))$. With both the treatment and instrument binary, there are four types of responses for the potential treatment. It is useful to define the compliance types explicitly:

$$T_i = \begin{cases} \text{never}-\text{taker} & \text{if } W_i(0) = W_i(1) = 0, \\ \text{complier} & \text{if } W_i(0) = 0, W_i(1) = 1, \\ \text{defier} & \text{if } W_i(0) = 1, W_i(1) = 0, \\ \text{always}-\text{taker} & \text{if } W_i(0) = W_i(1) = 1. \end{cases}$$

The labels never-taker, complier, defier and always-taker (e.g., Angrist, Imbens and Rubin, 1996) refer to the setting of a randomized experiment with noncompliance, where the instrument is the (random) assignment to the treatment and the endogenous regressor is an indicator for the actual receipt of the treatment. Compliers are in that case individuals who (always) comply with their assignment, that is, take the treatment if assigned to it and not take it if assigned to the control group. One cannot infer from the observed data (Z_i, W_i, Y_i) whether a particular individual is a complier or not. It is important not to confuse compliers (who comply with their actual assignment and would have complied with the alternative assignment) with individuals who are observed to comply with their actual assignment: that is, individuals who complied with the assignment they actually received $Z_i = W_i$. For such individuals we do not know what they would have done had their assignment been different, that is we do not know the value of $W_i(1 - Z_i)$.

Imbens and Angrist then invoke an additional assumption they refer to as *monotonicity*. Monotonicity requires that $W_i(1) \geq W_i(0)$ for all individuals, or that increasing the level of the instrument does not decrease the level of the treatment. This assumption is equivalent to ruling out the presence of defiers, and it is therefore sometimes referred to as the "no-defiance" assumption (Balke and Pearl, 1994; Pearl, 2000). Note that in the Bloom set up with one-sided noncompliance both always-takers and defiers are absent by assumption.

Under these two assumptions, independence of all four potential outcomes $(Y_i(0), Y_i(0), W_i(0), W_i(1))$ and the instrument Z_i, and monotonicity, Imbens and Angrist show that one can identify the average effect of the treatment for the subpopulation of compliers. Before going through their argument, it is useful to see why we cannot generally identify the average effect of the treatment for others subpopulations. Clearly, one cannot identify the average effect of the treatment for never-takers because they are never observed receiving the treatment, and so $\mathbb{E}[Y_i(1)|T_i = n]$ is not identified. Thus, only compliers are observed in both treatment groups, so only for this group is there any chance of identifying the average treatment effect. In order to understand the

positive component of the Imbens-Angrist result, that we can identify the average effect for compliers, it is useful to consider the subpopulations defined by instrument and treatment. Table 4 shows the information we have about the individual's type given the monotonicity assumption. Consider individuals with $(Z_i = 1, W_i = 0)$. Because of monotonicity such individuals can only be never-takers. Similarly, individuals $(Z_i = 0, W_i = 1)$ can only be always-takers. However, consider individuals with $(Z_i = 0, W_i = 0)$. Such individuals can be either compliers or never-takers. We cannot infer the type of such individuals from the observed data alone. Similarly, individuals with $(Z_i = 1, W_i = 1)$ can be either compliers or never-takers.

The intuition for the identification result is as follows. The first step is to see that we can infer the population proportions of the three remaining subpopulations, never-takers, alway-stakers and compliers (using the fact that the monotonicity assumption rules out the presence of defiers). Call these population shares $P_t = \text{pr}(T_i = t)$, for $t \in \{n, a, c\}$. Consider the subpopulation with $Z_i = 0$. Within this subpopulation we observe $W_i = 1$ only for always-takers. Hence the conditional probability of $W_i = 1$ given $Z_i = 0$ is equal to the population share of always-takers: $P_a = \text{pr}(W_i = 1|Z_i = 0)$. Similarly, in the subpopulation with $Z_i = 1$ we observe $W_i = 0$ only for never-takers. Hence the population share of never-takers is equal to the conditional probability of $W_i = 0$ given $Z_i = 1$: $P_n = \text{pr}(W_i = 0|Z_i = 1)$. The population share of compliers is then obtained by subtracting the population shares of never-takers and always-takers from one: $P_c = 1 - P_n - P_a$. The second step uses the distribution of Y_i given (Z_i, W_i). We can infer the distribution of $Y_i|W_i = 0, T_i = n$ from the subpopulation with $(Z_i, W_i) = (1, 0)$ since all these individuals are known to be never-takers. Then we use the distribution of $Y_i|Z_i = 0, W_i = 0$. This is a mixture of the distribution of $Y_i|W_i = 0, T_i = n$ and the distribution of $Y_i|W_i = 0, T_i = c$, with mixture probabilities equal to the relative population shares, $P_n/(P_c + P_n)$ and $P_c/(P_c + P_n)$, respectively. Since we already inferred the population shares of the never-takers and compliers as well as the distribution of $Y_i|W_i = 0, T_i = n$, we can back out the conditional distribution of $Y_i|W_i = 0, T_i = c$. Similarly we can infer the conditional distribution of $Y_i|W_i = 1, T_i = c$. The difference between the means of these two conditional distributions is the Local Average Treatment Effect (LATE, Imbens and Angrist, 1994):

$$\tau_{\text{late}} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i(0) = 0, W_i(1) = 1] = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = \text{complier}].$$

In practice one need not estimate the local average treatment effect by decomposing the mixture distributions directly. Imbens and Angrist show that LATE equals the standard instrumental variables estimand, the ratio of the covariance of Y_i and Z_i and the covariance of W_i and Z_i:

$$\tau_{\text{late}} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]} = \frac{\mathbb{E}[Y_i \cdot (Z_i - \mathbb{E}[Z_i])]}{\mathbb{E}[W_i \cdot (Z_i - \mathbb{E}[Z_i])]},$$

which can be estimated using two-stage-least-squares.

Earlier we argued that one cannot consistently estimate the average effect for either never-takers or always-takers in this setting. Nevertheless, we can still use the bounds approach from Manski (1990, 1995) to bound the average effect for the full population. To understand the nature of the bound, it is useful to decomponse the average effect τ_{pate} by compliance type (maintaining monotonicity, so there are no defiers):

$$\tau_{\text{pate}} = P_n \cdot \mathbb{E}[Y_i(1) - Y_i(0)|T_i = n] + P_a \cdot \mathbb{E}[Y_i(1) - Y_i(0)|T_i = a] + P_c \cdot \mathbb{E}[Y_i(1) - Y_i(0)|T_i = c].$$

The only quantities not consistently estimable are the average effects for never-takers and always-takers. Even for those we have some information. For example, we can write $\mathbb{E}[Y_i(1) - Y_i(0)|T_i = n] = \mathbb{E}[Y_i(1)|T_i = n] - \mathbb{E}[Y_i(0)|T_i = n]$. The second term we can estimate, and the data are completely uninformative about the first term. Hence if there are natural bounds on $Y_i(1)$ (for example, if the outcome is binary), we can use that to bound $\mathbb{E}[Y_i(1)|T_i = n]$, and then in turn use that to bound τ_{pate}. These bounds are tight. See Manski (1990), and Balke and Pearl (1994).

6.3.2 Multi-valued Instruments and Weighting Local Average Treatment Effects

The previous discussion was in terms of a single binary instrument. In that case there is no other average effect of the treatment that can be estimated consistently other than the local average treatment effect, τ_{late}. With a multi-valued instrument, or with multiple binary instruments (still maintaining the setting of a binary treatment), we can estimate a variety of local average treatment effects. Let $\mathbb{Z} = \{z_1, \ldots, z_K\}$ denote the set of values for the instruments. Initially we take the set of values to be finite. Then for each pair (z_k, z_l) with $\text{pr}(W_i = 1|Z_i = z_k) > \text{pr}(W_i = 1|Z_i = z_l)$ one can define a local average treatment effect:

$$\tau_{\text{late}}(z_k, z_l) = \mathbb{E}[Y_i(1) - Y_i(0)|W_i(z_l) = 0, W_i(z_k) = 1].$$

We can combine these to estimate any weighted average of these local average treatment effects:

$$\tau_{\text{late},\lambda} = \sum_{k,l} \lambda_{k,l} \cdot \tau_{\text{late}}(z_k, z_l).$$

Imbens and Angrist show that the standard instrumental variables estimand, using $g(Z_i)$ as an instrument for W_i, is equal to a particular weighted average:

$$\frac{\mathbb{E}[Y_i \cdot (g(Z_i) - \mathbb{E}[g(Z_i)])]}{\mathbb{E}[W_i \cdot (g(Z_i) - \mathbb{E}[g(Z_i)])]} = \tau_{\text{late},\lambda},$$

for a particular set of nonnegative weights as long as $\mathbb{E}[W_i|g(Z_i) = g]$ increases in g.

Heckman and Vytlacil (2005), and Heckman, Urzua, and Vytlacil (2006) study the case with a continuous instrument. They use an additive latent single index setup where the treatment received is equal to

$$W_i = 1\{h(Z_i) + V_i \geq 0\},$$

where $h(\cdot)$ is strictly monotonic, and the latent type V_i is independent of Z_i. In general, in the presence of multiple instruments, this latent single index framework imposes substantive restrictions.[9] Without loss of generality we can take the marginal distribution of V_i to be uniform. Given this framework, Heckman, Urzua and Vytlacil (2006) define the marginal treatment effect as a function of the latent type v of an individual,

$$\tau_{\text{mte}}(v) = \mathbb{E}[Y_i(1) - Y_i(0)|V_i = v].$$

[9] See Vytlacil (2002) for a discussion in the case with binary instruments where the latent index set up implies no loss of generality.

In the single continuous instrument case, $\tau_{\text{mte}}(v)$ is, under some differentiability and invertibility conditions, equal to a limit of local average treatment effects:

$$\tau_{\text{mte}}(v) = \lim_{z \downarrow h^{-1}(v)} \tau_{\text{late}}(h^{-1}(v), z).$$

A parametric version of this concept goes back to work by Björklund and Moffitt (1987). All average treatment effects, including the overall average effect, the average effect for the treated, and any local average treatment effect can now be expressed in terms of integrals of this marginal treatment effect, as shown in Heckman and Vytlacil (2005). For example, $\tau_{\text{pate}} = \int_0^1 \tau_{\text{mte}}(v)dv$. A complication in practice is that not necessarily all the marginal treatment effects can be estimated. For example, if the instrument is binary, $Z_i \in \{0, 1\}$, then for individuals with $V_i < \min(-h(0), -h(1))$, it follows that $W_i = 0$, and for these never-takers we cannot estimate $\tau_{\text{mte}}(v)$. Any average effect that requires averaging over such values of v is therefore also not point-identified. Moreover, average effects that can be expressed as integrals of $\tau_{\text{mte}}(v)$ may be identified even if some of the $\tau_{\text{mte}}(v)$ that are being integrated over are not identified. Again, in a binary instrument example with $\text{pr}(W_i = 1|Z_i = 1) = 1$, and $\text{pr}(W_i = 1|Z_i = 0) = 0$, the average treatment effect τ_{pate} is identified, but $\tau_{\text{mte}}(v)$ is not identified for any value of v.

6.4 Regression Discontinuity Designs

Regression discontinuity methods have been around for a long time in the psychology and applied statistics literature, going back to the early sixties. For discussions and references from this literature, see Thistlewaite and Campbell, 1960; Trochim, 2001; Shadish, Cook, and Campbell, 2002; Cook, 2008. Except for some important foundational work by Goldberger (1972a,b), it is only recently that these methods have attracted much attention in the economics literature. For some of the recent applications, see Van der Klaauw (2002), Lee (2008), Angrist and Lavy (1999), DiNardo and Lee (1994), Chay and Greenstone (2005), Card, Mas and Rothstein (2006), Lee, Moretti, and Butler (2004), Ludwig and Miller (2007), McEwan and Shapiro (2007), Black (1999), Chen and VanderKlaauw (2008), Jin and Leslie (2003), and VanderKlaauw (2008). Key theoretical and conceptual contributions include the interpretation of estimates for fuzzy regression discontinuity designs allowing for general heterogeneity of treatment effects (Hahn, Todd and Van der Klaauw 2001, HTV from hereon), adaptive estimation methods (Sun, 2005), methods for bandwidth selection tailored to the RD setting, (Ludwig and Miller, 2005; Imbens and Kalyanaraman, 2008) and various tests for discontinuities in means and distributions of non-affected variables (Lee, 2007; McCrary, 2007). For recent reviews in the economics literature, see Van Der Klaauw (2007), Imbens and Lemieux (2007), and Lee and Lemieux (2008).

