Estimation of Treatment Effects under Essential Heterogeneity

James Heckman

University of Chicago and American Bar Foundation

Sergio Urzua

University of Chicago

Edward Vytlacil

Columbia University

March 20, 2006

1 Introduction

The objective of this document is to describe different estimation techniques that allow the computation of treatment effecs in the context of models with essential heterogeneity. We provide a FORTRAN code (*MTE exe)* that implements these techniques, and an example based on the Generalized Roy Model.

2 The Illustrative Model

Assume that the economic model generating the data for potential outcomes is of the form:

$$Y_1 = \alpha_1 + \varphi + X\beta_1 + U_1 \quad (1)$$
$$Y_0 = \alpha_0 + X\beta_0 + U_0 \quad (2)$$

where vector X represents the observed variables (regressors), U_0 and U_1 represent the unobservables in the potential outcome equations, and φ represents the benefit associated with the treatment ($D = 1$). The assumptions of linearity and additive separability are not intrinsic to the model of essential heterogeneity and are used in this supplement just to illustrate the estimation method we propose. The individuals decide whether or not to receive the treatment ($D = 1$ or $D = 0$) based on a latent variable I

$$I = \gamma' Z - V \quad (3)$$

where Z and V represent observables and unobservables respectively. Thus, we can define a binary variable D indicating the treatment status

$$D = \begin{cases} 1 \text{ if } I > 0 \\ 0 \text{ if } I \leq 0 \end{cases} \quad (4)$$

Finally, we assume that the error terms in the model are not independent even conditioning on the observables, i.e. $U_1 \not\perp\!\!\!\perp U_0 \not\perp\!\!\!\perp V | X_1, X_2$.

Equations (1)-(4) can be interpreted as the Generalized Roy Model (Heckman and Vytlacil, 2001a).

3 The Marginal Treatment Effect, Treatment Parameters and IV estimates.

The marginal treatment effect in the model of the previous section is

$$E(Y_1 - Y_0|X = x, V = v) \tag{5}$$

and it represents the benefits of treatment when $V = v$.

Notice that without lost of generality we can consider the following

$$\gamma' Z > V \iff F_V(\gamma' Z) > F_V(V)$$

where $F_V()$ is the cumulative distribution function of V. Then, if $P(Z)$ and U_D denote $F_V(\gamma' Z)$ and $F_V(V)$ respectively, we have that the choice model can be re-written as

$$P(Z) > U_D.$$

and (5) can be defined

$$E(Y_1 - Y_0|X = x, U_D = u_D)$$

The TT, TUT, ATE and IV estimators can be constructed as a weighted averages of the MTE (Heckman and Vytlacil, 2001a,b, 2005). In particular, if $\Delta_J^{IV}(x)$ denotes the IV estimator obtained by using the instrument J we have that:

$$\Delta_J^{IV}(x) = \int MTE\,(x, u_d)\,\omega_J(x, u_d)du_d$$

where

$$\omega_J(x, u_d) = \frac{(E(J|P(Z) > u_d, X = x) - E(J|X = x))\Pr(P(Z) > u_d|X = x)}{Cov(J, D|X = x)}$$

and $\omega_J(x, 0) = \omega_J(x, 1) = 0$ and $\int \omega_J(x, u_d)dF_{U_d}(u_d) = 1$.

Likewise,

$$\Delta^{TT}(x) = \int MTE\,(x, u_d)\,\omega_{TT}(x, u_d)du_d$$

where $\Delta^{TT}(x)$ represents the treatment on the treated conditional on $X = x$. Similar expressions exist for the other treatment parameters. These estimators depend on the particular value of X (x). In order to eliminate this dependence we need to integrate X out, so that we can define the unconditional estimators as

$$\Delta_J^{IV} = \int \Delta_J^{IV}(x)dF_X(x)$$

and

$$\Delta^{TT} = \int \Delta^{TT}(x)dF_{X|D=1}(x).$$

4 The Estimation of the Propensity Score and The Identification of the Relevant Support

The first step in the computation of the MTE is to estimate the probability of participation or propensity score, $\Pr(D = 1|Z = z) = P(z)$. This probability can be estimated using different methods. In this document, we assume

$V \sim N(0,1)$ and thus estimate $P(z)$ using a probit model. Let $\widehat{\gamma}$ denote the estimated value of γ in equation (3).[1] The predicted value of the propensity score (conditional on $Z = z$), $\widehat{P}(z)$, is then computed as $\widehat{P}(z) = \Pr(\widehat{\gamma}Z > V|Z = z) = \Phi(\widehat{\gamma}z)$ where Φ represents the cumulative distribution function of a standard normal random variable.

The predicted values of the propensity score allow us to define the values of u_D over which the *MTE* can be identified. In particular, as emphasized by Heckman and Vytlacil (2001a), identification of the *MTE* depends critically on the support of the propensity score. The larger the support of the propensity score, the bigger the set over which the *MTE* can be identified.

In order to define the relevant support we first estimate the frequencies of the predicted propensity scores in the samples of treated ($D = 1$) and untreated ($D = 0$) individuals. These frequencies are computed using simple histograms, and in both subsamples the same grid of values of $\widehat{P}(z)$ (Γ) specifies the number of points at which the histogram is to be evaluated.[2]

Let $\mathcal{P}_l$ denote the set of evaluation points (coming from the grid) such that

$$\mathcal{P}_l = \{p \in \Gamma | \epsilon < \Pr(\widehat{P}(z) = p | D = l) < 1 - \epsilon\} \text{ with } l = 0, 1 \text{ and } \epsilon > 0$$

so $\mathcal{P}_l$ represents the set of values of p for which we compute frequencies in the range $(0, 1)$ using the subsample of individuals declaring $D = l$ ($l = 0, 1$). Notice that the extreme values 0 and 1 are excluded from $\mathcal{P}_l$. Finally, if we denote by $\mathcal{P}$ the set of evaluation points used to define the relevant support of the propensity score, we have that

$$\mathcal{P} = \mathcal{P}_0 \bigcap \mathcal{P}_1 = \{p \in \Gamma | \epsilon < \Pr(\widehat{P}(z) = p | D = 0) < 1 - \epsilon \text{ and } \epsilon < \Pr(\widehat{P}(z) = p | D = 1) < 1 - \epsilon\}.$$

for $\epsilon > 0$. Therefore, the *MTE* is defined only for those evaluations of $\widehat{P}(z)$ for which we obtain positive frequencies for both subsamples.

In practice, after identifying the relevant or common support of the propensity score, it is necessary to adjust the sample. In particular, the observations for which $\widehat{P}(z)$ is contained in the common support are kept. The rest of the sample is dropped. From this point on, our analysis refers to the resulting sample.

5 Different Approaches to Estimate the Marginal Treatment under Essential Heterogeneity

Using the expression (1)-(4), it is easy to show that

$$E\left(Y|X = x, P\left(Z\right) = p\right) = \alpha + \beta_0 x + \left(\left(\beta_1 - \beta_0\right) x\right) p + K(p) \tag{6}$$

where $P(Z)$ represents the propensity score or probability of selection, p is a particular evaluation value of the propensity score and

$$K(p) = \varphi p + E(U_0 | P(Z) = p) + E\left(U_1 - U_0 | D = 1, P\left(Z\right) = p\right) p. \tag{7}$$

Equations (6) and (7) are the cornerstones in the approaches we start presenting now.

[1] Our code allows the utilization of non-parametric probit, linear probability model, and standard probit model in estimating the propensity score.

[2] In practice we set $\Gamma = \{0.01, 0.02, ..., 0.98, 0.99\}$.

