NBER WORKING PAPER SERIES

USING MATCHING, INSTRUMENTAL VARIABLES AND CONTROL FUNCTIONS TO ESTIMATE ECONOMIC CHOICE MODELS

James Heckman
Salvador Navarro-Lozano

Working Paper 9497
http://www.nber.org/papers/w9497

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2003

This research was supported by NSF SES-0099195, NIH HD34958-04 and the American Bar Foundation. Navarro-Lozano acknowledges financial support from CONACYT, Mexico. We thank Alberto Abadie, Pedro Carneiro, Michael Lechner and Costas Meghir for helpful comments. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Using Matching, Instrumental Variables and Control Functions
to Estimate Economic Choice Models
James Heckman and Salvador Navarro-Lozano
NBER Working Paper No. 9497
February 2003
JEL No. C31

ABSTRACT

This paper investigates four topics. (1) It examines the different roles played by the propensity score (probability of selection) in matching, instrumental variable and control functions methods. (2) It contrasts the roles of exclusion restrictions in matching and selection models. (3) It characterizes the sensitivity of matching to the choice of conditioning variables and demonstrates the greater robustness of control function methods to misspecification of the conditioning variables. (4) It demonstrates the problem of choosing the conditioning variables in matching and the failure of conventional model selection criteria when candidate conditioning variables are not exogenous.

James J. Heckman
Department of Economics
The University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
jjh@uchicago.edu

Salvador Navarro-Lozano
Department of Economics
The University of Chicago
1126 East 59th Street
Chicago, IL 60637
snavarro@uchicago.edu

1 Introduction

The method of matching has become popular in evaluating social programs because it is easy to understand and easy to apply. It uses observed explanatory variables to adjust for differences in outcomes unrelated to treatment that give rise to selection bias. Propensity score matching as developed by Rosenbaum and Rubin (1983) is particularly simple to apply. The propensity score is the probability that an agent takes treatment. If the analyst knows (without having to estimate) the probability that a person takes treatment, and the assumptions of matching are fulfilled, he can condition on that known probability and avoid selection in means and marginal distributions. This choice probability also plays a central role in econometric selection models based on the principle of control functions (Heckman, 1980; Heckman and Robb, 1986, reprinted 2000; Heckman and Hotz, 1989; Ahn and Powell, 1993) and in instrumental variable models (see e.g. Heckman and Vytlacil, 1999, 2001, 2003 or Heckman, 2001).

The multiple use of the propensity score in different statistical methods has given rise to some confusion in the applied literature.[1] This paper seeks to clarify the different assumptions that justify the propensity score in selection, matching and instrumental variables methods. We develop the following topics:

1. We orient the discussion of the selection of alternative estimators around the economic theory of choice. We compare the different roles that the propensity score plays in three widely used econometric methods, and the implicit economic assumptions that underlie applications of these methods.

2. Conventional matching methods do not distinguish between excluded and included variables.[2] We show that matching breaks down when there are variables that predict the choice of treatment perfectly whereas control function methods take advantage of exclusion restrictions and use the information available from perfect prediction to obtain identification. Matching assumes away the possibility of perfect prediction while selection models rely on this property in limit sets.

3. We define the concepts of "relevant" information and "minimal relevant" information, and distinguish agent and analyst information sets. We state clearly what information is required to identify different treatment parameters. In particular we show that when the analyst does not have access to the "minimal relevant" information, matching estimates of different treatment parameters are biased. Having more information, but not all of the "minimal relevant" information, can increase the bias compared to having less information. Enlarging the analyst's information set with variables that do not belong in the relevant information set may either increase or decrease the bias from matching. Because the method of control functions explicitly models omitted relevant variables, rather than

assuming that there are none, it is more robust to omitted conditioning variables.

4. The method of matching offers no guidance as to which variables to include or exclude in conditioning sets. Such choices can greatly affect inference. There is no support for the commonly used rules of selecting matching variables by choosing the set of variables that maximizes the probability of successful prediction into treatment or by including variables in conditioning sets that are statistically significant in choice equations. This weakness is shared by many econometric procedures but is not fully appreciated in recent applications of matching which apply these selection rules when choosing conditioning sets.

To simplify the exposition, throughout this paper we consider a one-treatment, two-outcome model. Our main points apply more generally.

2 A Prototypical Model of Economic Choice

To focus the discussion, and interpret the implicit assumptions underlying the different estimators presented in this paper, we present a benchmark model of economic choice. For simplicity we consider two potential outcomes (Y_0, Y_1). $D = 1$ if Y_1 is selected. $D = 0$ if Y_0 is selected. Agents pick their outcome based on utility maximization. Let V be utility. We write

$$V = \mu_V(Z, U_V) \quad D = 1 \ (V > 0), \tag{1}$$

where the Z are factors (observed by the analyst) determining choices, U_V are the unobserved (by the analyst) factors determining choice and 1 is an indicator function ($1(A) = 1$ if A is true; $1(A) = 0$ otherwise). We consider differences between agent information sets and analyst information sets in Section (6).

Potential outcomes are written in terms of observed variables (X) and unobserved (by the analyst) outcome-specific variables

$$Y_1 = \mu_1(X, U_1) \tag{2a}$$

$$Y_0 = \mu_0(X, U_0). \tag{2b}$$

We assume throughout that U_0, U_1, U_V are (absolutely) continuous random variables and that all means are finite. The individual level treatment effect is

$$\Delta = Y_1 - Y_0.$$

More familiar forms of (1), (2a) and (2b) are additively separable:

$$V = \mu_V(Z) + U_V \quad E(U_V) = 0 \tag{1'}$$

$$Y_1 = \mu_1(X) + U_1 \quad E(U_1) = 0 \tag{2a'}$$

$$Y_0 = \mu_0(X) + U_0 \quad E(U_0) = 0. \tag{2b'}$$

Additive separability is not strictly required in matching, or most versions of selection (control function) models. However, we use the additively separable representation throughout most of this paper because of its familiarity noting when it is a convenience and when it is an essential part of a method.

The distinction between X and Z is crucial to the validity of many econometric procedures. In matching as conventionally formulated there is no distinction between X and Z. The roles of X and Z in alternative estimators are explored in this paper.

3 Parameters of Interest in this Paper

There are many parameters of interest that can be derived from this model if $U_1 \neq U_0$ and agents use some or all of the U_0, U_1 in making their decisions (see Heckman and Robb, 1985, 1986; Heckman, 1992; Heckman, Smith and Clements, 1997 ; Heckman and Vytlacil, 2001 and Heckman, 2001). Here we focus on certain means because they are traditional. As noted by Heckman and Vytlacil (2000) and Heckman (2001), the traditional means do not answer many interesting economic questions.

The traditional means are:

$$\begin{aligned} ATE &: E(Y_1 - Y_0|X) \text{ (Average Treatment Effect)} \\ TT &: E(Y_1 - Y_0|X, D = 1) \text{ (Treatment on the Treated)} \\ MTE &: E(Y_1 - Y_0|X, Z, V = 0) \text{ (Marginal Treatment Effect).} \end{aligned}$$

The MTE is the marginal treatment effect introduced into the evaluation literature by Björklund and Moffitt (1987). It is the average gain to persons who are indifferent to participating in sector 1 or sector 0 given X, Z. These are persons at the margin, defined by X and Z. Heckman and Vytlacil (1999, 2000) show how the MTE can be used to construct all mean treatment parameters, including the policy relevant treatment parameters, under the conditions specified in their papers.

4 The Selection Problem

Let $Y = DY_1 + (1-D)Y_0$. Samples generated by choices have the following means which are assumed to be known:

$$E(Y|X, Z, D = 1) = E(Y_1|X, Z, D = 1)$$

and

$$E(Y|X, Z, D = 0) = E(Y_0|X, Z, D = 0)$$

for outcomes of Y_1 for participants and the outcomes of Y_0 for non-participants, respectively. In addition, choices are observed so that in large samples $\Pr(D = 1|X, Z)$, *i.e.*, the probability of choosing treatment is known. From the means we can integrate out Z given X and D to construct

$$E(Y_1|X, D = 1) \text{ and } E(Y_0|X, D = 0).$$

The biases from using the difference of these means to construct various counterfactuals are, for the three parameters studied in this paper:

$$\begin{aligned} \text{Bias } TT &= [E(Y|X, D = 1) - E(Y|X, D = 0)] - [E(Y_1 - Y_0|X, D = 1)] \\ &= [E(Y_0|X, D = 1) - E(Y_0|X, D = 0)]. \end{aligned}$$

In the case of additive separability

$$\text{Bias } TT = E[U_0|X, D = 1] - E[U_0|X, D = 0].$$

For ATE,

$$\text{Bias } ATE = E[Y|X, D = 1] - E(Y|X, D = 0) - [E(Y_1 - Y_0|X)].$$

In the case of additive separability

$$\text{Bias } ATE = [E(U_1|X, D = 1) - E(U_1|X)] - [E(U_0|X, D = 0) - E(U_0|X)].$$

For MTE,

$$\begin{aligned} \text{Bias } MTE &= E(Y|X, Z, D = 1) - E(Y|X, Z, D = 0) - E(Y_1 - Y_0|X, Z, V = 0) \\ &= [E(U_1|X, Z, D = 1) - E(U_1|X, Z, V = 0)] - [E(U_0|X, Z, D = 0) - E(U_0|X, Z, V = 0)] \end{aligned}$$

in the case of additive separability. The MTE is defined for a subset of persons indifferent between the two sectors and so is defined for X and Z. The bias is the difference between average U_1 for participants and marginal U_1 minus the difference between average U_0 for nonparticipants and marginal U_0. Each of these terms is a bias which can be called a selection bias.

5 How Different Methods Solve the Bias Problem

In this section we consider the identification conditions that underlie matching, control functions and instrumental variable methods to identify the three parameters using the data on mean outcomes. We start with the method of matching.

5.1 Matching

The method of matching as conventionally formulated makes no distinction between X and Z. Define the conditioning set as $W = (X, Z)$. The strong form of matching advocated by Rosenbaum and Rubin (and numerous predecessor papers) assumes that

$$(Y_1, Y_0) \perp\!\!\!\perp D|W \tag{M-1}$$

and

$$0 < \Pr(D = 1|W) = P(W) < 1, \tag{M-2}$$

where "$\perp\!\!\!\perp$" denotes independence given the conditioning variables after "|". Condition (M-2) implies that the treatment parameters can be defined for all values of W (*i.e.*, for each W, in very large samples there are observations for which we observe a Y_0 and other observations for which we observe a Y_1). Rosenbaum and Rubin show that under (M-1) and (M-2)

$$(Y_1, Y_0) \perp\!\!\!\perp D|P(W). \tag{M-3}$$

This reduces the dimensionality of the matching problem. They assume that P is known.[3] Under these assumptions, conditioning on P eliminates all three biases defined in section (4) because

$$\begin{aligned} E\left(Y_1|D = 0, P\left(W\right)\right) &= E\left(Y_1|D = 1, P\left(W\right)\right) = E\left(Y_1|P\left(W\right)\right) \\ E\left(Y_0|D = 1, P\left(W\right)\right) &= E\left(Y_0|D = 0, P\left(W\right)\right) = E\left(Y_0|P\left(W\right)\right). \end{aligned}$$

Thus for TT we can identify $E\left(Y_0|D = 1, P\left(W\right)\right)$ from $E\left(Y_0|D = 0, P\left(W\right)\right)$. In fact, we only need the weaker condition $Y_0 \perp\!\!\!\perp D|P(W)$ to remove the bias[4] because $E\left(Y_1|P\left(W\right), D = 1\right)$ is known, and only $E\left(Y_0|P\left(W\right), D = 1\right)$ is unknown. From the observed conditional means we can form ATE. Observe that since $ATE = TT$ for all X, Z under (M-1) and (M-2), the average person equals the marginal person, conditional on W, and there is no bias in estimating MTE.[5] The strong implicit assumption that the marginal participant in a program gets the same return as the average participant in the program, conditional on W, is an unattractive implication of these assumptions (see Heckman, 2001 and Heckman

and Vytlacil, 2003). The method assumes that all of the dependence between U_V and (U_1, U_0) is eliminated by conditioning on W:

$$U_V \perp\!\!\!\perp (U_1, U_0)|W.$$

This motivates the term "selection on observables" introduced in Heckman and Robb (1985; 1986, reprinted 2000).

Assumption (M-2) has the unattractive feature that if the analyst has too much information about the decision of who takes treatment so that $P(W) = 1$ or 0 the method breaks down because people cannot be compared at a common W. The method of matching assumes that, given W, some unspecified randomization device allocates people to treatment.

Introducing the distinction between X and Z allows the analyst to overcome the problem of perfect prediction if there are some variables Z not in X so that, for certain values of these variables, and for each X either $P(X, Z) = 1$ or $P(X, Z) = 0$. If P is a nontrivial function of Z (so $P(X, Z)$ varies with Z for all X) and X can be varied independently of Z,[6] and outcomes are defined solely in terms of X, this difficulty with matching disappears and treatment parameters can be defined for all values of X in its support (see Heckman, Ichimura and Todd, 1997).

Offsetting the disadvantages of matching, the method of matching with a known conditioning set that produces (M-1) does not require separability of outcome or choice equations, exogeneity of conditioning variables, exclusion restrictions or adoption of specific functional forms of outcome equations. Such features are common in conventional selection (control function) methods and conventional IV formulations although recent work in semiparametric estimation relaxes many of these assumptions, as we note below. Moreover, the method does not strictly require (M-1). One can get by with weaker mean independence assumptions:

$$\begin{aligned} E(Y_1|W, D = 1) &= E(Y_1|W) \\ E(Y_0|W, D = 0) &= E(Y_0|W), \end{aligned} \tag{M-1$'$}$$

in the place of the stronger (M-1) conditions. However, if (M-1$'$) is involved, the assumption that we can replace W by $P(W)$ does not follow from the analysis of Rosenbaum and Rubin, and is an additional new assumption.

In the recent literature, the claim is sometimes made that matching is "for free" (see, e.g., Gill and Robins, 2001). The idea is that since $E(Y_0|D = 1, W)$ is not observed, we might as well set it to

$E(Y_0|D=0,W)$, an implication of (M-1). This argument is correct so far as data description goes. Matching imposes just-identifying restrictions and in this sense –at a purely empirical level– is as good as any other just-identifying assumption in describing the data.