The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor (the forcing variable X_i) being on either side of a common threshold. This generates a discontinuity, sometimes of size one, in the conditional probability of receiving the treatment as a function of this particular predictor. The forcing variable is often itself associated with the potential outcomes, but this association is assumed to be smooth. As a result any discontinuity of the conditional distribution of

the outcome as a function of this covariate at the threshold is interpreted as evidence of a causal effect of the treatment. The design often arises from administrative decisions, where the incentives for individuals to participate in a program are rationed for reasons of resource constraints, and clear transparent rules, rather than discretion, by administrators are used for the allocation of these incentives.

It is useful to distinguish between two general settings, the Sharp and the Fuzzy Regression Discontinuity designs (e.g., Trochim, 1984, 2001; HTV; Imbens and Lemieux, 2007; Lee and Lemieux, 2008; Van der Klaauw, 2007).

6.4.1 The Sharp Regression Discontinuity Design

In the Sharp Regression Discontinuity (SRD) design, the assignment W_i is a deterministic function of one of the covariates, the forcing (or treatment-determining) variable X_i:

$$W_i = 1[X_i \geq c],$$

where $1[\cdot]$ is the indicator function, equal to one if the even in brackets is true and zero otherwise. All units with a covariate value of at least c are in the treatment group (and participation is mandatory for these individuals), and all units with a covariate value less than c are in the control group (members of this group are not eligible for the treatment). In the SRD design we focus on estimation of

$$\tau_{\text{srd}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]. \tag{32}$$

(Naturally, if the treatment effect is constant, then $\tau_{\text{srd}} = \tau_{\text{pate}}$.) Writing this expression as $\mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = c]$, we focus on identification of the two terms separately. By design there are no units with $X_i = c$ for whom we observe $Y_i(0)$. To estimate $\mathbb{E}[Y_i(w)|X_i = c]$ without making functional form assumptions, we exploit the possibility of observing units with covariate values arbitrarily close to c.[10] In order to justify this averaging we make a smoothness assumption that the two conditional expectations $\mathbb{E}[Y_i(w)|X_i = x]$, for $w = 0, 1$, are continuous in x. Under this assumption, $\mathbb{E}[Y_i(0)|X_i = c] = \lim_{x\uparrow c} \mathbb{E}[Y_i(0)|X_i = x] = \lim_{x\uparrow c} \mathbb{E}[Y_i|X_i = x]$, implying that

$$\tau_{\text{srd}} = \lim_{x\downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x\uparrow c} \mathbb{E}[Y_i|X_i = x],$$

where this expression uses the fact that W_i is a deterministic function of X_i (a key feature of the SRD). The statistical problem becomes one of estimating a regression function nonparametrically at a boundary point. We discuss the statistical problem in more detail in Section 6.4.4.

[10] Although in principle the first term in the difference in (32) would be straightforward to estimate if we actually observe individuals with $X_i = x$, with continuous covariates we also need to estimate this term by averaging over units with covariate values close to c.

6.4.2 The Fuzzy Regression Discontinuity Design

In the Fuzzy Regression Discontinuity (FRD) design, the probability of receiving the treatment need not change from zero to one at the threshold. Instead the design only requires a discontinuity in the probability of assignment to the treatment at the threshold:

$$\lim_{x \downarrow c} \text{pr}(W_i = 1 | X_i = x) \neq \lim_{x \uparrow c} \text{pr}(W_i = 1 | X_i = x).$$

In practice the discontinuity needs to be sufficiently large that typically it can be seen easily in simple graphical analyses. These discontinuities can arise if incentives to participate in a program change discontinuously at a threshold, without these incentives being powerful enough to move all units from nonparticipation to participation.

In this design we look at the ratio of the jump in the regression of the outcome on the covariate to the jump in the regression of the treatment indicator on the covariate as an average causal effect of the treatment. Formally, the functional of interest is

$$\tau_{\text{frd}} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[W_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i | X_i = x]}.$$

HTV exploit the instrumental variables connection to interpret the fuzzy regression discontinuity design when the effect of the treatment varies by unit. They define complier to be units whose participation is affected by the cutoff point. That is, a complier is someone with a value of the forcing variable Xi close to c, and who would participate if c were choose to be just below X_i, and not participate if c were choosen to be just above X_i. HTV then exploit that structure to show that in combination with a monotonicity assumption,

$$\tau_{\text{frd}} = \mathbb{E}[Y_i(1) - Y_i(0) | \text{unit } i \text{ is a complier and } X_i = c].$$

The estimand τ_{frd} is an average effect of the treatment, but only averaged for units with $X_i = c$ (by regression discontinuity), and only for compliers (people who are affected by the threshold). Clearly the analysis generally does not have much external validity. It is only valid for the subpopulation who is complier at the threshold, and it is only valid for the subpopulation with $X_i = c$. Nevertheless, the FRD analysis may do well in terms of internal validity.

It is useful to compare the RD method in this setting with standard methods based on unconfoundedness. In contrast to the Sharp RD case, an unconfoundedness-based analysis is possible in the Fuzzy RD setting because some treated observations will have $X_i \leq c$, and some control observations will have $X_i \geq c$. Ignoring the FRD setting – that is, ignoring the discontinuity in $\mathbb{E}[W_i | X_i = x]$ at $x = c$ – and acting as if unconfoundedness holds, would lead to estimating the average treatment effect at $X_i = c$ based on the expression

$$\tau_{\text{unconf}} = \mathbb{E}[Y_i | X_i = c, W_i = 1] - \mathbb{E}[Y_i | X_i = c, W_i = 0],$$

which equals $\mathbb{E}[Y_i(1) - Y_i(0) | X_i = c]$ under unconfoundedness. In fact, under unconfoundedness one can estimate the average effect $\mathbb{E}[Y_i(1) - Y_i(0) | X_i = x]$ at values of x different from c. However, an interesting result is that if unconfoundedness holds, the FRD also estimates

$\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]$, as long as the potential outcomes have smooth expectations as a function of the forcing variable around $x = c$. A special case of this is discussed in HTV, who assume only that treatment is unconfounded with respect to the individual-specific gain. Therefore, in principle, there are situations where even if one believes that unconfoundedness holds, one may wish to use the FRD approach. In particular, even if we maintain unconfoundedness, a standard analysis based on τ_{unconf} can be problematic because the potential discontinuities in the regression functions (at $x = c$) under the FRD design invalidate the usual statistical methods that treat the regression functions as continuous at $x = c$.

Although unconfoundedness in the FRD setting is possible, its failure makes it difficult to interpret τ_{unconf}. By contrast, provided monotocity holds, the FRD parameter, τ_{frd}, identifies the average treatment effect for compliers at $x = c$. In other words, approaches that exploit the FRD nature of the design identify an interesting parameter both when unconfoundedness holds and in a leading case (monoticity) when unconfoundedness fails.

6.4.3 Graphical Methods

Graphical analyses are typically an integral part of any RD analysis. RD designs suggest that the effect of the treatment of interest can be measured by the value of the discontinuity in the conditional expectation of the outcome as a function of the forcing variable at a particular point. Inspecting the estimated version of this conditional expectation is a simple yet powerful way to visualize the identification strategy. Moreover, to assess the credibility of the RD strategy, it can be useful to inspect additional graphs, as discussed below in Section 6.4.5. For strikingly clear examples of such plots, see Lee, Moretti, and Butler (2004), Lalive (2008), and Lee (2008).

The main plot in a Sharp RD setting is a histogram-type estimate of the average value of the outcome by the forcing variable. For some binwidth h, and for some number of bins K_0 and K_1 to the left and right of the cutoff value, respectively, construct bins $(b_k, b_{k+1}]$, for $k = 1, \ldots, K = K_0 + K_1$, where $b_k = c - (K_0 - k + 1) \cdot h$. Then calculate the number of observations in each bin, and the average outcome in the bin:

$$N_k = \sum_{i=1}^{N} 1_{b_k \leq X_i \leq b_{k+1}}, \qquad \overline{Y}_k = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Y_i \cdot 1\{b_k \leq X_i \leq b_{k+1}\}.$$

The key plot is that of the $\overline{Y}_k$, for $k = 1, \ldots, K$ against the mid point of the bins, $\tilde{b}_k = (b_k + b_{k+1})/2$. The question is whether around the threshold c (by construction on the edge of one of the bins) there is any evidence of a jump in the conditional mean of the outcome. The formal statistical analyses discussed below are essentially just sophisticated versions of this, and if the basic plot does not show any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates with statistically and substantially significant magnitudes.

In addition to inspecting whether there is a jump at this value of the covariate, one should inspect the graph to see whether there are any other jumps in the conditional expectation of Y_i given X_i that are comparable in size to, or larger than, the discontinuity at the cutoff value. If so, and if one cannot explain such jumps on substantive grounds, it would call into question the interpretation of the jump at the threshold as the causal effect of the treatment.

In order to optimize the visual clarity it is recommended to calculate averages that are not smoothed across the cutoff point c. In addition, it is recommended not to artificially smooth on either side of the threshold in a way that implies that the only discontinuity in the estimated regression fuction is at c. One should therefore use non-smooth methods such as the histogram type estimators described above rather than smooth methods such as kernel estimators.

In a Fuzzy RD setting one should also calculate

$$\overline{W}_k = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Y_i \cdot 1\{b_k \leq X_i \leq b_{k+1}\},$$

and plot the $\overline{W}_k$ against the bin centers $\tilde{b}_k$, in the same way as described above.

6.4.4 Estimation and Inference

The object of interest in regression discontinuity designs is a difference in two regression functions at a particular point (in the sharp RD case), and the ratio of two differences of regression functions (in the fuzzy RD case). These estimands are identified without functional form assumptions, and in general one might therefore like to use nonparametric regression methods that allow for flexible functional forms. Because we are interested in the behavior of the regression functions around a single value of the covariate, it is attractive to use local smoothing methods such as kernel regression rather than global smoothing methods such as sieves or series regression because the latter will generally be sensitive to behavior of the regression function away from the threshold. Local smoothing methods are generally well understood (e.g., Stone, 1977; Bierens, 1987; Härdle, 1991; Pagan and Ullah, 1999). For a particular choice of the kernel, $K(\cdot)$, e.g., a rectangular kernel $K(z) = 1_{-h \leq z \leq h}$, or a Gaussian kernel $K(z) = \exp(-z^2/2)/\sqrt(2\pi)$, the regression function at x, $m(x) = \mathbb{E}[Y_i|X_i = x]$ is estimated as

$$\hat{m}(x) = \sum_{i=1}^{N} Y_i \cdot \lambda_i, \qquad \text{with weights } \lambda_i = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right)}.$$

An important difference with the primary focus in the nonparametric regression literature is that in the RD setting we are interested in the value of the regression functions at boundary points. Standard kernel regression methods do not work well in such cases. More attractive methods for this case are local linear regression (Fan and Gijbels, 1996; Porter, 2003), where locally a linear regression function, rather than a constant regression function, is fitted. This leads to an estimator for the regression function at x equal to

$$\hat{m}(x) = \hat{\alpha}, \qquad \text{where } \left(\hat{\alpha}, \hat{\beta}\right) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \lambda_i \cdot \left(Y_i - \alpha - \beta \cdot (X_i - x)\right)^2,$$

with the same weights λ_i as before. In that case the main remaining choice concerns the bandwidth, denoted by h. Suppose one uses a rectangular kernel, $K(z) = 1_{-h \leq z \leq z}$ (and typically the results are relatively robust with respect to the choice of the kernel). The choice of bandwidth then amounts to to dropping all observations such that $X_i \notin [c-h, c+h]$. The question becomes how to choose the bandwidth h.

Most standard methods for choosing bandwidths in nonparametric regression, including both cross-validation and plug-in methods, are based on criteria that integrate the squared error over the entire distribution of the covariates: $\int_z (\hat{m}(z) - m(z))^2 f_X(z) dz$. For our purposes this criterion does not reflect the object of interest. We are specifically interested in the regression function at a single point, moreover, this point is always a boundary point. Thus we would like to choose h to minimize $\mathbb{E}[(\hat{m}(c) - m(c))^2]$ (using the data with $X_i \leq c$ only, or using the data with $X_i \geq c$ only). If the density of the forcing variable is high at the threshold, a bandwidth selection procedure based on global criteria may lead to a bandwidth that is much larger than is appropriate.