5.1 The Parametric Approach

This approach uses the parametric form of the marginal treatment effect under the assumption of joint normality for the error terms. In particular, we add to the model presented in section 2 the following assumption:

$$(U_0, U_1, V) \sim N(0, \Sigma) \tag{8}$$

where Σ represents the variance and covariance matrix. In what follows we denote by σ_V^2 the variance of V, σ_i^2 the variance of U_i (with $i = 0, 1$), $\sigma_{V,i}$ the covariance between U_i and V (with $i = 0, 1$), and $\sigma_{i,j}$ the covariance between U_i and U_j (with $i \neq j$).

Therefore, we can write

$$\Pr(D = 1, Z) = \Pr\left(\frac{V}{\sigma_V} < \frac{Z\gamma}{\sigma_V}\right) = \Phi\left(\frac{Z\gamma}{\sigma_V}\right) = P(Z)$$

so

$$\frac{Z\gamma}{\sigma_V} = \Phi^{-1}(P(Z)).$$

where Φ represents the cumulative distribution function of a standard normal random variable and Φ^{-1} its inverse.

Additionally, by using assumption (8) we can obtain:

$$\begin{aligned} E(Y_1|D = 1, X, Z) &= \alpha_1 + \varphi + X\beta_1 + E(U_1|V < Z\gamma) \\ &= \alpha_1 + \varphi + X\beta_1 + \rho_1 E\left(\frac{V}{\sigma_V}|\frac{V}{\sigma_V} < \frac{Z\gamma}{\sigma_V}\right) \\ &= \alpha_1 + \varphi + X\beta_1 + \rho_1\left(-\frac{\phi\left(\frac{Z\gamma}{\sigma_V}\right)}{\Phi\left(\frac{Z\gamma}{\sigma_V}\right)}\right) \end{aligned}$$

or, in terms of the propensity score

$$E(Y_1|D = 1, X, P(Z)) = E(Y_1|D = 1, X, P(Z)) = \alpha_1 + \varphi + X\beta_1 + \rho_1\left(-\frac{\phi\left(\Phi^{-1}(P(Z))\right)}{P(Z)}\right)$$

where $\rho_1 = \frac{\sigma_{V,1}}{\sigma_V}$ and $\phi(.)$ denotes the density function associated with a standard normal random variable. Likewise,

$$E(Y_1|D = 0, X, P(Z)) = \alpha_1 + \varphi + X\beta_1 + \rho_1\frac{\phi\left(\Phi^{-1}(P(Z))\right)}{1 - P(Z)}$$

and

$$E(Y_1|Z\gamma - V = 0, X, P(Z)) = \alpha_1 + \varphi + X\beta_1 + \rho_1\Phi^{-1}(P(Z)).$$

Analogous expression can be obtained for $E(Y_0|D = 0, X, P(Z))$, $E(Y_0|D = 1, X, P(Z))$, and $E(Y_0|Z\gamma - V = 0, X, P(Z))$.

Notice that, as mentioned in Section 3, without lost of generality, we can re-write the decision rule as follows

$$D = \begin{cases} 1 \text{ if } Z\gamma - V > 0 \\ 0 \text{ if } Z\gamma - V \leq 0 \end{cases} \iff D = \begin{cases} 1 \text{ if } P(Z) - U_D > 0 \\ 0 \text{ if } P(Z) - U_D \leq 0 \end{cases}$$

where in this case $U_D = \Phi(V/\sigma_V)$. Therefore, the marginal treatment in this case can be defined as:

$$MTE(X = x, U_D = u_D) = \beta_1' x - \beta_0' x + (\rho_1 - \rho_0)\Phi^{-1}(u_D)$$

The following algorithm is used in the computation of the *MTE* under the parametric approach.

Step 1: We estimate $E(Y|X=x,D=1,P(Z)=p)$ and $E(Y|X=x,D=0,P(Z)=p)$ using the expressions

$$\begin{aligned} E(Y|X &= x,D=1,P(Z)=p) = \alpha_1+\varphi+X\beta_1+\rho_1\left(-\frac{\phi(\Phi^{-1}(p))}{p}\right) \\ E(Y|X &= x,D=0,P(Z)=p) = \alpha_0+X\beta_0X+\rho_0\frac{\phi(\Phi^{-1}(p))}{(1-p)} \end{aligned}$$

where we use the following facts

$$\begin{aligned} E(U_1|X &= x,D=1,P(Z)=p) = E(U_1|X=x,Z'\gamma>V,P(Z)=p) = -\rho_1\frac{\phi(\Phi^{-1}(p))}{p} \\ E(U_0|X &= x,D=0,P(Z)=p) = E(U_0|X=x,Z'\gamma<V,P(Z)=p) = \rho_0\frac{\phi(\Phi^{-1}(p))}{(1-p)}. \end{aligned}$$

Step 2: With the estimated values $\widehat{\alpha}_0, \widehat{\alpha_1+\varphi}, \widehat{\beta}_0, \widehat{\beta}_1, \widehat{\rho}_1$ and $\widehat{\rho}_0$ and by using the estimated propensity score (under the full support condition) we compute:

$$\widehat{MTE}_{Par}\left(X=x,U_D=u_D\right) = \widehat{\alpha_1+\varphi}-\widehat{\alpha}_0+x'\left(\widehat{\beta}_1-\widehat{\beta}_0\right)+\left(\widehat{\rho}_1-\widehat{\rho}_0\right)\Phi^{-1}\left(u_D\right).$$

5.2 Relaxing the Assumption of Normality: Using a Polynomial of the Propensity Score.

We can approximate the function $K(p)$ in equations (6) and (7) by a polynomial of p.[3] Thus, if ϑ denote the degree of the polynomial we obtain:

$$E\left(Y|X=x,P\left(Z\right)=p\right) = \alpha_0+x'\beta_0+\left(\alpha_1+\varphi-\alpha_0\right)p+\left(x'\left(\beta_1-\beta_0\right)\right)p+\sum_{i=1}^{\vartheta}\phi_i p^i \tag{9}$$

and, consequently, the estimator of the *MTE* is

$$\frac{\partial E\left(Y|X=x,P\left(Z\right)=p\right)}{\partial p} = \left(\alpha_1+\varphi-\alpha_0\right)+x'\left(\beta_1-\beta_0\right)+\sum_{i=1}^{\vartheta}i\phi_i p^{i-1}$$

Therefore, the following algorithm can be used to compute the *MTE*.

Step 1: First, fit the model

$$Y = \alpha_0+X'\beta_0+\left(X'\left(\beta_1-\beta_0\right)\right)P(Z)+\sum_{i=1}^{\vartheta}\phi_i P(Z)^i+\xi$$

where we assume $E(\xi|X=x,P(Z)=p)=0$. Notice that the term ϕ_1 includes $\left(\alpha_1+\varphi-\alpha_0\right)$.

Step 2: With the parameters values found in step 1 compute

$$\widehat{MTE}_{Pol}\left(X=x,U_D=p\right) = \widehat{\kappa}+x'(\widehat{\beta_1-\beta_0})+\sum_{i=1}^{\vartheta}i\widehat{\phi}_i p^{i-1}$$

where $\kappa=\left(\alpha_1+\varphi-\alpha_0\right)+\phi_1$.

[3] Intuitively, this idea is based on a series approximation of the conditional expectation.