However, the implied economic restrictions are not "for free". Imposing that, conditional on X and Z, the marginal person is the same as the average person is a strong and restrictive feature of these assumptions and is not a "for free" assumption in terms of economic content.[7]

5.2 Control Functions

The principle motivating the method of control functions is different. (See Heckman, 1980 and Heckman and Robb, 1985, 1986, reprinted 2000, where this principle was developed). Like matching, it works with conditional expectations of (Y_1, Y_0) given (X, Z and D). Conventional applications of the control function method assume additive separability which is not required in matching. Strictly speaking, additive separability is not required in the application of control functions either.[8] What is required is a model relating the outcome unobservables to the observables, including the choice of treatment. The method of matching assumes that, conditional on the observables (X, Z), the unobservables are independent of D.[9] For the additively separable case, control functions are based on the principle of modeling the conditional expectations given X, Z and D :

$$\begin{aligned} E(Y_1|X,Z,D=1) &= \mu_1(X) + E(U_1|X,Z,D=1) \\ E(Y_0|X,Z,D=0) &= \mu_0(X) + E(U_0|X,Z,D=0). \end{aligned}$$

The idea underlying the method of control functions is to explicitly model the stochastic dependence of the unobservables in the outcome equations on the observables. This is unnecessary under the assumptions of matching because conditional on (X, Z) there is no dependence between (U_1, U_0) and D. Thus, if one can model $E(U_1|X,Z,D=1)$ and $E(U_0|X,Z,D=0)$ and these functions can be independently varied against $\mu_1(X)$ and $\mu_0(X)$ respectively, one can identify $\mu_1(X)$ and $\mu_0(X)$ up to constant terms.[10] Nothing in the method intrinsically requires that X, Z, or D be stochastically independent of U_1 or U_0, although conventional methods often assume that $(U_1, U_0, U_V) \perp\!\!\!\perp (X, Z)$.

If we assume that $U_1, U_V \perp\!\!\!\perp (X, Z)$ and adopt $(1')$ as the choice model,

$$E(U_1|X,Z,D=1) = E(U_1|U_V \geq -\mu_V(Z)) = K_1(P(X,Z)),$$

so the control function only depends on $P(X, Z)$. By similar reasoning, if $U_0, U_V \perp\!\!\!\perp (X, Z)$,

$$E(U_0|X,Z,D=0) = E(U_0|U_V < -\mu_V(Z)) = K_0(P(X,Z))$$

and the control function only depends on the propensity score. The key assumption needed to represent the control function solely as a function of $P(X,Z)$ is thus

$$(U_1, U_0, U_V) \perp\!\!\!\perp (X, Z).$$

Under these conditions

$$\begin{aligned} E(Y_1|X,Z,D=1) &= \mu_1(X) + K_1(P(X,Z)) \\ E(Y_0|X,Z,D=0) &= \mu_0(X) + K_0(P(X,Z)) \end{aligned}$$

with $\lim_{P\to 1} K_1(P) = 0$ and $\lim_{P\to 0} K_0(P) = 0$ where it is assumed that Z can be independently varied for all X, and the limits are obtained by changing Z while holding X fixed.[11] These limit results just say that when the probability of being in a sample in one there is no selection bias.

If $K_1(P(X,Z))$ can be independently varied from $\mu_1(X)$ and $K_0(P(X,Z))$ can be independently varied from $\mu_0(X)$, we can identify $\mu_1(X)$ and $\mu_0(X)$ up to constants. If there are limit sets $\mathbb{Z}_0$ and $\mathbb{Z}_1$ such that $\lim_{Z\to\mathbb{Z}_0} P(X,Z) = 0$ and $\lim_{Z\to\mathbb{Z}_1} P(X,Z) = 1$, then we can identify the constants, since in those limit sets we identify $\mu_1(X)$ and $\mu_0(X)$.[12] Under these conditions, it is possible to nonparametrically identify all three treatment parameters:

$$ATE = \mu_1(X) - \mu_0(X)$$

$$\begin{aligned} TT &= \mu_1(X) - \mu_0(X) + E(U_1 - U_0|X,Z,D=1) \\ &= \mu_1(X) - \mu_0(X) + K_1(P(X,Z)) + \left(\tfrac{1-P}{P}\right) K_0(P(X,Z)) \end{aligned} \quad [13]$$

$$\begin{aligned} MTE &= \mu_1(X) - \mu_0(X) + \frac{\partial\left[E(U_1 - U_0|X,Z,D=1)\,P(X,Z)\right]}{\partial(P(X,Z))} \\ &= \mu_1(X) - \mu_0(X) + \frac{\partial\left[P(X,Z)\left\{K_1(P(X,Z) + \frac{1-P}{P}K_0(P(X,Z)\right\}\right]}{\partial(P(X,Z))}. \end{aligned}$$

Unlike the method of matching, the method of control functions allows the marginal treatment effect to be different from the average treatment effect or from treatment on the treated. Although conventional practice is to derive the functional forms of $K_0(P)$, $K_1(P)$ by making distributional assumptions (e.g., normality, see Heckman, Tobias and Vytlacil (2001)), this is not an intrinsic feature of the method and there are many non normal and semiparametric versions of this method (see Heckman and Vytlacil, 2003 for a survey).

In its semiparametric implementation, the method of control functions requires an exclusion restriction (a Z not in X) to achieve nonparametric identification.[14] The method of matching does not. The method of

control functions requires that $P(X,Z)=1$ and $P(X,Z)=0$ to achieve full nonparametric identification. The conventional method of matching excludes this case. Both methods require that treatment parameters can only be defined on a common support:

$$\text{support}\,(X|D=1)\cap\text{support}\,(X|D=0)$$

A similar requirement is imposed on the generalization of matching with exclusion restrictions introduced in Heckman, Ichimura and Todd (1997). Exclusion, both in matching and selection models, makes it more likely to satisfy this condition.

In the method of control functions, $P(X,Z)$ is a conditioning variable used to predict U_1 conditional on D and U_0 conditional on D. In the method of matching, it is used to generate stochastic independence between (U_0,U_1) and D. In the method of control functions, as conventionally applied, $(U_0,U_1)\perp\!\!\!\perp(X,Z)$, but this is not intrinsic to the method.[15] This assumption plays no role in matching if the correct conditioning set is known (*i.e.*, one that satisfies (M-1) and (M-2)). However, as noted in section (6.6), exogeneity plays a key role in the selection of conditioning variables. The method of control functions does not require that $(U_0,U_1)\perp\!\!\!\perp D|(X,Z)$, which is a central requirement of matching. Equivalently, the method of control functions does not require

$$(U_0,U_1)\perp\!\!\!\perp U_V|(X,Z)$$

whereas matching does. Thus matching assumes access to a richer set of conditioning variables than is assumed in the method of control functions.

The method of control functions is more robust than the method of matching, in the sense that it allows for outcome unobservables to be dependent on D even conditioning on (X,Z), and it models this dependence, whereas the method of matching assumes no such dependence. Matching is thus a special case of the method of control functions in which under assumptions (M-1) and (M-2),

$$\begin{aligned}E(U_1|X,Z,D=1) &= E(U_1|X,Z)=E(U_1|P(W))\\ E(U_0|X,Z,D=0) &= E(U_0|X,Z)=E(U_0|P(W)).\end{aligned}$$

In the method of control functions in the case when $(X,Z)\perp\!\!\!\perp(U_0,U_1,U_V)$

$$\begin{aligned}E(Y|X,Z,D) &= E(Y_1|X,Z,D=1)D+E(Y_0|X,Z,D=0)(1-D)\\ &= \mu_0(X)+(\mu_1(X)-\mu_0(X))D+E(U_1|X,Z,D=1)D+E(U_0|P(X,Z),D=0)(1-D)\\ &= \mu_0(X)+(\mu_1(X)-\mu_0(X))D+E(U_1|P(X,Z),D=1)D+E(U_0|P(X,Z),D=0)(1-D)\\ &= \mu_0(X)+[\mu_1(X)-\mu_0(X)+K_1(P(X,Z))-K_0(P(X,Z))]D+K_0(P(X,Z)).\end{aligned}$$

Under assumptions (M-1) and (M-2) of the method of matching, we may write

$$E\left(Y|P\left(W\right),D\right)=\mu_0\left(P\left(W\right)\right)+\left[\left(\mu_1\left(P\left(W\right)\right)-\mu_0\left(P\left(W\right)\right)\right)+E\left(U_1|P\left(W\right)\right)-E\left(U_0|P\left(W\right)\right)\right]D+\left\{E\left(U_0|P\left(W\right)\right)\right\}.$$

Notice that

$$E\left(Y|P\left(W\right),D\right)=\mu_0\left(P\left(W\right)\right)+\left[\mu_1\left(P\left(W\right)\right)-\mu_0\left(P\left(W\right)\right)\right]D,$$

since $E\left(U_1|P\left(W\right)\right)=E\left(U_0|P\left(W\right)\right)=0$.

The treatment effect is identified from the coefficient on D. Condition (M-2) guarantees that D is not perfectly predictable by W so the variation in D identifies this parameter. Since $\mu_1\left(P\left(W\right)\right)-\mu_0\left(P\left(W\right)\right)=ATE$ and $ATE=TT=MTE$, the method of matching identifies all of the mean treatment parameters. Under the assumptions of matching, when means of Y_1 and Y_0 are the parameters of interest, the bias terms vanish. They do not in the more general case considered by the method of control functions. This is the mathematical counterpart of the randomization implicit in matching: conditional on W or $P\left(W\right)$, (U_1,U_0) are random with respect to D. The method of control functions allows them to be nonrandom with respect to D. In the absence of functional form assumptions, it requires an exclusion restriction to separate out $K_0\left(P\left(X,Z\right)\right)$ from the coefficient on D. Matching produces identification without exclusion restrictions whereas identification with exclusion restrictions is a central feature of the control function method in the absence of functional form assumptions.

The fact that the control function approach is more general than the matching approach is implicitly recognized in the work of Rosenbaum (1995) and Robins (1997). Their sensitivity analyses for matching when there are unobserved conditioning variables are, in their essence, sensitivity analyses using control functions.[16]

Tables 1 and 2 perform sensitivity analysis under different assumptions about the parameters of the underlying selection model. In particular, we assume that the data are generated by the model of equations $(1')$, $(2a')$ and $(2b')$ and that

$$\begin{aligned}\left(U_1,U_0,U_V\right)' &\sim N\left(0,\Sigma\right)\\ corr\left(U_j,U_V\right) &= \rho_{jV}\\ var\left(U_j\right) &= \sigma_j^2;\quad j=\left\{0,1\right\}.\end{aligned}$$

Using the formulae derived in the Appendix, we can write the biases of section (4) as

$$\begin{aligned}\text{Bias } TT\left(P\left(Z\right)=p\right) &= \sigma_0\rho_{0V}M(p)\\ \text{Bias } ATE\left(P\left(Z\right)=p\right) &= M(p)\left[\sigma_1\rho_{1V}\ \left(1-p\right)+\sigma_0\rho_{0V}\ p\right]\end{aligned}$$

where $M(p) = \frac{\phi\left(\Phi^{-1}(1-p)\right)}{p(1-p)}$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a standard normal random variable and p is the propensity score. We assume that $\mu_1 = \mu_0$ so that the true average treatment effect is zero.

We simulate the bias for different values of the ρ_{jV} and σ_j. The results in the tables show that, as we let the variances of the outcome equations grow, the value of the bias that we obtain can become substantial. With large variances there are large biases. With larger correlations come larger biases. These tables demonstrate the greater generality of the control function approach given the assumption of separability between model and errors. Even if the correlation between the observables and the unobservables $\left(\rho_{jV}\right)$ is small, so that one might think that selection on unobservables is relatively unimportant, we still get substantial biases if we do not control for relevant omitted conditioning variables. Only for special values of the parameters do we avoid the bias by matching. These examples also demonstrate that sensitivity analyses can be conducted for control function models even when they are not fully identified.

5.3 Instrumental Variables

Both the method of matching and the method of control functions work with $E(Y|X,Z,D)$ and $\Pr(D=1|X,Z)$. The method of instrumental variables works with $E(Y|X,Z)$ and $\Pr(D=1|X,Z)$. There are two versions of the method of instrumental variables: (a) conventional linear instrumental variables and (b) local instrumental variables (LIV) (Heckman and Vytlacil, 1999, 2000, 2003; Heckman, 2001). LIV is equivalent to a semiparametric selection model (See Vytlacil, 2002). It is an alternative way to implement the principle of control functions. $LATE$ (Imbens and Angrist, 1994) is a special case of LIV under the conditions we specify below.

We first consider the conventional method of instrumental variables. In this framework, $P(X,Z)$ arises less naturally than it does in the matching and control function approaches. Z is the instrument and $P(X,Z)$ is a function of the instrument.

Rewrite the model of equations $(2a')$ and $(2b')$ as

$$\begin{aligned} Y &= DY_1 + (1-D)Y_0 \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X) + U_1 - U_0)D + U_0 \\ &= \mu_0(X) + \Delta(X)D + U_0 \end{aligned}$$

where $\Delta(X) = \mu_1(X) - \mu_0(X) + U_1 - U_0$. When $U_1 = U_0$, this is a conventional IV model with D correlated with U_0. Standard instrumental variables conditions apply and $P(X,Z)$ is a valid instrument if:

$$E(U_0|P(X,Z),X) = E(U_0|X) \tag{IV-1}$$

and

$$\Pr\left(D=1|X,Z\right) \text{ is a nontrivial function of } Z \text{ for each } X. \tag{IV-2}$$

When $U_1 \neq U_0$ but $D \perp\!\!\!\perp (U_1 - U_0)|X$ (or alternatively $U_V \perp\!\!\!\perp (U_1 - U_0)|X$) then the same two conditions identify

$$\begin{aligned} ATE &= E\left(Y_1 - Y_0|X\right) = E\left(\Delta\left(X\right)|X\right) \\ TT &= E\left(Y_1 - Y_0|X, D=1\right) = E\left(Y_1 - Y_0|X\right) \\ &= MTE \end{aligned}$$

and marginal equals average conditional on X and Z. The requirement that $D \perp\!\!\!\perp (U_1 - U_0)|X$ is strong and assumes that agents do not participate in the program on the basis of *any* information about unobservables in gross gains (Heckman and Robb, 1985, 1986; Heckman, 1997).

