Recent Developments in the Econometrics of Program Evaluation*

Guido W. Imbens† Jeffrey M. Wooldridge‡

First Draft: February 2005,
This Draft: October 2005

Abstract

JEL Classification: C14, C21, C52
Keywords:

*Financial support for this research was generously provided through NSF grants SES 0136789 and 0452590.

†Department of Economics, and Department of Agricultural and Resource Economics, University of California at Berkeley, 330 Giannini Hall, Berkeley, CA 94720-3880, and NBER. Electronic correspondence: imbens@econ.berkeley.edu, http://elsa.berkeley.edu/users/imbens/.

‡Department of Economics, Michigan State University. Electronic correspondence: wooldri1@msu.edu, http://www.msu.edu/ec/faculty/wooldridge/wooldridge.html

1 Introduction

Many empirical questions in economics and other social sciences center around the causal effect of programs or policies. In the last two decades much research has been done on the econometric and statistical analysis of the effect 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 a historical overview of the area and give references to more technical research.

The problem studied in this literature is that of evaluating 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, firms, or county, state or country administrations, but in other disciplines where these 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, 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 some comparison of 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 since 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 treatment we will therefore need to compare different units receiving the different levels of the treatment, that is either 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 (1986), Heckman and Robb (1984), Lalonde (1986) and Fraker and Maynard (). Motivated by the applications to the evaluation of labor market programs in observational settings the focus in the econometric literature is largely on endogeneity issues. Individuals who choose to enroll in a training program are likely to be different from those who choose not to enroll, possibly even after adjusting for observed covariates. As a result many of the advances concerned the use of 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.

The statistics literature starts from a different perspective. This literature originates in the analysis of randomized experiments by Fisher (1925) and Neyman (1923). Then from the early seventies Rubin (1973a,b, 1974, 1977, 1978) in a series of papers formulated the now dominant approach to analysis of causal effects in observational studies. Rubin's approach views causal statements as comparisons of so-called potential outcomes, pairs of outcomes defined for the

same unit given different exposures to the treatment of interest. Models are developed for the pair of potential outcomes rather than only for the observed outcome. Rubin then argues that the key issue is the assignment mechanism: what determines which level of the treatment a particular unit receives, and how is this related to the potential outcomes? Rubin's formulation of the evaluation problem or the problem of causal inference, labeled the "Rubin Causal Model" by Holland (1986), is now standard in both the statistics and econometrics literature.

A key role in the analysis of this problem is played by the assignment mechanism. Of particular importance is the relation between the assignment and the potential outcomes. The simplest case for analysis is that where assignment to treatment is randomized, and thus unrelated to the potential outcomes. In such classical randomized experiments it is straightforward to obtain attractive estimators for the average effect of the treatment. However, such experiments are rare in economics. In practice we typically analyze data from observational studies. In that case there is an important special case, variously referred to as unconfoundedness, exogeneity, ignorability, or selection on observables. All cases refer to some form of the assumption that adjusting treatment and control group for differences in observed covariates removes all biases in comparisons between treated and control units. This case is by now fairly well understood. The semiparametric efficiency bound has been calculated and various efficient estimators have been proposed. Without such an assumption there is no general approach. Various methods have been proposed for special cases. In this review we will discuss several of them. One approach (Rosenbaum and Rubin, 1983) consists of sensitivity analyses where robustness to limited departures from unconfoundedness are investigated. A second approach developed by Manski (1990, 2003) consists of bounds analyses where ranges of estimands consistent with the data and the limited assumptions the researcher is willing to make are estimated. A third approach, instrumental variables, relies on the presence of additional treatments that satisfy specific 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 approach, known as the regression discontinuity method has a long tradition in statistics but has recently been revived in the economics literature through work by VanderKlaauw (), Hahn, Todd, and VanderKlaauw (), Lee (), and Porter (). 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. A functional form free version of this is developed in Athey and Imbens (2005).

In this review we will discuss in detail some of new methods that have been developed in this literature. We will pay particular attention to the practical implications of these methods. By now we think that the literature has matured to the extent that it has much to offer to the empirical researcher. Although the evaluation problem is one where identification problems are important, there is now a much better understanding of assumptions are useful to consider, as well as a better set of methods for inference given various sets of assumptions.

We will largely limit this review to settings with binary treatments. This is in keeping

with the literature which has largely focused on this case. There are some extensions of these methods to multivalued and even continuous treatments but the work in this area is ongoing.

The running example we will use throughout the paper is that of a jobmarket training program. Such programs have been among the leading applications in the economics literature, starting with Ashenfelter (1978) and including Lalonde (1986) as an influential study. In such settings a number of individuals enroll or not in a training program, with labor market outcomes, e.g., earnings or employment status as the main outcome of interest. The 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 will be an important component of any analysis. In addition to observing participation status and outcome measures such as subsequent earnings or labor market status we may observe individual background characteristics such as education levels and age, as well as prior labor market histories including 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 Interactions, Potential Outcomes and the Assignment Mechanism

Suppose we wish to analyze a job training program using data for N individuals, indexed by $i = 1, \ldots, N$. Some of them have enrolled in the training program. Others either choose not to, or were not eligible to enroll in the program. We will use the indicator W_i to indicate whether individual i enrolled in the training program, with $W_i = 0$ indicating that individual i did not, and $W_i = 1$ indicating that individual i did enroll in the program.

2.1 Potential Outcomes

For each individual we postulate the existence of two potential outcomes, denoted by $Y_i(0)$ and $Y_i(1)$ for individual i. The first, $Y_i(0)$ denotes the outcome that would be realized by individual i if this individual did not participate in the program. Similarly, $Y_i(1)$ denotes the outcome that would be realized by individual i if this individual did participate in the program. Since individual i can either participate or not participate in the program, only one of these potential outcomes can be realized. 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. The preceeding discussion implies that

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

Table 1 illustrates that the essence of the causal inference problem is a missing data problem. We took five observations from an experimental evaluation of a labor market program in Los Angeles (part of the GAIN programs). Three of the individuals entered the program. For those

individuals we observe $Y_i(1)$, the earnings given the training, but not the earnings without the training. For the remaining two individuals we do not observe $Y_i(1)$ but we observe $Y_i(0)$, the earnings without the training program.

This distinction between the pair of potential outcomes $(Y_i(0), Y_i(1))$ and the realized outcome Y_i is very useful and has become the standard in statistical and econometric analyses of treatment effects. We will offer some comments on it. In the particular setting of estimation of treatment effects in observational studies that we are discussing in this review the framework was introduced and forcefully advocated in a series of papers by Rubin (1973, 1975, 1978). He points out the advantages of this set up for all studies of causality and made it a central part of his approach. The framework does, however, have important predecessors in a variety of other settings.

Most directly, in the context of randomized experiments the potential outcome framework was used previously by Neyman (19, translated in 198?) to derive the properties of estimators and confidence intervals under repeated sampling. It also has important antecedents in econometrics. Specifically, it is interesting to compare this distinction between potential and realized outcomes in Rubin's approach to the original work in econometrics on simultaneous equations models. Haavelmo (1943) 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. Although in many respects clearer than the subsequent literature, Haavelmo's distinction between realized and potential prices got blurred in the textbook discussion of simultaneous equations. In such discussions the starting point is often the general formulation of $Y\Gamma + XB = U$ for $N \times K$ matrices of realized outcomes Y, $N \times L$ matrices of exogenous covariates X and an $N \times K$ matrix of unobserved components U.

Potential outcomes are also used explicitly in labor market settings by Roy (195?). Roy models individual choosing from a set professions. Individuals know what their earnings would be in each of these professions and choose the one that maximizes their earnings. Here we see the explicit use of the potential outcomes, combined with a specific selection/assignment mechanism, namely choosing the one with the highest potential outcome.

The first advantage of the potential outcome set up is that it allows us to define causal effects before specifying or considering the assignment mechanism and without making functional form 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 comparisons. Such definitions do not require us to take a stand on whether the effect is constant or varies accross the population, or whether the assignment is endogenous or exogenous. In contrast, in terms of the realized outcomes the causal effects are difficult to define. Typically 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 are. The potential outcome approach separates all these issues and allows the researcher to define the causal effect

of interest without considering statistical properties of the outcomes or assignment.

The second advantage is that it links the analysis of causal effects to explicit manipulations. Considering the two potential outcomes forces the researcher to think about the scenario under which they could be observed, that is to consider what experiment could reveal the causal effects. Doing so clarifies the interpretation of causal effects. Let us elaborate this with a couple of examples. First consider the causal effects of ethnicity or race. Simply comparing economic outcomes by ethnicity is difficult to interpret as one can never be sure that all relevant differences between the two groups other than ethnicity have been adjusted for. Any resulting difference are ambiguous. Are they the causal effect of childhood experiences or the result of genetic differences? One can obtain much clearer causal interpretations by linking comparisons to specific manipulations. A recent example is the study by Betrand and Mullainathan (200?) who compare call back rates for job applications submitted with names that suggest African-American or Caucasian ethnicity. In that case there is a clear manipulation, namely 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 height (). What do the differences in, for example, earnings by ratings on a beauty scale represent? One interpretation is that they represent causal effects of plastic surgery? Such a manipulation would make the difference potentially causal, but it appears unclear whether cross-sectional correlations between beauty and earnings in the general population represent causal effects of plastic surgery. A third example is known as Lord's paradox in statistics. Lord () considered the evaluation of the effect of a new diet in a college dorm on student's weight. Weight was recorded for two groups, men and women, both at the start of the new year prior to the introduction of the new diet and at the end of the year after the diet was introduced. Lord shows that under some configuration of the data ... Rubin and Holland () argue that the difficulty in interpreting the results stems from the absence of data on any alternative treatment: all students were exposed to the new diet and inferring an effect of the diet is therefore completely dependent on strong assumptions regarding the outcome under the old diet. Assumptions one may justify ...

A third advantage is that separates the modelling of the potential outcomes from that of the assignment mechanism. Modelling the realized outcome is complicated by the fact that it depends both on the potential outcomes and on 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 outcome approach is that it allows us to formulate 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 (condi-

tional) independence assumptions involving the potential outcomes. Assessing their validity requires the researcher to consider the dependence structure if all potential outcomes where observed. In contrast, the traditional approach where the realized outcomes are modelled often formulates the critical assumptions in terms of residuals from 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 both functional form assumptions and substantive assumptions. This bundling makes these assumptions more difficult to evaluate.

A fifth advantage of the potential outcome approach is that it clarifies where the uncertainty comes from in the estimators considered. Even if we observe the entire (finite) population (and so we can estimate population averages without uncertainty), causal effects will always be uncertain because for each unit at least one of the two potential outcomes is not observed. We may still use superpopulation arguments to justify approximations to the finite sample distributions, but we do not requires such arguments to justify the existence of uncertainty about the causal effect.

In this review we will largely consider settings where there is no interference between units and no multiple versions of the treatment. The combination of these assumptions is what Rubin called SUTVA (stable unit treatment value assumption). The first part corresponds to ruling out interactions between units. In economic terms, we are looking at a partial equilibrium analysis, not a general equilibrium. This is obviously unrealistic in some settings, but in many cases it may be reasonable to assume that general equilibrium effects are small and that we can therefore ignore effects that receipt of the treatment for one individual has on outcomes for another individual.

2.2 The Assignment Mechanism

The second ingredient of Rubin's approach is the assignment mechanism. This is defined as the conditional probability of receiving the treatment as a function of potential outcomes and covariates. We distinguish three classes of assignment mechanisms, in order of increasing complexity of analysis. The first case is that of randomized experiments. In randomized experiments the probability of assignment to either 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 $M < N$ randomly choosen units are assigned to the treatment and the remaining $N - M$ units receive the control treatment. There are important variations on this 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. However, there are in practice few experiments in economics and most of them are of the completely randomized experiment variety so we shall mostly limit our discussion to this type of experiment. 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 discussion the analysis of randomized experiments, including more powerful randomization based methods for inference, in Section **??**.