There are few attempts to formalize to standardize the choice of a bandwidth for such cases. Ludwig and Miller (2005) and Imbens and Lemieux (2007) discuss some cross-validation methods that target more directly the object of interest in RD designs. Assuming the density of X_i is continuous at c, and that the conditional variance of Y_i given X_i is continuous and equal to σ^2 at $X_i = c$, Imbens and Kalyanaram (2008) show that the optimal bandwidth depends on the second derivatives of the regression functions at the threshold and has the form

$$h_{\text{opt}} = N^{-1/5} \cdot C_K \cdot \sigma^2 \cdot \left(\frac{\frac{1}{p} + \frac{1}{1-p}}{\lim_{x \downarrow c} \left(\frac{\partial^2 m}{\partial x^2}(x) \right)^2 + \lim_{x \uparrow c} \left(\frac{\partial^2 m}{\partial x^2}(x) \right)^2} \right)^{1/5},$$

where p is the fraction of observations with $X_i \geq c$, and C_K is a constant that depends on the kernel. For a rectangular kernel $K(z) = 1_{-h \leq z \leq h}$, the constant equals $C_K = 2.70$. Imbens and Kalyanaram propose and implement a plug in method for the bandwidth.[11]

If one uses a rectangular kernel, and given a choice for the bandwidth, estimation for the sharp and fuzzy RD designs can be based on ordinary least squares and two stage least squares, respectively. If the bandwidth goes to zero sufficiently fast, so that the asymptotic bias can be ignored, one can also base inference on these methods. See HTV and Imbens and Lemieux (2007).

6.4.5 Specification Checks

There are two important concerns in the application of RD designs, be they sharp or fuzzy. These concerns can sometimes be assuaged by investigating various implications of the identification argument underlying the regression discontinuity design.

A first concern about RD designs is the possibility of other changes at the same threshold value of the covariate. For example, the same age limit may affect eligibility for multiple programs. If all the programs whose eligibility changes at the same cutoff value affect the outcome of interest, an RD analysis may mistakenly attribute the combined effect to the treatment of interest. The second concern is that of manipulation by the individuals of the covariate value that underlies the assignment mechanism. The latter is less of a concern when the forcing variable is a fixed, immutable characteristic of an individual such as age. It is a particular concern when eligibility criteria are known to potential participants and are based on variables

[11] Code in Matlab and Stata for calculating the optimal bandwidth is available on their website.

that are affected by individual choices. For example, if eligibility for financial aid depends on test scores that are graded by teachers who know the cutoff values, there may be a tendency to push grades high enough to make students eligible. Alternatively if thresholds are known to students they may take the test multiple times in an attempt to raise their score above the threshold.

There are two sets of specification checks that researchers can typically perform to at least partly assess the empirical relevance of these concerns. Although the proposed procedures do not directly test null hypotheses that are required for the RD approach to be valid, it is typically difficult to argue for the validity of the approach when these null hypotheses do not hold. First, one may look for discontinuities in average value of the covariates around the threshold. In most cases, the reason for the discontinuity in the probability of the treatment does not suggest a discontinuity in the average value of covariates. Finding a discontinuity in other covariates typically casts doubt on the assumptions underlying the RD design. Specifically, for covariates Z_i, the test would look at the difference

$$\tau_Z = \lim_{x \downarrow c} \mathbb{E}[Z_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Z_i|X_i = x].$$

Second, McCrary (2008) suggests testing the null hypothesis of continuity of the density of the covariate that underlies the assignment at the threshold, against the alternative of a jump in the density function at that point. A discontinuity in the density of this covariate at the particular point where the discontinuity in the conditional expectation ocurs is suggestive of violations of the no-manipulation assumption. Here the focus is on the difference

$$\tau_{f(x)} = \lim_{x \downarrow c} f_X(x) - \lim_{x \uparrow c} f_X(x).$$

In both cases a substantially and statistically significant difference in the left and right limits suggest that there may be problems with the RD approach. In practice more useful than formal statistical tests are graphical analyses of the type discussed in Section 6.4.3 where histogram-type estimates of the conditional expectation of $\mathbb{E}[Z_i|X_i = x]$ and of the marginal density $f_X(x)$ are graphed.

6.5 Difference-in-Differences Methods

Since the seminal work by Ashenfelter (1978) and Ashenfelter and Card (1985), the use of Difference-In-Differences (DID) methods has become widespread in empirical economics. Influential applications include Card (1990), Meyer, Viscusi and Durbin (1995), Card and Krueger (1993), Eissa and Liebman (1996), Blundell, Duncan and Meghir (1998), and many others. The simplest setting is one where outcomes are observed for units observed in one of two groups, in one of two time periods. Only units in one of the two groups, in the second time period, are exposed to a treatment. There are no units exposed to the treatment in the first period, and units from the control group are never observed to be exposed to the. The average gain over time in the non-exposed (control) group is substracted from the gain over time in the exposed (treatment) group. This double differencing removes biases in second period comparisons between the treatment and control group that could be the result from permanent differences between

those groups, as well as biases from comparisons over time in the treatment group that could be the result of time trends unrelated to the treatment. We discuss here the conventional set up, and recent work on inference (Bertrand, Duflo and Mullainathan, 2004; Hansen, 2007a,b; Donald and Lang, 2007), as well as the recent extensions by Athey and Imbens (2006) who develop a functional form free version of the difference in difference methodology, and Abadie, Diamond and Hainmueller (2007), who develop a method for constructing an artificial control group from multiple non-exposed groups.

6.5.1 Repeated Cross-sections

The standard model for the DID approach is as follows. Individual i belongs to a group, $G_i \in \{0, 1\}$ (where group 1 is the treatment group), and is observed in time period $T_i \in \{0, 1\}$. For $i = 1, \ldots, N$, a random sample from the population, individual i's group identity and time period can be treated as random variables. In the standard DID model we can write the outcome for individual i in the absence of the intervention, $Y_i(0)$ as

$$Y_i(0) = \alpha + \beta \cdot T_i + \gamma \cdot G_i + \varepsilon_i, \tag{33}$$

with unknown parameters α, β, and γ. We ignore the potential presence of other covariates, which introduce no special complications. The second coefficient in this specification, β, represents the time component common to both groups. The third coefficient, γ, represents a group-specific, time-invariant component. The fourth term, ε_i, represents unobservable characteristics of the individual. This term is assumed to be independent of the group indicator and have the same distribution over time, i.e. $\varepsilon_i \perp\!\!\!\perp (G_i, T_i)$, and is normalized to have mean zero.

An alternative set up leading to the same estimator allows for a time-invariant individual-specific fixed effect, γ_i, potentially correlated with G_i, and models $Y_i(0)$ as

$$Y_i(0) = \alpha + \beta \cdot T_i + \gamma_i + \varepsilon_i, \tag{34}$$

See, e.g., Angrist and Krueger (2000). This generalization of the standard model does not affect the standard DID estimand, and it will be subsumed as a special case of the model we propose.

The equation for the outcome without the treatment is combined with an equation for the outcome given the treatment: $Y_i(1) = Y_i(0) + \tau_{\text{did}}$. The standard DID estimand is under this model equal to

$$\begin{aligned}
\tau_{\text{did}} &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \\
&= \Big(\mathbb{E}[Y_i|G_i = 1, T_i = 1] - \mathbb{E}[Y_i|G_i = 1, T_i = 0]\Big) \\
&\quad - \Big(\mathbb{E}[Y_i|G_i = 0, T_i = 1] - \mathbb{E}[Y_i|G_i = 0, T_i = 0]\ \Big).
\end{aligned} \tag{35}$$

In other words, the population average difference over time in the control group ($G_i = 0$) is subtracted from the population average difference over time in the treatment group ($G_i = 1$) to remove biases associated with a common time trend unrelated to the intervention.

We can estimate τ_{did} simply using least squares methods on the regression function for the observed outcome,

$$Y_i = \alpha + \beta_1 \cdot T_i + \gamma_1 \cdot G_i + \tau_{\text{did}} \cdot W_i + \varepsilon_i,$$

where the treatment indicator W_i is equal to the interaction of the group and time indicators, $I_i = T_i \cdot G_i$. Thus the treatment effect is estimated through the coefficient on the interaction between the indicators for the second time period and the treatment group. This leads to

$$\hat{\tau}_{\text{did}} = \Big(\overline{Y}_{11} - \overline{Y}_{10}\Big) - \Big(\overline{Y}_{01} - \overline{Y}_{00}\Big),$$

where $\overline{Y}_{gt} = \sum_{i|G_i=g,T_i=t} Y_i/N_{gt}$ is the average outcome among units in group g and time period t.

6.5.2 Multiple Groups and Multiple Periods

With multiple time periods and multiple groups we can use a natural extension of the two-group two-time-period model for the outcome in the absence of the intervention. Let T and G denote the number of time periods and groups respectively. Then:

$$Y_i(0) = \alpha + \sum_{t=1}^{T} \beta_t \cdot 1_{T_i=t} + \sum_{g=1}^{G} \gamma_g \cdot 1_{G_i=g} + \varepsilon_i \tag{36}$$

with separate parameters for each group and time period, γ_g and β_t, for $g = 1, \ldots, G$ and $t = 1, \ldots, T$, where the initial time period coefficient and first group coefficient have implicitly been normalized to zero. This model is then combined with the additive model for the treatment effect, $Y_i(1) = Y_i(0) + \tau_{\text{did}}$, implying that the parameters of this model can still be estimated by ordinary least squares based on the regression function

$$Y_i = \alpha + \sum_{t=1}^{T} \beta_t \cdot 1_{T_i=t} + \sum_{g=1}^{G} \gamma_g \cdot 1_{G_i=g} + \tau_{\text{did}} \cdot I_i + \varepsilon_i, \tag{37}$$

where I_i is now an indicator for unit i being in a group and time period that was exposed to the treatment.

This model with more than two time periods, or more than two groups, or both, imposes testable restrictions on the data. For example, if group g_1 and g_2 are both not exposed to the treatment in periods t_1 and t_2, under this model the double difference

$$\Big(\overline{Y}_{g_2,t_2} - \overline{Y}_{g_2,t_1}\Big) - \Big(\overline{Y}_{g_1,t_2} - \overline{Y}_{g_1,t_1}\Big),$$

should estimate zero, which can be tested using conventional methods – this possibility is exploited in the next subsection. In the two-period, two-group setting there are no testable restrictions on the four group/period means.

6.5.3 Standard Errors in the Multiple Group and Multiple Period Case

Recently there has been attention called to the concern that ordinary least square standard errors for the DID estimator may not be accurate in the presence of correlations between outcomes within groups and between time periods. This is a particular case of clustering where the regressor of interest does not vary within clusters. See for a general discussion Moulton (1990), Moulton and Randolph (1989), and Wooldridge (2002a). The specific problem has been analyzed recently by Donald and Lang (2007), Bertrand, Duflo and Mullainathan (2004), and Hansen (2007a,b).