6 The *LIV* Estimator (Semiparametric Method 1)

Heckman and Vytlacil (2001,2005) show that

$$\Delta^{LIV}(x, u_D) = \left. \frac{\partial E(Y|X = x, P(Z) = p)}{\partial p} \right|_{p=u_D} = \Delta^{MTE}(x, u_D).$$

This expression indicates that in general the computation of the *MTE* involves the estimation of the partial derivative of the expectation of the outcome Y (conditional on $X = x$ and $P(Z) = p$) with respect to p. This is the method of local instrumental variables introduced in Heckman and Vytlacil (2001). However, since we are considering the linear and separable version of the model of essential heterogeneity, we can use equations (6) and (7) to show that

$$\left. \frac{\partial E(Y|X = x, P(Z) = p)}{\partial p} \right|_{p=u_D} = x'(\beta_1 - \beta_0) + \left. \frac{\partial K(p)}{\partial p} \right|_{p=u_D} \tag{10}$$

Thus, in order to compute the *MTE* we need to estimate values for $(\beta_1 - \beta_0)$ and $\frac{\partial K(p)}{\partial p}$. Notice that without additional assumptions, the estimation of this last term requires the utilization of nonparametric techniques.

Different approaches can be used in the estimation of (10). The following steps describe the first semiparametric approach described in this document.

Step 1 We first estimate the coefficients β_0 and $(\beta_1 - \beta_0)$ in (6) using a semi-parametric version of the double residual regression procedure.[4] In order to do so, we start by fitting a local linear regression (*LLR*) of each regressor in (6) on the predicted propensity score $\widehat{P}(z)$. Notice that if n_X represents the number of variables in X, this step involves the estimation of $2 \times n_X$ local linear regressions. This is because equation (6) also contains terms of the form $X_k \widehat{P}(z)$ for $k = 1, ..., n_X$. We use the k-th regressor in (6), X_k, to illustrate the *LLR* procedure. Let $X_k(j)$ and $\widehat{P}(z(j))$ denote the values of the k-th regressor and predicted propensity score for the j-th individual, respectively, the latter evaluated at $z(j)$ (the observed value for the individual). The estimation of the *LLR* of X_k on $\widehat{P}(z)$ requires obtaining the values of $\{\theta_0(p), \theta_1(p)\}$ for a set of values of p contained in the support of $\widehat{P}(z)$ such that

$$\{\theta_0(p), \theta_1(p)\} = \arg \min_{\{\theta_0, \theta_1\}} \left\{ \sum_{j=1}^{N} (X_k(j) - \theta_0 - \theta_1(\widehat{P}(z(j)) - p))^2 \Psi((\widehat{P}(z(j)) - p)/h) \right\}$$

where $\Psi(.)$ and h represent the kernel function and the bandwidth, respectively and where θ_0 and θ_1 are parameters.[5] In practice, we use the set of all values of $\widehat{P}(z)$ to define the set of evaluation points (p) in the *LLR*. This allows us to estimate the predicted value of X_k for each individual in the sample.[6]

Let $\widehat{X}_k(j)$ denote the predicted value of X_k for the j-th individual. This procedure is repeated for each of the $2 \times n_X$ regressors in the outcome equations.

Step 2 Given the predicted values of the $2 \times n_X$ regressors $\widehat{X}_k$ $(k = 1, ..., 2 \times n_X)$, we now generate the residual for each regressor k and person j,

$$\widehat{e}_{X_k}(j) = X_k(j) - \widehat{X}_k(j) \quad \text{with } k = 1, \ldots, , 2 \times n_X.$$

[4] In the textbook case $Y = \lambda_1 X_1 + \lambda_2 X_2 + \epsilon$ where ϵ is assumed independent of X_1 and X_2, a double residual regression procedure estimates λ_2 using two stages. In the first stage, the estimated residuals of regressions of Y on X_2 and X_1 on X_2 are computed. Let ε_Y and ε_{X_1} denote these estimated residuals. In the second stage, λ_2 is estimated from the regression of ε_Y on ε_{X_1}.

[5] The selection of optimal bandwidth is an extensively studied issue in the nonparametric literature. In the code utilized in this paper two procedures computing optimal bandwidth in the context of local regressions are implemented. The first one is the standard leave-one-out crossvalidation procedure. The second procedure is the refined bandwidth selector described in Section 4.6 of Fan and Gijbels (1996). Our code allows the utilization of three different kernel functions: Epanechnikov, Gaussian and Biweight kernel functions.

[6] An alternative could be to use $\mathcal{P}$ as the set of evaluation points. In this case, in order to compute the predicted value of X_k for each individual, it would be necessary to replace his value of the predicted propensity score.by the closest value in $\mathcal{P}$.

We denote by $\widehat{e}_{X_k}$ the vector of residuals $(\widehat{e}_{X_k}(1), \widehat{e}_{X_k}(2), ..., \widehat{e}_{X_k}(N))'$, and by $\widehat{e}_X$ the matrix of residuals such that its k-th column contains the vector $\widehat{e}_{X_k}$.

Step 3 As in the standard double residual regression procedure, we also need to estimate a *LLR* of Y on $\widehat{P}(z)$. The same procedure as the one described in Step 1 is used in this case. Let $\widehat{Y}(j)$ denote the resulting predicted value of outcome Y for the j-th individual.

Step 4 With $\widehat{Y}(j)$ in hand, we generate the residual associated with outcome Y for each person j,

$$\widehat{e}_Y(j) = Y(j) - \widehat{Y}(j)$$

Following the notation used before, we denote by $\widehat{e}_Y$ the vector of residuals $(\widehat{e}_Y(1), ..., \widehat{e}_Y(N))'$.

Step 5 Finally, we can estimate the values of β_0 and $(\beta_1 - \beta_0)$ in (6) from a regression of $\widehat{e}_Y$ on $\widehat{e}'_X$. More specifically,

$$\left[\widehat{\beta}_0, \widehat{(\beta_1 - \beta_0)}\right] = [\widehat{e}'_X \widehat{e}_X]^{-1} [\widehat{e}'_X \widehat{e}_Y].$$

Heckman et al. (1998) use a similar double residual regression argument to characterize the selection bias in a semiparametric setup that arises from using nonexperimental data.

Step 6 From equation (10) we observe that after obtaining the estimated value of $(\beta_1 - \beta_0)$, only $\partial K(p)/\partial p$ remains to be estimated. However, with the estimated values of β_0 and $(\beta_1 - \beta_0)$ in hand, this term can be estimated using standard nonparametric techniques. To see why, notice that we can write

$$\widetilde{Y} = K(P(Z)) + \widetilde{v} \tag{11}$$

where $\widetilde{Y} = Y - X'\widehat{\beta}_0 - \left(X'\widehat{(\beta_1 - \beta_0)}\right) P(Z)$ and, as before, we assume $E(\widetilde{v}|P(z), X) = 0$. Then, it is clear from (11) that the problem reduces to the estimation of $\partial K\left(\widehat{P}(z)\right) \Big/ \partial \widehat{P}(z)$, where $K\left(\widehat{P}(z)\right)$ can be interpreted as the conditional expectation $E\left(\widetilde{Y}|P(Z) = \widehat{P}(z)\right)$.

Step 7 Let $\widehat{\vartheta}_1(p)$ denote the nonparametric estimator of $\partial K(p)/\partial p$. Notice that we define this estimator as a function of p instead of $\widehat{P}(z)$. This is because, unlike the case of the *LLR* estimators described in Step 1, we now use a subset of values of $\widehat{P}(z)$ to define the set of points (p) on which our estimator is evaluated. In particular, we use the set $\mathcal{P}$ to define this set of evaluation points. As shown above, $\mathcal{P}$ contains the values of $\widehat{P}(z)$ for which we obtain positive frequencies in both the $D = 0$ and $D = 1$ samples. Thus, $\widehat{\vartheta}_1(p)$ is computed as

$$\{\vartheta_0(p), \vartheta_1(p)\} = \arg \min_{\{\vartheta_0, \vartheta_1\}} \left\{ \sum_{j=1}^{N} (\widetilde{Y}(j) - \vartheta_0 - \vartheta_1(\widehat{P}(z(j)) - p))^2 \Psi((\widehat{P}(z(j)) - p)/h) \right\}$$

where as before $\Psi(.)$ and h represent the kernel function and the bandwidth, respectively.