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

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

However, the assignment probabilities are no longer assumed to be a known function of the covariates. We refer to this assignment mechanism as *unconfounded assignment*, following Rosenbaum and Rubin () who first articulate the assumption in this potential outcome form. The assumption that assignment mechanism or variations on it apply are also referred to by various other labels. These include *selection on observables* (Heckman, Ichimura, Smith and Todd), *exogeneity* (Manski, Sandefur, Powers and XXX, 19XX), *conditional independence* (Lechner (), Cameron and Trivedi (2005)). Although the analysis of data with such assignment mechanisms is not as simple as that of randomized experiments, there are now many methods available for this case. We will review them in detail in Section 5.

The third class of assignment mechanisms contains all remaining assignment mechanisms with some dependence on potential outcomes. 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 sofar. 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 will ignore them in this discussion. Instead we focus on cases where the dependende of the assignment mechanism on the potential outcomes creates substantive problems for the analysis. Here is no general solution. There are a number of special cases that are by now well understood, and we will discuss these in some detail in Section **??**. 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.

2.3 Interactions and General Equilibrium Effects

In most of the literature it is assumed that treatments received by one individual do not affect outcomes for another unit. Only the treatment applied to the specific individual are assumed to potentially affect this individual's outcomes. In the statistics literature this assumption is referred to as the Stable-Unit-Treatment-Value-Assumption (SUTVA, Rubin, XXXX). 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 of the biomedical applications. Whether one individual receives an active treatment for a stroke or not is unlikely to have a substantial impact on health outcomes for another 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 very cognizant of potential problems with this assumption. In order to minimize leaking of fertilizer applied to one plot into an adjacent plot they 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.

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 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 employment. Such concerns have rarely been addressed in the recent program evaluation literature. Exceptions include Heckman and Lochner (XXXX) who provide some theoretical calculations to assess the potential biases from ignoring general equilibrium effects.

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

On 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 no interaction between local labor markets or classrooms. Such aggregation is likely to make the no-interaction assumption more plausible, but it comes at the expense of reduced precision.

An alternative approach 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. 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 relevant literature in economics is that on social interactions. This literature has been growing rapidly in the last decade, following the early work by Manski (1993). See Manski

() and Brock and Durlauf () for surveys. Empirical work includes (Moving to Opportunity, Katz, Liebman and Kling XXXX, Roommates, Sacerdote XXXX, early Case-Katz paper, Krueger classrooms, Graham variance) Many identification questions remain.

3 Estimands, Hypotheses, and Internal Versus External Validity

In this section we discuss some of the estimands that researchers have focused on and some of the null hypotheses that have been tested.

3.1 The Traditional Estimands: The Average Treatment Effect and the Average Effect on the Treated

The econometric literature has largely focused on a couple of estimands. The two most prominent ones both rely on a superpopulation perspective. The sample of size N is viewed as a random sample from a large (super-)population. Interest is in this superpopulation. The two most popular ones are the Average Treatment Effect (ATE), the population expectation of the unit-level causal effect $Y_u(1) - Y_i(0)$:

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

and the average over the subpopulation of treated units:

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

3.2 Subpopulation and Weighted Average Treatment Effects

In addition to the two leading estimands τ_{ate} and τ_{att} there are other possible estimands. Two general classes include average causal effects for subpopulations and weighted average causal effects. Let $\tau_{\mathcal{A}}$ denote the average causal effect for the subpopulation with $X_i \in \mathcal{A}$:

$$\tau_{\mathcal{A}} = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i \in \mathcal{A}\right],$$

and let τ_{g} denote the weighted average causal effect with weight function $g(x)$:

$$\tau_{\text{g}} = \mathbb{E}\left[g(X_i) \cdot (Y_i(1) - Y_i(0))\right] \Big/ \mathbb{E}\left[g(X_i)\right].$$

In both cases we can also define sample versions of these estimands. Often the motivation for these estimands is not so much that the subpopulations they refer to are of intrinsic interest. Rather, it may be that it is much easier to precisely estimate average effects for these subpopulations compared to others. Instead of only reporting an imprecisely estimated average effect for the overall population it may be informative to augment this with a precise estimate for the average effect of some subpopulation. Such 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 formulation of the estimands becomes more cumbersome. 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 generally two approaches. One is to focus on bounds for well-defined estimands such as the average effect τ_{ate}. Manski (1990, xxxx) 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 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.

3.3 Quantile and Distributional Treatment Effects

Firpo (2005) Bitler, M., J. Gelbach, and H. Hoynes (2002)

3.4 Other Interesting Estimands

argmax of causal relation (Flores, 2005)

3.5 Sample Average Treatment Effects

There is a somewhat subtle issue that we may wish to separate the extrapolation to the super-population from inference for the sample at hand. This suggest that an alternative is to first focus on the average causal effect for the sample,

$$\tau^{\text{s}}_{\text{ate}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right),$$

and the average over the subsample of treated units:

$$\tau^{\text{s}}_{\text{att}} = \frac{1}{N_1}\sum_{i|W_i=1}\left(Y_i(1) - Y_i(0)\right).$$

If the effect of the treatment or intervention is constant ($Y_i(1) - Y_i(0) = \tau$ for some constant τ) all the estimands are obviously identical. However, if there is heterogeneity in the effect of the treatment, the estimands are all distinct. The difference between τ_{ate} and $\tau^{\text{s}}_{\text{ate}}$ is relatively subtle. Most estimators that are attractive for one of the two are also attractive estimators for the other, and we therefore do not have to be particularly concerned with the distinction between the two at the estimation stage. There is an important difference between the two estimands at the inference stage, however. In general we can estimate the sample average treatment effect $\tau^{\text{s}}_{\text{ate}}$ more precisely than the population average treatment effect τ_{ate}. When

one estimates the variance one therefore needs to be explicit about whether one is interested in the population or sample average treatment effect. We will return to this issue in more detail in Section 5

The motivation for the almost exclusive focus on the overall average effect and the average effect for the treated is somewhat limited. Take a job market program. The overall average effect would be the parameter of interest if the policy under consideration is a mandatory exposure to the treatment. Typically only limited groups are targeted by most of the current programs. Similarly the average effect for the treated would be informative about the effect of eliminating the program, but it would generally not correspond to an extension of the program to other jurisdictions even if such areas would be similar in terms of populations, unless the extension would leave the take up rate unchanged.

3.6 Testing

Abadie??

The literature on testing in these models is relatively limited. Since many of the estimators are asymptotically normally distributed with zero asymptotic bias standard confidence intervals can be used for testing hypotheses that any of these estimands are equal to zero. However, there are other interesting hypotheses to consider. 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 outcomes. In addition, in cases where there is not sufficient information to obtain precise inferences for the average causal effect τ_{ate}, it may still be possible to establish where 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 that may differ in terms of some observable characteristics.

4 Randomized Experiments

Randomized experiments have a long tradition in statistics. In this literature they are often viewed as the only plausible approach to establishing causality. A first comment concerns the fact that even randomized experiments rely heavily on substantive knowledge. It is only once the researcher is willing to rule out interactions between units that randomization can establish causal effects. In the absence of knowledge about the lack or nature of interactions even randomization cannot solve the identification problems required for establishing causality.

In the economics 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. (external validity)

Early social experiments in economics were the Seattle and Denver Income Maintenance Experiments (SIME and DIME).

More recently a number of experimental evaluations of labor market programs have been conducted. (NSW, Lalonde, 1986) GAIN, JOBS. These experiments have been extremely useful not merely in establishing the effects of particular programs, but also in providing testing grounds for new evaluations methods. We return to this literature in Section 5.11.

Although the actual number of experiments conducted in economics is relatively small, experimental methods are still every important. Many of the nonexperimental methods are directly based on these methods, essentially making assumptions such that some of the experimental methods are applicable. We therefore now review two of the most important experimental methods, Fisher's method for exact tests and Neyman's repeated sampling approach. Especially the former is still very relevant for current analyses of experimental data even though it is in practice very rarely used.

4.1 Fisher's Randomization Inference

Fisher (XXXX, XXXX) studied agricultural experiments. He was interested in testing hypotheses regarding the effect of treatments. The setting is one with N units, indexed by i. The aim is to provide inferences for this finite population. There may be a superpopulation that this population is drawn from in a random fashion, but that is not essential and it is not used in the analysis: the analysis is solely focused on inference for the sample in hand. The inference is both nonparametric in the sense that it does not make functional form assumptions regarding the effects, and exact in the sense that it does not rely on large sample approximations. In other words, the p-values coming out this analysis are valid irrespective of the sample size.

The typical null hypothesis in Fisher's framework is that there is no effect of the treatment for any unit in this population:

$$H_0 \; Y_i(0) = Y_i(1), \; \forall i = 1, \ldots, N,$$

against the alternative that for some units there is a non-zero effect:

$$H_a \; \exists i \text{ such that } Y_i(0) \neq Y_i(1).$$

It is not essential 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. Thus a more general null hypothesis could be that $Y_i(0) = Y_i(1) + c$ for some c, or that $Y_i(0) = Y_i(1) + c_i$ for some set of c_i. Importantly this framework *cannot* accomodate null hypotheses such as the hypothesis that the *average* effect of the treatment is zero, or $\sum_i (Y_i(1) - Y_i(0))/N = 0$. 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 first was only of "academic"interest). Putting this argument about its ultimate relevance aside, Fisher's test is a very powerful tool for establishing whether a treatment has any effect. It is also not essential in this framework that the probabilities of assignment to the treatment group are equal for all units. What is essential is that all the probabilities of assignment are known to the researcher, so the distribution of any particular assignment vector is known. They may differ by covariates or in some other way, but as long as they are known the Fisher framework applies.

The attraction of Fisher's framework is that under the null 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*, a function of the realized values of $(Y_i, W_i)_{i=1}^N$ generated by the randomization. So, suppose the statistic is the average difference between treated and control outcomes, $T = \bar{y}_1 - \bar{y}_0$, where $\bar{y}_w = \sum_i W_i Y_i / N_w$. Now supose we had assigned a different set of units to the treatment. Denote the vector of treatment assignments by $\tilde{\mathbf{W}}$. Under the null hypothesis we can deduce what the value of the statistic would have been in that case. Call this $T(\tilde{\mathbf{W}})$. We can do this 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 $T(\mathbf{W})$. This distribution generated by the randomization of the treatment assignment is referred to as the *randomization distribution*. If the realized statistic, given the actual assignments, is unusual in this distribution, that is if it is far out in the tail, as measured by its p-value, the null hypothesis is rejected.

This is a very powerful tool. It does not rely on functional form assumptions (it is full nonparametric), and is exact even in finite samples. 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 controlled by the researcher: 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, as long as the sample size is moderately large. This approach to testing is much more interesting with other choices for the statistic. For example, as advocated by Rosenbaum in a series of papers, (XXXX, XXXX), a particularly 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 by treatment status. The test is still exact, but now much more robust to outliers.

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 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 are from the GAIN programs in California. The four locations here 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 (WIN) or second (GAIN) year following the program. We compare p-values 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 with exact p-values based on randomization inference using either the difference in average outcomes by treatment status or the difference in average ranks by treatment status as the statistic. The results are in Table **??**.