The analytically more interesting case arises when $U_1 \neq U_0$ and $D \not\perp\!\!\!\perp (U_1 - U_0)$. To identify ATE, we require

$$E\left(U_0 + D\left(U_1 - U_0\right)|P\left(X,Z\right),X\right) = E\left(U_0 + D\left(U_1 - U_0\right)|X\right) \tag{IV-3}$$

and condition (IV-2) (Heckman and Robb, 1985, 1986; Heckman, 1997). To identify TT, we require

$$\begin{aligned} & E\left(U_0 + D\left(U_1 - U_0\right) - E\left(U_0 + D\left(U_1 - U_0\right)|X\right)|P\left(X,Z\right),X\right) \\ = \; & E\left(U_0 + D\left(U_1 - U_0\right) - E\left(U_0 + D\left(U_1 - U_0\right)|X\right)|X\right) \end{aligned}$$

and condition (IV-2). No simple conditions exist to identify the MTE using linear instrumental variables methods in the general case where $D \not\perp\!\!\!\perp (U_1 - U_0)|X, Z$ (Heckman and Vytlacil, 2000, 2003 characterize what conventional IV estimates in terms of a weighted average of MTEs).

The conditions required to identify ATE using P as an instrument, may be written in the following alternative form:

$$\begin{aligned} & E\left(U_0|P\left(X,Z\right),X\right) + E\left(U_1 - U_0|D=1,P\left(X,Z\right),X\right)P\left(X,Z\right) \\ = \; & E\left(U_0|X\right) + E\left(U_1 - U_0|D=1,X\right)P\left(X,Z\right) \end{aligned}$$

If $U_1 = U_0$ (everyone with the same X responds to treatment in the same way) or $(U_1 - U_0) \perp\!\!\!\perp D|P\left(X,Z\right),X$ (people do not participate in treatment on the basis of unobserved gains), then these conditions are satisfied. In general, the conditions are not satisfied by economic choice models, except under special cancellations that are not generic. If Z is a determinant of choices, and $U_1 - U_0$ is in the agent's choice set (or is correlated only partly with information in the agent's choice set), then this condition is not likely to be satisfied.

These identification conditions are fundamentally different from the matching and control function identification conditions. In matching, the essential condition for means is

$$\begin{aligned} E\left(U_0|X, D=0, P\left(X,Z\right)\right) &= E\left(U_0|X, P\left(X,Z\right)\right) \text{ and} \\ E\left(U_1|X, D=1, P\left(X,Z\right)\right) &= E\left(U_1|X, P\left(X,Z\right)\right) \end{aligned}$$

These require that, conditional on $P\left(X,Z\right)$ and X, U_1 and U_0 are mean independent of U_V (or D). When $\mu_1\left(W\right)$ and $\mu_0\left(W\right)$ are the conditional means of Y_1 and Y_0 respectively, these terms are zero.

The method of control functions models and estimates this dependence rather than assuming it vanishes. The method of linear instrumental variables requires that the composite error term $U_0 + D\left(U_1 - U_0\right)$ be mean independent of Z (or $P\left(X,Z\right)$), given X. Essentially, the conditions require that the dependence of U_0 and $D\left(U_1 - U_0\right)$ on Z vanish through conditioning on X. Matching requires that U_1 and U_0 are independent of D given (X,Z). These conditions are logically distinct. One set of conditions does not imply the other set. Conventional IV in the general case does not answer a well posed economic question (see Carneiro, Heckman and Vytlacil, 2001).

Local instrumental variables methods developed by Heckman and Vytlacil (1999, 2000, 2003) estimate all three treatment parameters in the general case where $(U_1 - U_0) \not\perp\!\!\!\perp D|\left(X,Z\right)$ under the following additional conditions

$$\mu_D\left(Z\right) \qquad \text{is a non-degenerate random variable given } X \qquad \text{(LIV-1)}$$

(existence of an exclusion restriction)

$$(U_0, U_1, U_V) \perp\!\!\!\perp Z|X \qquad \text{(LIV-2)}$$

$$0 < \Pr\left(D|X\right) < 1 \qquad \text{(LIV-3)}$$

$$\text{Support } P\left(D|\left(X,Z\right)\right) = [0,1] \qquad \text{(LIV-4)}$$

Under these conditions

$$\frac{\partial E(Y|X,P(Z))}{\partial(P(Z))} = MTE\left(X, P\left(Z\right), V=0\right). \ ^{17}$$

Only (LIV-1) - (LIV-3) are required to identify this parameter.

As demonstrated by Heckman and Vytlacil (1999, 2000, 2003) and Heckman (2001), over the support of (X,Z), MTE can be used to construct (under LIV-4) or bound (in the case of partial support) ATE and TT. Policy relevant treatment effects can be defined, $LATE$ is a special case of this method. Table

3 summarizes the alternative assumptions used in matching, control functions and instrumental variables to identify treatment parameters. For the rest of the paper, we discuss matching, the topic of this special issue. We first turn to consider the informational requirements of matching.

6 The Informational Requirements of Matching and the Bias When They are not Satisfied

This section considers the informational requirements for matching.[18] We introduce five distinct information sets and establish relationships among them: (1) An information set that satisfies conditional independence (M-1), $\sigma(I_{R^*})$, a "relevant" information set; (2) the minimal information set needed to satisfy conditional independence (M-1), $\sigma(I_R)$, the "minimal relevant" information set; (3) the information set available to the agent at the time decisions to participate are made, $\sigma(I_A)$; (4) the information available to the economist $\sigma(I_{E^*})$ and (5) the information used by the economist $(\sigma(I_E))$. We will define the random variables generated by these sets as $I_{R^*}, I_R, I_A, I_{E^*}, I_E$ respectively.[19]

After defining these information sets, we show the biases that result when econometricians use information other than the relevant information set. More information does not necessarily reduce the bias in matching. Standard algorithms for selecting conditioning variables are not guaranteed to pick the relevant conditioning variables or reduce bias compared to conditioning sets not selected by these algorithms.

First we define the information sets more precisely.

Definition 1 *We say that $\sigma(I_{R^*})$ is a **relevant information set** if its associated random variable, I_{R^*}, satisfies (M-1) so*

$$(Y_1, Y_0) \perp\!\!\!\perp D | I_{R^*}$$

Definition 2 *We say that $\sigma(I_R)$ is a **minimal relevant information set** if it is the intersection of all sets $\sigma(I_{R^*})$. The associated random variable I_R is the minimum amount of information that guarantees that (M-1) is satisfied.*

If we define the minimal relevant information set as one that satisfies conditional independence, it might not be unique. If the set $\sigma(I_{R1})$ satisfies the conditional independence condition, then the set $\sigma(I_{R1}, Q)$ such that $Q \perp\!\!\!\perp (Y_1, Y_0) \mid I_{R1}$ would also guarantee conditional independence. For this reason, we define the relevant information set to be the minimal; *i.e.*, to be the intersection of all such sets.

Definition 3 *The agent's information set, $\sigma(I_A)$, is defined by the information I_A used by the agent when choosing among treatments. Accordingly, we call I_A the **agent's information**.*

Definition 4 *The econometrician's* ***full information set****,* $\sigma(I_{E^*})$*, is defined as* ***all*** *of the information* ***available*** *to the econometrician,* I_{E^*}*.*

Definition 5 *The econometrician's* ***information set****,* $\sigma(I_E)$*, is defined by the information* ***used*** *by the econometrician when analyzing the agent's choice of treatment,* I_E*.*

Only three restrictions are imposed on the structure of these sets: $\sigma(I_R) \subseteq \sigma(I_{R^*}), \sigma(I_R) \subseteq \sigma(I_A)$ and $\sigma(I_E) \subseteq \sigma(I_{E^*})$.[20] The first we have already discussed. The second one requires that the minimal relevant information set must be part of the information the agent uses when deciding whether to take treatment. The third requires that the information used by the econometrician must be part of the information he observes. Other than these obvious orderings, the econometrician's information set may be different from the agent's or the relevant information set. The econometrician may know something the agent doesn't know since typically he is observing events after the decision is made. At the same time, there may be private information known to the agent. The matching assumptions (M-1) or (M-3) imply that

$$\sigma(I_R) \subseteq \sigma(I_E)$$

so that the econometrician uses the minimal relevant information set.

In order to have a concrete example of these information sets and their associated random variables, we assume that the economic model generating the data is a generalized Roy model of the form

$$\begin{aligned} V &= Z\gamma + U_V \qquad \text{where} \\ U_V &= \alpha_{V1} f_1 + \alpha_{V2} f_2 + \varepsilon_V \\ D &= 1 \text{ if } V \geq 0, \ = 0 \text{ otherwise} \end{aligned}$$

and

$$\begin{aligned} Y_1 &= \mu_1 + U_1 \quad \text{where } U_1 = \alpha_{11} f_1 + \alpha_{12} f_2 + \varepsilon_1, \\ Y_0 &= \mu_0 + U_0 \quad \text{where } U_0 = \alpha_{01} f_1 + \alpha_{02} f_2 + \varepsilon_0, \end{aligned}$$

where $(f_1, f_2, \varepsilon_V, \varepsilon_1, \varepsilon_0)$ are assumed to be mean zero random variables that are mutually independent of each other and Z so that all the correlation among the elements of (U_0, U_1, U_V) is captured by $f = (f_1, f_2)$.[21] We keep implicit any dependence on X which may be general. The minimal relevant information for this model when the factor loadings are not zero ($\alpha_{ij} \neq 0$) is

$$I_R = \{f_1, f_2\}.$$

The agent's information set may include different variables. If we assume that $\varepsilon_0, \varepsilon_1$ are shocks to outcomes not known to the agent at the time decisions are made, the agent's information is

$$I_A = \{f_1, f_2, Z, \varepsilon_V\}.$$

Under perfect certainty on the part of the agent

$$I_A = \{f_1, f_2, Z, \varepsilon_V, \varepsilon_1, \varepsilon_0\}.$$

In either case, all of the information available to the agent is not required to obtain conditional independence (M-1). All three information sets guarantee conditional independence, but only the first is minimal relevant.

The observing economist may know some variables not in I_A, I_{R^*} or I_R but may not know all of the variables in I_R. In the following subsections, we address the question of what happens when the matching assumption that $\sigma(I_E) \supseteq \sigma(I_R)$ does not hold. That is, we analyze what happens to the matching bias as the amount of information used by the econometrician is changed. In order to get closed form expressions for the biases of the treatment parameters we add the additional assumption that

$$(f_1, f_2, \varepsilon_V, \varepsilon_1, \varepsilon_0) \sim N(0, \Sigma),$$

where Σ is a matrix with $\left(\sigma^2_{f_1}, \sigma^2_{f_2}, \sigma^2_{\varepsilon_V}, \sigma^2_{\varepsilon_1}, \sigma^2_{\varepsilon_0}\right)$ in the diagonal and zero in all the non-diagonal elements. This assumption links matching models to conventional normal selection models. We next analyze various cases.

6.1 The economist uses the minimal relevant information: $\sigma(I_R) \subseteq \sigma(I_E)$

We begin by analyzing the case in which the information used by the analyst is $I_E = \{Z, f_1, f_2\}$, so that the econometrician has access to the relevant information set and it is larger than the minimal relevant information set. In this case it is straightforward to show that matching identifies all of the mean treatment parameters with no bias. The matching estimator is

$$E(Y_1|D=1, I_E) - E(Y_0|D=0, I_E) = \mu_1 - \mu_0 + (\alpha_{11} - \alpha_{01}) f_1 + (\alpha_{12} - \alpha_{02}) f_2$$

and all of the treatment parameters collapse to this same expression since, conditional on knowing f there is no selection because $(\varepsilon_1, \varepsilon_0) \perp\!\!\!\perp U_V$. Recall that $I_R = \{f_1, f_2\}$ and the economist needs less information to achieve (M-1).

The analysis of Rosenbaum and Rubin (1983) tells us that knowledge of (Z, f_1, f_2) and knowledge of $P(Z, f_1, f_2)$ are equivalent so that matching on the propensity score also identifies all of the treatment

parameters. If we write the propensity score as

$$P\left(I_E\right)=\Pr\ \left(\frac{\varepsilon_V}{\sigma_{\varepsilon_V}}>\frac{-Z\gamma-\alpha_{V1}f_1-\alpha_{V2}f_2}{\sigma_{\varepsilon_V}}\right)=1-\Phi\left(\frac{-Z\gamma-\alpha_{V1}f_1-\alpha_{V2}f_2}{\sigma_{\varepsilon_V}}\right)=p,$$

the event $\left(V\underset{>}{\overset{\leqq}{}}0,P\left(f,Z\right)=p\right)$ can be written as $\frac{\varepsilon_V}{\sigma_{\varepsilon_V}}\underset{>}{\overset{\leqq}{}}\Phi^{-1}\left(1-p\right)$, where Φ is the cdf of a standard normal random variable and ϕ is its density and $f=\left(f_1,f_2\right)$. The population matching condition is

$$\begin{aligned}
& E\left(Y_1|D=1,P\left(I_E\right)=p\right)-E\left(Y_0|D=0,P\left(I_E\right)=p\right)\\
= \ & \mu_1-\mu_0+E\left(U_1|D=1,P\left(I_E\right)=p\right)-E\left(U_0|D=0,P\left(I_E\right)=p\right)\\
= \ & \mu_1-\mu_0+E\left(U_1|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}}>\Phi^{-1}\left(1-p\right)\right)-E\left(U_0|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}}\leq\Phi^{-1}\left(1-p\right)\right)\\
= \ & \mu_1-\mu_0
\end{aligned}$$

and it is equal to all of the treatment parameters since

$$E\left(U_1|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}}>\Phi^{-1}\left(1-p\right)\right)=\frac{Cov\left(U_1,\varepsilon_V\right)}{\sigma_{\varepsilon_V}}M_1\left(p\right)$$

and

$$E\left(U_0|\frac{\varepsilon_V}{\sigma_{\varepsilon_V}}\leq\Phi^{-1}\left(1-p\right)\right)=\frac{Cov\left(U_0,\varepsilon_V\right)}{\sigma_{\varepsilon_V}}M_0\left(p\right),$$

where

$$\begin{aligned}
M_1(p) &= \frac{\phi(\Phi^{-1}(1-p))}{p}\\
M_0(p) &= -\frac{\phi(\Phi^{-1}(1-p))}{1-p}
\end{aligned}$$

As a consequence of the assumptions about mutual independence of the errors

$$Cov\left(U_i,\varepsilon_V\right)=Cov\left(\alpha_{i1}f_1+\alpha_{i2}f_2+\varepsilon_i,\varepsilon_V\right)=0,\quad i=0,1.$$

In the context of this model, the case considered in this subsection is the one matching is designed to solve. Even though a selection model generates the data, the fact that the information used by the econometrician includes the minimal relevant information makes matching equivalent to the selection model. We can estimate the treatment parameters with no bias since, as a consequence of the assumptions made $\left(U_1,U_0\right)\perp\!\!\!\perp D|\left(f,Z\right)$, which is exactly what matching requires. The minimal relevant information set is even smaller. We only need to know $\left(f_1,f_2\right)$ to secure this result, and we can define the propensity score solely in terms of f_1 and f_2, and the Rosenbaum-Rubin result still goes through.