Step 8 The *LIV* estimator of the *MTE* is finally computed as follows:

$$\Delta^{LIV}(x, u_D) = \widehat{(\beta_1 - \beta_0)}'x + \left.\widehat{\frac{\partial K(p)}{\partial p}}\right|_{p=u_D} = \widehat{MTE}(x, u_D)$$

and is evaluated over the set of p's contained in $\mathcal{P}$.

7 Adding Structure (Semiparametric Method 2).

A different approach, that combines the ideas of the two previous methods, can also be used in the estimation of the *MTE*. This approach has two stages. The first stage is closely related with the polynomial approximation method presented in section 5.2, whereas the second stage uses some of the ideas behind the method in section 6.

The method is the following. First, a control function approach is used to estimate the coefficients in $E\left(Y|X=x,P\left(Z\right)=p\right)$. In particular, based on the expression (6), we use a polynomial of degree d to approximate the function $K(p)$ (see expression (9) in Section 5.2). With these coefficients in hand, and using the same logic as in the step 6 in the previous method, we construct a residualized version of the outcome. Finally, by fitting a nonparametric regression of this new outcome on the propensity score, and by computing its derivative we obtain the *MTE*. These two last steps are similar to the steps 6,7 and 8 in the previous section.

Formally, the method can be described by the following algorithm.

Step 1: We fit the model

$$Y=\alpha_0+X\beta_0+\left(X\left(\beta_1-\beta_0\right)\right)P\left(Z\right)+\sum_{i=1}^{d}\psi_i P\left(Z\right)^i+\varsigma$$

where we assume $E(\varsigma|X=x,P(Z)=p)=0$. Notice that the term ψ_1 includes $\left(\alpha_1+\varphi-\alpha_0\right)$.

Step 2. With the estimated values of α_0, β_0 and $(\beta_1-\beta_0)$ in hand, we compute a residualized version of our outcome. This new variable allows us to apply standard nonparametric techniques in the estimation of $K(p)$. To see why, notice that we can write:

$$\widetilde{Y}=K(P\left(Z\right))+\widetilde{v}$$

where $\widetilde{Y}=Y-\widehat{\alpha}_0-X'\widehat{\beta}_0-\left(X'\widehat{(\beta_1-\beta_0)}\right)P\left(Z\right)$ and as before, we assume $E(\widetilde{v}|P\left(Z\right))=0$. Therefore, the problem reduces to the estimation of $K(p)$, where $K(p)$ can be interpreted as the conditional expectation $E(\widetilde{Y}|P\left(Z\right)=p)$.

Step 3. Standard nonparametric techniques for the estimation of local derivatives are used in the estimation of

$$\frac{\partial E(\widetilde{Y}|P\left(Z\right)=p)}{\partial p}=\frac{\partial K(p)}{\partial p}$$

Our code allows the utilization of local polynomial of higher order to approximate $K(p)$, and so the derivative is computed accordingly to the selected order.

Step 4: Finally, the *LIV* estimator of the *MTE* in this case is

$$\widehat{MTE}_{Non2}\left(x=x,U_D=u_D\right)=x'\widehat{(\beta_1-\beta_0)}+\left.\widehat{\frac{\partial K(p)}{\partial p}}\right|_{p=u_D}.$$

8 The Weights

The IV weights. Let J be the instrument. For simplicity we assume that J is a scalar. Then, the *IV* weight is:

$$\omega_J(x,u_D)=\frac{\left(E(J|P(Z)>u_D,X=x)-E(J|X=x)\right)\Pr(P(Z)>u_D|X=x)}{Cov(J,D|X=x)} \tag{12}$$

see Heckman and Vytlacil (2001a) and Heckman et al. (2006) for a derivation of this expression.

In order to compute the weight:

Step 1 We approximate $\widehat{E}(J|X=x)$ using a linear projection, i.e. we assume $J=\lambda' X+V$ where $E(V|X=x)=0$, so $\widehat{E}(J|X=x)=\widehat{\lambda}' x$.

Step 2 For each value of u_D we generate the auxiliary indicator function $I\left[P(Z)>u_D\right]$ which is equal to 1 if the argument of the function is true and 0 otherwise.

Step 3 We use linear projections to estimate $E(J|X=x,P(Z)>u_D)$. More precisely, we use OLS to estimate the equation $J(u_d)=\lambda'_{J(u_D)}X+V$ using only the observations for which $I\left[P(Z)>u_D\right]=1$. Since we assume $E(V|X=x,P(Z)>u_D)=0$, then $\widehat{E}(J|X=x,P(Z)>u_D)=\widehat{\lambda}'_{J(u_D)}x$.

Step 4 Since $\Pr(P(Z)>u_D|X=x)=\Pr(I\left[P(Z)>u_D\right]=1|X=x)$ we use a probit model (for each value of u_D) to estimate this probability. Let $\widehat{\Pr}(P(Z)>u_D|X=x)$ denote the estimated probability.

Step 5 We repeat steps 2, 3 and 4 for each value of $u_{D.}$.

Step 6 With $\widehat{E}(J|X=x)$, $\widehat{E}(J|X=x,P(Z)>u_D)$ and $\widehat{\Pr}(P(Z)>u_D|X=x)$ in hand we can compute the numerator of (12). In order to get the denominator, we use the fact that

$$\int \omega_J(x,u_D)du_D=\frac{1}{Cov(J,D|X=x)}\int\left(E(J|P(Z)>u_D,X=x)-E(J|X=x)\right)\Pr(P(Z)>u_D|X=x)du_D=1$$

so with the numerator in hand, it is straightforward to obtain the value of the covariance (conditional on X).

The Treatment Parameter weights. We use the Treatment on the Treated (TT) parameter to illustrate the computation of the treatment parameter weights. The TT weight is:

$$\omega_{TT}(x,u_D)=\frac{\Pr(P(Z)>u_D|X=x)}{\int\Pr(P(Z)>u_D|X=x)du_D}$$

and consequently, we can use $\widehat{\Pr}(P(Z)>u_D|X=x)$ to estimate the $\omega_{TT}(x,u_D)$. As in the case of $\omega_J(x,u_D)$, with the estimated value of $\widehat{\Pr}(P(Z)>u_D|X=x)$ in hand, we can directly obtain the value for $\int\Pr(P(Z)>u_D|X=x)du_D$, using the fact $\int\omega_{TT}(x,u_D)du_D=1$.

Therefore, provided with $\widehat{MTE}\left(x,u_d\right)$ and estimated values for the weights we can can compute $\widehat{\Delta}^{IV}_J(x)$ and $\widehat{\Delta}^{TT}(x)$. Finally, since these estimators depend on the particular value of X (x) we can integrate X out to obtain $\widehat{\Delta}^{IV}_J$ and $\widehat{\Delta}^{TT}$.

9 Optimal Bandwidth and Kernel Functions

When applying nonparametric techniques is necessary to consider the selection of the bandwidth and the kernel function. Cross-validation and data-driven procedures of bandwidth selection are frequently used in the applied nonparametric literature. We implement the standard leave-one-out crossvalidation procedure, as well as the global optimal bandwidth selection method recommended by Fan and Gijbels (1996).

In term of the kernel function, our code allows us to use three different kernel functions: Gaussian or Normal kernel, Biweight Kernel, and Epanechnikov kernel.