In all eight cases the simple t-test gives almost exactly the same p-values as the exact p-values based on the randomization inference. This is not surprising given the reasonably large sample sizes, ranging from xxx to xxx. However, the p-values for the rank tests are fairly different in many of the cases, leading to substantively difference conclusions. 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 0.025 (ranks) and in the GAIN (San Diego) evaluation the p-values 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, and those with positive earnings have a fairly dispersed distribution. In those cases rank-based tests are likely to have much more power against alternatives that shift the distribution towards higher earnings.

In addition to testing, Fisher's approach can also be used to construct confidence intervals. These are constructed as sets of treatment effects that are not rejected using the tests described above. For example, assuming a constant treatment effect τ we can test the null hypothesis that $\tau = c$ for a range of values of c. The confidence set is then the range of c that do not get rejected by the data. This part of Fisher's approach is less attractive as it is rare that if we reject the null hypothesis of a zero treatment effect we would still wish to maintain the assumption of a constant treatment effect.

4.2 Neyman's Repeated Sampling Inference

The second approach to analyzing randomized experiments is due to Neyman (). Neyman's approach to analyzing data from randomized experiments focuses on estimating the average treatment effect

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right).$$

Note that this is again a finite sample perspective without the need for random sampling from a superpopulation. In addition Neyman is interested in obtaining measures of uncertainty for any estimators of τ, based on the randomization distribution. Neyman's methods are mainly useful because they provide a formal justification for least squares estimators in this setting without resorting to functional form or distributional assumptions.

Neyman suggests the simple difference in means,

$$\hat{\tau} = \frac{1}{N_t}\sum_{i=1}^{N} W_i \cdot Y_i - \frac{1}{N_c}\sum_{i=1}^{N}(1 - W_i) \cdot Y_i,$$

as an estimator for the average treatment effect τ. He shows this is unbiased under the randomization distribution.

Lengthy calculations show that the exact variance of this estimator is

$$\mathrm{Var}(\hat{\tau}) = \frac{1}{(N-M)(N-1)} \sum_{i=1}^{N} \Big(Y_i(1) - \bar{Y}(1) \Big)^2 + \frac{1}{M(N-1)} \sum_{i=1}^{N} \Big(Y_i(1) - \bar{Y}(1) \Big)^2$$

$$- \frac{1}{N(N-1)} \sum_{i=1}^{N} \Big(Y_i(1) - Y_i(0) - \tau \Big)^2.$$

There are a couple of important aspects of this expression for the variance. First, the last term, the variance of the unit-level treatment effect cannot be estimated. It is therefore typically ignored in the estimation of the variance. However, this term is nonpositive, so we can sign the bias from omitting it. This implies that estimates of the variance are typically overestimates, leading to conservative confidence intervals. For the other two components there are unbiased estimators available, in the standard form of the sample variance of the treated and control samples, divided by their subsample sizes. In large samples these can be used to construct confidence intervals.

This estimator and the resulting variance are identical to the standard least squares estimator and the heteroskedasticity-consistent variance estimator based on the regression function for the observed outcomes:

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i.$$

However, it should be noted that justification is quite different. The justification for Neyman's procedures stems exclusively from the randomization. There are no appeals to independence of ε_i and W_i. In fact, if we define $\varepsilon_i = Y_i - \bar{Y}(0) - (\bar{Y}(1) - \bar{Y}(0)) \cdot W_i$, such independence will typically not hold. However, by construction such a residual will be uncorrelated with W_i, thereby making the least squares estimator unbiased.

5 Selection on Observables or Unconfoundedness

Estimation of causal effects under unconfoundedness assumptions is most closely related to standard multiple regression analysis with a rich set of controls. In the case of a binary treatment W, in effect we assume that we have good enough predictors of treatment, contained in the covariates X, such that treatment is independent of the counterfactual outcomes, $Y(0)$ and $Y(1)$, conditional on X. For reasons we shall see, the elements of X should not themselves be affected by treatment. Typically they are determined prior to treatment.

Given the variables X, which are also know as *pre-treatment* and sometimes *exogenous* variables, we can define conditional treatment effects. The treatment effect conditional on $X = x$ is

$$\tau(x) = \mathrm{E}[Y(1) - Y(0) | X = x];$$

in other words, $\tau(x)$ is the average treatment effect for that slice of the population having observed covariates x. The average treatment effect across the population is

$$\tau = \mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[\tau(X)],$$

where the second equality is a simple implication of the law of iterated expectations. Often a researcher's goal is to estimate the effect of treatment only on those receiving treatment, defined as

$$\tau_T = \mathrm{E}[Y(1) - Y(0)|W = 1].$$

In this section, we present assumptions under which the treatment parameters in (), (), and () are identified, provide the smallest asymptotic variances achievable for estimating τ and τ_T, and discuss the common methods of estimation.

5.1 Identification

Without further assumptions, we cannot estimate any of the treatment effect parameters defined earlier. A key assumption in the causal effects literature the following notion of unconfoundedness, introduced by Rosenbaum and Rubin (1983):

Unconfoundedness: Conditional on X, $(Y(0), Y(1))$ is independent of W, which is often written as

$$(Y(0), Y(1)) \perp W \mid X,$$

where the symbol "$\perp$" means "independent of."

We can also write unconfoundedness as $\mathrm{D}(Y(0), Y(1)|W, X) = \mathrm{D}(Y(0), Y(1)|X)$or $\mathrm{D}(W|Y(0), Y(1), X) = \mathrm{D}(W$ where $\mathrm{D}(\cdot|\cdot)$ denotes conditional distribution. The first statement has led to the phrase "treatment is ignorable conditional on X." The second has led to the name "selection on observables" because treatment, W, does not depend on the unobservables in $Y(0), Y(1)$ once we know the observables in X.

The unconfoundedness assumption can be controversial, as it assumes that treatment is not affected by observables that can influence the responses once we control for X. 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 $Y_i(0) = \alpha + \beta' X_i + U_i$, where U_i contains the unobservables affecting the response in the absense of treatment. Then, letting $Y_i = (1 - W_i)Y_i(0) + W_i Y_i(1)$ as before, we can write

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

and unconfoundedness holds if and only if U_i is independent of W_i, conditional on X_i.

(Something here on economic models that imply unconfoundedness?)

The second assumption used to identify treatment effects is that for all possible outcomes on X, there are treated and control units. We call this the "overlap" assumption:

Overlap: For all x,

$$0 < P(W = 1|X = x) < 1.$$

The probability in () is known as the *propensity score*, which we denote

$$e(x) = P(W = 1|X = x).$$

Because we have a random sample on (W, X), we can estimate $e(x)$, 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. As we will see, for estimating the average treatment on the treated, we could allow for units that have zero probability of being treated, but, in practice, that generality is not especially useful.

When we combine unconfoundedness with overlap we arrive at Rosenbaum and Rubin's (1983) assumption of *strong ignorability*. There are various ways to establish identification of $\tau(x)$, τ, and τ_t under strong ignorability. Perhaps the easiest is to note that $\tau(x)$ is identified because

$$\begin{aligned} \tau(x) &\equiv \mathrm{E}[Y(1)|X = x] - \mathrm{E}[Y(0)|X = x] \\ &= \mathrm{E}[Y(1)|W = 1, X = x] - \mathrm{E}[Y(0)|W = 0, X = x], \end{aligned} \tag{5.1}$$

where the second equality follows by unconfoundedess: $\mathrm{E}[Y(w)|W, X]$ does not depend on W. By definition, we observe data on $Y(1)$ whenever $W = 1$ and we observe data on $Y(0)$ whenever $W = 0$. Thus, provided we observe both treated and untreated units with covariates x – as we assume in the overlap assumption – we can estimate both terms on the right hand side of (z). Given that we can identify $\tau(x)$ for all x, we can identify

$$\tau = \mathrm{E}[\tau(X)] \tag{(a)}$$

and

$$\tau_T = \mathrm{E}[\tau(X)|W = 1], \tag{(b)}$$

where (b) follows under ignorability. Because we can consistently estimate $\tau(x)$ by, say, $\hat{\tau}(x)$, equation (a) suggests a simple averaging estimator, $\hat{\tau} = N^{-1}\sum_{i=1}^{N}\hat{\tau}(X_i)$ and (b) suggests averaging $\hat{\tau}(X_i)$ only over the treated units. Before we discuss estimators in detail, it is important to know how precisely we can expect to estimate τ and τ_T.

5.2 Efficiency Bounds

Let $\sigma_0(X) = \text{Var}(Y(0)|X)$ and $\sigma_1(X) = \text{Var}(Y(1)|X)$. Then, from Hahn (1998), the lower bounds for asymptotic variances of $\sqrt{N}$-consistent estimators are

$$\text{E}\left[\frac{\sigma_1^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} + (\tau(X)-\tau)^2\right] \qquad ((\text{aa}))$$

and

$$\text{E}\left[\frac{e(X)\sigma_1^2(X)}{p} + \frac{e(X)^2\sigma_0^2(X)}{p^2(1-e(X))} + \frac{(\tau(X)-\tau_T)^2 e(X)}{p^2}\right] \qquad ((\text{ab}))$$

for τ and τ_T, respectively, where $p = \text{E}[e(X)]$ is the unconditional treatment probability. These lower bounds are obtained under the assumption that the propensity score, $e(\cdot)$, is unknown. As shown by Hahn (1998), knowing the propensity score does not affect the variance lower bound for estimating τ, but it does change the lower bound for estimating τ_T. Only rarely do we know the propensity score, and so (ab) is the usually relevant formula for estimating τ_T.

Having displayed these lower bounds, the natural question is: Are there estimators that achieve these lower bounds that do not require parametric models for the conditional means or the propensity score? The answer is yes, and we now consider different classes of estimators in turn.

5.3 Regression Methods

Let

$$\mu_0(x) = \text{E}[Y(0)|X=x] \text{ and } \mu_1(x) = \text{E}[Y(1)|X=x],$$

so that $\tau(x) = \mu_1(x) - \mu_0(x)$. Under the unconfoundedness assumption, $\mu_0(x) = \text{E}[Y(0)|W = 0, X = x]$ and $\mu_1(x) = \text{E}[Y(1)|W = 1, X = x]$, 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 of τ is

$$\hat{\tau}_{reg} = N^{-1}\sum_{i=1}^{N}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]. \qquad ((\text{q}))$$

Note that although $\hat{\mu}_1$ and $\hat{\mu}_0$ are obtained on the appropriate subsamples, we average the differences in predicted values across all N observations. That is, we

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

$$\mu_0(x) = \alpha_0 + \beta_0' x,\ \mu_1(x) = \alpha_1 + \beta_1' x$$

(or we replace x with general functions of x). Then $\hat{\tau} = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\bar{X}$. This estimator is also obtained from the coefficient on the treatment indicator W_i in the regression Y_i on $1, W_i, X_i, W_i \cdot (X_i - \bar{X})$. This last regression is especially attractive because if we ignore estimation of the mean of X by the sample average, then a valid standard error for $\hat{\tau}$ is easily obtained by the usual OLS standard error or the heteroskedasticity-robust form. [Typically, the distortion in the standard errors caused by ignoring the estimation error in $\bar{X}$ is small.]

If one is to use parametric models it makes sense to choose the regression functions to reflect the nature of the response. For example, if Y is a binary response (such as an employment indicator), logit and probit models, with the covariates entering the linear index in flexible ways, are natural. If Y is a count variable (such as number of times an individual is arrested during a year) or a nonnegative continuous variable, an exponential function is sensible, and the parameters can be estimated using a variety of quasi-likelihood methods [for example, those based on the Poisson and exponential distributions – see Wooldridge (2002, Chapter 19)]. If Y is a corner solution response (say, annual wage income), one might use Tobit models and then obtain the conditional mean functions, $\mathrm{E}[Y|W = 0, X]$ and $\mathrm{E}[Y|W = 1, X]$, that are implied by the Tobit model – see Wooldridge (2002, Chapter 16) Including flexible functions of X is a simple way to estimate $\mu_0(\cdot)$ and $\mu_1(\cdot)$ that are logically consistent with the nature of Y.