The starting point of these analyses is a particular structure on the error term ε_i:

$$\varepsilon_i = \eta_{G_i,T_i} + \nu_i,$$

where ν_i is an individual-level idiosyncratic error term, and η_{gt} is a group/time specific component. The unit level error term ν_i is independent accross all units, $\mathbb{E}\left[\nu_i \cdot \nu_j\right] = 0$ if $i \neq j$ and $\mathbb{E}\left[\nu_i^2\right] = \sigma_\nu^2$. Now suppose we also assume that $\eta_{gt} \sim \mathcal{N}(0, \sigma_\eta^2)$, and all the η_{gt} are independent. Let us focus initially on the two-group, two-time-period case. With a large number of units in each group and time period, $\overline{Y}_{gt} \to \alpha + \beta_t + \gamma_g + 1_{g=1,t=1} \cdot \tau_{\text{did}} + \eta_{gt}$, so that

$$\hat{\tau}_{\text{did}} = \left(\overline{Y}_{11} - \overline{Y}_{10}\right) - \left(\overline{Y}_{01} - \overline{Y}_{00}\right) \longrightarrow \tau_{\text{did}} + (\eta_{11} - \eta_{10}) - (\eta_{01} - \eta_{00}) \sim \mathcal{N}(\tau_{\text{did}}, 4 \cdot \sigma_\eta^2).$$

Thus, in this case with two groups and two time periods, the conventional DID estimator is not consistent. In fact, no consistent estimator exists because there is no way to eliminate the influence of the four unobserved components η_{gt}. In this two-group, two-time-period case the problem is even worse than the absence of a consistent estimator, because one cannot even establish whether there is a clustering problem: the data are not informative about the value of σ_η^2. If we have data from more than two groups or from more than two time periods, we can typically estimate σ_η^2, and thus, at least under the normality and independence assumptions for η_{gt}, construct confidence intervals for τ_{did}. Consider, for example, the case with three groups, and two time periods. If groups $G_i = 0, 1$ are both not treated in the second period, then $(\overline{Y}_{11} - \overline{Y}_{10}) - (\overline{Y}_{01} - \overline{Y}_{00}) \sim \mathcal{N}(0, 4 \cdot \sigma_\eta^2)$, which can be used to obtain an unbiased estimator for σ_η^2. See Donald and Lang (2007) for details.

Bertrand, Duflo and Mullainathan (2004) and Hansen (2007a,b) focus on the case with multiple (more than two) time periods. In that case we may wish to relax the assumption that the η_{gt} are independent over time. Note that with data from only two time period there is no information in the data that allows one to establish the absence of independence over time. The typical generalization is to allow for a autoregressive structure on the η_{gt}, for example,

$$\eta_{gt} = \alpha \cdot \eta_{gt-1} + \omega_{gt},$$

with a serially uncorrelated ω_{gt}. More generally, with T time periods, one can allow for an autoregressive process of order $T - 2$. Using simulations and real data calculations based on data for 50 states and multiple time periods Bertrand, Duflo and Mullainathan (2004) show that corrections to the conventional standard errors taking into account the clustering and

autoregressive structure make a substantial difference. Hansen (2007a,b) provides additional large sample results under sequences where the number of time periods increases with the sample size.

6.5.4 Panel Data

Now suppose we have panel data, in the two period, two group case. Here we have N individuals, indexed $i = 1, \ldots, N$, for whom we observe $(G_i, Y_{i0}, Y_{i1}, X_{i0}, X_{i1})$, where G_i is, as before, group membership, X_{it} is the covariate value for unit i at time t, and Y_{it} is the outcome for unit i at time t.

One option is to proceed with estimation exactly as before, essentially ignoring the fact that the observations in different time periods come from the same unit. We can now interpret the estimator as the OLS estimator based on the regression function for the difference outcomes:

$$Y_{i1} - Y_{i0} = \beta + \tau_{\text{did}} \cdot G_i + \epsilon_i,$$

which leads to the double difference as the estimator for τ_{did}: $\hat{\tau}_{\text{did}} = (\overline{Y}_{11} - \overline{Y}_{10}) - \left(\overline{Y}_{01} - \overline{Y}_{00}\right)$. This estimator is identical to that discussed in the context of repeated cross-sections, and so does not exploit directly the panel nature of the data.

A second, and very different, approach with panel data, which does exploit the specific features of the panel data, would be to assume unconfoundedness given lagged outcomes. Let us look at the differences between these two approaches in a simple setting, without covariates, and assuming linearity. In that case the DID approach suggests the regression of $Y_{i1} - Y_{i0}$ on the group indicator, leading to $\hat{\tau}_{\text{did}}$. The unconfoundedness assumption would suggest the regression of the difference $Y_{i1} - Y_{i0}$ on the group indicator and the lagged outcome Y_{i0}:

$$Y_{i1} - Y_{i0} = \beta + \tau_{\text{unconf}} \cdot G_i + \delta \cdot Y_{i0} + \epsilon_i.$$

While it appears that the analysis based on unconfoundedness is necessarily less restrictive because it allows a free coefficient in Y_{i0}, this is not the case. The DID assumption implies that adjusting for lagged outcomes actually compromises the comparison because Y_{i0} may in fact be correlated with ϵ_i. In the end, the two approaches make fundamentally different assumptions. One needs to choose between them based on substantive knowledge. When the estimated coefficient on the lagged outcome is close to zero, obviously there will be little difference between the point estimates. In addition, using the formula for omitted variable bias in least squares estimation, the results will be very similar if the average outcomes in the treatment and control groups are similar in the first period. Finally, note that in the repeated cross-section case the choice between the DID and unconfoundedness approaches did not arise because the unconfoundedness approach is not feasible: it is not possible to adjust for lagged outcomes when we do not have the same units available in both periods.

As a practical matter, the DID approach appears less attractive than the unconfoundedness-based approach in the context of panel data. It is difficult to see how making treated and control units comparable on lagged outcomes will make the causal interpretation of their difference less credible, as suggested by the DID assumptions.

6.5.5 The Changes-in-Changes Model

Now we return to the setting with two groups, two time periods, and repeated cross-sections. Athey and Imbens (2006) generalize the standard model in several ways. They relax the additive linear model by assuming that, in the absence of the intervention, the outcomes satisfy

$$Y_i(0) = h_0(U_i, T_i), \tag{38}$$

with $h_0(u,t)$ increasing in u. The random variable U_i represents all unobservable characteristics of individual i, and (38) incorporates the idea that the outcome of an individual with $U_i = u$ will be the same in a given time period, irrespective of the group membership. The distribution of U_i is allowed to vary across groups, but not over time within groups, so that $U_i \perp T_i \mid G_i$. Athey and Imbens call the resulting model that changes-in-changes (CIC) model.

The standard DID model in (33) adds three additional assumptions to the CIC model, namely

$$U_i - \mathbb{E}[U_i|G_i] \perp G_i \qquad \text{(additivity)} \tag{39}$$

$$h_0(u,t) = \phi(u + \delta \cdot t), \qquad \text{(single index model)} \tag{40}$$

for a strictly increasing function $\phi(\cdot)$, and

$$\phi(\cdot) \text{ is the identity function.} \qquad \text{(identity transformation).} \tag{41}$$

In the CIC extension, the treatment group's distribution of unobservables may be different from that of the control group in arbitrary ways. In the absence of treatment, all differences between the two groups can be interpreted as coming from differences in the conditional distribution of U given G. The model further requires that the changes over time in the distribution of each group's outcome (in the absence of treatment) arise solely from the fact that $h_0(u,0)$ differs from $h_0(u,1)$, that is, the relation between unobservables and outcomes changes over time. Like the standard model, the Athey-Imbens approach does not rely on tracking individuals over time. Although the distribution of U_i is assumed not to change over time within groups, the model does not make any assumptions about whether a particular individual has the same realization U_i in each period. Thus, the estimators derived by Athey and Imbens will be the same whether one observes a panel of individuals over time or a repeated cross-section. Just as in the standard DID approach, if one only wishes to estimate the effect of the intervention on the treatment group, no assumptions are required about how the intervention affects outcomes.

The average effect of the treatment for the second period treatment group is $\tau_{\text{cic}} = \mathbb{E}[Y_i(1) - Y_i(0)|G_i = 1, T_i = 1]$. Because the first term of this expression is equal to $\mathbb{E}[Y_i(1)|G_i = 1, T_i = 1] = \mathbb{E}[Y_i|G_i = 1, T_i = 1]$, it can be estimated directly from the data. The difficulty is in estimating the second term. Under the assumptions of monotonicity of $h_0(u,t)$ in u, and conditional independence of T_i and U_i given G_i, Athey and Imbens show that in fact the full distribution of $Y(0)$ given $G_i = T_i = 1$ is identified through the equality

$$F_{Y_{11}}(y) = F_{Y_{10}}(F^{-1}_{Y_{00}}(F_{Y_{01}}(y))), \tag{42}$$

where $F_{Y_{gt}}(y)$ denotes the distribution function of Y_i given $G_i = g$ and $T_i = t$. The expected outcome for the second period treatment group under the control treatment is

$$\mathbb{E}[Y_i(0)|G_i = 1, T_i = 1] = \mathbb{E}\left[F_{01}^{-1}\left(F_{00}(Y_{i10})\right)\right].$$

To analyze the counterfactual effect of the intervention on the control group, Athey and Imbens assume that, in the presence of the intervention,

$$Y_i(1) = h_1(U_i, T_i)$$

for some function $h_1(u, t)$ that is increasing in u. That is, the effect of the treatment at a given time is the same for individuals with the same $U_i = u$, irrespective of the group. No further assumptions are required on the functional form of h_1, so the treatment effect, equal to $h_1(u, 1) - h_0(u, 1)$ for individuals with unobserved component u, can differ across individuals. Because the distribution of the unobserved component U can vary across groups, the average return to the policy intervention can vary across groups as well.

6.5.6 The Abadie-Diamond-Hainmueller Artificial Control Group Approach

Abadie, Diamond, and Hainmueller (2007) develop an alternative approach to the setting with multiple control groups. See also Abadie and Gardeazabal (2003). Here we discuss a simple version of their approach, with $T + 1$ time periods, and $G + 1$ groups, one treated in the final period, and G not treated in either period. The Abadie-Diamond-Hainmueller idea is to construct an artificial control group that is more similar to the treatment group in the initial period than any of the control groups on their own. Let $G_i = G$ denote the treated group, and $G_i = 0, \ldots, G - 1$ denote the G control groups. The outcome for the second period treatment group in the absence of the treatment will be estimated as a weighted average of period T outcomes in the G control groups,

$$\hat{\mathbb{E}}[Y_i(0)|T_i = T, G_i = G] = \sum_{g=0}^{G-1} \lambda_g \cdot \overline{Y}_{gT},$$

with weights λ_g satisfying $\sum_{g=0}^{G-1} \lambda_g = 1$, and $\lambda_g \geq 0$. The weights are chosen to make the weighted control group resemble the treatment group prior to the treatment. That is, the weights λ_g are chosen to minimize the difference between the treatment group and the weighted average of the control groups prior to the treatment, namely,

$$\left\| \begin{array}{c} \overline{Y}_{G0} - \sum_{g=0}^{G-1} \lambda_g \cdot \overline{Y}_{g0} \\ \vdots \\ \overline{Y}_{G,T-1} - \sum_{g=0}^{G-1} \lambda_g \cdot \overline{Y}_{g,T-1} \end{array} \right\|,$$

where $\|\cdot\|$ denotes a measure of distance. One can also add group level covariates to the criterion to determine the weights. These group-level covariates may be averages of individual level covariates, or quantiles of the distribution of within group covariates. The idea is that the future path of the artificial control group, consisting of the λ-weighted average of all the control

groups, mimics the path that would have been observed in the treatment group in the absence of the treatment. Applications in Abadie, Diamond and Hainmueller (2007) to estimation of the effect of smoking legislation in California, and the effect of reunification on West Germany are very promising.

7 Multi-valued and Continuous Treatments

Most of the recent econometric program evaluation literature has focused on the case with a binary treatment. As a result this case is now understood much better than it was a decade or two ago. However, much less is known about settings with multi-valued, discrete or continuous treatments. Such cases are common in practice. Social programs are rarely homogenous. Typically individuals are assigned to various activities and regimes, often sequentially, and tailored to their specific circumstances and characteristics.

To provide some insight into the issues arising in settings with multivalued treatments we discuss in this review five separate cases. First, the simplest setting where the treatment is discrete and one is willing to assume unconfoundedness of the treatment assignment. In that case straightforward extensions of the binary treatment case can be used to obtain estimates and inferences for causal effects. Second, we look at the case with a continuous treatment under unconfoundedness. In that case the definition of the propensity score requires some modification, but many of the insights from the binary treatment case still carry over. Third, we look at the case where units can be exposed to a sequence of binary treatments. For example, an individual may remain in a training program for a number of periods. In each period the assignment to the program is assumed to be unconfounded, given permanent characteristics and outcomes up to that point. In the last two cases we briefly discuss multi-valued endogenous treatments. In the fourth case we look at settings with a discrete multi-valued treatment in the presence of endogeneity. In the final case we allow the treatment to be continuous. The last two cases tie in closely with the simultaneous equations literature, where, somewhat separately from the program evaluation literature, there has been much recent work on nonparametric identification and estimation. Especially in the discrete case many of the results in this literature are negative in the sense that without unattractive restrictions on heterogeneity or functional form few objects of interest are point-identified. Some of the literature has turned towards establishing bounds. This is an area with much ongoing work and considerable scope for further research.

