11

Marginal treatment effects

In this chapter, we review policy evaluation and Heckman and Vytlacil's [2005, 2007a] (*HV*) strategy for linking marginal treatment effects to other average treatment effects including policy-relevant treatment effects. Recent innovations in the treatment effects literature including dynamic and general equilibrium considerations are mentioned briefly but in-depth study of these matters is not pursued. *HV*'s marginal treatment effects strategy is applied to the regulated report precision setting introduced in chapter 2, discussed in chapter 10, and continued in the next chapter. This analysis highlights the relative importance of probability distribution assignment to unobservables and quality of instruments.

11.1 Policy evaluation and policy invariance conditions

Heckman and Vytlacil [2007a] discuss causal effects and policy evaluation. Following the lead of Bjorklund and Moffitt [1987], *HV* base their analysis on marginal treatment effects. *HV*'s marginal treatment effects strategy combines the strengths of the treatment effect approach (simplicity and lesser demands on the data) and the Cowles Commission's structural approach (utilize theory to help extrapolate results to a broader range of settings). *HV* identify three broad classes of policy evaluation questions.

(P-1) Evaluate the impact of historically experienced and documented policies on outcomes via counterfactuals. Outcome or welfare evaluations may be objective (inherently ex post) or subjective (may be ex ante or ex post). P-1 is an *inter-*

nal validity problem (Campbell and Stanley [1963]) — the problem of identifying treatment parameter(s) in a given environment.

(P-2) Forecasting the impact of policies implemented in one environment by extrapolating to other environments via counterfactuals. This is the *external validity* problem (Campbell and Stanley [1963]).

(P-3) Forecasting the impact of policies never historically experienced to various environments via counterfactuals. This is the most ambitious policy evaluation problem.

The study of policy evaluation frequently draws on some form of policy invariance. Policy invariance allows us to characterize outcomes without fully specifying the structural model including incentives, assignment mechanisms, and choice rules. The following policy invariance conditions support this relaxation.[1]

(PI-1) For a given choice of treatment, outcomes are invariant to variations in incentive schedules or assignment mechanisms. PI-1 is a strong condition. It says that randomized assignment or threatening with a gun to gain cooperation has no impact on outcomes for a given treatment choice (see Heckman and Vytlacil [2007b] for evidence counter to the condition).

(PI-2) The actual mechanism used to assign treatment does not impact outcomes. This rules out general equilibrium effects (see Abbring and Heckman [2007]).

(PI-3) Utilities are unaffected by variations in incentive schedules or assignment mechanisms. This is the analog to (PI-1) but for utilities or subjective evaluations in place of outcomes. Again, this is a strong condition (see Heckman and Vytlacil [2007b] for evidence counter to the condition).

(PI-4) The actual mechanism used to assign treatment does not impact utilities. This is the analog to (PI-2) but for utilities or subjective evaluations in place of outcomes. Again, this rules out general equilibrium effects.

It's possible to satisfy (PI-1) and (PI-2) but not (PI-3) and (PI-4) (see Heckman and Vytlacil [2007b]). Next, we discuss marginal treatment effects and begin the exploration of how they unify policy evaluation.

Briefly, Heckman and Vytlacil's [2005] local instrumental variable (*LIV*) estimator is a more ambitious endeavor than the methods discussed in previous chapters. *LIV* estimates the marginal treatment effect (*MTE*) under standard *IV* conditions. *MTE* is the treatment effect associated with individuals who are indifferent between treatment and no treatment. Heckman and Vytlacil identify weighted distributions (Rao [1986] and Yitzhaki [1996]) that connect *MTE* to a variety of other treatment effects including *ATE*, *ATT*, *ATUT*, *LATE*, and policy-relevant treatment effects (*PRTE*).

MTE is a generalization of *LATE* as it represents the treatment effect for those individuals who are indifferent between treatment and no treatment.

$$MTE = E\left[Y_1 - Y_0 \mid X = x, V_D = v_D\right]$$

[1]Formal statements regarding policy invariance are provided in Heckman and Vytlacil [2007a].

Or, the marginal treatment effect can alternatively be defined by a transformation of unobservable V by $U_D = F_{V|X}(V)$ so that we can work with $U_D \sim Unif[0,1]$

$$MTE = E[Y_1 - Y_0 \mid X = x, U_D = u_D]$$

11.2 Setup

The setup is the same as the previous chapters. We repeat it for convenience. Suppose the *DGP* is
outcome equations:

$$Y_j = \mu_j(X) + V_j, j = 0, 1$$

selection equation:

$$D^* = \mu_D(Z) - V_D$$

observable response:

$$\begin{aligned} Y &= DY_1 + (1 - D)Y_0 \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X))D + V_0 + (V_1 - V_0)D \end{aligned}$$

where

$$D = \begin{matrix} 1 & D^* > 0 \\ 0 & otherwise \end{matrix}$$

and Y_1 is (potential) outcome with treatment while Y_0 is the outcome without treatment. The outcomes model is the Neyman-Fisher-Cox-Rubin model of potential outcomes (Neyman [1923], Fisher [1966], Cox]1958], and Rubin [1974]). It is also Quandt's [1972] switching regression model or Roy's income distribution model (Roy [1951] or Heckman and Honore [1990]).