Let $m_0(x, \delta_0)$ and $m_1(x, \delta_1)$ be general parametric models of $\mu_0(\cdot)$ and $\mu_1(\cdot)$; as a practical matter, m_0 and m_1 would have the same structure but with different parameters. Assuming that we have consistent, $\sqrt{N}$-asymptotically normal estimators $\hat{\delta}_0$ and $\hat{\delta}_1$, the asymptotic variance of

$$\hat{\tau}_{reg} = N^{-1} \sum_{i=1}^{N} [m_1(X_i, \hat{\delta}_1) - m_0(X_i, \hat{\delta}_0)] \qquad ((\mathrm{ca}))$$

can be estimated using the delta method. From Wooldridge (2002, Problem 12.12), it can be shown that

$$\begin{aligned} \mathrm{Avar}\sqrt{N}(\hat{\tau}_{reg} - \tau) &= \mathrm{E}\{[m_1(X_i, \delta_1) - m_0(X_i, \delta_0) - \tau]^2\} + \mathrm{E}[\nabla_{\delta_0} m_0(X_i, \delta_0)] V_0 \mathrm{E}[\nabla_{\delta_0} m_0(X_i, \delta_0)]' \\ &\quad + \mathrm{E}[\nabla_{\delta_1} m_1(X_i, \delta_1)] V_1 \mathrm{E}[\nabla_{\delta_1} m_1(X_i, \delta_1)]', \end{aligned} \qquad ((\mathrm{cb}))$$

where V_0 is the asymptotic variance of $\sqrt{N}(\hat{\delta}_0 - \delta_0)$ and similarly for V_1. Not surprisingly, using more efficient estimators of δ_0 and δ_1 results in more efficient estimation of τ. Each of

the quantities in (cc) is easy to estimate by replacing expectations with sample averages and replacing unknown parameters with estimators. As always, the asymptotic standard error is obtained by taking the square root and dividing by $\sqrt{N}$. Alternatively, a resampling method, such as the bootstrap, can be used – see Horowitz (2004).

Esimation of τ_T is similarly straightforward. The general form of the estimator is

$$\hat{\tau}_{T,reg} = N_1^{-1} \sum_{i=1}^{N} W_i[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$$

where $N_1 = \sum_{i=1}^{N} W_i$ is the number of treated units in the sample.

Because $\hat{\tau}_{reg}$ and $\hat{\tau}_{T,reg}$ are essentially averages of fitted values, different methods for estimating $\mu_0(\cdot)$ and $\mu_1(\cdot)$ may lead to similar estimates of average treatment effects. For example, if Y is a binary outcome, it is often the case that the linear probability model and nonlinear models such as logit and probit provide very similar estimated probabilities for covariates near the mean of the covariate distribution. Unfortunately, there is little simulation or empirical evidence (??) that applies directly to treatment effect estimation. Still, one can imagine that, in some situations, parametric approaches can lead to inconsistency in estimating the treatment effects, in which case nonparametric approaches can be valuable. Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) consider kernel and local kernel methods for estimate the regression functions, although some modification is needed for discrete covariates in X. Recently, Qi and Racine (2004) have proposed kernel methods that can be used when covariates include both continuous and discrete variables. 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 effectsm where the parameter of interest is effectively the average of a regression function, and not the entire function. See Imbens (2004) for more discussion. Unfortunately, estimators based on local smoothing have not been shown to attain the variance efficiency bound.

An alternative to local smoothing methods are global smoothing methods, such as series estimators (which can be obtained from linear regression) and sieve estimators (which include standard nonlinear functional forms such as the logistic and exponential). Such estimators are parametric for a given sample size, and so they are easy to understand and computation is relatively simple. 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. But Imbens, Newey, and Ridder (2004) have shown that linear series estimators can achieve the variance lower bounds for estimating τ and τ_T. [Hahn (1998) uses series estimators in a somewhat different way – which requires more nonparametric estimation that just the two conditional means – and his estimator also is asymptotically efficient.] We conjecture that nonlinear sieve estimators, where the regression function can be of the logistic, exponential, or other attractive forms, can be shown to be asymptotically efficient, but this has yet to be worked out.

A practically important issue is estimating the asymptotic variance of estimators based on

nonparametric estimation of μ_0 and μ_1. If the estimator has been shown to be asymptotically efficient then (aa) and (ab) can be used to estimate the asymptotic variances. Using these formulas is at best inconvenient, especially because we only need $\mathrm{E}[Y(1)|X]$ and $\mathrm{E}[Y(0)|X]$ to estimate τ or τ_T. If we use, say, (aa), then we must also estimate $\mathrm{Var}[Y(1)|X]$, $\mathrm{Var}[Y(0)|X]$,.and the propensity score, $\mathrm{P}(W = 1|X)$. Presumably we would estimate these nonparametrically because we estimated the mean functions nonparametrically. Sieve estimators for the variances and propensity score can be used, but we would want to ensure nonnegative variances and propensity score estimates in the unit interval. While such sieve estimators are known to be consistent under general circumstances, details about when the variance and propensity score estimators converge at a fast enough rate so that their estimation error can be ignored do not appear to be available.

In some cases for sieve estimators, it is valid to use the asymptotic variance obtained by treating the estimation problem as a standard parametric problem, and using a variance estimate such as (cc). Not much is known about the bootstrap in these situations; see Imbens (2004) for further discussion.

5.4 Methods Based on the Propensity Score

Much of the empirical literature that estimates treatment effects under ignorability of treatment or unconfoundedness relies heavily on the propensity score, rather than estimating regression functions. There are probably a few reasons for this, in addition to the appeal of propensity score methods as being somewhat exotic – at least compared with regression methods. First, estimating the propensity score requires only a single parametric or nonparametric estimation. The regression methods described in the previous subsection require estimation of $E(Y|W = 0, X)$ and $E(Y|W = 1, X)$. Second, because the propensity score is a binary outcome, we have a good idea about the kinds of methods that work well for estimating $E(W|X) = P(W = 1|X)$. For regression methods, depending on the nature of Y we might not have much information about how well nonparametric methods, particular sieve methods, work. For example, Y might have both discrete and continuous characteristics. Third, attractive propensity score methods have been developed that achieve the variance lower bound. Nevertheless, it is important to understand that the methods we describe in this section use the same unconfoundedness and overlap assumptions that we imposed for the regression methods. In short, propensity score methods are valid in the context of selection on observables.

There are three ways that the propensity score is used in estimating causal effects. One approach, matching, is discussed in Section xx. We discuss the other two methods in this section, propensity score weighting and using the propensity score in regression.

Estimators of τ and τ_T based on propensity score weighting are simple to motivate. Recall that $\tau = E[Y(1)] - E[Y(0)]$. Because $WY = WY(1)$, we have

$$
\begin{aligned}
E\left[\frac{WY}{e(X)}\right] &= E\left[\frac{WY(1)}{e(X)}\right] = E\left\{E\left[\frac{WY(1)}{e(X)}|X\right]\right\} = E\left[\frac{E(W|X)E(Y(1)|X)}{e(X)}\right] \\
&= E\left[\frac{e(X)E(Y(1)|X)}{e(X)}\right] = E[E(Y(1)|X] = E[Y(1)],
\end{aligned}
$$

where the second and final inequalities follow by iterated expectations and the third equality holds by unconfoundedness. In other words, weighting the treated population by the inverse of the propensity score recovers the expected response under treatment. A similar calculation shows

$$
E\left[\frac{(1-W)Y}{1-e(X)}\right] = E[Y(0)],
$$

and together these imply

$$
\tau = E\left[\frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)}\right] = E\left\{\frac{[W-e(X)]Y}{e(X)[1-e(X)]}\right\} \tag{(dd)}
$$

where the second equality holds by simple algebra. Equation (dd) suggests an obvious estimator of τ:

$$
\tilde{\tau}_{weight} = N^{-1}\sum_{i=1}^{N}\left[\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right], \tag{(de)}
$$

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

In most applications, $\tilde{\tau}$ is not feasible because it depends on the propensity score function $e(\cdot)$, which is typically unknown. A more surprising drawback is that, even if we know $e(\cdot)$, $\tilde{\tau}_{weight}$ does not achieve the efficiency bound given in (). It turns out to be better to estimate the propensity score, particularly if one uses a nonparametric estimator. 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 lower 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 is restricted by the sample size. In the second step, one obtains

$$
\hat{\tau}_{weight} = N^{-1}\sum_{i=1}^{N}\left[\frac{W_iY_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right].
$$

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. Estimating the asymptotic variance requires estimating (aa), which we already discussed. A conservative variance estimator is obtained by ignoring the estimation error in $\hat{e}(\cdot)$, and this may be acceptable in some cases.

Ichimura and Linton (2001) studied $\hat{\tau}_{weight}$ 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. More recent, Li, Racine, and Wooldridge (2005) consider kernel estimation and all for discrete as well as continuous covariates. The estimator proposed by Li, Racine and Wooldridge achieves the variance lower bound.

A different way of using the estimated propensity score is to use in in place of the covariates in regression analysis. The motivation is given by the following fundamental lemma due to Rosenbaum and Rubin (1983).

Lemma 1: Under the assumption of unconfoundedness,

$$(Y(0), Y(1)) \perp W \mid e(X). \qquad ((\text{dg}))$$

A simple proof of Lemma 1 can be found in Imbens (2004). It says that if unconfoundedness holds conditional on X, then it holds conditional only on the propensity score as well. An important implication is that all covariate effects can be accounted for by conditioning solely on the propensity score. (This is also implicit in the estimators based on propensity score weighting.)

An immediate implication of (dg) is

$$E[Y(w)|W, e(X)] = E[Y(w)|e(X)] \equiv \nu_w[e(X)], w = 0, 1,$$

where $\nu_w(e) = E([Y(w)|e(X) = e]$. We can estimate ν_0 and ν_1 by regressions using the untreated and treated subsamples, respectively. Let $\hat{e}(\cdot)$ be an estimator of the propensity score. Then we can estimate $v_0(e)$ and $\nu_1(e)$ very generally by using kernel or series estimation on the estimated propensity score, something which is feasible because 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

$$\hat{\tau}_{regprop} = N^{-1} \sum_{i=1}^{N} \{\hat{v}_1[\hat{e}(X_i)] - \hat{v}_0[\hat{e}(X_i)]\},$$

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 estimator based on the propensity score does not achieve the efficiency bound.

Because of its simplicity, regression on just a linear function of the propensity score has been widely used in applied work. Allowing for different linear functions we would run separate regressions $W_i = 0$ and $W_i = 1$:

$$Y_i \text{ on } 1, \hat{e}(X_i) \text{ for } W_i = 0 \text{ and } Y_i \text{ on } 1, \hat{e}(X_i) \text{ for } W_i = 1,$$

which gives fitted values $\hat{\alpha}_0 + \hat{\gamma}_0\hat{e}(X_i)$ and $\hat{\alpha}_1 + \hat{\gamma}_1\hat{e}(X_i)$, respectively. If ν_0 and v_1 are linear in e, a consistent estimator of τ is

$$\hat{\tau}_{regprop} = N^{-1}\sum_{i=1}^{N}[(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\hat{e}(X_i)].$$

The variance of this estimator should account for estimation of the propensity score but the estimator that ignores the estimation error can be shown to be conservative in the case where $\hat{e}(x)$ is a consistent estimator obtained from a parametric model.