7.1 Multi-valued Discrete Treatments with Unconfounded Treatment Assignment

If there are a few different levels of the treatment, rather than just two, essentially all of the methods discussed before go through in the unconfoundedness case. Suppose, for example, that the treatment can be one of three leves, say $W_i \in \{0, 1, 2\}$. In order to estimate the effect of treatment level 2 relative to treatment level 1, one can simply put aside the data for units exposed to treatment level 0 if one is willing to assume unconfoundedness. More specifically, one can estimate the average outcome for each treatment level conditional on the covariates,

$\mathbb{E}[Y_i(w)|X_i = x]$, using data on units exposed to treatment level w, and average these over the (estimated) marginal distribution of the covariates, $\hat{F}_X(x)$. In practice, the overlap assumption may more likely to be violated with more than two treatments. For example, with three treatments, it may be that no units are exposed to treatment level 2 if X_i is in some subset of the covariate space. The insights from the binary case directly extend to this multiple (but few) treatment case. If the number of treatments is relatively large, one may wish to smooth across treatment levels in order to improve precision of the inferences.

7.2 Continuous Treatments with Unconfounded Treatment Assignment

In the case where the treatment taking on many values, Imbens (2000), Lechner (2001), Hirano and Imbens (2004), and Flores (2005) extended some of the propensity score methodology under unconfoundedness. The key maintained assumption is that adjusting for pre-treatment differences removes all biases, and thus solves the problem of drawing causal inferences. This is formalized by using the concept of weak unconfoundedness, introduced by Imbens (2000). Assignment to treatment W_i is weakly unconfounded, given pre-treatment variables X_i, if

$$W_i \perp\!\!\!\perp Y_i(w) \,\Big|\, X_i,$$

for all w. Compare this to the stronger assumption made by Rosenbaum & Rubin (1983) in the binary case:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i,$$

which requires the treatment W_i to be independent of the entire set of potential outcomes. Instead, weak unconfoundedness requires only pairwise independence of the treatment with each of the potential outcomes. A similar assumption is used in Robins (1995). The definition of weak unconfoundedness is also similar to that of "missing at random"(Rubin, 1976; Little and Rubin, 1987) in the missing data literature.

Although in substantive terms the weak unconfoundedness assumption is not very different from the assumption used by Rosenbaum and Rubin (1983), it is important that one does not need the stronger assumption to validate estimation of the expected value of $Y_i(w)$ by adjusting for X_i: under weak unconfoundedness, we have $\mathbb{E}[Y_i(w)|X_i] = \mathbb{E}[Y_i(w)|W_i = w, X_i] = \mathbb{E}[Y_i|W_i = w, X_i]$, and expected outcomes can then be estimated by averaging these conditional means: $\mathbb{E}[Y_i(w)] = \mathbb{E}[\mathbb{E}[Y_i(w)|X_i]]$. In practice, it can be difficult to estimate $\mathbb{E}[Y_i(w)]$ in this manner when the dimension of X_i is large, or if w takes on many values, because the first step requires estimation of the expectation of $Y_i(w)$ given the treatment level and all pre-treatment variables. It was this difficulty that motivated Rosenbaum and Rubin (1983) to develop the propensity score methodology.

Imbens (2000) introduces the generalized propensity score for the multiple treatment case. It is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables:

$$r(w, x) \equiv \text{pr}(W_i = w|X_i = x).$$

In the continuous case, where, say, W_i takes values in the unit interval, $r(w,x) = F_{W|X}(w|x)$. Suppose assignment to treatment W_i is weakly unconfounded given pre-treatment variables X_i. Then, by the same argument as in the binary treatment case, assignment is weakly unconfounded given the generalised propensity score, as $\delta \to 0$,

$$1\{w-\delta \leq W_i \leq w+\delta\} \perp\!\!\!\perp Y_i(w) \;\Big|\; r(w, X_i),$$

for all w. This is the point where using the weak form of the unconfoundedness assumption is important. There is, in general, no scalar function of the covariates such that the level of the treatment W_i is independent of the set of potential outcomes $\{Y_i(w)\}_{w\in[0,1]}$, unless additional structure is imposed on the assignment mechanism; see for example, Joffe and Rosenbaum (1999).

Because weak unconfoundedness given all pretreatment variables implies weak unconfoundedness given the generalised propensity score, one can estimate average outcomes by conditioning solely on the generalised propensity score. If assignment to treatment is weakly unconfounded given pre-treatment variables X, then two results follow. First, for all w,

$$\beta(w,r) \equiv \mathbb{E}[Y_i(w)|r(w,X_i)=r] = \mathbb{E}[Y_i|W_i=w, r(W_i,X_i)=r],$$

which can be estimated using data on Y_i, W_i, and $r(W_i, X_i)$. Second, the average outcome given a particular level of the treatment, $\mathbb{E}[Y_i(w)]$, can be estimated by appropriately averaging $\beta(w,r)$:

$$\mathbb{E}[Y_i(w)] = \mathbb{E}[\beta(w, r(w,X_i))].$$

As with the implementation of the binary treatment propensity score methodology, the implementation of the generalised propensity score method consists of three steps. In the first step the score $r(w,x)$ is estimated. With a binary treatment the standard approach (Rubin and Rosenbaum, 1984; Rosenbaum, 1995) is to estimate the propensity score using a logistic regression. More generally, if the treatments correspond to ordered levels of a treatment, such as the dose of a drug or the time over which a treatment is applied, one may wish to impose smoothness of the score in w. For continuous W_i, Hirano and Imbens (2004) use a lognormal distribution. In the second step the conditional expectation $\beta(w,r) = \mathbb{E}[Y_i|W_i=w, r(W_i,X_i)=r]$ is estimated. Again, the implementation may be different in the case where the levels of the treatment are qualitatively distinct than in the case where smoothness of the conditional expectation function in w is appropriate. Here, some form of linear or nonlinear regression may be used. In the third step the average response at treatment level w is estimated as the average of the estimated conditional expectation, $\hat{\beta}(w, r(w,X_i))$, averaged over the distribution of the pre-treatment variables, $X_1, \ldots, X_N$. Note that to get the average $\mathbb{E}[Y_i(w)]$, the second argument in the conditional expectation $\beta(w,r)$ is evaluated at $r(w,X_i)$, not at $r(W_i,X_i)$.

7.2.1 Dynamic Treatments with Unconfounded Treatment Assignment

Multiple-valued treatments can arise because at any point in time individuals can be assigned to multiple different treatment arms, or because they can be assigned sequentially to different

treatments. Gill and Robins (2001) analyze this case where they assume that at any point in time an unconfoundedness assumption holds. Lechner and Miquel (2005) (see also Lechner, Miquel and Wunsch, 2006) study a related case, where again a sequential unconfoundedness assumption is maintained to identify the average effects of interest. These methods hold great promise, but until now there have been few substantive applications.

7.3 Multi-valued Discrete Endogenous Treatments

In settings with general heterogeneity in the effects of the treatment the case with more than two treatment levels is considerably more challenging than the binary case. There are few studies investigating identification in these settings. Angrist and Imbens (1995) study the interpretation of the standard instrumental variable estimand, the ratio of the covariances of outcome and instrument and treatment and instrument. They show that in general, with a valid instrument, the instrumental variables estimand can still be interpreted as an average causal effect, but with a complicated weighting scheme. There are essentially two levels of averaging going on. First, at each level of the treatment we can only get the average effect of a unit increase in the treatment for compliers at that level. In addition there is averaging over all levels of the treatment, with the weights equal to the proportion of compliers at that level.

Imbens (2007) studies in more detail the case where the endogenous treatment takes on three values, and shows the limits to identification in the case with heterogenous treatment effects.

7.4 Continuous Endogenous Treatments

Somewhat surprisingly, there are many more results for the case with continuous endogenous treatments than for the discrete case. Much of the focus has been on triangular systems, with a single unobserved component of the equation determining the treatment:

$$W_i = h(Z_i, \eta_i),$$

where η_i is scalar, and an essentially unrestricted outcome equation:

$$Y_i = g(W_i, \varepsilon_i),$$

where ε_i may be a vector. Chernozhukov and Hansen (2005), Imbens and Newey (2006), and Chesher (2003) study various forms of this set up. Imbens and Newey (2006) show that if $h(z, \eta)$ is strictly monotone in η, then one can identify average effects of the treatment subject to support conditions on the instrument. They suggest a control function approach to estimation. First η is normalized to have a uniform distribution on $[0, 1]$ (e.g., Matzkin, 2003). Then η_i is estimated as $\hat{\eta}_i = \hat{F}_{W|Z}(W_i|Z_i)$. In the second stage Y_i is regressed nonparametrically on X_i and $\hat{\eta}_i$. Chesher (2003) studies local versions of this problem.

When the treatment equation has an additive form, say $W_i = h_1(Z_i) + \eta_i$, where η_i is independent of Z_i, Blundell and Powell (2003) derive nonparametric control function methods for estimating the average structural function, $\mathbb{E}[g(w, \varepsilon_i)]$.

8 Conclusion

Over the last two decades there has been a proliferation of the literature on program evaluation. This includes theoretical econometrics work, as well as empirical work. Important features of the modern literature are the convergence of the statistical and econometric literatures, with the Rubin potential outcomes framework now the dominant framework. The modern literature has stressed the importance of relaxing functional form and distributional assumptions, and has allowed for general heterogeneity in the effects of the treatment. This has led to renewed interest in identification questions, leading to unusual and controversial estimands such as the local average treatment effect (Imbens and Angrist, 1994), as well as to the literature on partial identication (Manski, 1990). It has also borrowed heavily from the semiparametric literature, using both efficiency bound results (Hahn, 1998) and methods for inference based on series and kernel estimation (Newey, 1994ab). It has by now matured to the point that is is of great use for practitioners.

References

Abadie, A. (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variables Models," *JASA*, 97, 284-292, 231-263.

Abadie, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231-263.

Abadie, A., (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72(1), 1-19.

Abadie, A., and J. Gardeazabal (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93(1), 112-132.

Abadie, A., A. Diamond, and J. Hainmueller, (2007), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," Unpublished Manuscript, Harvard University.

Abadie, A., and G. Imbens, (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235-267.

Abadie, A., and G. Imbens, (2007), "On the Failure of the Bootstrap for Matching Estimators," forthcoming *Econometrica*.

Abadie, A., and G. Imbens, (2008a), "Estimation of the Conditional Variance in Paired Experiments," forthcoming, *Annales d'Economie et de Statistique*.

Abadie, A., and G. Imbens, (2008b), "Bias Corrected Matching Estimators for Average Treatment Effects," unpublished manuscript, Harvard University.

Abadie, A., D. Drukker, H. Herr, and G. Imbens, (2003), "Implementing Matching Estimators for Average Treatment Effects in STATA," *The Stata Journal*, 4(3), 290-311.

Abadie, A., J. Angrist, and G. Imbens, (2002), "Instrumental Variables Estimation of Quantile Treatment Effects," *Econometrica*. Vol. 70, No. 1, 91-117.

Abbring, J., and G. van den Berg, (2003), "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 71(5): 1491-1517.

Abbring, J., and J. Heckman, (2007), "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation, "in J. Heckman and E. Leamer eds. *Handbook of Econometrics*, vol. 6B, Chapter 72, 5144-5303. New York: Elsevier Science.

Andrews, D., and G. Soares, (2007), "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," Unpublished Manuscript, Department of Economics, Yale University.

Angrist, J., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

Angrist, J., (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66(2), 249-288.

Angrist, J., (2004), "Treatment Effect Heterogeneity in Theory and Practice," *Economic Journal*.

Angrist, J., E. Bettinger and M. Kremer, (2005), "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia,", *American Economic Review*, forthcoming.

Angrist, J., K. Graddy and G. Imbens, (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *Review of Economics Studies* 67(3):499-527.

Angrist, J., G. Imbens and D. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J., and A. Krueger, (1991), "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics*, 106, 979-1014.

Angrist, J. D. and A. B. Krueger (2000), "Empirical Strategies in Labor Economics," in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

Angrist, J., and K. Lang, (2004), "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program", *American Economic Review*, 94(5), 1613-1634.

Angrist, J. D., and J. Hahn, (2004) "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," forthcoming, *Review of Economics and Statistics*.