10 Example

Our example considers the following parametrization of the model:

$$\begin{aligned} U_1 &= \sigma_1 \epsilon \\ U_0 &= \sigma_0 \epsilon \\ V &= \sigma_V^* \epsilon \end{aligned}$$

$$\epsilon \sim N(0,1) \tag{13}$$

$$\sigma_1 = 0.012, \sigma_0 = -0.05, \sigma_V^* = -1 \tag{14}$$

$$\alpha_0 = 0.02, \alpha_1 = 0.04$$

$$\beta_0 = [\beta_{10}, \beta_{20}] = [0.5, 0.1], \beta_1 = [\beta_{11}, \beta_{21}] = [0.8, 0.4]$$

$$\varphi = 0.2$$

$$\gamma_0 = 0.2, \gamma_1 = 0.3, \gamma_2 = 0.1$$

In the case of the independent variables we assume:

$$\begin{aligned} X_1 &\sim N(-2,4),\ X_2 \sim N(2,4) \\ Z_1 &\sim N(-1,9),\ Z_2 \sim N(1,9) \end{aligned}$$

where $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp Z_1 \perp\!\!\!\perp Z_2$. Since we only observe Y_1 or Y_0 for each individual, but not both, we construct the observed outcome Y as

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

This framework allows us to compute the value of the different treatment parameters, as well as the IV estimator, using exact forms. In particular, if we denote by $\Pr(D=1|Z=z) = P(z)$ the propensity score we have that

$$\Pr(D=1,Z) = \Pr\left(\frac{V}{\sigma_V} < \frac{Z\gamma}{\sigma_V}\right) = \Phi\left(\frac{Z\gamma}{\sigma_V}\right) = P(Z)$$

so

$$\frac{Z\gamma}{\sigma_V} = \Phi^{-1}(P(Z)).$$

where Φ represents the cumulative distribution function of a standard normal random variable and Φ^{-1} its inverse.

Additionally, by using assumption (13) we have

$$\begin{aligned} E(Y_1|D=1,X,Z) &= \alpha_1 + \varphi + \beta_{11}X_1 + \beta_{12}X_2 + E(U_1|V < Z\gamma) \\ &= \alpha_1 + \varphi + \beta_{11}X_1 + \beta_{12}X_2 + \rho_1 E\left(\frac{V}{\sigma_V}|\frac{V}{\sigma_V} < \frac{Z\gamma}{\sigma_V}\right) \\ &= \alpha_1 + \varphi + \beta_{11}X_1 + \beta_{12}X_2 + \rho_1\left(-\frac{\phi\left(\frac{Z\gamma}{\sigma_V}\right)}{\Phi\left(\frac{Z\gamma}{\sigma_V}\right)}\right) \end{aligned}$$

or, in terms of the propensity score

$$E(Y_1|D=1,X,P(Z)) = \alpha_1 + \varphi + \beta_{11}X_1 + \beta_{12}X_2 + \rho_1\left(-\frac{\phi\left(\Phi^{-1}(P(Z))\right)}{P(Z)}\right)$$

where $\rho_1 = \frac{\sigma_V^* \sigma_1}{|\sigma_V^*|} = -\sigma_1$ and ϕ denotes the density function associated with a standard normal random variable. Likewise,

$$E\left(Y_1|D=0,X,P\left(Z\right)\right) = \alpha_1 + \varphi + \beta_{11}X_1 + \beta_{12}X_2 + \rho_1 \frac{\phi\left(\Phi^{-1}\left(P\left(Z\right)\right)\right)}{1-P\left(Z\right)}$$

and

$$E\left(Y_1|Z\gamma - V = 0,X,P\left(Z\right)\right) = \alpha_1 + \varphi + \beta_{11}X_1 + \beta_{12}X_2 + \rho_1 \Phi^{-1}\left(P\left(Z\right)\right).$$

Analogous expression can be obtained for $E\left(Y_0|D=0,X,P\left(Z\right)\right)$, $E\left(Y_0|D=1,X,P\left(Z\right)\right)$, and $E\left(Y_0|Z\gamma - V = 0,X,P\left(Z\right)\right)$.

Notice that, without lost of generality, we can re-write the decision rule as follows

$$D = \begin{cases} 1 \text{ if } Z\gamma - V > 0 \\ 0 \text{ if } Z\gamma - V \leq 0 \end{cases} \iff D = \begin{cases} 1 \text{ if } P\left(Z\right) - U_D > 0 \\ 0 \text{ if } P\left(Z\right) - U_D \leq 0 \end{cases} \iff$$

where $U_D = \Phi(V/\sigma_V)$.

Therefore, the marginal treatment can be written as:

$$MTE\left(X=x,U_D=u_D\right) = \left(\alpha_1 - \alpha_0\right) + \varphi + x'\left(\beta_1 - \beta_0\right) + \left(\rho_1 - \rho_0\right)\Phi^{-1}\left(u_D\right)$$

and provided with the values of $\alpha_0, \alpha_1, \beta_0, \beta_1, \rho_0$, and ρ_1 we can compute the MTE in this case.

The IV weights can be easily simulated as well. In particular, since $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp Z_1 \perp\!\!\!\perp Z_2$ we have that

$$\omega_J(x,u_D) = \frac{\left(E(J|P(Z) > u_d) - E(J)\right)\Pr(P(Z) > u_D)}{Cov(J,D)}$$

where J represents the instrument (either Z_1, Z_2 or a function of them), so since

$$\Pr(P(Z) > u_D) = \Pr(Z\gamma > \Phi^{-1}\left(u_D\right))$$

and Z_1 and Z_2 are normally distributed, we have that

$$\Pr(P(Z) > u_D) = 1 - \Phi_{Z\gamma}(\Phi^{-1}\left(u_D\right))$$

where $\Phi_{Z\gamma}()$ represents the cdf of a normal distribution with mean 0 $(= 0.2 - 0.3 + 0.1)$ and variance 1.6 $(= 0.3^2 \times 16 + 0.1^2 \times 16)$. In order to give an explicit expression for $E(J|P(Z) > u_d)$ let's assume that $J = Z_1$. Then,

$$\begin{aligned}
E(J|P(Z) > u_D) &= E(Z_1|Z\gamma > \Phi^{-1}\left(u_D\right)) \\
&= E(Z_1|Z\gamma > \Phi^{-1}\left(u_D\right)) \\
&= E(Z_1) + \frac{Cov(Z_1, Z\gamma)}{Var(Z\gamma)} E(Z\gamma|Z\gamma > \Phi^{-1}\left(u_D\right)) \\
&= E(Z_1) + \frac{Cov(Z_1, Z\gamma)}{\sqrt{Var(Z\gamma)}} \frac{\phi\left(\frac{\Phi^{-1}(u_D)}{\sqrt{Var(Z\gamma)}}\right)}{1 - \Phi\left(\frac{\Phi^{-1}(u_D)}{\sqrt{Var(Z\gamma)}}\right)} \\
&= E(Z_1) + \frac{Cov(Z_1, Z\gamma)}{\sqrt{Var(Z\gamma)}} \left(\frac{\phi\left(\frac{\Phi^{-1}(u_D)}{\sqrt{Var(Z\gamma)}}\right)}{1 - \Phi\left(\frac{\Phi^{-1}(u_D)}{\sqrt{Var(Z\gamma)}}\right)}\right)
\end{aligned}$$

so

$$\begin{aligned}\omega_{Z_1}(u_D) &= \frac{\left(\frac{Cov(Z_1,Z\gamma)}{\sqrt{Var(Z\gamma)}}\left(\frac{\phi\left(\frac{\Phi^{-1}(u_D)}{\sqrt{Var(Z\gamma)}}\right)}{1-\Phi\left(\frac{\Phi^{-1}(u_D)}{\sqrt{Var(Z\gamma)}}\right)}\right)\right)\left(1-\Phi_{Z\gamma}(\Phi^{-1}(u_D))\right)}{Cov(J,D)} \\ &= \frac{\frac{Cov(Z_1,Z\gamma)}{\sqrt{Var(Z\gamma)}}\phi\left(\frac{\Phi^{-1}(u_D)}{\sqrt{Var(Z\gamma)}}\right)}{Cov(J,D)}\end{aligned}$$

and the denominator can be estimated using its sample analogue. Notice that in this case the weights do not depend on X, so

$$\begin{aligned}\Delta^{IV}_{Z_1} &= \int\int_0^1 MTE(x,u_D)\,\omega_{Z_1}(u_D)du_D dF_X(x) \\ &= \int_0^1\int MTE(x,u_D)\,dF_X(x)\omega_{Z_1}(u_D)du_D \\ &= \int_0^1 \left[(\alpha_1-\alpha_0)+\varphi+\overline{x}'(\beta_1-\beta_0)+(\rho_1-\rho_0)\Phi^{-1}(u_D)\right]\omega_{Z_1}(u_D)du_D \\ &= (\alpha_1-\alpha_0)+\beta+\overline{x}'(\beta_1-\beta_0)+(\rho_1-\rho_0)\int_0^1\Phi^{-1}(u_D)\,\omega_{Z_1}(u_D)du_D\end{aligned}$$