An even simply strategy is to assume the slopes γ_0 and γ_1 are the same. Then $\hat{\tau}_{regprop}$ is just the coefficient on W_i in the regression

$$Y_i \text{ on } 1, W_i, \hat{e}(X_i) \; i = 1, ..., N. \qquad ((\text{dh}))$$

The appeal of (dh) is that a univariate function of the covariates, $\hat{e}(X_i)$, controls for selection into treatment. At least nomically this estimator appears to impose strong assumptions on $E[Y(w)|e(X)]$ – assumptions that are not needed for the propensity score-weighted estimator, $\hat{\tau}_{weight}$. Interestingly, the estimator from (dh) does have some robustness. Wooldridge (200xx) shows that if the conditional variance of the propensity score, $e(X)[1 - e(X)]$, is uncorrelated with the unit-specific treatment effect, $Y(1) - Y(0)$, then the simple estimator from (dh) is consistent for τ. This includes the case of constant treatment effect, $\tau = Y(1) - Y(0)$, but could be approximately true more generally because $e(X)[1-e(X)]$ is not a monotonic function of $e(X)$. (That is, even if $Y(1) - Y(0)$ is correlated with $e(X)$, it could be uncorrelated with $e(X)[1 - e(X)]$.)

On balance, estimators that use the propensity score in regression seem have little to offer compared with weighting by the propensity score. Each requires estimation of the propensity score but the weighted estimator requires simply plugging into (); no further nonparametric estimator, or functional form restrictions, are needed. Plus, the propensity score regression estimators are not asymptotically efficient while the propensity score weighted estimators are.

5.5 Methods Combining Regression and Propensity Score Weighting

In the previous two subsections we described methods for estimating average causal effects based on two strategies: the first is based on estimating $E[Y(w)|X]$ for $w = 0, 1$, and the

second is based on estimating the propensity score $e(X) = P(W = 1|X)$. For each approach we have discussed estimators that achieve the asymptotic efficiency bound (as well as estimators that do not achieve the bound). If we have large sample sizes relative to the dimension of X, we might think our nonparametric estimators of the conditional means or propensity score are sufficient to invoke the asymptotic efficiency results described above. In other cases, 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. Still, 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.

Here is the idea. As in Section x.x, let $m_0(x, \delta_0)$ and $m_1(x, \delta_1)$ be parametric functions for $E[Y(0)|X = x]$ and $E[Y(1)|X = x]$, respectively, and let $G(x, \gamma)$ be a parametric model for the propensity score. We assume that $0 < G(x, \gamma) < 1$ for all x and γ, as is true for all common parametric models of binary responses. In the first step we estimate γ by maximum likelihood and obtain the estimated propensity scores as $\hat{e}(X_i) = G(X_i, \hat{\gamma})$. In the second step, we use linear or nonlinear regression, or a quasi-likelihood method, where we weight the objective function by the inverse probability of treatment or non-treatment. For example, to estimate $E[Y(1)|X = x]$, we might solve the weighted least squares problem

$$\min_{\delta_1} \sum_{i=1}^{N} W_i[Y_i - m_1(X_i, \delta_1)]^2/\hat{e}(X_i); \qquad ((\text{ga}))$$

for $E[Y(0)|X = x]$, we simply weight the $W_i = 0$ observations by $1/[1 - \hat{e}(X_i)]$. Given the estimated conditional mean functions, we estimate the ATE, τ, using the expression for $\hat{\tau}_{reg}$ *exactly* as in equation (ca). But why should we weight (ga) by the inverse propensity score when we did not use such weighting in Section x.x? The answer is in two parts, and it produces the "double robustness" result due to " Scharfstein, Rotnitzky, and Robins (1999).

First, suppose that $m_1(x, \delta_1)$ is correctly specified for $E[Y(1)|X = x]$, which means that there exits a value, say δ_1^*, such that $E[Y(1)|X = x] = m_1(x, \delta_1^*)$. Then, as discussed in the treatment effect context by Wooldridge (2005), weighting the objective function by any nonnegative function of X_i does not affect consistency of least squares, or 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 GMT (1984).] In fact, even if $G(x, \gamma)$ is misspecified for $e(x)$, the binary response MLE $\hat{\gamma}$ still has a well-defined probability limit, say γ^*, and the IPW estimator that uses weights $1/G(X_i, \hat{\gamma})$ is asymptotically equivalent to the estimator that uses weights $1/G(X_i, \gamma^*)$ [see Wooldridge (2005, xx)]. It does not matter that $G(x, \gamma^*) \neq P(W = 1|X = x)$. This is the first part of the double robustness result: if the parametric conditional means for $E[Y(w)|X = x], w = 0, 1$ are correctly specified, the model $G(x, \gamma)$ can be arbitrarily misspecified for $P(W = 1|X = x)$. Equation (ca) still consistently estimates τ.

When the conditional means are correctly specified, weighting by $1/G(x, \gamma^*)$ is known to

actually hurt in terms of terms of asymptotic efficiency in the leading cases. In particular, under nominal second moment assumptions – such as homoskedasticity of $Var[Y(1)|X = x]$ in the context of least squares – the IPW estimator of δ_1^* is actually less efficient than the unweighted estimator; see Wooldridge (2005, xx).. The same holds for any quasi-MLE in the linear exponential family when the nominal second moment assumption holds. Therefore, we should only weight if it offers a robustness advantage for estimating τ. For somewhat subtle reasons, this turns out to be the case, but we must be careful in choosing functional forms for $m_w(x, \delta_w)$ along with appropriate estimation methods.

The second part of the double robustness result assumes that $G(x, \gamma)$ is correctly specified for the propensity score, so that $G(x, \gamma^*) = e(x)$, but allows the conditional mean functions, $m_w(x, \delta_w)$, to be misspecified. The result is easiest to describe when $m_w(x, \delta_w)$ are linear in parameters, so $m_w(x, \delta_w) = \alpha_w + h_w(x)\lambda_w$, $w = 0, 1$, where $h_w(x)$ is a vector of known functions (which would probability be the same for the treatment and control estimation, but need not be). If these linear functions are misspecified then we cannot use $\tau = E\{E[Y(1)|X] - E[Y(0)|X]\}$ to argue that (ca) is consistent for τ. Neverthless, it is still possible that $E[Y(w)] = E[m_w(X, \delta_w^*)]$ for $w = 0, 1$, where δ_w^* denotes the probability limit from the weighted least squares estimation. In fact, if $\alpha_w^* + h_w(X)\lambda_w^*$ is the linear projection of $Y(w)$ on $h_w(X)$, then α_w^* and λ_w^* are chosen so that $E[Y(w)] = \alpha_w^* + E[h_w(X)]\lambda_w^*$. [This is simply the population analog of the well-known in-sample result that the least squares residuals always average to zero whenever an intercept is included in the regression.] Conveniently, the IPW estimator with a correctly specified model of the propensity score always consistently estimates the parameters in the population linear projection. If we do not weight by the inverse of the propensity score, and the conditional mean is not linear in $h_w(x)$, then the estimators do not necessarily converge to the parameters in the population linear projection. The bottom line is this: even if the mean functions are not linear in the functions $h_w(x)$, $\hat{\tau}_{reg}$ in (ca) is consistent for τ provided $\hat{\delta}_0 = (\hat{\alpha}_0, \hat{\lambda}_0)$ is obtained using weights $1/[1-G(X_i, \hat{\gamma})]$, $\hat{\delta}_1 = (\hat{\alpha}_1, \hat{\lambda}_1)$ is obtained using weights $1/G(X_i, \hat{\gamma})$, and $G(x, \gamma)$ is correctly specified for $e(x)$. 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 $E[Y(w)|X = x]$ 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 $E[Y(w)|X = x]$ 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, $E[Y(w)] = E[m_w(X, \delta_w^*)]$, 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.

If $Y(w)$ is a binary response, or a fractional response (with outcomes in $[0, 1]$), the logistic function combined with the binary response (quasi-) MLE has the requisite properties. It is well known that, provided a constant is included among the covariates, the residuals from logit

estimation always sum to zero. The population analog is $E[Y(w)] = E[\Lambda(\alpha_w^* + h_w(X)\lambda_w^*)]$, where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ is the logistic function and $(\alpha_w^*, \lambda_w^*)$ are the values that set the expected score of the logit quasi-log-likelihood to zero, whether or not $E[Y(w)|X] = \Lambda(\alpha_w^* + h_w(X)\lambda_w^*)$ and regardless of the nature of $Y(w)$. These are the quantities consistently estimated by IPW, provided the propensity score is correctly specified; see Wooldridge (2005).

If $Y(w)$ is nonnegative, an exponential mean function is natural. We combine this with the Poisson quasi-likelihood function, and then $E[Y(w)] = E[\exp(\alpha_w^* + h_w(X)\lambda_w^*)]$ where, again, $(\alpha_w^*, \lambda_w^*)$ set the expected score of the Poisson quasi-log-likelihood to zero. Note that we use the Poisson quasi-MLE even if $Y(w)$ is continuous, or has continuous and discrete characteristics. Other estimators combined with the exponential functional form, sush as least squares or the gamma quasi-MLE, will not produce the double robustness result. See Wooldridge (2005) for more discussion.

Once we estimate τ as in (ca), using IPW estimators for $\hat{\delta}_1$ and $\hat{\delta}_0$, how should we obtain a standard error? The fomula (xx) is still appropriate, but how should we estimate V_0 and V_1? Fortunately, for all estimators described here, asymptotic variances estimates are easily obtained. Naturally, we should use variance estimators that are valid under either of the double robustness scenarios. There are two possibilities. First, we can ignore the sampling variation in $\hat{\gamma}$ and obtain the usual "sandwich" estimators $\hat{A}_w^{-1}\hat{B}_w\hat{A}_w^{-1}, w = 0, 1$, where $\hat{A}_w$ is the average weighted Hessian and $\hat{B}_w$ is the average weighted outer product of the score. The sandwich estimators from the weighted problems are very convenient because they are reported routinely by packages that support inverse probability weighting. Plus, if the conditional mean models are correctly specified, inserting the sandwich estimators into (xx) produces a consistent asymptotic variance of $\hat{\tau}_{reg}$. If the mean functions are misspecified but the propensity score function is correctly specified, then the usual sandwich estimators lead to an overestimate of the asymptotic variance of $\hat{\tau}_{reg}$. This is because, as shown in Wooldridge (2005), estimating the propensity score via maximum likelihood reduces the asymptotic variance of the IPW estimator, at least when the conditional mean is misspecified. [This result is closely related to the Hirano, Imbens, and Ridder (2003) result that a more efficient estimator is obtained by estimating the propensity score rather than using the known propensity score.]

A second possibilitiy is to adjust the matrices $\hat{B}_w$ by first netting out the score from the first-stage propensity score estimation from the weighted scores (for $w = 0$ and $w = 1$) before forming the outer product; see Wooldridge (2005, xx) for details. This adjustment has no effect asymptotically if the conditional means are correctly specified, but it appropriately reduces the standard errors of the means are misspecified and the propensity score model is correctly specified.

Table xx summarizes the appropriate conditional mean functions and corresponding estimation methods that can be used to exploit double robustness.

Table xx

Response Variable, $Y(w)$	Continuous on Real Line	Binary or Fractional	Nonn
Conditional Mean Function, $E[Y(w)\|X = x]$	Linear in Parameters	Logistic	Expo
Estimation (Weighted)	Least Squares	Binary Response QMLE	Poiss

5.6 Matching

5.7 Testing

5.8 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. This was highlighted in an influential paper by Dehejia and Wahba (1999) who re-analyzed the data 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 public use data sets, using both the Panel Study of Income Dynamics (PSID) and the Current Population Survey (PSID). The NSW program targeted individuals who were very disadvantaged in the labor market, with very poor labor market histories. As a result they were very different from the raw comparison groups constructed by Lalonde. Lalonde partially addressed this 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 look 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. Although this is certainly a sensible starting point, inspecting differences one covariate at a time is not sufficient. 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. A more direct way of assessing the overlap in covariate distributions is therefore to inspect histograms of the estimated propensity score by treatment status.