Angrist, J., and V. Lavy (1999), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics*, Vol. CXIV, 1243.

Angrist, J., and S. Pischke (2008), *Mostly Harmless Econometrics: An Empiricists' Companion*, Princeton University Press, Princeton, NJ.

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47-57.

Ashenfelter, O., and D. Card, (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, 67, 648-660.

Athey, S., and G. Imbens (2006), "Identification and Inference in Nonlinear Difference-In-Differences Models," *Econometrica*, 74(2).

Athey, S., and S. Stern, (1998), "An Empirical Framework for Testing Theories About Complementarity in Organizational Design," NBER working paper 6600.

Attanasio, O., C. Meghir, C. and A. Santiago (2001), "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to evaluate Progresa," UCL Mimeo.

Austin, P., (2008a), "A Critical Appraisal of Propensity-score Matching in the Medical Literature Between 1996 and 2003,"*Statistics in Medicine*, 27(12): 2037-2049.

Austin, P., (2008b), "Rejoinder: Discussion of 'A Critical Appraisal of Propensity-score Matching in the Medical Literature Between 1996 and 2003',"*Statistics in Medicine*, 27(12): 2066-2069.

Balke, A., and J. Pearl, (1994), "Nonparametric Bounds of Causal Effects from Partial Compliance Data," Technical Report R–199–J, Computer Science Department, University of California, Los Angeles.

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

Barnow, B.S., G.G. Cain and A.S. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in *Evaluation Studies* , vol. 5, ed. by E. Stromsdorfer and G. Farkas. San Francisco: Sage.

Becker, S., and A. Ichino, (2002), "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, 2(4): 358-377.

Behncke, S., M. Froelich, and M. Lechner (2006) "Statistical Assistance for Programme Selection - For a Better Targeting of Active Labour Market Policies in Switzerland," Working Paper No. 2007-05, University of St Gallen Law School.

Beresteanu, A., and F. Molinari, (2006), "Asymptotic Properties for a Class of Partially Identified Models," Unpublished Manuscript, Department of Economics, Cornell University.

Bertrand, M., and S. Mullainathan, (2004): "Are Emily and Brandon more Employable than Latoya and Tyrone? Evidence on Racial Discrimination in the Labor Market from a Large Randomized Experiment," *American Economic Review*, .

Bertrand, M., E. Duflo, and S. Mullainathan, (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, Vol 119(1), 249-275.

Besley, T., and A. Case, (2000), "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," *Economic Journal* v110, n467 (November): F672-94.

Bierens, H. (1987), "Kernel Estimators of Regression Functions," *Advances in Econometrics*, Fifth Worldcongress, Vol 1, Bewley (ed.), 99-144.

Bitler, M., J. Gelbach, and H. Hoynes (2002) "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," unpublished paper, Department of Economics, University of Maryland.

Björklund, A. and R. Moffitt, (1987), "The Estimation of Wage Gains and Welfare Gains in Self–Selection Models," *Review of Economics and Statistics*, Vol. LXIX, 42–49.

Black, S., (1999), "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, Vol. CXIV, 577.

Bloom, H., (1984), "Accounting for No–shows in Experimental Evaluation Designs," *Evaluation Review*, 8(2) 225–246.

Bloom, H., (ed.), (2005), *Learning More from Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation, New York, New York.

Blundell, R., M. Costa Dias, C. Meghir, and J. Van Reenen, (2001), "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program," Working paper WP01/20, IFS.

Blundell, R. and M. Costa-Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics," Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Blundell, R., A. Duncan and C. Meghir, (1998), "Estimating Labour Supply Responses Using Tax Policy Reforms," *Econometrica,* 6 (4), 827-861.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir, (2002) "Changes in the Distribution of Male and Female Wages Accounting for the Employment Composition," unpublished paper, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, United Kingdom.

Blundell, R., and T. MaCurdy, (2000): "Labor Supply," *Handbook of Labor Economics*, O. Ashenfelter and D. Card, eds., North Holland: Elsevier, 2000, 1559-1695.

Brock, B., and S. Durlauf., (2000), "Interaction-Based Models," NBER Technical Working Paper 258.

Bruhn, M., and D. McKenzie, (2008), "In Pursuit of Balance: Randomization in Practice in Development Economics ", Unpublished Manuscript, World Bank.

Caliendo, M., (2006), *Microeconometric Evaluation of Labour Market Policies*, Springer Verlag, Berlin.

Canay, I. (2007), "Empirical Likelihood Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," Unpublished Manuscript, Department of Economics, Northwestern University.

Card, D., (1990), "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43, 245-257.

Card, D., (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69, 1127-1160.

Card, D., C. Dobkin and N. Maestas, 2004, The Impact of Nearly Universal Insurance Coverage on Health Care Utilization and Health: Evidence from Medicare, NBER Working Paper No. 10365.

Card, D., and D. Hyslop, (2005): "Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare Leavers," *Econometrica* 73(6).

Card, D., and A. Krueger, (1993), "Trends in Relative Black-White Earnings Revisited," *American Economic Review*, vol. 83, no. 2, 85-91.

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), 772-784.

Card, D., and P. Levine, (1994), "Unemployment Insurance Taxes and the Cyclical Properties of Employment and Unemployment," *Journal of Public Economics*, vol. 53, 1-29.

Card, D., and B. McCall, (1996), "Is Workers' Compensation Covering Uninsured Medical Costs? Evidnece from the 'Monday Effect'," *Industrial and Labor Relations Review*, vol. 49, 690-706.

CARD, D., AND P. ROBBINS, (1996), "Do Financial Incentives Encourage Welfare Receipients to Work? Evidence from a Randomized Evaluation of the Self-Sufficiency Project," NBER working paper No 5701.

CARD, D., AND D. SULLIVAN, (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica*, vol. 56, no. 3, 497-530.

CASE, A., AND L. KATZ, (1991), "The Company You Keep, The Effects of Family and Neighbordhood Disadvantaged Families," NBER Working Paper 3705.

CHAMBERLAIN, G. (1986), "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, Vol 32(2), 189-218.

CHAY, K., AND M. GREENSTONE, (2005) "Does Air Quality Matter? Evidence from the Housing Market," *Journal of Political Economy*, 103(2).

CHEN, S., AND W. VANDERKLAAUW (2008) "The work disincentive effects of the disability insurance program in the 1990s," *Journal of Econometrics*, Vol 142(2): 757-784.

CHEN, X., (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models, "in J. Heckman and E. Leamer eds. Handbook of Labor Economics, vol. 6B, Chapter 76, 5549-5632. New York: Elsevier Science.

CHEN, X., H. HONG, AND . TAROZZI, (2005), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," unpublished working paper, Department of Economics, New York University.

CHERNOZHUKOV, V., AND C. HANSEN, (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245-261.

CHERNOZHUKOV, V., H. HONG AND E. TAMER, (2005), "Parameter Set Inference in a Class of Econometric Models," forthcoming, *Econometrica*.

CHESHER, A., (2003), "Identification in Nonseparable Models," *Econometrica* 71(5), 1405-1441.

COCHRAN, W., (1968) "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies", *Biometrics* 24, 295-314.

COCHRAN, W., AND D. RUBIN (1973) "Controlling Bias in Observational Studies: A Review" *Sankhya*, 35, 417-46.

COOK, T., (2008), "Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics, *Journal of Econometrics*. Vol 142(2):636-654

CRUMP, R., V. J. HOTZ, V. J., G. IMBENS, AND O.MITNIK, (2008a), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," forthcoming *Biometrika*.

CRUMP, R., V. J. HOTZ, V. J., G. IMBENS, AND O.MITNIK, (2008b), "Nonparametric Tests for Treatment Effect Heterogeneity," forthcoming, *Review of Economics and Statistics*.

DAVISON, A., AND D. HINKLEY (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, UK.

DEHEJIA, R., (2002) "Was there a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data", *Journal of Business and Economic Statistics* 21(1): 1-11.

DEHEJIA, R., AND S. WAHBA, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94, 1053-1062.

DEHEJIA, R. (2003) "Practical Propensity Score Matching: A Reply to Smith and Todd," forthcoming *Journal of Econometrics*.

DEHEJIA, R. (2005) "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141-173.

DIAMOND, A., AND J. SEKHON, J., (2008), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," Mimeo, Department of Political Science, University of California at Berkeley.

DOKSUM, K., (1974), "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *Annals of Statistics*, 2, 267-277.

Donald, S. and K. Lang, (2007), "Inference with Difference in Differences and Other Panel Data," forthcoming, *Review of Economics and Statistics*, Vol. 89(2): 221-233.

Duflo, E., (2001), "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review 91, 795-813.*

Duflo, E., and R. Hanna, (2006), "Monitoring Works: Getting Teachers to Come to School," NBER Working Paper 11880.

Duflo, E., R. Glennester, and M. Kremer, (2007), "Using Randomization in Development Economics Research: A Toolkit," *Handbook of Development Economics*, forthcoming.

Efron, B., and R. Tibshirani, (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Eissa, N., and J. Liebman, (1996): "Labor Supply Response to the Earned Income Tax Credit," *Quarterly Journal of Economics*, vol 111(2): 605-37.

Engle, R., D. Hendry, and J.-F. Richard, (1983) "Exogeneity," *Econometrica*, 51(2): 277-304.

Fan, J. and I. Gijbels, (1996), *Local Polynomial Modelling and Its Applications* (Chapman and Hall, London).

Firpo, S. (2006), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259-276.

Fisher, R. A., (1925), *The Design of Experiments*, 1st ed, Oliver and Boyd, London.

Fitzgerald, J., P. Gottschalk, and R. Moffitt, (1998), "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics", *Journal of Human Resources* 33, 251–299.

Flores, C. (2005), PhD Thesis, Chapter 2, Department of Economics, University of California, Berkeley.

Fraker, T., and R. Maynard, (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs", *Journal of Human Resources*, Vol. 22, No. 2, p 194–227.

Friedlander, D., and P. Robins, (1995), "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods", *American Economic Review*, Vol. 85, p 923–937.

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

Frölich, M. (2002), "A Note on the Role of the Propensity Score for Estimating Average Treatment Effects," *Econometric Reviews* Vol 23(2): 167-174.

Gill, R., and J. Robins, J., (2001), "Causal Inference for Complex Longitudinal Data: The Continuous Case," *Annals of Statistics*, 29(6): 1785-1811.

Glaeser, E., B. Sacerdote, and J. Scheinkman, (1996), "Crime and Social Interactions," *Quarterly Journal of Economics*, 111(2), 507-548.

Goldberger, A. (1972a), "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations,"Unpublished manuscript, Madison, WI.

Goldberger, A. , (1972b), "Selection Bias in Evaluating Treatment Effects: The case of interaction,"Unpublished manuscript, Madison, WI.

Gozalo, J., and O. Linton (2003), "Conditional Independence Restrictions: Testing and Estimation," unpublished manuscript, London School of Economics.

Graham, B. (2006), "Identifying Social Interactions Through Conditional Variance Restrictions,"Department of Economics, University California at Berkeley.

Graham, B., G. Imbens, and G. Ridder (2006), "Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis,"Department of Economics, University California at Berkeley.

Gu, X., and P. Rosenbaum, (1993), "Comparison of Multivariate Matching Methods: Structures, Distances and Algorithms", *Journal of Computational and Graphical Statistics*, 2, 405-20.

HAHN, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

HAHN, J., P. TODD, AND W. VANDERKLAAUW, (2000), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1): 201-209.

HAM, J., AND R. LALONDE, (1996) "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training," *Econometrica*, 64: 1.

HAMERMESH, D., AND J. BIDDLE (1994) "Beauty and the Labor Market," *American Economic Review*, 84(5): 1174-1194.

HANSEN, B., (2008), "The Essential Role of Balance Tests in Propensity-score Matched Observational Studies: Comments on 'A Critical Appraisal of Propensity-score Matching in the Medical Literature Between 1996 and 2003' by Peter Austin, *Statistics in Medicine*,"*Statistics in Medicine*, 27(12): 2037-2049.

HANSEN, C., (2007a), "Asymptotic Properties of Robust Variance Matrix Estimator for Panel Data when T is Large ," *Journal of Econometrics.*

HANSEN, C., (2007b), "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects," *Journal of Econometrics.*

HANSON, S., AND A. SUNDERAM (2008), "The Variance of Average Treatment Eect Estimators in the Presence of Clustering,"Department of Economics, Harvard University.