6.2 The Economist does not Use All of the Minimal Relevant Information

Now, suppose that the information used by the econometrician is

$$I_E = \{Z\}$$

but there is selection on the unobservable (to the analyst) f_1, f_2, *i.e.*, the factor loadings α_{ij} are all non zero. Recall that we assume that Z and the f are independent. In this case the event $\left(V \lesseqgtr 0, P(Z) = p\right)$ is

$$\frac{\alpha_{V1} f_1 + \alpha_{V2} f_2 + \varepsilon_V}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} \lesseqgtr \Phi^{-1}(1-p).$$

Using the analysis presented in the Appendix, the bias for the different treatment parameters is given by

$$\text{Bias } TT(P(Z) = p) = \beta_0 M(p), \tag{3}$$

where $M(p) = M_1(p) - M_0(p)$.

$$\begin{aligned} \text{Bias } ATE(P(Z) = p) &= M(p)\left[\beta_1(1-p) + \beta_0 p\right] \\ &= \beta_0 M(p)\left[p + \frac{\beta_1}{\beta_0}(1-p)\right];\ \beta_0 \neq 0 \end{aligned} \tag{4}$$

$$\text{Bias } MTE(P(Z) = p) = M(p)\left[\beta_1(1-p) + \beta_0 p\right] - \Phi^{-1}(1-p)\left[\beta_1 - \beta_0\right] \tag{5}$$

where

$$\begin{aligned} \beta_1 &= \frac{\alpha_{V1}\alpha_{11}\sigma_{f_1}^2 + \alpha_{V2}\alpha_{12}\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} \\ \beta_0 &= \frac{\alpha_{V1}\alpha_{01}\sigma_{f_1}^2 + \alpha_{V2}\alpha_{02}\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2 \sigma_{f_1}^2 + \alpha_{V2}^2 \sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}}. \end{aligned}$$

It is not surprising that matching on variables that exclude the relevant conditioning variables produces bias. The advantage of working with a closed form expression for the bias is that it allows us to answer questions about the *magnitude* of this bias under different assumptions about the information available to the analyst, and to present some simple examples. We next use expressions (3), (4) and (5) as benchmarks against which to compare the relative size of the bias when we enlarge the econometrician's information set beyond Z.

6.3 Adding information to the Econometrician's Information Set I_E: Using Some but not All the Information from the Minimal Relevant Information Set I_R

Suppose next that the econometrician uses more information but not all of the information in the minimal relevant information set. Possibly, the data set assumed in the preceding section is augmented or else the

econometrician decides to use information previously available. In particular, assume that

$$I'_E = \{Z, f_2\}.$$

Under conditions 1, 2 and 3 presented below the biases for the treatment parameters of section (6.2) are reduced by changing the conditioning set in this way. We define expressions comparable to β_1 and β_0 for this case:

$$\begin{aligned}\beta'_1 &= \frac{\alpha_{V1}\alpha_{11}\sigma^2_{f_1}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1}+\sigma^2_{\varepsilon_V}}} \\ \beta'_0 &= \frac{\alpha_{V1}\alpha_{01}\sigma^2_{f_1}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1}+\sigma^2_{\varepsilon_V}}}.\end{aligned}$$

Then, we just compare the biases under the two cases using formulae (3) - (5) suitably modified but keeping p fixed.

Condition 1 *The bias produced by using matching to estimate TT is smaller in absolute value for any given p when the new information set $\sigma(I'_E)$ is used if*

$$|\beta_0| > \left|\beta'_0\right|.$$

Condition 2 *The bias produced by using matching to estimate ATE is smaller in absolute value for any given p when the new information set $\sigma(I'_E)$ is used if*

$$\left|\beta_1(1-p)+\beta_0 p\right| > \left|\beta'_1(1-p)+\beta'_0 p\right|.$$

Condition 3 *The bias produced by using matching to estimate MTE is smaller in absolute value for any given p when the new information set $\sigma(I'_E)$ is used if*

$$\left|M(p)\left[\beta_1(1-p)+\beta_0 p\right]-\Phi^{-1}(1-p)\left[\beta_1-\beta_0\right]\right| > \left|M(p)\left[\beta'_1(1-p)+\beta'_0 p\right]-\Phi^{-1}(1-p)\left[\beta'_1-\beta'_0\right]\right|.$$

Proof. These are straightforward applications of formulae (3)-(5), modified to account for the different covariance structure produced by the information structure assumed in this Section (replacing β_0 with β'_0, β_1 with β'_1). ■

It is important to notice that we condition on the same p in deriving these expressions.

These conditions do not always hold. In general, whether or not the bias will be reduced by adding additional conditioning variables depends on the relative importance of the additional information in both the outcome equations and on the signs of the terms inside the absolute value.

Consider whether Condition (1) is satisfied and assume $\beta_0 > 0$ for all α_{02}, α_{V2}. Then $\beta_0 > \beta_0'$ if

$$\beta_0 = \frac{\alpha_{V1}\alpha_{01}\sigma_{f_1}^2 + \left(\alpha_{V2}^2\right)\left(\frac{\alpha_{02}}{\alpha_{V2}}\right)\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} > \frac{\alpha_{V1}\alpha_{11}\sigma_{f_1}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \sigma_{\varepsilon_V}^2}} = \beta_0'.$$

When $\left(\frac{\alpha_{02}}{\alpha_{V2}}\right) = 0$, clearly $\beta_0 < \beta_0'$. Adding information to the conditioning set increases bias. We can vary $\left(\frac{\alpha_{02}}{\alpha_{V2}}\right)$ holding all other parameters constant. A direct computation shows that

$$\frac{\partial \beta_0}{\partial\left(\frac{\alpha_{02}}{\alpha_{V2}}\right)} = \frac{\alpha_{V2}^2\sigma_{f_2}^2}{\sqrt{\alpha_{V1}^2\sigma_{f_1}^2 + \alpha_{V2}^2\sigma_{f_2}^2 + \sigma_{\varepsilon_V}^2}} > 0.$$

As α_{02} increases, there is some critical value α_{02}^* beyond which $\beta_0 > \beta_0'$.

If we assumed that $\beta_0 < 0$ however, the exact opposite conclusion would hold and the conditions would be harder to meet as the relative importance of the new information is increased. Similar expressions can be derived for ATE and MTE in which the direction of the effect depends on the signs of the terms in the absolute value.

Figures 1, 2 and 3 illustrate the point that adding some but not all information from the minimal relevant set might increase the bias for all treatment parameters. In these figures we let the variances of the factors and the error terms be equal to one and set

$$\begin{aligned}
\alpha_{01} &= \alpha_{V1} = \alpha_{V2} = 1 \\
\alpha_{02} &= \alpha_{12} = 0.1 \\
\alpha_{11} &= 2
\end{aligned}$$

so that we have a case in which the information being added is relatively unimportant in terms of outcomes.

The fact that the bias might increase when adding some but not all information from I_R is a feature that is not shared by the method of control functions. Since the method of control functions models the stochastic dependence of the unobservables in the outcome equations on the observables, changing the variables observed by the econometrician to include f_2 does not generate bias, it only changes the control function used. That is, by adding f_2 we simply change the control function from

$$\begin{aligned}
K_1\left(P\left(Z\right) = p\right) &= \beta_1 M_1(p) \\
K_0\left(P\left(Z\right) = p\right) &= \beta_0 M_0(p)
\end{aligned}$$

to

$$\begin{aligned}
K_1'\left(P\left(Z, f_2\right) = p\right) &= \beta_1' M_1(p) \\
K_0'\left(P\left(Z, f_2\right) = p\right) &= \beta_0' M_0(p)
\end{aligned}$$

but do not generate any bias. This is a major advantage of this method. It controls for the bias of the omitted conditioning variables by modelling it. Of course, if the model for the bias is not valid, neither is the correction for the bias. Matching evades this problem by assuming that the analyst always knows the correct conditioning variables and they satisfy (M-1).

6.4 Adding information to the econometrician's information set: using proxies for the relevant information

Suppose that instead of knowing some part of the minimal relevant information set, such as f_2, the analyst has access to a proxy for it.[22] In particular, assume that he has access to a variable $\widetilde{Z}$ that is correlated with f_2 but that is not the full minimal relevant information set. That is, define the econometrician's information to be

$$\widetilde{I}_{E^*} = \left\{ Z, \widetilde{Z} \right\}.$$

and suppose that he uses it so $\widetilde{I}_E = \widetilde{I}_{E^*}$. In order to obtain closed form expressions for the biases we further assume that

$$\begin{aligned} \widetilde{Z} &\sim N\left(0, \sigma^2_{\widetilde{Z}}\right) \\ corr\left(\widetilde{Z}, f_2\right) &= \rho, \text{ and } \widetilde{Z} \perp\!\!\!\perp (\varepsilon_0, \varepsilon_1, \varepsilon_V, f_1). \end{aligned}$$

We define expressions comparable to β and β' :

$$\begin{aligned} \widetilde{\beta}_1 &= \frac{\alpha_{11}\alpha_{V1}\sigma^2_{f_1} + \alpha_{12}\alpha_{V2}\left(1-\rho^2\right)\sigma^2_{f_2}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \alpha^2_{V2}\sigma^2_{f_2}\left(1-\rho^2\right) + \sigma^2_{\varepsilon_V}}} \\ \widetilde{\beta}_0 &= \frac{\alpha_{01}\alpha_{V1}\sigma^2_{f_1} + \alpha_{02}\alpha_{V2}\left(1-\rho^2\right)\sigma^2_{f_2}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \alpha^2_{V2}\sigma^2_{f_2}\left(1-\rho^2\right) + \sigma^2_{\varepsilon_V}}}. \end{aligned}$$

By substituting I'_E for $\widetilde{I}_E$ and β'_j for $\widetilde{\beta}_j$ $(j = 0, 1)$ into Conditions (1), (2) and (3) of section (6.3) we obtain equivalent results for this case. Whether $\widetilde{I}_E$ will be bias reducing depends on how well it spans I_R and on the signs of the terms in the absolute values.

In this case, however, there is another parameter to consider: the correlation between $\widetilde{Z}$ and f_2. If $|\rho| = 1$ we are back to the case of $\widetilde{I}_E = I'_E$ because $\widetilde{Z}$ is a perfect proxy for f_2. If $\rho = 0$ we are essentially back to the case analyzed in section (6.3). Since we know that the bias might either increase or decrease when f_2 is used as a conditioning variable but f_1 is not, we know that it is not possible to determine whether the bias increases or decreases as we change the correlation between f_2 and $\widetilde{Z}$. That is, we know

that going from $\rho = 0$ to $|\rho| = 1$ might change the bias in any direction. Use of a better proxy in this correlational sense may produce a more biased estimate.

>From the analysis of section (6.3), it is straightforward to derive conditions under which the bias generated when the econometrician's information is $\widetilde{I}_E$ is smaller than when it is I'_E. That is, it can be the case that knowing *the proxy* variable $\widetilde{Z}$ is *better* than knowing the actual variable f_2. Take again the treatment on the treated case as a simple example (*i.e.,* Condition (1)). The bias is reduced when $\widetilde{Z}$ is used instead of f_2 if

$$\left| \frac{\alpha_{01}\alpha_{V1}\sigma^2_{f_1} + \alpha_{02}\alpha_{V2}\left(1-\rho^2\right)\sigma^2_{f_2}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \alpha^2_{V2}\sigma^2_{f_2}\left(1-\rho^2\right) + \sigma^2_{\varepsilon_V}}} \right| < \left| \frac{\alpha_{01}\alpha_{V1}\sigma^2_{f_1}}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \sigma^2_{\varepsilon_V}}} \right|.$$

Figures 4, 5 and 6 use the same example of the previous section to illustrate the two points being made here. Namely, that *using* a *proxy* for an unobserved relevant variable *might increase the bias*. On the other hand, it *might* be *better* in terms of bias to use a *proxy* than to use the actual variable, f_2.

6.5 The case of a discrete treatment

The points that we have made so far do not strictly depend on all of the assumptions we have made to produce simple examples. In particular, we require neither normality nor additive separability of the outcomes. The proposition that if the econometrician's information set includes all the minimal relevant information, matching identifies the correct treatment, is true more generally provided that any additional extraneous information used is "exogenous" in a sense to be precisely defined in the next section. In this subsection, we present a simple analysis of a discrete treatment that does not rely on either normality or separability of outcome equations.[23]

Suppose that outcomes (Y_j) are binary random variables generated by the following model:

$$\begin{aligned} Y^*_j &= \mu_j + U_j \\ U_j &= \alpha_{j1}f_1 + \alpha_{j2}f_2 + \varepsilon_j \ , \ j = 0, 1 \\ Y_j &= 1 \text{ if } Y^*_j \geq 0, \ = 0 \text{ otherwise,} \end{aligned} \tag{6}$$

where $j = 1$ corresponds to treatment and $j = 0$ corresponds to no treatment. People receive treatment according to the rule

$$\begin{aligned} V &= \mu_V + U_V \\ U_V &= \alpha_{V_1}f_1 + \alpha_{V_2}f_2 + \varepsilon_V \\ D &= 1 \text{ if } V \geq 0, \ = 0 \text{ otherwise;} \end{aligned} \tag{7}$$

and we assume that

$$f_1 \perp\!\!\!\perp f_2 \perp\!\!\!\perp \varepsilon_0 \perp\!\!\!\perp \varepsilon_1 \perp\!\!\!\perp \varepsilon_V.$$

Each of these error components has a zero mean, the observed outcome is either zero or one and is given by

$$Y = DY_1 + (1 - D) Y_0.$$

An example of such a model arises when we observe whether a person is working or not and when the probability of being employed might be different if the person has participated in a training program.