Dehejia and Wahba (1999) focus on the average effect for the treated, They suggest dropping all control units with an estimated propensity score lower than the smallest value, or larger than the largest value, for the estimated propensity score among the treated units. Formally, they first estimate the propensity score. Let the estimated propensity score for unit i be $\hat{e}(X_i)$. Then let $\bar{e}_1$ be the minimum of the $\hat{e}(X_i)$ among treated units and let $\overline{e}_1$ be the maximum of the $\hat{e}(X_i)$ among treated units. DW then drop all control units such that $\hat{e}(X_i) < \bar{e}_1$ or $\hat{e}(X_i) > \overline{e}_1$.

Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998) also focus on the average effect for the treated. They propose discarding units with covariate values at which the estimated density is below some threshold. The precise method is as follows.[1] First they estimate the propensity score $\hat{e}(x)$. Next, they estimate the density of the estimated propensity score in both treatment arms. Let $\hat{f}_w(e)$ denote the estimated density of the estimated propensity score. The specific estimator they use is a kernel estimator

$$\hat{f}_w(e) = \frac{1}{N_w \cdot h} \sum_{i|W_i=w} K\left(\frac{\hat{e}(X_i) - e}{h}\right),$$

[1] See Heckman, Ichimura and Todd (1997) and Smith and Todd (2005) for details, and Ham, Li and Reagan (2005) for an application of this method.

with bandwidth h.[2] First HIT discard observations with $\hat{f}_0(\hat{e}(X_i))$ or $\hat{f}_1(\hat{e}(X_i))$ exactly equal to zero leaving J observations. Observations with the estimated denstiy equal to zero may exist when the kernel has finite support. Smith and Todd for example use a quadratic kernel with $K(u) = (u^2 - 1)^2$ for $|u| \leq 1$ and zero elsewhere. Next, they fix a quantile q (Smith and Todd use $q = 0.02$). Among the J observations with positive densities they rank the $2J$ values of $\hat{f}_0(\hat{e}(X_i))$ and $\hat{f}_1(\hat{e}(X_i))$. They then drop units i with $\hat{f}_0(\hat{e}(X_i))$ or $\hat{f}_1(\hat{e}(X_i))$ less than or equal to c_q, where c_q is the largest real number such that

$$\frac{1}{2J}\sum_{i=1}^{J}\Big(1\{\hat{f}_0(\hat{e}(X_i)) < c_q\} + 1\{\hat{f}_1(\hat{e}(X_i)) < c_q\}\Big) \leq q.$$

Ho, Imai, King and Stuart (2004) propose combining any specific parametric procedure that the researcher may wish to employ with a nonparametric first stage in which the units are matched to the closest unit of the opposite treatment. This typically leads to a data set that is much more balanced in terms of covariate distributions between treated and control. It therefore thus reduces sensitivity of the parametric model to specific modelling decisions such as the inclusion of covariates or functional form assumptions.

All these methods tend to make the estimators more robust to specification decisions. However, few formal results are available on the properties of these procedures.

Crump, Hotz, Imbens and Mitnik suggest a systematic way of doing this. They consider estimators for the average treatment effect for the subpopulation with $X \in \mathcal{A}$. They then choose the set $\mathcal{A}$ to minimize the asymptotic variance of the efficient estimator. Under some conditions (mainly homoskedasticity), they show that the optimal set has the form

$$\mathcal{A}^* = [\alpha, 1-\alpha],$$

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

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

The vertical lines in Figures 1-3 give the values of α for the three GAIN sample comparisons. For the San Diego versus Alameda comparison the cutoff value is estimated to be 0.088. For the San Diego versus Riverside comparison the cutoff value is 0.139 and for the experimental Los Angeles data the cutoff point is 0.123. In the first comparison this leads to dropping 426 observations out of 1755. According to the efficiency bound calculations this 25% reduction in sample size reduces the variance to 22% of the asymptotic variance for the original average treatment effect. In comparison, these calculations suggest that there is essentially no gain from dropping observations from the San Diego versus Riverside comparison, nor from the Los Angeles experimental evaluation.

A potentially troubling feature of this approach is that it changes what is being estimated.

[2] In their application Smith and Todd (2005) use Silverman's rule of thumb to choose the bandwidth.

5.9 Assessing the Unconfoundedness Assumption

5.10 Selection of Covariates

5.11 Comparing Methods

Lalonde (1986) used the data from the NSW experiment in combination with non-experimental comparison groups to assess how well a set of econometric evaluation methods would recover the experimental estimates.

6 Selection on Unobservables

6.1 Bounds

In a series of papers Manski (1990, XXXX XXXX) has developed a general framework for inference in settings where the parameters of interest are not identified. The 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*. Here we start off by discussing this 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, $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$. We can decompose this as

$$\tau = \mathbb{E}[Y_i(1)|W_i=1]\cdot\Pr(W_i=1) + \mathbb{E}[Y_i(1)|W_i=0]\cdot\Pr(W_i=0)$$
$$- \left(\mathbb{E}[Y_i(0)|W_i=1]\cdot\Pr(W_i=1) + \mathbb{E}[Y_i(0)|W_i=0]\cdot\Pr(W_i=0)\right).$$

Of the eight components of this expression we can estimate six. The data are not informative regarding the remaining two, $\mathbb{E}[Y_i(1)|W_i=0]$ and $\mathbb{E}[Y_i(0)|W_i=1]$. From the fact that the outcome is binary we can deduce that these two conditional expectations must lie inside the interval $[0,1]$, but we cannot say anymore without additional assumptions. This implies that without additional assumptions we can be sure that

$$\tau \in [\mathbb{E}[Y_i(1)|W_i=1]\cdot\Pr(W_i=1) - \Pr(W_i=1) - \mathbb{E}[Y_i(0)|W_i=0]\cdot\Pr(W_i=0),$$
$$\mathbb{E}[Y_i(1)|W_i=1]\cdot\Pr(W_i=1) + \Pr(W_i=0) - \mathbb{E}[Y_i(0)|W_i=0]\cdot\Pr(W_i=0)].$$

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.

In this specific case the bounds are not particularly informative. The width of the bounds 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 individual.) More generally, we may be able to add some assumptions short of making the type of assumption like unconfoundedness that gets us back to the identified case. With such assumptions we may be able to tighten the bounds and obtain informative results without the strong assumptions considered elsewhere. The presence of covariates does not directly narrow these bounds if we maintain that within

subpopulations homogenous in the covariates one cannot be sure about the relation between potential outcomes and covariates. However, the presence of covariates increases the scope for additional assumptions that may tighten the bounds.

This discussion has solely 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. Imbens and Manski (2004) and Chernozhukov, Hong and Tamer (2005) discuss construction of confidence intervals in settings with partial identification. Imbens and Manski (2004) focus on confidence sets for the parameter of interest (τ in this case). 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 1.95 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 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. It depends on the context what the object of interest is.

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 developed by Manski (XXXX), as well as approaches relying on alternative assumptions such as instrumental variables. However, there is an important alternative that has not received much attention in the economics literature. Instead of completely relaxing the unconfoundedness assumption the idea is to relax it slightly. More specifically, violations of unconfoundedness correspond to the presence of unobserved covariates that are both correlated with the potential outcomes and the treatment indicator. The amount of bias these violations can induce depends on the strength of these correlations. Sensitivity analyses investigate results obtained under unconfoundedness can be changed substantially or even overturned by modest violations of the unconfoundedness assumption.

To be specific, consider a job training program with voluntary enrollment. Suppose that we have detailed labor market histories. We may still 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. 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 year, 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. The sensitivity analyses developed by Rosenbaum and Rubin (XXXX) formalizes this and provides a tool for making such assessments.

The second approach to relaxing the unconfoundedness assumption does not drop the assumption completely. Instead it allows for moderate deviations from unconfoundedness and assesses how much the implied average treatment effects change. We consider two specific approaches here. The first one originates with Rosenbaum and Rubin (1984). In this approach the existence of a binary unobserved covariate is postulated. Conditional on this unobserved covariate unconfoundedness holds, but since it is not observed one cannot use the standard methods. Assumptions are made regarding the correlation between this covariate and the potential outcomes and between this covariate and the treatment assignment. If these correlations would need to be particularly strong in order to change the results from the analysis based on unconfoundedness the results are viewed as robust. Imbens (2003) applies this sensitivity analysis to data from labor market training programs. The second approach is associated with work by Rosenbaum. Similar to the Rosenbaum-Rubin approach this 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.

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 covariate U_i:

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

Rosenbaum and Rubin (1984) then consider both the conditional distribution of the potential outcomes given observed and unobserved covariates and and conditional probability of assignment given observed and unobserved covariates. Rather than attempting to estimate both these conditional distributions the idea 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 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 this is done is fully parametric settings. First we specify the conditional distribution of U_i given X_i to be binomial with $\Pr(U_i = 1|X_i) = \Pr(U_i = 1|X_i) = 1/2$.

For example, one may specify the assignment probability as a logistic regression function,

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

We cannot estimate α_0, α_1 and α_2 from data on (W_i, X_i) alone. However, if we specify a value for α_2, it is straightforward to estimate α_0 and α_1.

Similarly, we may specify

$$Y_i(w)|X_i, U_i \sim \mathcal{N}\left(\beta_{w0} + \beta_{w1}' X_i + \beta_{w2} \cdot U_i, \sigma_w^2\right).$$

Again the aim is not to estimate all parameters β_0, β_1, and β_2. Rather, we specify β_2 and then estimate β_0 and β_1 using maximum likelihood methods.

Given the estimated and specified parameters we can then derive the value of the average treatment effect. Finally we vary the specified parameters β_2 and α_2 over some range of reasonable values and see whether this changes the value of τ substantially.

Rosenbaum and Rubin (XXXX) apply this in a setting with binary outcomes. They fix the marginal distribution of the unobserved covariate to be binary with $p = \Pr(U = 1)$, and assume independence of U and X. They specify a logist distribution for the treatment assignment:

$$\Pr(W = 1|X, U) = \frac{\exp(\alpha_0 + \alpha_1' X + \alpha_2 \cdot U)}{\exp(\alpha_0 + \alpha_1' X + \alpha_2 \cdot U)}.$$

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

$$\Pr(Y(w) = 1|X, U) = \frac{\exp(\beta_{w0} + \beta_{w1}' X + \beta_{w2} \cdot U)}{\exp(\beta_{w0} + \beta_{w1}' X + \beta_{w2} \cdot U)}.$$

The average treatment effect can be expressed in terms of the parameters of this model and the distribution of the observable covariates:

$$\tau = \frac{1}{N} \sum_{i=1}^{N} p \left(\frac{\exp(\beta_{10} + \beta_{11}' X + \beta_{12})}{\exp(\beta_{10} + \beta_{11}' X + \beta_{12})} - \frac{\exp(\beta_{00} + \beta_{01}' X + \beta_{02})}{\exp(\beta_{00} + \beta_{01}' X + \beta_{02})} \right)$$

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

Here we have integrated out the unobserved covariate U. We do not know the parameters (p, α, β). We divide these into two sets. First the sensitivity parameters $(p, \alpha_2, \beta_{02}, \beta_{12})$, and then the remaining parameters $(\alpha_0, \alpha_1, \beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})$. The plan is to fix the first set of parameters and estimate the others by maximum likelihood, and then translate this into an estimate for τ. We do this for reasonable sets of values for the first set of sensitivity parameters and obtain a set of values for τ.