HÄRDLE, W. (1991), *Applied Nonparametric Regression* Cambridge University Press, Cambridge, UK.

HECKMAN, J. (1990), "Varieties of Selection Bias," *American Economic Review*, Papers and Proceedings, 80, 313-318.

HECKMAN, J., AND J. HOTZ, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association.*, Vol. 84, No. 804, 862-874.

HECKMAN, J., AND R. ROBB, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

HECKMAN, J., J. SMITH, AND N. CLEMENTS, (1997), "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts", *Review of Economic Studies*, Vol 64, 487-535.

HECKMAN, J., H. ICHIMURA, AND P. TODD, (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program,"*Review of Economic Studies* 64, 605-654.

HECKMAN, J., H. ICHIMURA, AND P. TODD, (1998), "Matching as an Econometric Evaluation Estimator,"*Review of Economic Studies* 65, 261–294.

HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD, (1998), "Characterizing Selection Bias Using Experimental Data,"*Econometrica* 66, 1017-1098.

HECKMAN, J., R. LALONDE, AND J. SMITH (2000), "The Economics and Econometrics of Active Labor Markets Programs,"in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

HECKMAN, J., L. LOCHNER, AND TABER, (1999), "Human Capital Formation and General Equilbrium Treatment Effects: A Study of Tax and Tuition Policy,"*Fiscal Studies* 20(1), 25-40.

HECKMAN, J., S. URZUA, AND E. VYTLACIL, (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity,"*Review of Economics and Statistics*, 88(3): 389-432.

HECKMAN, J. AND E. VYTLACIL, (2006), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation,"*Econometrica*, 73(3): 669-738.

HECKMAN, J., AND E. VYTLACIL (2007a), "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation,"in J. Heckman and E. Leamer eds. *Handbook of Econometrics*, vol. 6B, Chapter 70, 4779-4874. New York: Elsevier Science.

Heckman, J., and E. Vytlacil (2007b), "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments, "in J. Heckman and E. Leamer eds. Handbook of Econometrics, vol. 6B, Chapter 71, 4875-5143. New York: Elsevier Science.

Hill, J., (2008), "Discussion of Research Using Propensity-score Matching: Comments on 'A Critical Appraisal of Propensity-score Matching in the Medical Literature Between 1996 and 2003' by Peter Austin, *Statistics in Medicine*,"*Statistics in Medicine*, 27(12): 2055-2061.

Hirano, K., and G. Imbens (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Hear Catherization,"*Health Services anf Outcomes Research Methodology*, 2, 259-278.

Hirano, K., and G. Imbens (2004). "The propensity score with continuous treatments,"*Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*: 73 - 84 (A. Gelman & X.L. Meng, Eds.). New York: Wiley.

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,"*Econometrica*, 71(4): 1161-1189. July

Hirano, K., and J. Porter, (2005), "Asymptotics for Statistical Decision Rules,"Working Paper, Dept of Economics, University of Wisconsin.

Holland, P., (1986), "Statistics and Causal Inference," (with discussion), *Journal of the American Statistical Association*, 81, 945-970.

Horowitz, J., (2002), "The Bootstrap,"*Handbook of Econometrics*, Vol. 5, Heckman and Leamer (eds.), Elsevier, North Holland.

Horowitz, J., and C. Manski, (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data,"*Journal of the American Statistical Association.*

Horvitz, D., and D. Thompson, (1952), "A Generalization of Sampling Without Replacement from a Finite Universe,"*Journal of the American Statistical Association*, 47, 663-685.

Hotz, J., C. Mullin, and S. Sandrers, (1997), "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analysis of Teenage Childbearing,"*Review of Economic Studies* 64(4), 575-603.

Hotz, V. J., G. Imbens, and J. Klerman, (2006), "The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program,"*Journal of Labor Economics*.

Hotz J., G. Imbens, and J. Mortimer (2005), "Predicting the Efficacy of Future Training Programs Using Past Experiences,"*Journal of Econometrics*, 125(1-2), 241-270.

Iacus, S., G. King, and G. Porro, (2008), "Matching for Causal Inference Without Balance Checking,"Unpublished Manuscript, IQSS, Harvard University.

Ichimura, H., and O. Linton, (2005), "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators."in D. Andrews and J. Stock (eds.), *Identification and Inference for Econometric Models*, Chapter 8, 149-170, Cambridge University Press, Cambridge.

Ichimura, H., and C. Taber, (2000), "Direct Estimation of Policy Effects", unpublished manuscript, Department of Economics, Northwestern University.

Ichimura, H., and P. Todd (2007), "Implementing Nonparametric and Semiparametric Estimators, "in J. Heckman and E. Leamer eds. Handbook of Econometrics, vol. 6B, Chapter 74, 5369-5468. New York: Elsevier Science.

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

Imbens, G. (2003), "Sensivity to Exogeneity Assumptions in Program Evaluation,"*American Economic Review*, Papers and Proceedings, 93(2), 126-132.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,"*Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G. (2007): "Nonadditive Models with Endogenous Regressors," in *Advances in Econometrics*, Blundell, Newey and Persson (eds.), Cambridge University Press.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G., G. King, D. McKenzie, and G. Ridder, (2008), "On the Benefits of Stratification in Randomized Experiments," unpublished manuscript, Department of Economics, Harvard University.

Imbens, G., and T. Lemieux, (2008) "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics* Vol 142(2):615-635.

Imbens, G., and W. Newey (2002) "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity", NBER Technical Working Paper 285.

Imbens, G., W. Newey and G. Ridder, (2003), "Mean-squared-error Calculations for Average Treatment Effects," unpublished manuscript, Department of Economics, UC Berkeley.

Imbens, G., and K. Kalyanaraman, (2008), "Optimal Bandwidth Selection in Regression Discontinuity Designs," unpublished manuscript, Department of Economics, Harvard University.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, Vol 72, No. 6, 1845-1857.

Imbens, G. W., and D. B. Rubin, (1997a), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance," *Annals of Statistics*, Vol. 25, No. 1, 305–327.

Imbens, G. W., and D. B. Rubin, (1997b): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models,," *Review of Economic Studies*, 64, 555-574.

Imbens, G., D. Rubin, and B. Sacerdote, (2001), "Estimating the Effect of Unearned Income on Labor Supply, Earnings, Savings and Consumption: Evidence from a Survey of Lottery Players," *American Economic Review* 91, 778-794.

Jin, G., and P. Leslie, (2003), "The Effect of Information on Product Quality: Evidence from Restaurants Hygiene Grade Cards," *The Quarterly Journal of Economics*, 118(2) 409-451.

Kitagawa, T., (2008), "Identification Bounds of the Local Average Treatment Effect," Unpublished Manuscript, Department of Economics, Brown University.

Kling, J., J. Liebman, and L. Katz (2007): "Experimental Analysis of Neigborhood Effects," *Econometrica*, 75(1), 83-119.

Lalive, R., (2008), "How do extended benefits affect unemployment duration? A regression discontinuity approach", *Journal of Econometrics*,Vol 142(2): 785-806.

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604-620.

Lechner, M, (1999), "Earnings and Employment Effects of Continuous Off-the-job Training in East Germany After Unification," *Journal of Business and Economic Statistics*, 17(1), 74-90

Lechner, M, (2002a), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *Review Economics and Statistics*, 84(2): 205-220, May.

Lechner, M, (2002b), "Some Practical Issues in the Evaluation of Heterogenuous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society*, Series A, 165: 659–82.

Lechner, M., (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in Lechner and Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*, Heidelberg, Physica.

Lechner, M., (2004), "Sequential Matching Estimation of Dynamic Causal Effects," Discussion Paper 2004-06, Department of Economics, University of St Gallen.

Lechner, M., and R. Miquel, (2005), "Identification of Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions," Discussion Paper 2005-17, Department of Economics, University of St Gallen.

Lee, D, (2001), "The Electoral Advantage of Incumbency and the Voter's Valuation of Political Experience: A Regression Discontinuity Analysis of Close Elections," unpublished manuscript, Department of Economics, University of California.

Lee, D.S., (2008), "Randomized Experiments from Non-random Selection in U.S. House Elections", *Journal of Econometrics*,Vol 142(2): 675-697.

Lee, D.S. and D. Card, (2008), "Regression Discontinuity Inference with Specification Error", *Journal of Econometrics*, Vol 142(2): 655-674.

Lee, D.S. and T. Lemieux, 2008, "", Unpublished Manuscript.

Lee, D.S., Moretti, E., and M. Butler, 2004, Do Voters Affect or Elect Policies? Evidence from the U.S. House, *Quarterly Journal of Economics* 119, 807-859.

Lee, M.-J., (2005a), *Micro-Econometrics for Policy, Program, and Treatment Effects* Oxford University Press, Oxford.

Lee, M.-J., (2005b), "Treatment Effect and Sensitivity Analysis for Self-selected Treatment and Selectively Observed Response," unpublished manuscript, School of Economics and Social Sciences, Singapore Management University.

Lehman, E., (1974), *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco.

Lemieux, T., and K. Milligan, (2008), "Incentive effects of social assistance: A regression discontinuity approach," *Journal of Econometrics*,Vol 142(2): 807-828.

Ludwig, J., and D. Miller, (2005), Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, NBER working paper 11702.

Ludwig, J., and D. Miller, (2007), Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, Quarterly Journal of Economics

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C., (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531-542.

Manski, C. (1995), *Identification Problems in the Social Sciences*, Cambridge, Harvard University Press.

Manski, C., (2000a), "Economic Analysis of Social Interactions," *Journal of Economic Perspectives*, 14(3), 115-136.

Manski, C., (2000b), "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treaetment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics*, 95, 415-442.

Manski, C., (2001), "Designing Programs for Heterogenous Populations: The Value of Covariate Information," *American Economic Review Papers and Proceedings*, 91, 103-1-6.

Manski, C., (2002), "Treatment Choice Under Ambiguity Induced by Inferential Problems," *Journal of Statistical Planning and Inference*, 105, 67-82.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., (2004), "Statistical Treatment Rules for Heterogenous Populations," *Econometrica*, 72(4), 1221-1246.

Manski, C. (2005), *Social Choice with Partial Knowledge of Treatment Response*, Princeton, Princeton University Press.

Manski, C. (2008), *Identification for Prediction and Decision*, Princeton, Princeton University Press.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

Manski, C., and J. Pepper, (2004), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997-1010.

MATZKIN, R., (2003), "Nonparametric Estimation of Nonadditive Random Functions", *Econometrica* 71, 1339-1375.

MCCRARY, J., 2007, Testing for Manipulation of the Running Variable in the Regression Discontinuity Design, *Journal of Econometrics*, this issue.

MEALLI, F., G. IMBENS, S. FERRO, AND A. BIGGERI, (2004), "Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes", *Biostatistics* , 5(2), 207-222.

MCEWAN, P., AND J. SHAPIRO, (2007),The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates using exact Birth Dates," Unpublished Manuscript.

MEYER, B., K. VISCUSI AND D. DURBIN, (1995), "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, Vol. 85, No. 3, 322-340.

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

MORGAN, S. AND C. WINSHIP, (2007), *Counterfactuals and Causal Inference*, Cambridge University Press, Cambridge.

MOULTON, B., (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Ecnomics and Statistics*, 334-338.

MOULTON, B., AND W. RANDOLPH, (1989) "Alternative Tests of the Error Component Model," *Econometrica*, Vol. 57, No. 3, 685-693.

NEYMAN, J., (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,"translated in *Statistical Science*, (with discussion), Vol 5, No 4, 465–480, 1990.

NEWEY, W. (1994a). "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory* 10 (2): 233 - 253.

NEWEY, W. (1994b). "Series Estimation of Regression Functionals," *Econometric Theory* 10 (2): 1-28.

OLKEN, B. (2007), "Monitoring Corruption: Evidence from a Field Experiment in Indonesia,"*Journal of Political Economy* 115(2): 200-49.

PAGAN, A. AND A. ULLAH, (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

PAKES, A., J. PORTER, K. HO, AND J. ISHII (2006), "Moment Inequalities and Their Application,"Unpublished Manuscript.

PEARL, J., (2000), *Causality: Models, Reasoning and Inference*, Cambridge, Cambridge University Press.