Finally, the weights associated with the treatment on the treated are

$$\omega_{TT}(xu_D)=\frac{\Pr(P(Z)>u_D)}{\int\Pr(P(Z)>u_D)du_d}=\frac{1-\Phi_{Z\gamma}(\Phi^{-1}(u_D))}{\int 1-\Phi_{Z\gamma}(\Phi^{-1}(u_D))du_D}$$

and

$$\Delta^{TT}=(\alpha_1-\alpha_0)+\varphi+\overline{x}'(\beta_1-\beta_0)+(\rho_1-\rho_0)\int_0^1\Phi^{-1}(u_D)\,\omega_{TT}(xu_D)du_D$$

Analogous logic applies to the other treatment effects.

In what follows we consider a montecarlo experiment. We simulate 50 samples of size 5000 from the model described above, and we use our code to estimate its components. Then we compare these results with the actual values.

Table 1 presents the actual and estimated values for the probit model.

Table 2 presents the actual and estimated values for the different parameter in the each of the different approaches.

Table 3 presents the actual and estimated values for the treatment parameters.

Table 4 presents the results from IV in this case.

Figure 1 presents the estimated propensity score under the states ($D=1$ and $D=0$).

Figures 2 and 3 present the IV and treatment parameter weights, respectively.

Figure 4 presents the estimated marginal treatment effects obtained by using the different approaches.

References

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. New York: Chapman and Hall.

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd (1998, September). Characterizing Selection Bias Using Experimental Data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics*. forthcoming.

Heckman, J. J. and E. J. Vytlacil (2001a). Local Instrumental Variables. In C. Hsiao, K. Morimune, and J. L. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pp. 1–46. New York: Cambridge University Press.

Heckman, J. J. and E. J. Vytlacil (2001b, May). Policy-Relevant Treatment Effects. *American Economic Review 91*(2), 107–111.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Table 1. Coefficients in the Choice Model(Probit)

Coefficient	Actual Values	Estimated Values
γ_0	0.2	0.226
		(0.022)
γ_1	0.3	0.311
		(0.009)
γ_2	0.1	0.092
		(0.006)

Table 2. Coefficients in the Outcome Equations

Coefficient	Actual Values	Estimated Values			
		Parametric	*Polynomial*	*SemiParametric 1 (LIV)*	*SemiParametric 2*
α_0	0.02	0.019	0.014		
		(0.001)	(0.023)		
$\alpha_1 + \varphi$	0.24	0.239			
		(0.004)			
β_{10}	0.5	0.499	0.499	0.498	0.499
		(0.000)	(0.004)	(0.004)	(0.004)
β_{20}	0.1	0.099	0.101	0.1	0.101
		(0.000)	(0.004)	(0.004)	(0.004)
β_{11}	0.8	0.799			
		(0.000)			
β_{21}	0.4	0.4			
		(0.000)			
$\beta_{11} - \beta_{10}$	0.3		0.296	0.297	0.296
			(0.008)	(0.008)	(0.008)
$\beta_{21} - \beta_{20}$	0.3		0.299	0.3	0.299
			(0.009)	(0.009)	(0.009)
σ_1	0.012	0.012			
		(0.000)			
σ_0	-0.05	-0.049			
		(0.002)			
ϕ_1	-		0.641		
			(0.444)		
ϕ_2	-		-1.536		
			(1.956)		
ϕ_3	-		2.209		
			(2.957)		
ϕ_4	-		-1.119		
			(1.144)		

Table 3. Treatment Parameters

Treatment Parameter	Actual Values	Estimated Values			
		Parametric	*Polynomial*	*SemiParametric 1 (LIV)*	*SemiParametric 2*
Treatment on the Treated	0.254	0.258	0.279	0.261	0.261
		(0.025)	(0.043)	(0.034)	(0.034)
Treatment on the Untreated	0.185	0.189	0.128	0.158	0.158
		(0.018)	(0.042)	(0.033)	(0.033)
Average Treatment Effect	0.22	0.223	0.202	0.209	0.208
		(0.015)	(0.025)	(0.021)	(0.021)

Table 4. IV Estimates

Instrument	Actual Values	Estimated Values				
		TSLS	*Parametric*	*Polynomial*	*SemiParametric 1*	*SemiParametric 2*
Z_1	0.221	0.22	0.224	0.206	0.207	0.207
		(0.002)	(0.01)	(0.02)	(0.02)	(0.02)
Z_2	0.213	0.231	0.222	0.205	0.206	0.206
		(0.007)	(0.01)	(0.02)	(0.02)	(0.02)
$\Pr(D=1 \mid Z_1, Z_2)$	0.219	0.221	0.224	0.207	0.208	0.210
		(0.002)	(0.01)	(0.03)	(0.02)	(0.03)