The key question is how to choose the set of reasonable values for the sensitivity parameters. Imbens (2003) suggests linking the parameters to the effects of the observed covariates on assignment and potential outcomes. In order to make the effect sizes comparable the suggestion is 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.

6.2.2 Rosenbaum's Method for Sensitivity Analysis

Rosenbaum in a series of papers (Rosenbaum, XXXX) developed a slightly different approach. The major advantage of his approach is that it requires fewer tuning parameters than the Rosenbaum-Rubin approach. 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. Consider two units i and j with the same value for the covariates, $x_i = x_j$. If the unconfoundedness assumption 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 odds ratio:

$$\frac{1}{\Gamma} \leq \frac{e_i \cdot (1 - e_j)}{(1 - e_i) \cdot e_j} \leq \Gamma.$$

If $\Gamma = 1$ we are back in the setting with unconfoundedness. If we allow $\Gamma = \infty$ we get back to the bounds in the absence of the unconfoundedness assumption. 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.

Suppose that a test of the null hypothesis of no effect has a p-value of 0.0001 under the assumption of unconfoundedness. If 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 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.

6.2.3 What makes some studies more robust than others?

Consider four parameters: estimated effect, standard error, t-statistic, proportion of treated and controls, and sample size. (t is ratio of estimate and standard error).

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. We largely limit the discussion to the case with a binary endogenous variable. Early on in this literature the focus was 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. Typical is the paper by Heckman (1990). Heckman uses a latent index model for the binary endogenous regressor, with the latent index linear in the instrument. He shows that if the instrument takes on values on the whole real line the average effect of the treatment is identified. Key is that the instrument can be choosen to make the probability of either value of the binary endogenous regressor arbitrarily close to zero. This type of identification is sometimes referred to as *identification at infinity*. The practical usefulness of such identification results is fairly limited. Finding credible instruments is typically difficult enough without also requiring that they shift the probability of the treatment close to zero and one. In fact, the trend in the literature at that time towards instruments that were more credible on theoretical grounds made it even more difficult to find cases that approximately satisfied these support conditions. Imbens and Angrist (1994) got around this problem by changing the focus to average effects for subpopulations other than the that of the individuals receiving the treatment. Björklund, A. and R. Moffitt (1987). Bloom (1984)

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 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. They use $W_i(0)$ and $W_i(1)$ to denote the treatment levels 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 values for the treatment received 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 as it does not require the researcher to specify a regression function in order to define the residuals.

They then introduce a new variable, the 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 the pair $(W_i(0), W_i(1))$. With both the regressor and instrument binary the total number of types fully describing the set of responses to the various levels of the instrument is four. It is useful to list them explicitly:

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

The labels nevertaker, complier, defier and alwaystaker (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.

Imbens and Angrist then make an additional assumption that they refer to as monotonicity. It 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 both compliers and defiers, and it is therefore sometimes referred to as the "no-defiance" assumption (Balke and Pearl, 1994; Pearl, 2000).

Under these two assumptions, independence of the potential outcomes and the instrument and monotonicity, Imbens and Angrist show that one can identify the average effect of the treatment for the subpopulation of compliers. It is obvious that one cannot identify the average effect of the treatment for any of the other groups as neither never-takers nor always-takers are observed in both treatment arms, and defiers are ruled out by assumption. This does not imply, however, that one can identify the average effect for compliers. In order to see this result it is useful to consider the subpopulations defined by instrument and treatment. Table 3 shows the information we have about the individual's type given the monotonicity assumption.

Consider individuals with $(Z_i, W_i) = (1, 0)$. Because of monotonicity such individuals can only be nevertakers. However, consider now individuals with $(Z_i, W_i) = (0, 0)$. Such individuals can be either compliers or alwaystakers. We cannot infer the type of such individuals from the observed data alone.

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, nevertakers, alwaystakers and compliers (using the fact that the monotonicity assumption rules out the presence of defiers). Call these population shares P_t, for $t \in \{n, a, c\}$. Consider the subpopulation with $Z_i = 0$. Within this subpopulation we observe $W_i = 1$ only for alwaystakers. Hence the conditional probability of $W_i = 1$ given $Z_i = 0$ is equal to the population share of alwaystakers: $P_a = \Pr(W = 1|Z = 0)$. Similarly, in the subpopulation with $Z_i = 1$ we observe $W_i = 0$ only for nevertakers. Hence the population share of nevertakers is equal to the conditional probability of $W_i = 0$ given $Z_i = 1$: $P_n = \Pr(W = 0|Z = 1)$. The population share of compliers is then obtained by subtracting the population shares of nevertakers and alwaystakers from one. The second step uses the distribution of Y given (Z, W). 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 nevertakers. 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. Since we already inferred the population shares of the nevertakers and compliers as well as the distribution of $Y_i|W_i = 0, T_i = n$, we can obtain the conditional distribution of $Y_i|W_i = 0, T_i = c$. Similarly we can infer the the conditional distribution of $Y_i|W_i = 1, T_i = c$. The average difference between these two conditional distributions is the Local Average Treatment Effect (LATE, Imbens and Angrist, 1994):

$$\tau^{\text{LATE}} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i(0) = 0, W_i(1) = 1].$$

6.3.2 A Multi-valued Instrument

6.4 Regression Discontinuity Designs

Regression discontinuity methods have been around for a long time in the statistics literature, but more recently they have attracted a considerable amount of interest in the economics literature (VanderKlaauw, XXXX, Lee xxxx, Hahn, Todd and VanderKlaauw XXXX, Porter XXXX, Black ().

There are two settings, the sharp and the fuzzy regression discontinuity designs. In the first

the assignment is a deterministic function of one of the covariates:

$$W_i = 1\{X_i \geq c\}.$$

All units with a covariate value of at least c are assigned to the treatment group and all units with a covariate value less than c are assigned to the control group.

6.4.1 The Sharp Regression Discontinuity Design

6.4.2 The Fuzzy Regression Discontinuity Design

6.5 Difference-in-Differences Methods

Since the work by Ashenfelter and Card (XXXX) the use of difference in differences methods has become very widespread. The simplest set up is one where outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. The average gain in the second ("control" group) is substracted from the gain in the first, the "treatment" group. This 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 trends. We discuss here the standard set up as well as the recent extensions by Athey and Imbens (2005) who develop a functional form free version of the difference in difference methodology.

6.5.1 Repeated Cross-sections

The standard model for the DID design 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. Letting the outcome be Y_i, the observed data are the triple (Y_i, G_i, T_i).[3] Using the potential outcome notation advocated in the treatment effect literature by Rubin (1974, 1978), let Y_i^N denote the outcome for individual i if that individual does not receive the treatment, and let Y_i^I be the outcome for the same individual if he or she does receive the treatment. Thus, if I_i is an indicator for the treatment, the realized (observed) outcome for individual i is

$$Y_i = Y_i^N \cdot (1 - I_i) + I_i \cdot Y_i^I.$$

In the DID setting we consider, $I_i = G_i \cdot T_i$.

In the standard DID model the outcome for individual i in the absence of the intervention satisfies

$$Y_i^N = \alpha + \beta \cdot T_i + \gamma \cdot G_i + \varepsilon_i. \tag{6.3}$$

[3] In Sections ?? ?? and we discuss cases with exogenous covariates.

The second coefficient, β, represents the time component. The third coefficient, γ, represents a group-specific, time-invariant component.[4] The third 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. $\varepsilon_i \perp (G_i, T_i)$, and is normalized to have mean zero. The standard DID estimand is

$$\begin{aligned} \tau^{DID} &= \Big[\mathbb{E}[Y_i|G_i=1,T_i=1] - \mathbb{E}[Y_i|G_i=1,T_i=0]\Big] \\ &\quad -\Big[\mathbb{E}[Y_i|G_i=0,T_i=1] - \mathbb{E}[Y_i|G_i=0,T_i=0]\Big]. \end{aligned} \tag{6.4}$$

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.

Note that the full independence assumption $\varepsilon_i \perp (G_i, T_i)$ (e.g., Blundell and MaCurdy, 2000) is stronger than necessary for τ^{DID} to give the average treatment effect. One can generalize this framework and allow for general forms of heteroskedasticity by group or time by relaxing the assumption to only mean-independence (e.g. Abadie (2001)), or zero correlation between ε_i and (G_i, T_i). Our proposed model will nest the DID model with independence (which for further reference will be labeled the standard DID model), but not the DID model with mean-independence.[5]

The interpretation of the standard DID estimand depends on assumptions about how outcomes are generated in the presence of the intervention. It is often assumed that the treatment effect is constant across individuals, so that $Y_i^I - Y_i^N = \tau$. Combining this restriction with the standard DID model for the outcome without intervention this leads to a model for the realized outcome

$$Y_i = \alpha + \beta \cdot T_i + \gamma \cdot G_i + \tau \cdot I_i + \varepsilon_i.$$

More generally, the effect of the intervention might differ across individuals. Then, the standard DID estimand gives the average effect of the intervention on the treatment group.

6.5.2 Multiple Groups and Multiple Periods

6.5.3 Standard Errors in the Multiple Group and Multiple Period Case

Donald and Lang () Bertrand, Duflo and Mullainathan ().

[4] In some settings, it is more appropriate to generalize the model to allow for a time-invariant individual-specific fixed effect γ_i, potentially correlated with G_i. 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. See Section **??** for more discussion of panel data.

[5] The DID model with mean-independence assumes that, for a given scaling of the outcome, changes across subpopulations in the mean of Y_i have a structural interpretation, and as such are used in predicting the counterfactual outcome for the second-period treatment group in the absence of the treatment. In contrast, all differences across subpopulations in the other moments of the distribution of Y_i are ignored when making predictions. In the model we propose, all changes in the distribution of Y_i across subpopulations are given a structural interpretation and used for inference. Neither our model, nor the DID model with mean-independence, impose any restrictions on the data.

6.5.4 Panel Data

6.5.5 The Changes-in-Changes Model

Return to the setting with two groups and two time periods.

We propose to generalize the standard model in several ways. First, we assume that in the absence of the intervention, the outcomes satisfy

$$Y_i^N = h(U_i, T_i), \tag{6.5}$$

with $h(u,t)$ increasing in u. The random variable U_i represents the unobservable characteristics of individual i, and (6.5) 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 $U_i \perp T_i \mid G_i$. The standard DID model in (6.3) embodies three additional assumptions, namely

$$U_i - \mathbb{E}[U_i|G_i] \perp G_i \qquad \text{(additivity)} \tag{6.6}$$

$$h(u,t) = \phi(u + \delta \cdot t), \qquad \text{(single index model)} \tag{6.7}$$

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

$$\phi(\cdot) \text{ is the identity function.} \qquad \text{(identity transformation)} \tag{6.8}$$

Thus the proposed model nests the standard DID as a special case. The mean-independence DID model is not nested; rather, the latter model requires that changes over time in moments of the outcomes other than the mean are not relevant for predicting the mean of Y_i^N. Note also that in contrast to the standard DID model, our assumptions do not depend on the scaling of the outcome, for example whether outcomes are measured in levels or logarithms.[6]

A natural extension of the standard DID model might have been to maintain assumptions (6.6) and (6.7) but relax (6.8), to allow $\phi(\cdot)$ to be an unknown function.[7] Doing so would maintain an additive single index structure within an unknown transformation, so that

$$Y_i^N = \phi(\alpha + \gamma \cdot G_i + \delta \cdot T_i + \varepsilon_i). \tag{6.9}$$

However, this specification still imposes substantive restrictions, for example ruling out some models with mean and variance shifts both across groups and over time.[8]

In the proposed model, 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 attributed to 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

[6] To be precise, we say that a model is invariant to the scaling of the outcome if, given the validity of the model for Y, the same assumptions remain valid for any strictly monotone transformation of the outcome.