POLITIS, D., J. ROMANO,, AND M. WOLFF, (1999), *Subsampling*, Springer Verlag.

PORTER, J. (2003), "Estimation in the Regression Discontinuity Model,"Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison.

QUADE, D., (1982), "Nonparametric Analysis of Covariance by Matching", *Biometrics*, 38, 597-611.

ROBINS, J., AND Y. RITOV, (1997), "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models,"*Statistics in Medicine* 16, 285-319.

ROBINS, J.M., AND A. ROTNITZKY, (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data,"*Journal of the American Statistical Association*, 90, 122-129.

ROBINS, J.M., ROTNITZKY, A., ZHAO, L-P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data,"*Journal of the American Statistical Association*, 90, 106-121.

ROBINSON, P., (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 67: 645-662.

ROMANO, J., AND A. SHAIKH (2006a), "Inference for Identifiable Parameters in Partially Identified Econometric Models,"unpublished manuscript, Department of Statistics, Stanford University.

ROMANO, J., AND A. SHAIKH (2006b), "Inference for the Identied Setin Partially Identified Econometric Models,"unpublished manuscript, Department of Statistics, Stanford University.

ROSEN, A., (2005), "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," Unpublished Manuscript, Department of Economics, University College London.

ROSENBAUM, P., (1984a), "Conditional Permutation Tests and the Propensity Score in Observational Studies,"*Journal of the American Statistical Association*, 79, 565-574.

ROSENBAUM, P., (1984b), "The Consequences of Adjustment for a Concomitant Variable that has been Affected by the Treatment,"*Journal of the Royal Statistical Society*, Series A, 147, 656-666.

ROSENBAUM, P., (1989), "Optimal Matching in Observational Studies", *Journal of the American Statistical Association*, 84, 1024-1032.

ROSENBAUM, P., (1987), "The role of a second control group in an observational study", *Statistical Science*, (with discussion), Vol 2., No. 3, 292–316.

ROSENBAUM, P., (1995), *Observational Studies*, Springer Verlag, New York.

ROSENBAUM, P., (2002), "Covariance Adjustment in Randomized Experiments and Observational Studies,"*Statistical Science*, 17(3): 286-304.

ROSENBAUM, P., AND D. RUBIN, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

ROSENBAUM, P., AND D. RUBIN, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome,"*Journal of the Royal Statistical Society*, Ser. B, 45, 212-218.

ROSENBAUM, P., AND D. RUBIN, (1984), "Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score", *Journal of the American Statistical Association*, 79, 516-524.

ROSENBAUM, P., AND D. RUBIN, (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score,"*American Statistician*, 39, 33-38.

ROTNITZKY, A., AND J. ROBINS, (1995), "Semiparametric Regression Estimation in the Presence of Dependent Censoring,"*Biometrika*, Vol. 82, No. 4, 805-820.

ROY, A., (1951), "Some Thoughts on the Distribution of Earnings,"*Oxford Economics Papers*, 3, 135-146.

RUBIN, D., (1973a), "Matching to Remove Bias in Observational Studies", *Biometrics*, 29, 159-183.

RUBIN, D., (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies", *Biometrics*, 29, 185-203.

RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies,"*Journal of Educational Psychology*, 66, 688-701.

RUBIN, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate,"*Journal of Educational Statistics*, 2(1), 1-26.

RUBIN, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *A*nnals of Statistics, 6:34–58.

RUBIN, D., (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies,"*Journal of the American Statistical Association*, 74, 318-328.

RUBIN, D. B., (1990), "Formal Modes of Statistical Inference for Causal Effects,"*J*ournal of Statistical Planning and Inference, 25, 279-292.

RUBIN, D. (1997), "Estimating Causal Effects from Large Data Sets Using Propensity Scores,"*Annals of Internal Medicine*, 127, 757-763 .

RUBIN, D. (2006), *Matched Sampling for Causal Effects*, Cambridge University Press, Cambridge, UK.

RUBIN, D., AND N. THOMAS, (1992a), "Affinely Invariant Matching Methods with Ellipsoidal Distributions,"*Annals of Statistics* 20 (2) 1079-1093.

Rubin, D., and N. Thomas, (1992b), "Characterizing the effect of matching using linear propensity score methods with normal distributions," *Biometrika* 79 797-809.

Rubin, D., and N. Thomas, (1996), "Matching Using Estimated Propensity Scores,"*Biometrics* 52 249-264.

Rubin, D., and N. Thomas, (2000), "Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates,"*Journal of the American Statistical Association*.

Sacerdote, B., (2001, "Peer Effects with Random Assignment: Results for Dartmouth Roommates,"*Quarterly Journal of Economics*, 116(2), 681-704.

Scharfstein, Robins, and Rotnizky (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models,"with comments and rejoinder *Journal of the American Statistical Association*, 94(448), 1096-1146.

Schulz, T., (2001), "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program", Center Discussion Paper 834, Economic Growth Center.

Sekhon, J., (2008), "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R,"Forthcoming, *Journal of Statistical Software*.

Sekhon, J., and R. Grieve (2007), "A new non-parametric matching method for bias adjustment with applications to economic evaluations,"Mimeo, Department of Political Science, University of California at Berkeley.

Seifert, B., and T. Gasser (1996), "Finite-sample Variance of Local Polynomials: Analysis and Solutions,"*Journal of the American Statistical Association*, 91, 267-275.

Seifert, B., and T. Gasser (2000), "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, 9(2): 338-360.

Sianesi, B., (2001), "psmatch: propensity score matching in STATA", University College London, and Institute for Fiscal Studies.

Shadish, W., T. Campbell and D. Cook, (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton and Mifflin, Boston.

Smith, J. A. and P. E. Todd, (2001), "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods,"*American Economic Review*, Papers and Proceedings, 91:112-118.

Smith, J. A. and P. E. Todd, (2005), "Does Matching Address LaLonde's Critique of Nonexperimental Estimators,"*Journal of Econometrics*, 125(1-2), 305-353.

Stock, J., (1989), "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84(406): 567-575.

Stone, C., (1977), "Consistent Nonparametric Regression", (with discussion) *Annals of Statistics*, vol. 5, 595–645.

Stoye, J., (2007), "More on Confidence Intervals for Partially Identified Models,"Unpublished Manuscript, Department of Economics, New York University.

Stuart, E., (2008), "Developing Practical Recommendations for the Use of Propensity Scores: Discussion on 'A Critical Appraisal of Propensity-score Matching in the Medical Literature Between 1996 and 2003' by Peter Austin, *Statistics in Medicine*,"*Statistics in Medicine*, 27(12): 2062-2065.

Thistlewaite, D., and D. Campbell, (1960), "Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment,"*Journal of Educational Psychology* 51, 309-317.

Trochim, W., (1984), *Research Design for Program Evaluation; The Regression-discontinuity Design* (Sage Publications, Beverly Hills, CA).

Trochim, W., (2001), "Regression-Discontinuity Design,"in N.J. Smelser and P.B Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences* 19 (Elsevier North-Holland, Amsterdam) 12940-12945.

VanderKlaauw, W., (2002), "A Regression–discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment,"*International Economic Review*, 43(4): 1249-1287

VanderKlaauw, W., (2008a), "RegressionDiscontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22(2): 219-245.

VanderKlaauw, W., (2008b), "Breaking the link between poverty and low student achievement: An evaluation of Title I," *Journal of Econometrics*, Vol 142(2): 731-756.

Van Der Laan, M., and J. Robins, (2003) *Unified Methods for Censored Longitudinal Data and Causality*, Springer, Berlin.

Vytlacil, E., (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331-341.

Wooldridge, J., (1999), "Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples," *Econometrica*, 67, 1385-1406.

Wooldridge, J., (2002a), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Wooldridge, J., (2002b), "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal*, 1, 117-139.

Wooldridge, J., (2007), "Inverse Probability Weighted M-Estimators for General Missing Data Problems," forthcoming, *Journal of Econometrics*.

Zhao, Z., (2004), "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics and an Application," forthcoming, *review of economics and statistics*.

Table 1: P-values for Fisher Exact Tests: Ranks versus Levels

Program	Location	sample size controls	sample size treated	p-values t-test	p-values FET (levels)	p-values FET (ranks)
GAIN	Alameda	601	597	0.835	0.836	0.890
GAIN	Los Angeles	1400	2995	0.544	0.531	0.561
GAIN	Riverside	1040	4405	0.000	0.000	0.000
GAIN	San Diego	1154	6978	0.057	0.068	0.018
WIN	Arkansas	37	34	0.750	0.753	0.805
WIN	Baltimore	260	222	0.339	0.339	0.286
WIN	San Diego	257	264	0.136	0.137	0.024
WIN	Virginia	154	331	0.960	0.957	0.249

Table 2: Balance Improvements in the Lalonde Data (Dehejia-Wahba Sample)

	CPS Controls (15992)		NSW Treated 185		Normalized Difference Treated-Controls			
					All	$\hat{e}(X_i) \geq \underline{e}_1$	P-score	Maha
Covariate	mean	(s.d.)	mean	(s.d.)	(16177)	(6286)	(370)	(370)
age	33.23	(11.05)	25.82	(7.16)	-0.56	-0.25	-0.08	-0.16
education	12.03	(2.87)	10.35	(2.01)	-0.48	-0.30	-0.02	-0.09
married	0.71	(0.45)	0.19	(0.39)	-0.87	-0.46	-0.01	-0.20
nodegree	0.30	(0.46)	0.71	(0.46)	0.64	0.42	0.08	0.18
black	0.07	(0.26)	0.84	(0.36)	1.72	1.45	-0.02	0.00
hispanic	0.07	(0.26)	0.06	(0.24)	-0.04	-0.22	-0.02	0.00
earn '74	14.02	(9.57)	2.10	(4.89)	-1.11	-0.40	-0.07	0.00
earn '74 positive	0.88	(0.32)	0.29	(0.46)	-1.05	-0.72	-0.07	0.00
earn '75	13.65	(9.27)	1.53	(3.22)	-1.23	-0.35	-0.02	-0.01
earn '75 positive	0.89	(0.31)	0.40	(0.49)	-0.84	-0.54	-0.09	0.00

Table 3: Balance Improvements in the Lottery Data

	Losers (259)		Winners (237)		Normalized Difference Treated-Controls All	$0.0914 \leq \hat{e}(X_i) \leq 0.9086$
Covariate	mean	(s.d.)	mean	(s.d.)	(496)	(388)
Year Won	1996.4	(1.0)	1996.1	(1.3)	-0.19	-0.13
Tickets Bought	2.19	(1.77)	4.57	(3.28)	0.64	0.33
Age	53.2	(12.9)	47.0	(13.8)	-0.33	-0.19
Male	0.67	(0.47)	0.58	(0.49)	-0.13	-0.09
Years of Schooling	14.4	(2.0)	13.0	(2.2)	-0.50	-0.35
Working Then	0.77	(0.42)	0.80	(0.40)	0.06	-0.02
Earnings Year -6	15.6	(14.5)	12.0	(11.8)	-0.19	-0.10
Earnings Year -5	16.0	(15.0)	12.1	(12.0)	-0.20	-0.12
Earnings Year -4	16.2	(15.4)	12.0	(12.1)	-0.21	-0.15
Earnings Year -3	16.6	(16.3)	12.8	(12.7)	-0.18	-0.14
Earnings Year -2	17.6	(16.9)	13.5	(13.0)	-0.19	-0.15
Earnings Year -1	18.0	(17.2)	14.5	(13.6)	-0.16	-0.14
Pos Earnings Year -6	0.69	(0.46)	0.70	(0.46)	0.02	0.05
Pos Earnings Year -5	0.68	(0.47)	0.74	(0.44)	0.10	0.07
Pos Earnings Year -4	0.69	(0.46)	0.73	(0.44)	0.07	0.02
Pos Earnings Year -3	0.68	(0.47)	0.73	(0.44)	0.09	0.02
Pos Earnings Year -2	0.68	(0.47)	0.74	(0.44)	0.10	0.04
Pos Earnings Year -1	0.69	(0.46)	0.74	(0.44)	0.07	0.02

Table 4: Type by Observed Variables

		Z_i = 0	Z_i = 1
W_i	0	Nevertaker/Complier	Nevertaker
	1	Alwaystaker	Alwaystaker/Complier