Figure 1. Frequency of the Propensity Score by Treatment Status

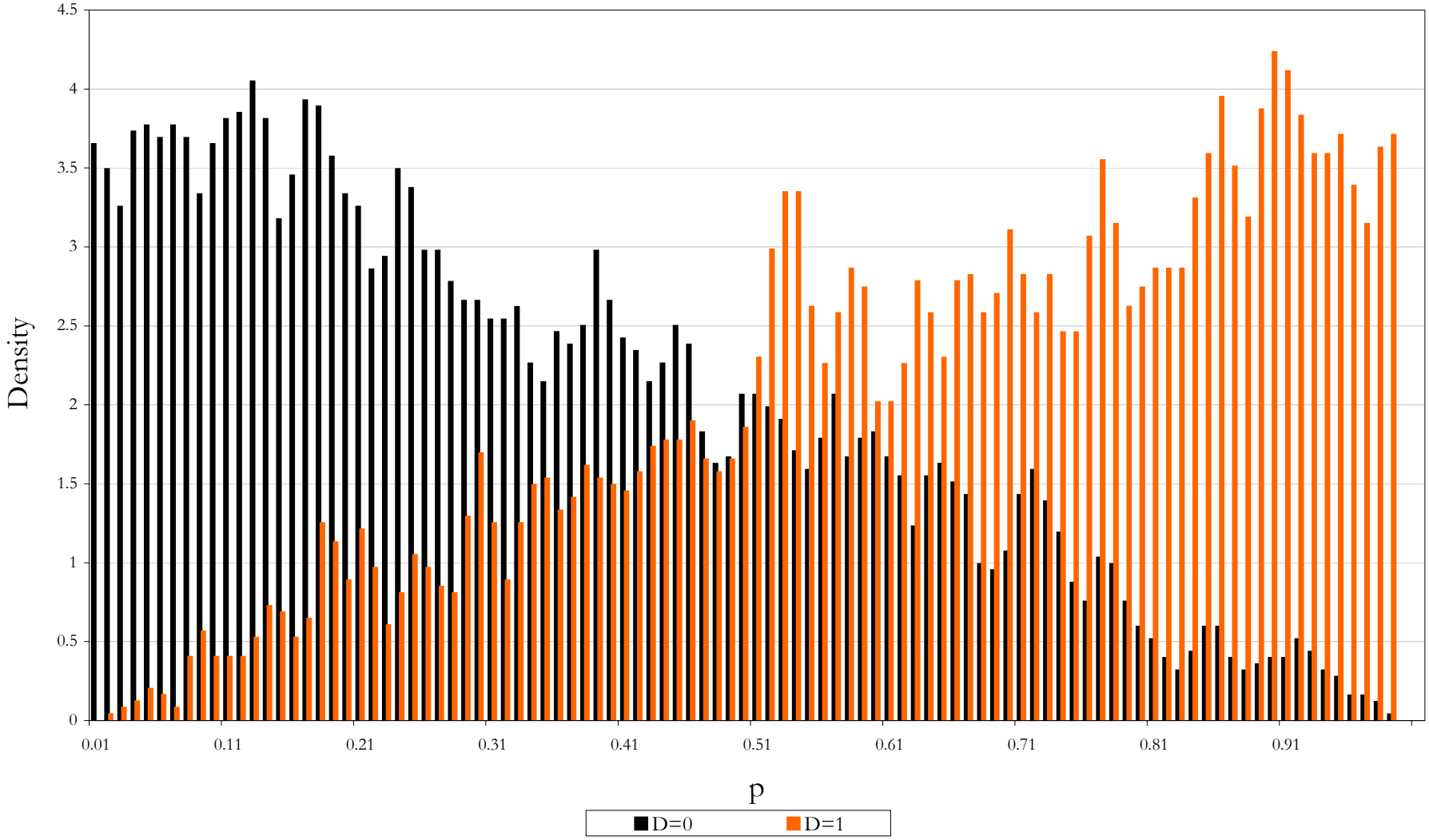
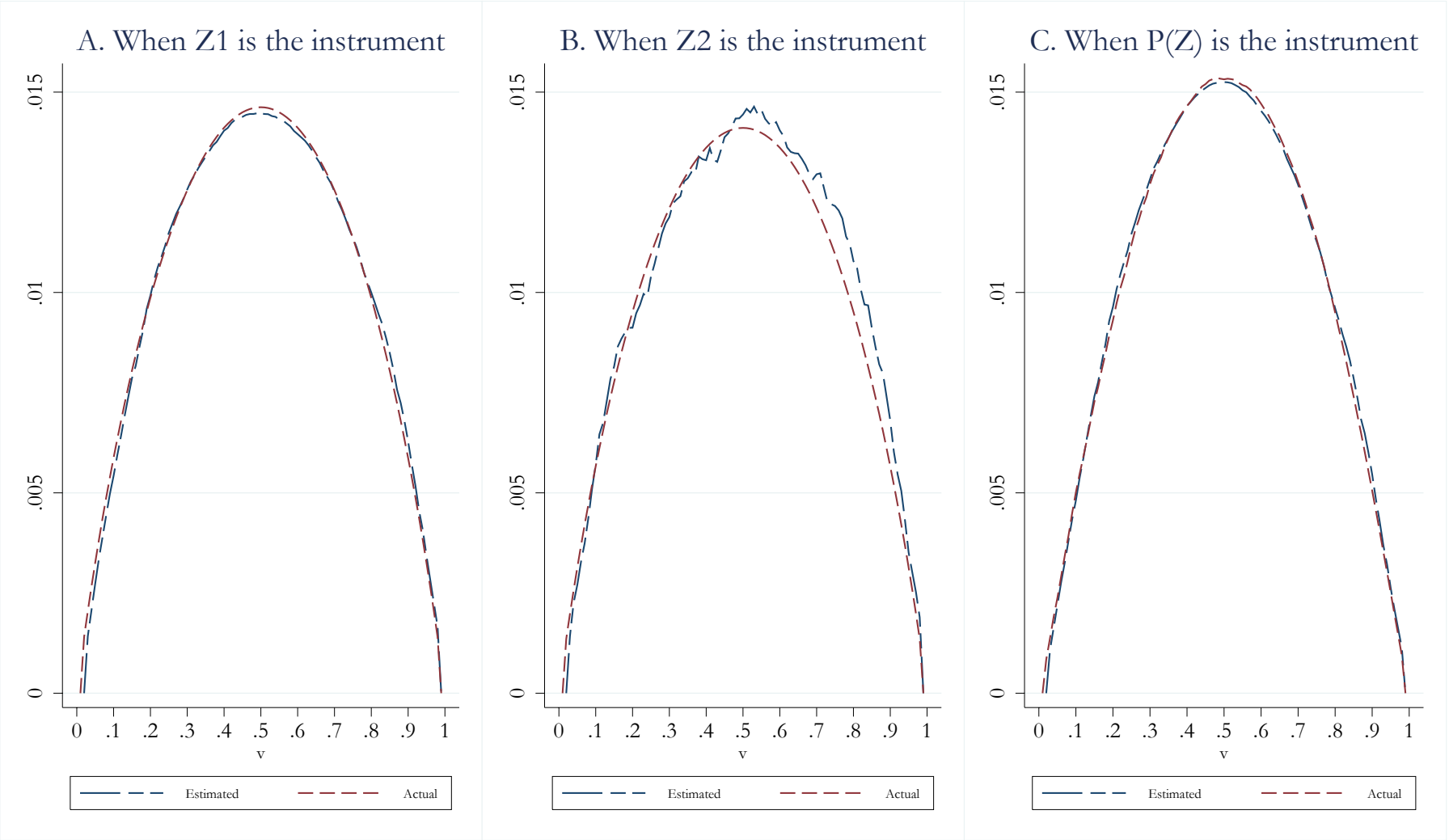
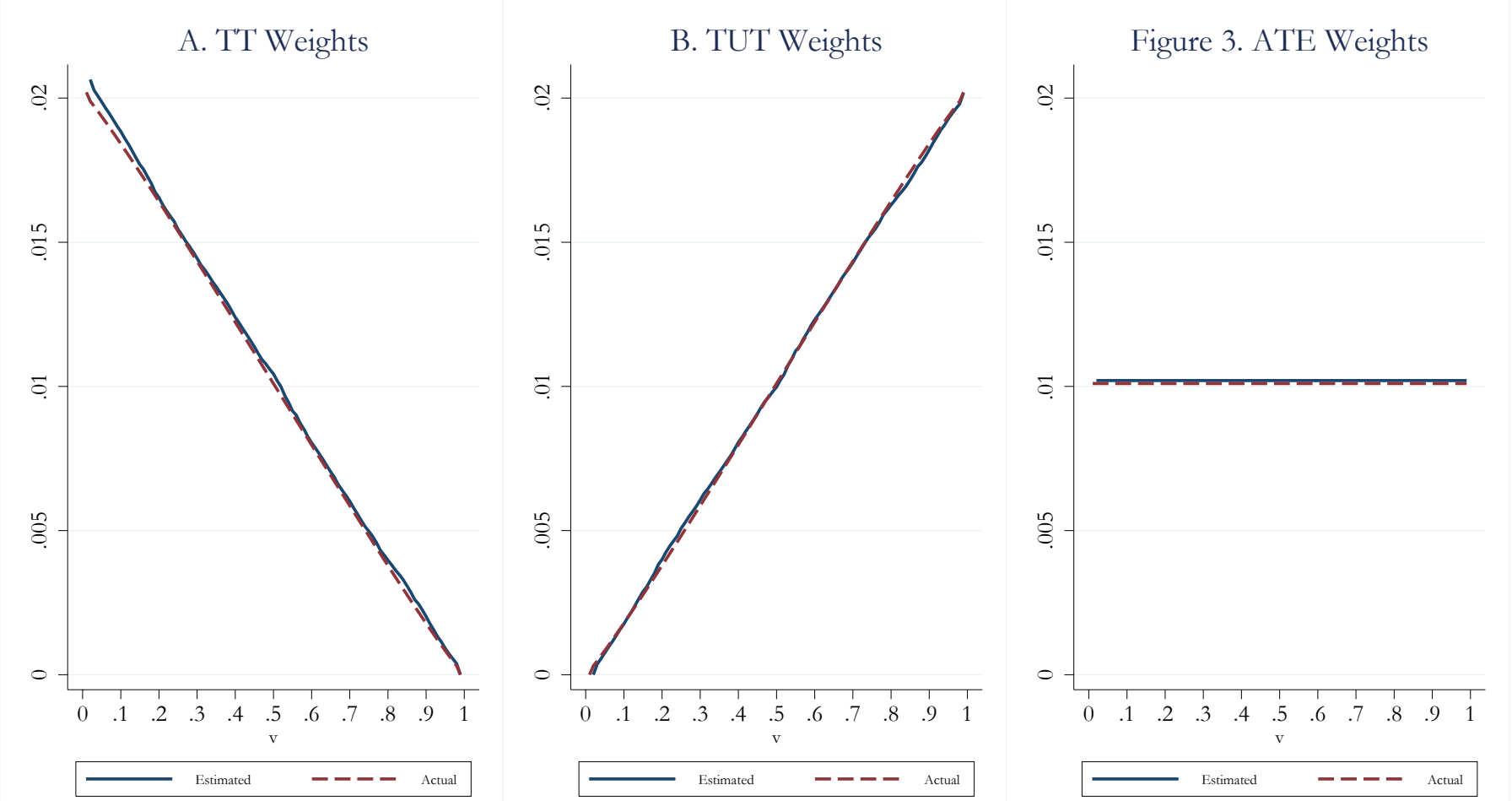