[7] Ashenfelter and Greenstone (2004) consider models where $\phi(\cdot)$ is a Box-Cox transformation with unknown parameter.

[8] For example, suppose that $Y_i^N = \alpha + \delta_1 \cdot T_i + (\gamma \cdot G_i + \varepsilon_i) \cdot (1 + \delta_2 \cdot T_i)$. In the second period there is a shift in the mean as well as unrelated shift in the variance, meaning the model is incompatible with 6.9.

outcome (in the absence of treatment) arise from the fact that $h(u,0)$ differs from $h(u,1)$, that is, the effect of the unobservable on outcomes changes over time. Like the standard model, our approach does not rely on tracking individuals over time; although the distribution of U_i is assumed not to change over time within groups, we do not make any assumptions about whether a particular individual has the same realization U_i in each period. Thus, the estimators we derive for our model will be the same whether we observe a panel of individuals over time or a repeated cross-section. We discuss alternative models for panel data in more detail in Section **??**.

Just as in the standard DID approach, if we only wish to estimate the effect of the intervention on the treatment group, no assumptions are required about how the intervention affects outcomes. To analyze the counterfactual effect of the intervention on the control group, we assume that in thepresence of the intervention,

$$Y_i^I = h^I(U_i, T_i)$$

for some function $h^I(u,t)$ 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^I, so the treatment effect, equal to $h^I(u,1) - h^N(u,1)$ 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.

Suppose that Assumptions **??**-**??** hold. Then the distribution of Y_{11}^N is identified, and

$$F_{Y^N,11}(y) = F_{Y,10}(F_{Y,00}^{-1}(F_{Y,01}(y))). \tag{6.10}$$

7 Multivalued and Continuous Treatments

Most of the recent econometric literature has focused on the case with a binary treatment. As a result we understand this case much better than we did a decade or two ago. However, we know much less about settings with multivalued and continuous treatments. Such cases are very common in practice. Social programs are rarely homogenous. Typically individuals are assigned to various activities and regimes 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 we have 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. In the fourth case we look at settings with a discrete multivalued treatment in the presence of endogeneity. In the final case we allow the treatment

to be continuous. The last two cases ty in closely with the simultaneous equations literature where somewhat separately from the program evaluation literature there has been much recent work on nonparametric identification. Especially in the discrete case many of the results here are negative in the sense that without unattractive restrictions on heterogeneity or functional form few objects of interest are identified. This is clearly still an area with much scope for further work.

7.1 Multivalued Discrete Treatments with Unconfoundedness Treatment Assignment

7.2 Continuous Treatments with Unconfoundedness Treatment Assignment

We are interested in the average causal effect of some treatment on some outcome. The treatment, denoted by W, takes on values in an interval $[0, 1]$. Associated with each unit i and each value of the treatment w there is a potential outcome, denoted by $Y_i(w)$. We are interested in average outcomes, $E\{Y(w)\}$, for all values of w, and in particular in differences of the form $E\{Y(v) - Y(w)\}$, the average causal effect of exposing all units to treatment v rather than treatment w. The average here is taken over the population of interest, which may be the population the sample is drawn from, or some subpopulation thereof. More generally we can look at average differences of functions of $Y(w)$ for different values of w, such as the distribution function of $Y(w)$ at a point. We observe, for each unit i in a random sample of size N drawn from a large population, the treatment W_i, the outcome associated with that treatment level $Y_i \equiv Y_i(W_i)$, and a vector of pre-treatment variables X_i.

The key assumption, maintained throughout the paper, is that adjusting for pre-treatment differences solves the problem of drawing causal inferences. This is formalised by using the concept of unconfoundedness. Assignment to treatment T is weakly unconfounded, given pre-treatment variables X, if

$$W_i \perp Y(w) \Big| X,$$

for all $w \in [0, 1]$. In the binary case Rosenbaum & Rubin (1983) make the stronger assumption that

$$W \perp (Y(0), Y(1)) \mid X,$$

which requires the treatment W 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, like the assumption used in Robins (1995). The definition of weak unconfoundedness is similar to that of 'missing at random' (Rubin, 1976; Little & Rubin, 1987, p. 14) in the missing data literature.

Although in substantive terms the weak unconfoundedness assumption is not very different from the assumption used by Rosenbaum & Rubin (1983), it is important that one does not need the stronger assumption to validate estimation of the expected value of $Y(w)$ by adjusting for X: $E[Y(w)|X] = E[Y(w)|W = w, X] = E[Y|W = w, X]$ Average outcomes can then be estimated by averaging these conditional means: $E[Y(w)] = E[E\{Y(w)|X\}\Big\}$. In practice it can

be difficult to estimate $E\{Y(w)\}$ in this manner when the dimension of X is large, because the first step requires estimation of the expectation of $Y(w)$ given the treatment level and all pre-treatment variables, and this motivated Rosenbaum & Rubin (1983) to develop the propensity score methodology.

The generalized propensity score is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables:

$$r(w,x) \equiv \Pr(W = w|X = x).$$

Suppose assignment to treatment W is weakly unconfounded given pre-treatment variables X. Then, by the same argument as in the binary treatment case, assignment is weakly unconfounded given the generalised propensity score:

$$D(w) \;\perp\; Y(w) \;\Big|\; r(w,X),$$

for all $w \in [0,1]$. 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 is independent of the set of potential outcomes $\{Y(w)\}_{w\in[0,1]}$. Such a scalar function may exist if additional structure is imposed on the assignment mechanism; see for example, Joffe & 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 $w \in [0,1]$,

$$\beta(w,r) \equiv E\{Y(w)|r(w,X) = r\} = E\{Y|W = w, r(W,X) = r\},$$

and second,

$$E\{Y(w)\} = E\{\beta(w,r(w,X))\}.$$

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 & Rosenbaum, 1984; Rosenbaum, 1995, p. 79) is to estimate the propensity score using a logistic regression. 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.

In the second step the conditional expectation $\beta(w,r) = E\{Y|W = w, r(W,X) = r\}$ is estimated. Again the implementation may be different in the case where the levels of the treatment qualitatively distinct than the case where smoothness of the conditional expectation function in w is appropriate.

In the third step the average response at treatment level w is estimated as the average of the estimated conditional expectation, $\hat{\beta}(w,r(w,X))$, averaged over the distribution of the

pre-treatment variables. Note that to get the average $E\{Y(w)\}$ the second argument in the conditional expectation $\beta(w,r)$ is evaluated at $r(w,X_i)$, not at $r(W_i,X_i)$.

As an alternative to the above implementation one can use the inverse of the generalised propensity score to weight the observations, using the following equality:

$$E\left\{\frac{Y\cdot D(w)}{r(W,X)}\right\} = E\{Y(w)\}.$$

It appears difficult to exploit smoothness of the outcome in the level of the treatment in this weighting approach. Similarly matching approaches where units are grouped in a way to allow causal comparisons within matches appear less well suited to the multi-valued treatment case.

7.2.1 Dynamic Treatments with Unconfounded Treatment Assignment

Robbins and Gill ()

Lechner ()

Heckman ()

7.3 Multivalued Discrete Endogenous Treatments

7.4 Continuous Endogenous Treatments with Endogenous Treatments

8 Conclusion

References

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

Abadie, Alberto, (2003b): "Semiparametric Difference-in-Differences Estimators," forthcoming, *Review of Economic Studies*.

Abadie, A., and G. Imbens, (2005), "Large Sample Properties of Matching Estimators for Average Treatment Effects," forthcoming, *Econometrica*.

Abadie, A., D. Drukker, H. Herr, and G. Imbens, (2003), "Implementing Matching Estimators for Average Treatment Effects in STATA," unpublished manuscript, department of economics, University of California, Berkeley.

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.

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

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

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. 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.

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 (2002), "Identification and Inference in Nonlinear Difference-In-Differences Models," NBER technical working paper # 280.

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

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.

Bertrand, M., E. Duflo, and S. Mullainathan, (2001): "How Much Should We Trust Differences-in-Differences Estimates?" Working Paper, MIT.

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

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.

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.

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.

Chernozhukov, V., and C. Hansen, (2001), "An IV Model of Quantile Treatment Effects," unpublished working paper, Department of Economics, MIT.

Chernozhukov, V., H. Hong and E. Tamer, (2004): "Parameter Set Inference in a Class of Econometric Models," unpublished manuscript, Department of Economics, Princeton University.

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.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2005), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand," unpublished manuscript, Department of Economics, UC Berkeley.

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*.

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, (2001): "Inference with Difference in Differences and Other Panel Data," unpublished manuscript, Boston University.

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

Engle, R., D. Hendry, and J.-F. Richard, (1974) "Exogeneity," *Econometrica*, 51(2): 277-304.

Firpo, S. (2003), "Efficient Semiparametric Estimation of Quantile Treatment Effects," PhD Thesis, Chapter 2, Department of Economics, University of California, Berkeley.

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. (2000), "Treatment Evaluation: Matching versus Local Polynomial Regression," Discussion paper 2000-17, Department of Economics, University of St. Gallen.

Frölich, M. (2002), "What is the Value of knowing the propensity score for estimating average treatment effects", Department of Economics, University of St. Gallen.

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

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.

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, (1984), "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.

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

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.

Hotz, V. J., G. Imbens, and J. Klerman, (2001), "The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program,"unpublished manuscript, Department of Economics, UCLA.

Hotz J., G. Imbens, and J. Mortimer (2003), "Predicting the Efficacy of Future Training Programs Using Past Experiences,"forthcoming, *Journal of Econometrics*.

Ichimura, H., and O. Linton, (2001), "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators."Institute for Fiscal Studies, cemmap working paper cwp04/01.

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

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.

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

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

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. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*.

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.

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, (2002), "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., (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.

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, 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.

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

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

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

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.

Pearl, J.,, (2000), *Casuality*, .

Politis, D., and J. Romano,, (1999), *Subsampling*, Springer Verlag.

Porter, J. (2003), "Estimation in the Regression Discontinuity Model," Unpublished Manuscript, Harvard University.

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.

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, , (xxxx),

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., and N. Thomas, (1992), "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20 (2) 1079-1093.

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, (2003), "Does Matching Address LaLonde's Critique of Nonexperimental Estimators," forthcoming in *Journal of Econometrics*.

Stock, J., (1989), "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84(406): 567-575.

Van Der Klaauw, W., (2002), "A Regression–discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment", *International Economic Review*, 43(4): 1249-1287

Wooldridge, J., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Wooldridge, J., (2002), "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification," *Econometrica*

Wooldridge, J., (2005),

Zhao, Z., (2004), "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics and an Application," forthcoming, *review of economics and statistics*.

Table 1: Five Observations from the GAIN Experiment in Los Angeles: Earnings in First Year Post-randomization

Individual	Potential Outcomes $Y_i(0)$	$Y_i(1)$	Actual Treatment	Observed Outcome Y_i
1	1,319	?	1	1,319
2	5,826	?	1	5,826
3	1,735	?	0	1,735
4	?	0	0	0
5	?	948	1	948

Table 2: P-values for Fisher Exact Tests: Ranks versus Levels

Program	Location	sample size controls	treated	t-test	p-values FET (levels)	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 3: Type by Observed Variables

		Z_i = 0	Z_i = 1
W_i	0	Nevertaker/Complier	Nevertaker
	1	Alwaystaker	Alwaystaker/Complier