The usual exclusion restriction and uniformity applies. That is, if instrument changes from z to z' then everyone either moves toward or away from treatment. Again, the treatment effects literature is asymmetric; heterogeneous outcomes are permitted but homogeneous treatment is required for identification of estimators. Next, we repeat the generalized Roy model — a useful frame for interpreting causal effects.

11.3 Generalized Roy model

Roy [1951] introduced an equilibrium model for work selection (hunting or fishing).[2] An individual's selection into hunting or fishing depends on his/her aptitude

[2]The *basic* Roy model involves no cost of treatment. The *extended* Roy model includes only observed cost of treatment. While the *generalized* Roy model includes both observed and unobserved cost of treatment (see Heckman and Vytlacil [2007a, 2007b]).

as well as supply of and demand for product of labor. A modest generalization of the Roy model is a common framing of self-selection that forms the basis for assessing treatment effects (Heckman and Robb [1986]).

Based on the *DGP* above, we identify the constituent pieces of the selection model.
Net benefit (or utility) from treatment is

$$\begin{aligned} D^* &= \mu_D(Z) - V_D \\ &= Y_1 - Y_0 - c(W) - V_c \\ &= \mu_1(X) - \mu_0(X) - c(W) + V_1 - V_0 - V_C \end{aligned}$$

Gross benefit of treatment is

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

Cost associated with treatment is[3]

$$c(W) + V_C$$

Observable cost associated with treatment is

$$c(W)$$

Observable net benefit of treatment is

$$\mu_1(X) - \mu_0(X) - c(W)$$

Unobservable net benefit of treatment is

$$-V_D = V_1 - V_0 - V_C$$

where the observables are $\begin{bmatrix} X & Z & W \end{bmatrix}$, typically Z contains variables not in X or W and W is the subset of observables that speak to cost of treatment.

11.4 Identification

Marginal treatment effects are defined conditional on the regressors X and unobserved utility V_D

$$MTE = E[Y_1 - Y_0 \mid X = x, V_D = v_D]$$

or transformed unobserved utility U_D.

$$MTE = E[Y_1 - Y_0 \mid X = x, U_D = u_D]$$

HV describe the following identifying conditions.

[3]The model is called the *original* or *basic* Roy model if the cost term is omitted. If the cost is constant ($V_C = 0$ so that cost is the same for everyone) it is called the *extended* Roy model.

Condition 11.1 *$\{U_0, U_1, V_D\}$ are independent of Z conditional on X (conditional independence),*

Condition 11.2 *$\mu_D(Z)$ is a nondegenerate random variable conditional on X (rank condition),*

Condition 11.3 *the distribution of V_D is continuous,*

Condition 11.4 *the values of $E[|Y_0|]$ and $E[|Y_1|]$ are finite (finite means),*

Condition 11.5 $0 < Pr(D = 1 \mid X) < 1$ *(common support).*

These are the base conditions for *MTE*. They are augmented below to facilitate interpretation.[4] Condition 11.7 applies specifically to policy-relevant treatment effects where p and p' refer to alternative policies.

Condition 11.6 *Let X_0 denote the counterfactual value of X that would be observed if D is set to 0. X_1 is defined analogously. Assume $X_d = X$ for $d = 0, 1$. (The X_D are invariant to counterfactual manipulations.)*

Condition 11.7 *The distribution of $(Y_{0,p}, Y_{1,p}, V_{D,p})$ conditional on $X_p = x$ is the same as the distribution of $(Y_{0,p'}, Y_{1,p'}, V_{D,p'})$ conditional on $X_{p\prime} = x$ (policy invariance of the distribution).*

Under the above conditions, *MTE* can be estimated by local *IV* (*LIV*)

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

where $P(Z) \equiv \Pr(D \mid Z)$. To see the connection between *MTE* and *LIV* rewrite the numerator of *LIV*

$$E[Y \mid X = x, P(Z) = p] = E[Y_0 + (Y_1 - Y_0) D \mid X = x, P(Z) = p]$$

by conditional independence and Bayes' theorem we have

$$E[Y_0 \mid X = x] + E[Y_1 - Y_0 \mid X = x, D = 1] \Pr(D = 1 \mid Z = z)$$

transforming V_D such that U_D is distributed uniform$[0, 1]$ produces

$$E[Y_0 \mid X = x] + \int_0^p E[Y_1 - Y_0 \mid X = x, U_D = u_D] \, du_D$$

Now, the partial derivative of this expression with respect to p evaluated at $p = u_D$ is

$$\left. \frac{\partial E[Y \mid X=x, P(Z)=p]}{\partial p} \right|_{p=u_D} = E[Y_1 - Y_0 \mid X = x, U_D = u_D]$$

[4]The conditions remain largely the same for *MTE* analysis of alternative settings including multilevel discrete treatment, continuous treatment, and discrete outcomes. Modifications are noted in the discussions of each.

Hence, LIV identifies MTE.