There are many ways in which the effect of treatment can be defined in this model. (see Aakvik, Heckman and Vytlacil, 2003) One way is given by the ratio of the probabilities of observing $Y_1 = 1$ given that the person receives treatment and the counterfactual probability of observing $Y_0 = 1$ given that the person chooses treatment but does not receive it. That is, the effect of treatment is given by:

$$\Delta_1 (I_E) = \frac{\Pr (Y_1 = 1, D = 1 | I_E)}{\Pr (Y_0 = 1, D = 1 | I_E)}.$$

A second definition works with odds ratios:

$$\Delta_2 (I_E) = \frac{\frac{\Pr(Y_1=1, D=1|I_E)}{\Pr(Y_1=0, D=1|I_E)}}{\frac{\Pr(Y_0=1, D=1|I_E)}{\Pr(Y_0=0, D=1|I_E)}}.$$

One could also work with logs:

$$\begin{aligned} \Delta_3 (I_E) &= \log (\Delta_1) \\ \Delta_4 (I_E) &= \log (\Delta_2). \end{aligned}$$

Under the null hypothesis of no effect of treatment $\Delta_1 = \Delta_2 = 1$. More generally these ratios can be either smaller or greater than one depending on whether there is a positive or negative effect of treatment. In order to fix ideas, we will call Δ_1 the effect of treatment under the understanding that equivalent results can be obtained for other definitions.

The econometrician measures the effect of treatment by "matching" the observed distributions according to some variables that he observes. Since Y_0 is only observed when $D = 0$ the analyst attempts to identify the effect of treatment by

$$\widehat{\Delta}_1 (I_E) = \frac{\Pr (Y_1 = 1, D = 1 | I_E)}{\Pr (Y_0 = 1, D = 0 | I_E)}.$$

The denominator replaces the desired probability $\Pr (Y_0 = 1, D = 1 | I_E)$ by the available information $\Pr (Y_0 = 1, D = 0|$

Let there be no real effect of treatment so that, in terms of the model given by equations (6) and (7) we

have that $\Delta_1 = 1$ and $\Delta_2 = 1$ so

$$\begin{aligned} \mu_1 &= \mu_0 = \mu \\ F_{U_1} &= F_{U_0} = F_U \end{aligned}$$

which can be generated by setting

$$\begin{aligned} \alpha_{11} &= \alpha_{01} = \alpha_1 \\ \alpha_{12} &= \alpha_{02} = \alpha_2 \\ F_{\varepsilon_1} &= F_{\varepsilon_0} = F_\varepsilon \end{aligned}$$

where F_X denotes the cdf of X.

We initially assume that the analyst has access to the minimal relevant information set and uses it. That is, we assume that

$$I_E = \{f_1, f_2\}.$$

In this case, in large samples the estimated effect of treatment is

$$\widehat{\Delta}_1(I_E) = \frac{\Pr(Y_1 = 1, D = 1 | f_1, f_2)}{\Pr(Y_0 = 1, D = 0 | f_1, f_2)} = \frac{\Pr(Y_1 = 1 | f_1, f_2)}{\Pr(Y_0 = 1 | f_1, f_2)} = \Delta_1(I_E).$$

Under the null of no treatment effect, $\Delta_1 = \Delta_2 = 1$. Conditioning on (f_1, f_2) removes any dependence on D, and we can replace the denominator of Δ_1 by $\Pr(Y_0 = 1, D = 0 | f_1, f_2)$. If we do not condition on information that contains the minimal relevant information set, this is no longer true. In general:

$$\Delta_1(I_E) = \frac{\Pr(Y_1 = 1, D = 1 | I_E)}{\Pr(Y_0 = 1, D = 1 | I_E)} \neq \frac{\Pr(Y_1 = 1, D = 1 | I_E)}{\Pr(Y_0 = 1, D = 0 | I_E)} = \widehat{\Delta}_1(I_E).$$

The biases can be substantial. Suppose that $I''_E = \{f_2\}$ and consider the following simulations. Assume that the true model is

$$\begin{aligned} \alpha_{11} &= \alpha_{01} = \alpha_{V1} = 1 \\ \alpha_{12} &= \alpha_{02} = 1 \\ \mu_1 &= \mu_0 = \mu_V = -1 \\ (\varepsilon_1, \varepsilon_0, \varepsilon_V, f_1, f_2) &\sim N(0, \Sigma) \end{aligned}$$

where Σ is the identity matrix. Values of α_{V2} are specified in the examples presented below. Given these assumptions, there is no effect of treatment so $\Delta_1 = 1$. In figures 7 and 8 we show what happens when the analyst uses the population counterpart to the matching estimator:

$$\widehat{\Delta}_1(I''_E) = \frac{\Pr(Y_1 = 1, D = 1 | I''_E)}{\Pr(Y_0 = 1, D = 0 | I''_E)}$$

to measure the effect of treatment. Figure 7 illustrates the case in which we assume that $\alpha_{V2} = 1$ whereas figure 8 shows the case of $\alpha_{V2} = -1$. In both cases matching does not estimate the true effect of treatment when the analyst uses information that does not contain the full minimal relevant information set. Furthermore, the discrepancy between the estimate and the true effect of treatment changes as we change the level of f_2 on which we are conditioning. Depending on the choice of f_2, we get either positive or negative estimated treatment effects. This result is again analogous to the continuous case result stating that matching estimates are biased when the analyst does not use the minimal relevant information set. Figures 9-14 show that equivalent results hold for the case in which the effect of treatment is defined by odds ratios

$$\Delta_2\left(I_E\right) = \frac{\frac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_1=0,D=1|I_E)}}{\frac{\Pr(Y_0=1,D=1|I_E)}{\Pr(Y_0=0,D=1|I_E)}}$$

and the analyst uses

$$\widehat{\Delta}_2\left(I_E''\right) = \frac{\frac{\Pr\left(Y_1=1,D=1|I_E''\right)}{\Pr\left(Y_1=0,D=1|I_E''\right)}}{\frac{\Pr\left(Y_0=1,D=0|I_E''\right)}{\Pr\left(Y_0=0,D=0|I_E''\right)}}$$

or the log versions of both $\widehat{\Delta}_1$ and $\widehat{\Delta}_2$.

6.6 On the use of model selection criteria to choose matching variables

We have just shown that adding more variables from the minimal relevant information set, but not all variables in it, may increase bias. There are no rigorously justified algorithms for identifying a relevant information set. Adding variables that are statistically significant in the treatment choice equation is not guaranteed to select a set of conditioning variables that satisfies condition (M-1). This is demonstrated by the analysis of section (6.3) that shows that adding f_2 when it determines D may increase bias. The existing literature (*e.g.*, Heckman, Ichimura and Todd, 1997) proposes other criteria based on selecting the set of variables that maximizes some goodness of fit criteria (λ) where a lower λ means a better fit. The intuition behind such criteria is that by using some measure of goodness of fit as a guiding principle one is using information relevant to the decision process. It is clear that knowing f_2 improves goodness of fit so that in general such a rule is deficient if f_1 is not known.

An implicit assumption underlying such procedures is that the added conditioning variables C are exogenous in the following sense

$$(Y_0, Y_1) \perp\!\!\!\perp D|I_E, C \tag{M-4}$$

where I_E is interpreted as the variables initially used as conditioning variables before C is added. Failure of exogeneity is a failure of (M-1), and matching estimators are biased.

In the literature, the use of such rules of thumb is justified in two different ways. Sometimes it is claimed that they provide a *relative* guide. Sets of variables with lower λ (better goodness of fit) are alleged to be better than sets of variables with higher λ in the sense that they generate lower biases. However, we have already shown that this is not true. We know that enlarging the analyst's information from $I_E = \{Z\}$ to $I'_E = \{Z, f_2\}$ will improve fit since f_2 is also in I_A. But, going from I_E to I'_E might increase the bias. So, it is not true that combinations of variables that decrease some measure of discrepancy λ necessarily reduce the bias. Table 4 illustrates this point using a normal example. Going from row 1 to row 2, adding f_2 improves goodness of fit and increases bias for all three treatment parameters, because (M-4) is violated.

A rule of thumb is sometimes invoked as an absolute standard against which to compare. The argument is as follows. The analyst asserts that there is a combination of variables I'' that satisfy (M-1) and hence produces zero bias and a value of $\lambda = \lambda''$ smaller than that of any other I. Now we know that conditioning on $\{Z, f_1, f_2\}$ generates zero bias. However, we can exclude Z and still get zero bias. Since Z is a determinant of D this shows immediately that the best fitting model does not necessarily identify the minimal relevant information set. In this example including Z is innocuous because there is still zero bias and the add conditioning variables satisfies (M-4). In general, such a rule is not innocuous. If goodness of fit is used as a rule to choose variables on which to match, there is no guarantee it produces a desirable conditioning set. If we include in the conditioning set variables C that violate (M-4), they may improve the fit of predicted probabilities but worsen bias.

We can always construct a collection of conditioning variables $\widetilde{\widetilde{I}}_E$ with a better fit and a *larger* bias than can be obtained from just conditioning on $\{f_1, f_2\}$. Let

$$\widetilde{\widetilde{I}}_E = \{Z, S\}$$

where

$$\begin{aligned} S &= V - Z\gamma + \eta \\ \eta &\sim N\left(0, \sigma_\eta^2\right) \\ \eta &\perp\!\!\!\perp \left(f_1, f_2, \varepsilon_0, \varepsilon_1, \varepsilon_V\right). \end{aligned}$$

The expressions for the biases are the same as in equations (3) - (5) using $\widetilde{\widetilde{\beta}}_j$ $(j = 0, 1)$ instead of β_j where:

$$\begin{aligned}
\widetilde{\widetilde{\beta}}_1 &= \frac{\pi\left(\alpha_{11}\alpha_{V1}\sigma^2_{f_1} + \alpha_{12}\alpha_{V2}\sigma^2_{f_2}\right)}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \alpha^2_{V2}\sigma^2_{f_2} + \sigma^2_{\varepsilon_V}}} \\
\widetilde{\widetilde{\beta}}_0 &= \frac{\pi\left(\alpha_{01}\alpha_{V1}\sigma^2_{f_1} + \alpha_{02}\alpha_{V2}\sigma^2_{f_2}\right)}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \alpha^2_{V2}\sigma^2_{f_2} + \sigma^2_{\varepsilon_V}}} \\
\pi &= \frac{\sigma_\eta}{\sqrt{\alpha^2_{V1}\sigma^2_{f_1} + \alpha^2_{V2}\sigma^2_{f_2} + \sigma^2_{\varepsilon_V} + \sigma^2_\eta}}.
\end{aligned}$$

In general, these expressions are not zero so that using propensity score matching will generate a bias. The source of the bias is the measurement error in S for V. Now, to prove that this combination of variables has a better fit all we need do is arbitrarily reduce σ^2_η. In particular, when $\sigma^2_\eta = 0$ we can perfectly predict D. That is, for

$$2\varepsilon > \sigma^2_\eta > \varepsilon > 0$$

then

$$\begin{aligned}
\lim_{\varepsilon \to 0} \Pr\left(D = 1 | V - Z\gamma + \eta, Z\right) &= 1 \text{ for } V > 0 \\
\lim_{\varepsilon \to 0} \Pr\left(D = 1 | V - Z\gamma + \eta, Z\right) &= 0 \text{ for } V < 0.
\end{aligned}$$

However, when the limit is attained assumption (M-2) is violated and matching breaks down. Making σ^2_η arbitrarily small, we can predict D arbitrarily well so we can always decrease λ enough to get a combination of variables with better fit for predicted probabilities and larger bias than a model that conditions only on the minimal relevant information f_1 and f_2.

Table 4 illustrates this point by generating two such variables (S_1, S_2) and showing that, by reducing σ^2_η, we are able to increase either of two goodness of fit criteria (the percentage of correct in sample predictions of D and the pseudo R^2) above those of the model with $I_E = I_R$. Adding a model based on S_2 and Z (bottom row) increases the successful prediction rate over the case when the true model is used (the model based on $\{Z, f_1, f_2\}$) but it is biased for all parameters and substantially biased for ATE and MTE.

The essential feature of this example is that the selected conditioning variables are endogenous with respect to the outcome equation (they violate (M-4)). If all candidate conditioning variables were restricted to be exogenous, our example could not be constructed. This underscores the importance of the econometric concept of endogeneity which is sometimes viewed as an inessential distinction in matching. Although it is irrelevant for *defining* parameters, it is essential when *selecting* conditioning variables.

7 Concluding remarks

This paper considers three main points regarding the use of the propensity score in econometric evaluation methods. The first point is that the economic and statistical assumptions required to justify the use of the propensity score are different in selection, matching and instrumental variables models. In general, one set of assumptions neither implies nor is implied by the other. In the case of additive separability of outcome equations, matching models are a special case of selection models that assumes that conditioning eliminates bias whereas control function methods model selection bias. Matching makes strong assumptions that are not required in the method of control functions. It assumes that conditional on observables the marginal return is the average return. One benefit of such strong assumptions is weaker assumptions about other features of the underlying economic model. Matching does not require separability of outcomes, exogeneity of regressors or exclusion restrictions provided valid conditioning sets are known.

The second main point is that the literature on matching provides no guidance on the choice of the conditioning variables that generate identification. We define the concept of the "minimum relevant" conditioning set that is assumed in matching. In general, it differs from the information set available to the analyst. Adding more "minimum relevant" variables but not all is not guaranteed to reduce bias and we offer examples of this point.

Our third main point is that the model selection criteria advocated to pick the variables in the conditioning set are not guaranteed to work. We offer examples where goodness of fit criteria advocated in the literature select conditioning sets that generate more bias than conditioning sets that are less successful in terms of model selection criterion. The methods work for choice among exogenous conditioning variables. This highlights the point that the econometric distinctions of exogeneity and endogeneity play crucial roles in the application of matching in the choice of conditioning sets.

The sensitivity of estimates obtained from matching to the choice of conditioning variables, the inability of the method to model omitted relevant conditioning variables and the lack of any clear rule for selecting conditioning variables should give pause to economists who embrace this method.[24] More robust methods based on the control function approach are more sensitive to problems of omitted conditioning variables. Recent semiparametric advances in the development of control functions make these procedures less vulnerable to the distributional assumptions that plagued the earlier literature on the topic (see Powell, 1994, and Heckman and Vytlacil, 2003).