Figure 2. Comparison of Actual and Estimated IV Weights Generalized Roy Model



Notes: In each figure we present the actual and estimated IV weigths. The actual IV weigths are computed using the structure of the Generalized Roy model described in the text. The estimated weigths on the other hand, are computed using the simulated data generated from the model. The sample size considered in this example is 5000 observations. A detailed description of the estimation method is presented in the text.

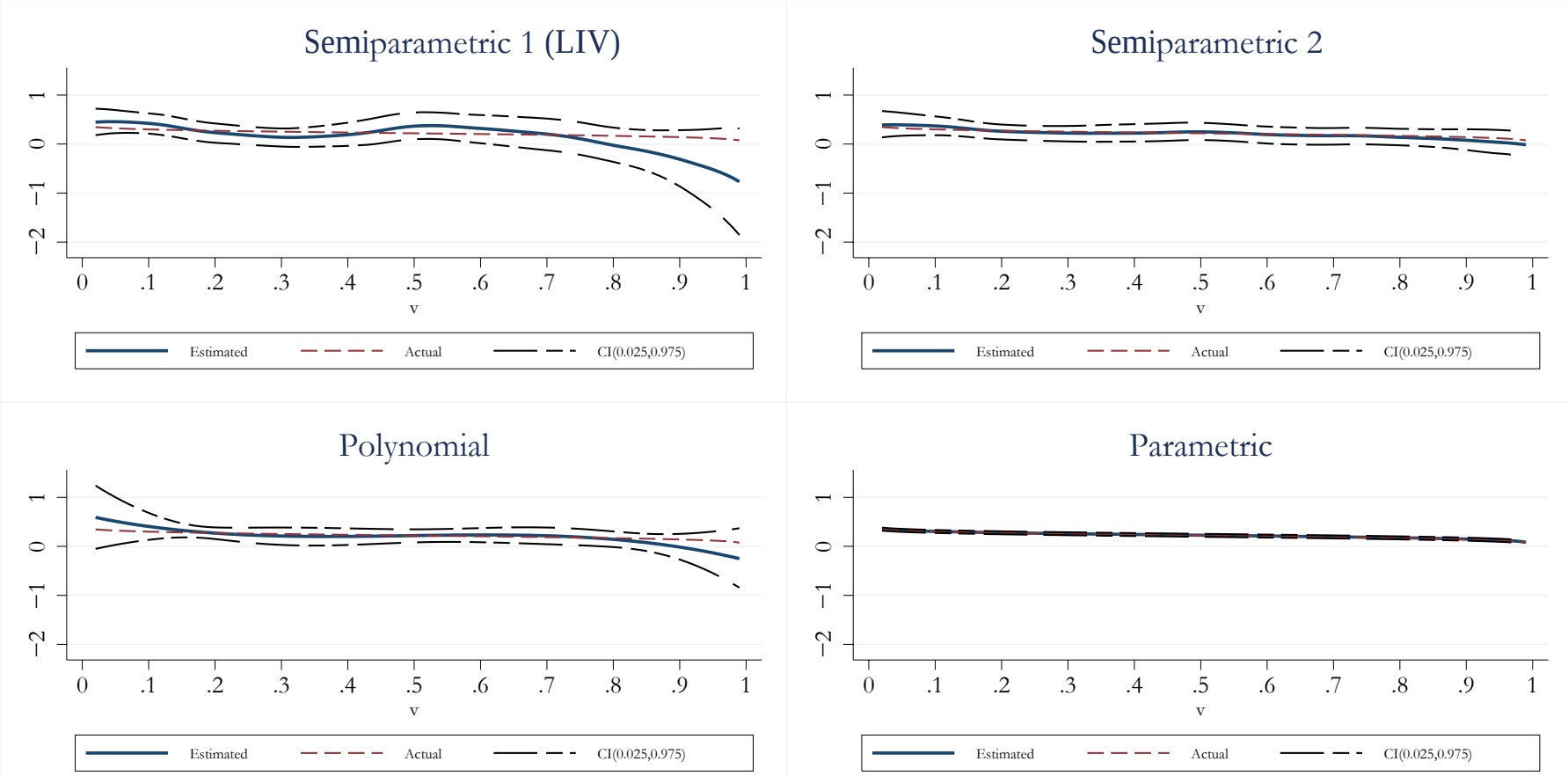
Figure 3. Comparison of actual and estimated Treatment Parameters Generalized Roy Model



Notes: In each figure we present the actual and estimated treatment parameter weights. Figure A depicts the comparison for the treatment on the treated (E(Y1−Y0|D=1)). Figure B presents the comparison for the treatment on the untreated (E(Y1−Y0|D=0)). Finally, Figure C presents the comparison for the average treatment effect (E(Y1−Y0)). In each case, the actual weigths are computed using the structure of the Generalized Roy model described in the text. The estimated weigths on the other hand, are computed using the simulated data generated from the model. The sample size considered in this example is 5000 observations. A detailed description of the estimation method is presented in the text.

Figure 4. Results using Different Approaches
Estimated Marginal Treatment Effects

5000 Observations



Notes: In each figure we present the actual MTE, its estimate, and the associated confidence interval. The confidence interval is computed using 50 bootstraps. The first figure (up−left) presents the LIV estimator (semiparametric 1). The second figure (up−right) presents the results from the second semiparametric approach presented in the text. The third figure (down−left) presents the results after approximating E(Y|P) by a fourth order polynomial. The last figure (down−right) presents the results after assuming a parametric representation of the MTE. This representation is obtained by assuming that the error terms in the model are normally distributed with means 0 and variances 1. A detailed description of the four procedures is presented in the text.