Appendix

Consider a general model of the form:

$$\begin{aligned} Y_1 &= \mu_1 + U_1 \\ Y_0 &= \mu_0 + U_0 \\ V &= \mu_V(Z) + U_V \\ D &= 1 \text{ if } V \geq 0, \ = 0 \text{ otherwise} \\ Y &= DY_1 + (1-D)Y_0. \end{aligned}$$

where

$$\begin{aligned} (U_1, U_0, U_V)' &\sim N(0, \Sigma) \\ var(U_i) &= \sigma_i^2 \\ cov(U_i, U_j) &= \sigma_{ij} \\ i = 0; j = 1 & \\ cov(U_1, V) &= \sigma_{1V} \\ cov(U_0, V) &= \sigma_{0V} \end{aligned}$$

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the pdf and the cdf of a standard normal random variable. Then, the propensity score for this model is given by:

$$\begin{aligned} \Pr(V > 0 | \mu_V(Z)) &= P(\mu_V(Z)) = \Pr(U_V > -\mu_V(Z)) = p \\ &= 1 - \Phi\left(\frac{-\mu_V(Z)}{\sigma_V}\right) = p \end{aligned}$$

so

$$\frac{-\mu_V(Z)}{\sigma_V} = \Phi^{-1}(1-p).$$

Since the event $\left(V \lessgtr 0, P(\mu_V(Z)) = p\right)$ can be written as

$$\begin{aligned} \frac{U_V}{\sigma_V} &\lessgtr -\frac{\mu_V(Z)}{\sigma_V} \\ \frac{U_V}{\sigma_V} &\lessgtr \Phi^{-1}(1-p) \end{aligned}$$

we can write the conditional expectations required to get the biases defined in Section (4) as a function of

p. For U_1 :

$$\begin{aligned} E\left(U_1|V>0, P\left(\mu_V\left(Z\right)\right)=p\right) &= \frac{\sigma_{1V}}{\sigma_V}E\left(\frac{U_V}{\sigma_V}|\frac{U_V}{\sigma_V}>\frac{-\mu_V\left(Z\right)}{\sigma_V}, P\left(\mu_V\left(Z\right)\right)=p\right) \\ &= \frac{\sigma_{1V}}{\sigma_V}E\left(\frac{U_V}{\sigma_V}|\frac{U_V}{\sigma_V}>\Phi^{-1}\left(1-p\right)\right) \\ &= \beta_1 M_1(p) \end{aligned}$$

$$\begin{aligned} E\left(U_1|V=0, P\left(\mu_V\left(Z\right)\right)=p\right) &= \frac{\sigma_{1V}}{\sigma_V}E\left(\frac{U_V}{\sigma_V}|\frac{U_V}{\sigma_V}=\frac{-\mu_V\left(Z\right)}{\sigma_V}, P\left(\mu_V\left(Z\right)\right)=p\right) \\ &= \frac{\sigma_{1V}}{\sigma_V}E\left(\frac{U_V}{\sigma_V}|\frac{U_V}{\sigma_V}=\Phi^{-1}\left(1-p\right), P\left(\mu_V\left(Z\right)\right)=p\right) \\ &= \beta_1\Phi^{-1}\left(1-p\right) \end{aligned}$$

where

$$\beta_1 = \frac{\sigma_{1V}}{\sigma_V}$$

Similarly for U_0 :

$$\begin{aligned} E\left(U_0|V>0, P\left(\mu_V\right)=p\right) &= \beta_0 M_1(p) \\ E\left(U_0|V<0, P\left(\mu_V\right)=p\right) &= \beta_0 M_0(p) \\ E\left(U_0|V=0, P\left(\mu_V\right)=p\right) &= \beta_0\Phi^{-1}\left(1-p\right) \end{aligned}$$

where

$$\beta_0 = \frac{\sigma_{0V}}{\sigma_V}.$$

and

$$\begin{aligned} M_1\left(p\right) &= \frac{\phi\left(\Phi^{-1}\left(1-p\right)\right)}{p} \\ M_0\left(p\right) &= -\frac{\phi\left(\Phi^{-1}\left(1-p\right)\right)}{\left(1-p\right)} \end{aligned}$$

are inverse Mills ratio terms.

Substituting these into the expressions for the biases

$$\begin{aligned} \text{Bias } TT\left(p\right) &= \beta_0 M_1(p) - \beta_0 M_0(p) \\ &= \beta_0 M(p) \end{aligned}$$

$$\begin{aligned} \text{Bias } ATE\left(p\right) &= \beta_1 M_1(p) - \beta_0 M_0(p) \\ &= M(p)\left(\beta_1\left(1-p\right)+\beta_0 p\right) \end{aligned}$$

$$\begin{aligned}\text{Bias } MTE &= \beta_1 M_1(p) - \beta_0 M_0(p) - \beta_1 \Phi^{-1}(1-p) + \beta_0 \Phi^{-1}(1-p) \\ &= M(p)\left(\beta_1(1-p) + \beta_0 p\right) - \Phi^{-1}(1-p)\left[\beta_1 - \beta_0\right].\end{aligned}$$

where

$$M(p) = M_1(p) - M_0(p) = \frac{\phi(\Phi^{-1}(1-p))}{p(1-p)}$$

Bibliography

1. Aakvik, Arild, James J. Heckman, and Edward Vytlacil, "Estimating Treatment Effects for Discrete Outcomes When Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs," *The Journal of Econometrics* (forthcoming, 2003).

2. Abadie, Alberto, (2002) "Semiparametric Difference-in-differences Estimators," Unpublished Manuscript, Harvard University.

3. Ahn, Hyungtaik and James L. Powell, "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *The Journal of Econometrics* 58:1-2 (1993), 3-29.

4. Andrews, Donald W.K. and Marcia M.A. Schafgans, "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies* 65:3 (1998), 497-518.

5. Björklund, Anders and Robert Moffitt, "The Estimation of Wage Gains and Welfare Gains in Self-selection," *The Review of Economics and Statistics* 69:1 (1987), 42-49.

6. Cameron, Stephen V. and James J. Heckman, "Life Cycle Schooling and Educational Selectivity: Models and Choice," *Journal of Political Economy* 106:2, (1998), 262-333

7. Carneiro, Pedro, "Heterogeneity in the Returns to Schooling: Implications for Policy Evaluation," Unpublished Ph.D. Thesis University of Chicago (2002).

8. Carneiro, Pedro, Karsten Hansen, and James J. Heckman, "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies," *Swedish Economic Policy Review* 8, (2001), 273-301

9. _____, "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," (forthcoming 2003) *International Economic Review*, May.

10. Carneiro, Pedro, James J. Heckman, and Edward Vytlacil, "Estimating the Return to Education When it Varies Among Individuals," Working paper, University of Chicago (2001).

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

12. Gerfin, Michael and Lechner, Michael, "A Microeconometric Evaluation of the Active Labor Market Policy in Switzerland," *The Economic Journal* 112 October, (2002), 854-893.

13. Gill, Richard D. and James M. Robins, "Causal inference for complex longitudinal data: the continuous case," *The Annals of Statistics* 29:6, (2001), 1-27.

14. Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66:2, (1998), 315-332.

15. Hansen, Karsten, James J. Heckman and Kathleen Mullen, "The Effect of Schooling and Ability on Achievement Test Scores," *The Journal of Econometrics* (forthcoming, 2003).

16. Heckman, James J., "Addendum to Sample Selection Bias as a Specification Error," in Ernst Stromsdorfer and George Farkas (Eds.), *Evaluation Studies Review Annual* Vol. 5, (Beverly Hills, CA: Sage Publications, 1980)

17. _____, "Varieties of Selection Bias," *American Economic Review* 80:2, (1990), 313-318.

18. _____, "Randomization and Social Policy Evaluation," in Charles Manski and Irwin Garfinkel (Eds.), *Evaluating Welfare and Training Programs* (Cambridge: Harvard University Press, 1992).

19. _____, "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *The Journal of Human Resources* 32:3, (1997), 441-462.

20. _____, "Detecting Discrimination," *Journal of Economic Perspectives* 12:2, (1998), 101-116.

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

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

23. Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, (1998), 1017 -1098.

24. Heckman, James J., Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *The Review of Economic Studies* 64:4, (1997), 605-654.

25. _____, "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies* 65:2, (1998), 261-294.

26. Heckman, James J. and Richard Robb, "Alternative Methods for Estimating The Impact of Interventions," In James J. Heckman and Burton Singer (Eds.), *Longitudinal Analysis of Labor Market Data* (Cambridge: Cambridge University Press, 1985).

27. _____, "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in Howard Wainer (Ed.), *Drawing Inferences from Self-selected Samples* (New Jersey: Lawrence Erlbaum Associates, 1986. Reprinted 2000).

28. Heckman, James J., Jeffrey Smith and Nancy Clements, "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies* 64:4, (1997), 487-535.

29. Heckman, James J., Justin Tobias and Edward Vytlacil, "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal* (2001), 68(2), 210-223..

30. Heckman, James J. and Edward Vytlacil, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences* 96, (1999), 4730-4734.

31. _____, "The Relationship Between Treatment Parameters within a Latent Variable Framework," *Economics Letters* 66:1, (2000), 33-39.

32. _____, "Local Instrumental Variables," in Cheng Hsiao, Kimio Morimune, and James Powell (Eds.), *Nonlinear statistical modeling :proceedings of the thirteenth International Symposium in Economic Theory and Econometrics : essays in honor of Takeshi Amemiya* (New York : Cambridge University Press, 2001).

33. _____, "Econometric Program Evaluation," in James Heckman and Edward Leamer (Eds.), *Handbook of Econometrics, Volume 5*, (Amsterdam: Elsevier, forthcoming 2003).

34. Imbens, Guido and Joshua Angrist "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62:2., (1994), 467-475.

35. LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76:4, (1986), 604-20.

36. Navarro-Lozano, Salvador, (2002) "The Importance of Being Formal: Testing for Segmentation in the Mexican Labor Market," Unpublished Manuscript, University of Chicago.

37. Olley, G. Steven and Ariel Pakes, "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica* 64:6 (1996), 1263-97.

38. Powell, James, "Estimation of Semiparametric Models," In Robert F. Engle and Daniel L. McFadden (Eds.), *Handbook of Econometrics Vol. 4* (Amsterdam, London and New York: Elsevier, North-Holland, 1994).

39. Robins, James M, "Causal Inference from Complex Longitudinal Data Latent Variable Modeling and Applications to Causality," in M. Berkane, (Ed), *Lecture Notes in Statistics* (New York: Springer Verlag, 1997).

40. Rosenbaum, Paul, *Observational Studies*, New York: Springer-Verlag. First edition 1995, second edition 2002.

41. Rosenbaum, Paul and Donald Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70:1, (1983), 41-55.

42. Smith, Jeffrey and Petra Todd, "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *American Economic Review* 91:2, (2001), 112-18.

43. Smith, Jeffrey and Petra Todd, "Is Matching the Answer to LaLonde's Critique of Nonexperimental Methods?," Forthcoming *Journal of Econometrics*.

44. Vijverberg, Wim, "Measuring the Unidentified Parameter of the Roy Model of Selectivity," *The Journal of Econometrics* 57:1-3, (1993), 69-89.

45. Vytlacil, Edward, "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica* 70:1, (2002), 331-341.

Notes

[1]See, *e.g.*, Olley and Pakes (1996) who confuse the use of the propensity score in matching and in control function methods.

[2]Heckman, Ichimura and Todd (1997) introduced this distinction into matching models.

[3]Papers that account for estimated P include Heckman, Ichimura and Todd (1997, 1998), and Hahn (1998).

[4]See Heckman, Ichimura and Todd (1997) and Abadie (2002).

[5]As demonstrated in Carneiro (2002), one can still distinguish marginal and average effects in terms of observables.

[6]The precise condition is that $Support\left(X|Z\right) = Support\left(X\right)$.

[7]As noted by Heckman, Ichimura, Smith and Todd (1998), if one seeks to identify $E\left(Y_1 - Y_0|D=1,W\right)$ one only needs to impose a weaker condition $\left(E\left(Y_0|D=1,W\right)\right) = E\left(Y_0|D=0,W\right)$ or $Y_0 \perp\!\!\!\perp D|W$ rather than (M-1). This imposes the assumption of no selection on levels of Y_0 (given W) and not the assumption of no selection on levels of Y_1 or change, as (M-1) does.

[8]Examples of nonseparable models are found in Cameron and Heckman (1998).

[9]Or mean independent in the case of mean parameters.

[10]Heckman and Robb (1985, 1986) introduce this general formulation of control functions. The identifiability requires that the members of the pairs $\left(\mu_1\left(X\right), E\left(U_1|X,Z,D=1\right)\right)$ and $\left(\mu_0\left(X\right), E\left(U_0|X,Z,D=0\right)\right)$ be "variation free" or "measurably separable" so that they can be independently varied against each other. See Heckman and Vytlacil (2003) for a precise statement of these conditions.

[11]More precisely, $Support\left(Z|X\right) = Support\left(Z\right)$. This is also the support condition used in the generalization of matching by Heckman, Ichimura and Todd (1997).

[12]This condition is sometimes called "identification at infinity." See Heckman (1990) or Andrews and Schafgans (1998).

[13]Since

$$\begin{aligned} E\left(U_0\right) &= 0 \\ &= E\left(U_0|D=1,Z\right)P\left(Z\right) + E\left(U_0|D=0,Z\right)\left(1-P\left(Z\right)\right) \\ E\left(U_0|D=1,Z\right) &= -\frac{\left(1-P\left(Z\right)\right)}{P\left(Z\right)}E\left(U_0|D=0,Z\right) = -\frac{\left(1-P\left(Z\right)\right)}{P\left(Z\right)}K_0\left(P\left(Z\right)\right) \end{aligned}$$

See Heckman and Robb (1986).

[14]For many common functional forms for the distributions of unobservables, no exclusion is required.

[15]Relaxing it, however, requires that the analyst model the dependence of the unobservables on the observables and that certain variation-free conditions are satisfied (See Heckman and Robb, 1985).

[16]See also Viverberg (1993) who does such a sensitivity analysis in a parametric model with an unidentified parameter.

[17]Proof:

$$\begin{aligned} E\left(Y|X,P\left(Z\right)\right) &= E\left(Y_1|D=1,X,P\left(Z\right)\right)P\left(Z\right) \\ &\quad +E\left(Y_0|D=0,X,P\left(Z\right)\right)\left(1-P\left(Z\right)\right) \\ &= \int_{-\infty}^{\infty}\int_{-P(Z)}^{\infty} y_1 f\left(y_1,U_V^*|X\right)dU_V^*dy_1 \\ &\quad +\int_{-\infty}^{\infty}\int_{-\infty}^{-P(Z)} y_0 f\left(y_0,U_V^*|X\right)dU_V^*dy_0 \end{aligned}$$

where $U_V^* = F_V\left(U_V\right)$. Thus

$$\begin{aligned} \frac{\partial E\left(Y|X,P\left(Z\right)\right)}{\partial P\left(Z\right)} &= E\left(Y_1-Y_0|X,U_V^*=-P\left(Z\right)\right) \\ &= MTE. \end{aligned}$$

[18]See also the discussion in Gerfin and Lechner (2002).

[19]We start with a primitive probability space (Ω,σ,P) with associated random variables I. We use minimal sigma algebras and assume the I are measurable with respect to these random variables.

[20]This formulation assumes that the agent makes the treatment decision. If not, then we mean by the agent, the decision maker.

[21]Models that take this form are known as factor models and have been applied in the context of selection by Aakvik, Heckman and Vytacil (2003), Carneiro, Hansen and Heckman (2001, 2003) Hansen, Heckman and Mullen (2003) and Navarro-Lozano (2002) among others.

[22]For example, the returns to schooling literature often uses different test scores, like AFQT or IQ, to proxy for missing ability variables.

[23]See Aakvik, Heckman and Vytlacil (2003) for an analysis of discrete treatment effects in a latent variables model. See also Heckman (1998) where this framework originates.

[24]A widely cited paper by Dehejia and Wahba (1999) claims that matching overcomes the sensitivity to estimators problem displayed by LaLonde (1986). Smith and Todd (2001, 2003) show that the Dehejia-Wahba results were manufactured by selectively discarding data from LaLonde's original sample and that when the full sample is used matching produces substantial biases. Matching does not solve the LaLonde sensitivity problem.

Table 1
Mean Bias for Treatment on the Treated

ρ_{0V}	Average Bias (σ_0=1)	Average Bias (σ_0=2)
-1.00	-1.7920	-3.5839
-0.75	-1.3440	-2.6879
-0.50	-0.8960	-1.7920
-0.25	-0.4480	-0.8960
0	0	0
0.25	0.4480	0.8960
0.50	0.8960	1.7920
0.75	1.3440	2.6879
1.00	1.7920	3.5839

$BIASTT = \rho_{0V} * \sigma_0 * M(p)$

$M(p) = \phi(\Phi^{-1}(p)) \; / \; [p*(1-p)]$

Table 2
Mean Bias for Average Treatment Effect*

(σ_0=1)

	ρ_{1V} (σ_1=1)								
ρ_{0V}	-1.00	-0.75	-0.50	-0.25	0	0.25	0.50	0.75	1.00
-1.00	-1.7920	-1.5680	-1.3440	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0
-0.75	-1.5680	-1.3440	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240
-0.50	-1.3440	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480
-0.25	-1.1200	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480	0.6720
0	-0.8960	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480	0.6720	0.8960
0.25	-0.6720	-0.4480	-0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200
0.50	-0.4480	-0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440
0.75	-0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440	1.5680
1.00	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440	1.5680	1.7920

	ρ_{1V} (σ_1=2)								
ρ_{0V}	-1.00	-0.75	-0.50	-0.25	0	0.25	0.50	0.75	1.00
-1.00	-2.6879	-2.2399	-1.7920	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960
-0.75	-2.4639	-2.0159	-1.5680	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200
-0.50	-2.2399	-1.7920	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960	1.3440
-0.25	-2.0159	-1.5680	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200	1.5680
0	-1.7920	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960	1.3440	1.7920
0.25	-1.5680	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200	1.5680	2.0159
0.50	-1.3440	-0.8960	-0.4480	0	0.4480	0.8960	1.3440	1.7920	2.2399
0.75	-1.1200	-0.6720	-0.2240	0.2240	0.6720	1.1200	1.5680	2.0159	2.4639
1.00	-0.8960	-0.4480	0	0.4480	0.8960	1.3440	1.7920	2.2399	2.6879

*Equal to the Mean Bias for the Marginal Treatment Effect

$BIASATE = \rho_{1V}*\sigma_1*M_1(p) - \rho_{0V}*\sigma_0*M_0(p)$

$BIASMTE = BIASATE - \Phi^{-1}(1-p)*(\rho_{1V}*\sigma_1 - \rho_{0V}*\sigma_0)$

$M_1(p) = \phi(\Phi^{-1}(p)) / p$

$M_0(p) = -\phi(\Phi^{-1}(p)) / [1-p]$

Table 3

Method	Exclusion Required?	Separability of Observables and Unobservables in Outcome Equations?	Functional Forms Required?	Marginal = Average? (Given X, Z)	Key Identification Condition for Means Assuming Separability (See text for full conditions)
Matching	No	No	No	Yes	$E(U_1\|X, D=1, Z) = E(U_1\|X, Z)$ $E(U_0\|X, D=0, Z) = E(U_0\|X, Z)$
Control Function	Yes (for nonparametric identification)	Conventional, but not required	Conventional, but not required	No	$E(U_0\|X, D=0, Z)$ and $E(U_1\|X, D=1, Z)$ can be varied independently of $\mu_0(X)$ and $\mu_1(X)$, respectively and intercepts can be identified through limit arguments
IV (conventional)	Yes	Yes	No	No (Yes in standard case)	$E(U_0 + D(U_1 - U_0)\|X, Z)$ $= E(U_0 + D(U_1 - U_0)\|X)$ (ATE) $E(U_0 + D(U_1 - U_0) - E(U_0 + D(U_1 - U_0)\|X)\|P(Z), X)$ $= E(U_0 + D(U_1 - U_0) - E(U_0 + D(U_1 - U_0)\|X)\|X)$ (TT)
LIV	Yes	No	No	No	$(U_0, U_1, U_V) \perp\!\!\!\perp Z\|X$

Table 4

Variables in Probit	Goodness of fit statistics		Average Bias		
	Correct in-sample prediction rate	Pseudo R^2	*TT*	*ATE*	*MTE*
Z	66.88%	0.1284	1.1380	1.6553	1.6553
Z, f_2	75.02%	0.2791	1.2671	1.9007	1.9007
Z, f_1, f_2	83.45%	0.4844	0.0000	0.0000	0.0000
Z, S_1	77.59%	0.3352	0.8603	1.2513	1.2513
Z, S_2	92.45%	0.7555	0.3156	0.4591	0.4591

Model:

$V = Z + f_1 + f_2 + \varepsilon_v$ $\varepsilon_v \sim N(0,1)$ $S_1 = V + u_1$ $u_1 \sim N(0,4)$

$Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1$ $\varepsilon_1 \sim N(0,1)$ $S_2 = V + u_2$ $u_2 \sim N(0,0.25)$

$Y_0 = f_1 + 0.1f_2 + \varepsilon_0$ $\varepsilon_0 \sim N(0,1)$

$f_1 \sim N(0,1)$

$f_2 \sim N(0,1)$

Figure 1
Bias for Treatment on the Treated
Special case: Adding relevant information f2 increases the bias

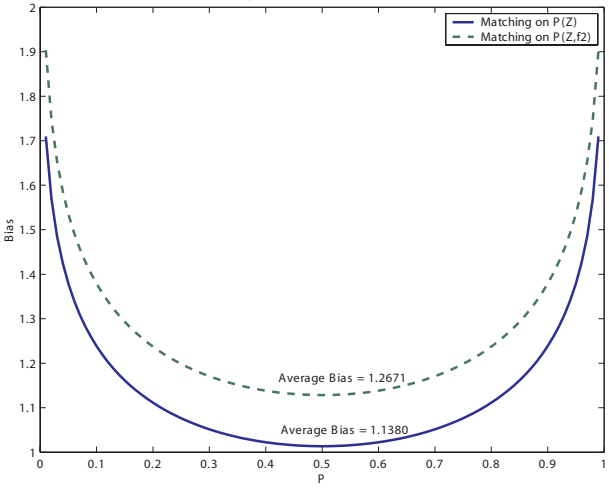


Figure 3
Bias for Marginal Treatment Effect
Special case: Adding relevant information f2 increases the bias

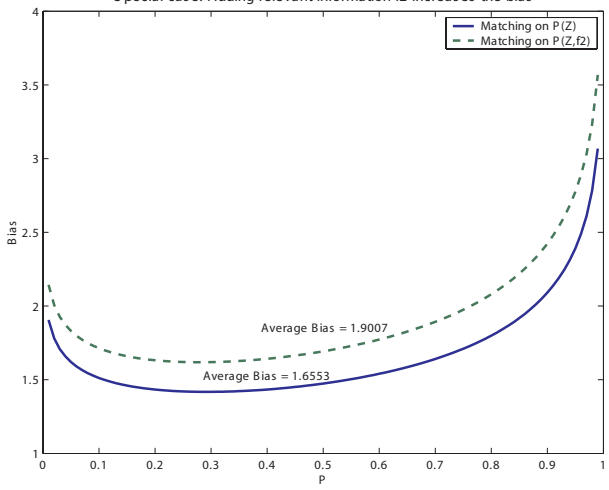


Figure 5
Bias for Average Treatment Effect
Special case: Adding irrelevant information $\tilde{Z}$ increases the bias
correlation($\tilde{Z}$,f2)=0.5

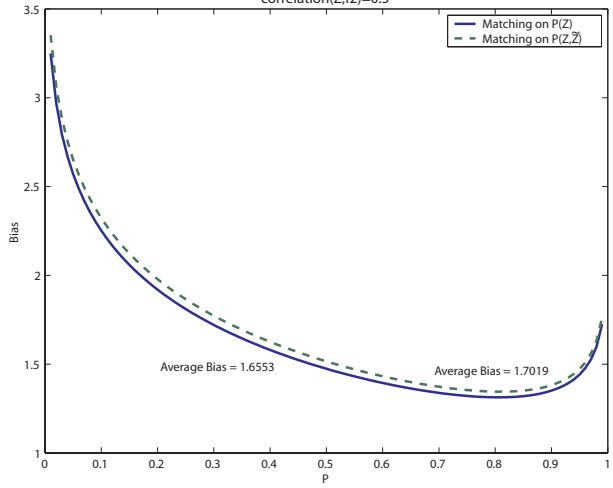


Figure 2
Bias for Average Treatment Effect
Special case: Adding relevant information f2 increases the bias

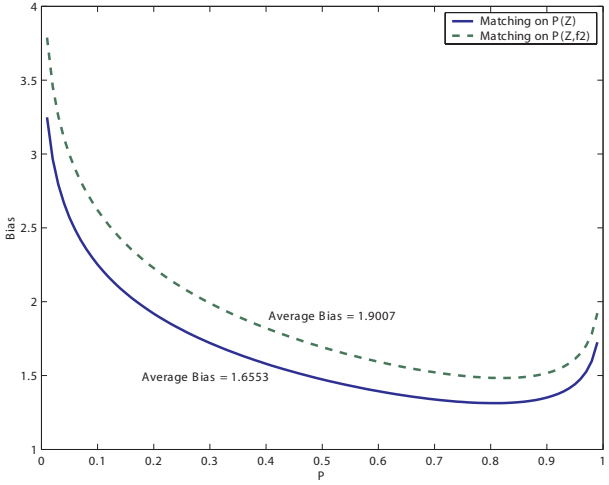


Figure 4
Bias for Treatment on the Treated
Special case: Adding irrelevant information $\tilde{Z}$ increases the bias
correlation($\tilde{Z}$,f2)=0.5

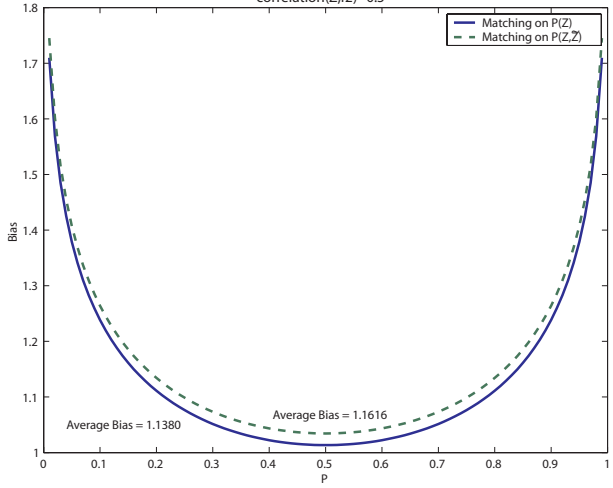
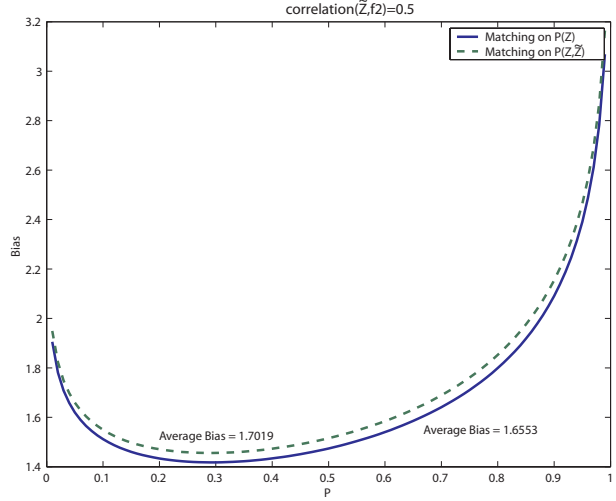


Figure 6
Bias for Marginal Treatment Effect
Special case: Adding irrelevant information $\tilde{Z}$ increases the bias
correlation($\tilde{Z}$,f2)=0.5



Model:

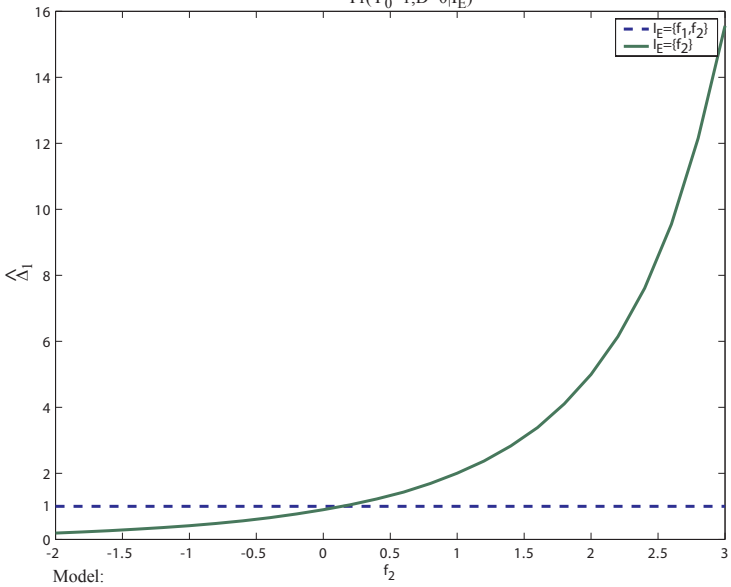
$V = Z + f_1 + f_2 + \varepsilon_V$ $\quad \varepsilon_V \sim N(0,1)$ $\quad f_1 \sim N(0,1)$

$Y_1 = 2f_1 + 0.1f_2 + \varepsilon_1$ $\quad \varepsilon_1 \sim N(0,1)$ $\quad f_2 \sim N(0,1)$

$Y_0 = f_1 + 0.1f_2 + \varepsilon_0$ $\quad \varepsilon_0 \sim N(0,1)$

Figure 7
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=1$

$$\hat{\Delta}_1 = \frac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_0=1,D=0|I_E)}$$

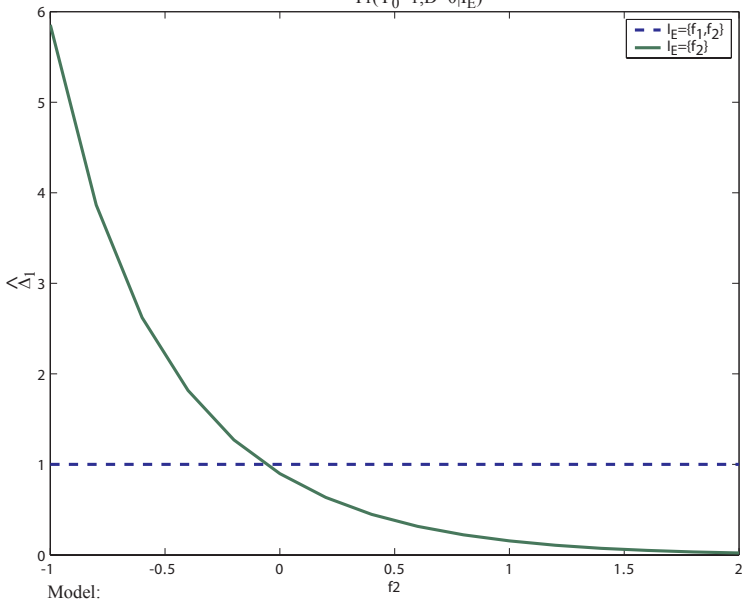


Model:

$V= -1+f_1+f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$
$D=1(V>0)$	

Figure 8
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=-1$

$$\hat{\Delta}_1 = \frac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_0=1,D=0|I_E)}$$

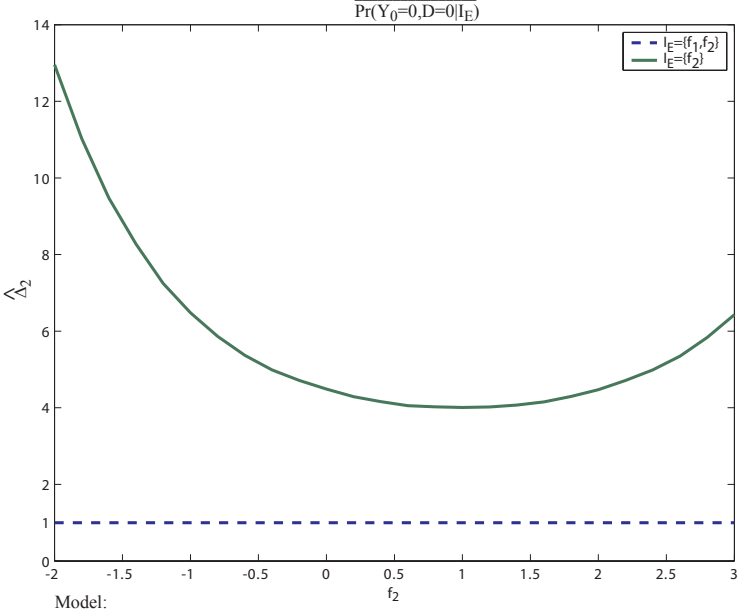


Model:

$V= -1+f_1-f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$

Figure 9
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=1$

$$\hat{\Delta}_2 = \frac{\dfrac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_1=0,D=1|I_E)}}{\dfrac{\Pr(Y_0=1,D=0|I_E)}{\Pr(Y_0=0,D=0|I_E)}}$$

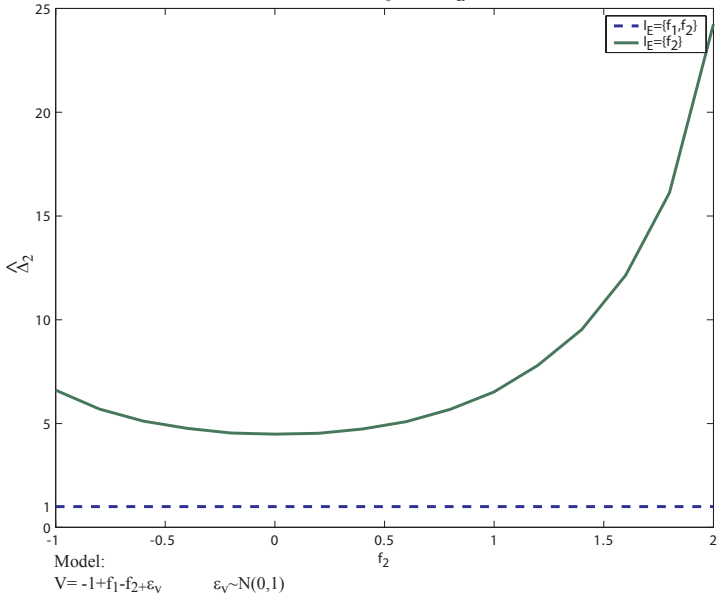


Model:

$V= -1+f_1+f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$
$D=1(V>0)$	

Figure 10
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=-1$

$$\hat{\Delta}_2 = \frac{\dfrac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_1=0,D=1|I_E)}}{\dfrac{\Pr(Y_0=1,D=0|I_E)}{\Pr(Y_0=0,D=0|I_E)}}$$

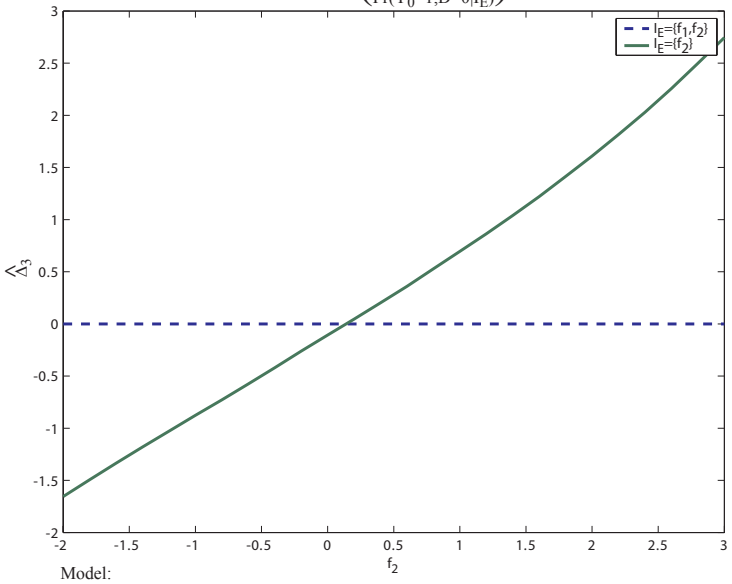


Model:

$V= -1+f_1-f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$
$D=1(V>0)$	

Figure 11
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=1$

$$\hat{\Delta}_3 = \text{Log}\left(\frac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_0=1,D=0|I_E)}\right)$$

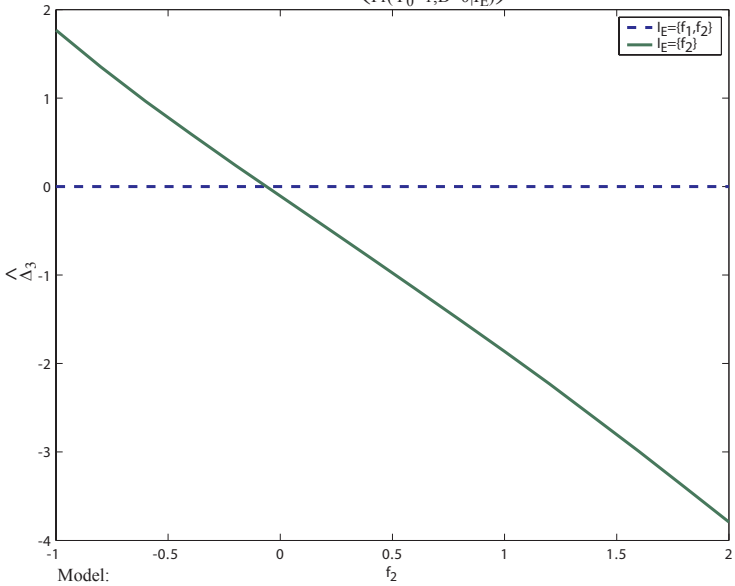


Model:

$V= -1+f_1+f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$
$D=1(V>0)$	

Figure 12
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=-1$

$$\hat{\Delta}_3 = \text{Log}\left(\frac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_0=1,D=0|I_E)}\right)$$

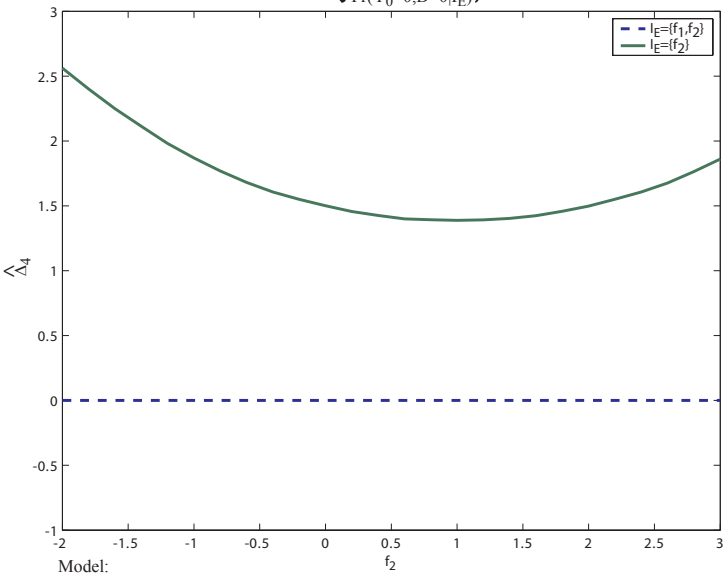


Model:

$V= -1+f_1-f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$
$D=1(V>0)$	

Figure 13
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=1$

$$\hat{\Delta}_4 = \text{Log}\left(\frac{\frac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_1=0,D=1|I_E)}}{\frac{\Pr(Y_0=1,D=0|I_E)}{\Pr(Y_0=0,D=0|I_E)}}\right)$$

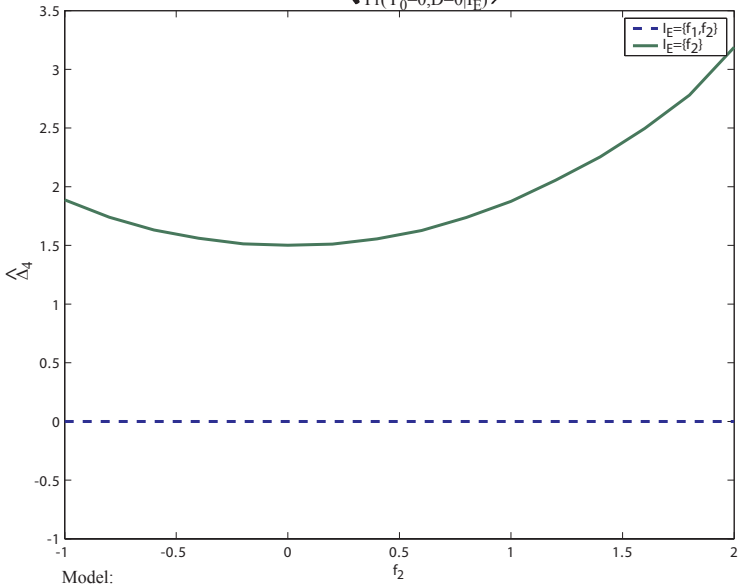


Model:

$V= -1+f_1+f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$
$D=1(V>0)$	

Figure 14
Estimated Effect of Treatment under Different Information Sets
No Effect of Treatment and $\alpha_{V2}=-1$

$$\hat{\Delta}_4 = \text{Log}\left(\frac{\frac{\Pr(Y_1=1,D=1|I_E)}{\Pr(Y_1=0,D=1|I_E)}}{\frac{\Pr(Y_0=1,D=0|I_E)}{\Pr(Y_0=0,D=0|I_E)}}\right)$$



Model:

$V= -1+f_1-f_2+\varepsilon_v$	$\varepsilon_v \sim N(0,1)$
$Y^*_1=-1+f_1+f_2+\varepsilon_1$	$\varepsilon_1 \sim N(0,1)$
$Y^*_0=-1+f_1+f_2+\varepsilon_0$	$\varepsilon_0 \sim N(0,1)$
$Y_1=1(Y^*_1>0)$	$f_1 \sim N(0,1)$
$Y_0=1(Y^*_0>0)$	$f_2 \sim N(0,1)$
$D=1(V>0)$	