With homogeneous response, *MTE* is constant and equal to *ATE*, *ATT*, and *ATUT*. With unobservable heterogeneity, *MTE* is typically a nonlinear function of u_D (where u_D continues to be distributed uniform$[0,1]$). The intuition for this is individuals who are less likely to accept treatment require a larger potential gain from treatment to induce treatment selection than individuals who are more likely to participate.

11.5 *MTE* connections to other treatment effects

Heckman and Vytlacil show that *MTE* can be connected to other treatment effects (*TE*) by weighted distributions $h_{TE}(\cdot)$ (Rao [1986] and Yitzhaki [1996]).[5] Broadly speaking and with full support

$$TE(x) = \int_0^1 MTE(x, u_D)\, h_{TE}(x, u_D)\, du_D$$

and integrating out x yields the population moment

$$Average(TE) = \int_0^1 TE(x)\, dF(x)$$

If full support exists, then the weight distribution for the average treatment effect is

$$h_{ATE}(x, u_D) = 1$$

Let f be the density function of observed utility $\tilde{W} = \mu_D(Z)$, then the weighted distribution to recover the treatment effect on the treated from *MTE* is

$$\begin{aligned} h_{TT}(x, u_D) &= \left[\int_{u_D}^1 f(p \mid X = x)\, dp\right] \frac{1}{E[p \mid X = x]} \\ &= \frac{\Pr\left(P\left(\tilde{W}\right) > u_D \mid X = x\right)}{\int_0^1 \Pr\left(P\left(\tilde{W}\right) > u_D \mid X = x\right) du_d} \end{aligned}$$

where $P\left(\tilde{W}\right) \equiv \Pr\left(D = 1 \mid \tilde{W} = w\right)$. Similarly, the weighted distribution to recover the treatment effect on the untreated from *MTE* is

$$\begin{aligned} h_{TUT}(x, u_D) &= \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E[1 - p \mid X = x]} \\ &= \frac{\Pr\left(P\left(\tilde{W}\right) \leq u_D \mid X = x\right)}{\int_0^1 \Pr\left(P\left(\tilde{W}\right) \leq u_D \mid X = x\right) du_d} \end{aligned}$$

[5] Weight functions are nonnegative and integrate to one (like density functions).

Figure 11.1 depicts *MTE* ($\Delta_{MTE}(u_D)$) and weighted distributions for treatment on treated $h_{TT}(u_D)$ and treatment on the untreated $h_{TUT}(u_D)$ with regressors suppressed.

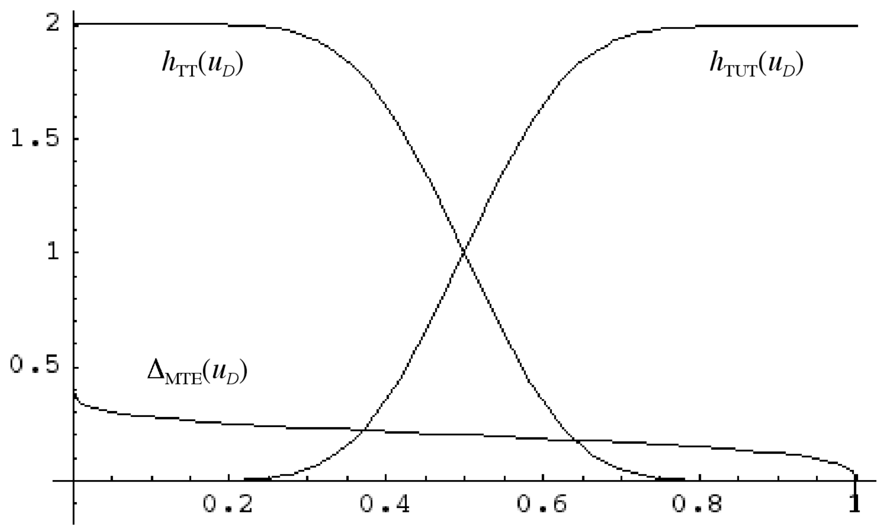

Figure 11.1: MTE and weight functions for other treatment effects

Applied work determines the weights by estimating

$$\Pr\left(P\left(\tilde{W}\right) > u_D \mid X = x\right)$$

Since $\Pr\left(P\left(\tilde{W}\right) > u_D \mid X = x\right) = \Pr\left(I\left[P\left(\tilde{W}\right) > u_D\right] = 1 \mid X = x\right)$ where $I[\cdot]$ is an indicator function, we can use our selection or choice model (say, probit) to estimate

$$\Pr\left(I\left[P\left(\tilde{W}\right) > u_D\right] = 1 \mid X = x\right)$$

for each value of u_D. As the weighted distributions integrate to one, we use their sum to determine the normalizing constant (the denominator). The analogous idea applies to $h_{TUT}(x, u_D)$.

However, it is rare that full support is satisfied as this implies both treated and untreated samples would be evidenced at all probability levels for some model of treatment (e.g., probit). Often, limited support means the best we can do is estimate a local average treatment effect.

$$LATE(x) = \frac{1}{u' - u} \int_u^{u'} MTE(x, u_D)\, du_D$$

In the limit as the interval becomes arbitrarily small *LATE* converges to *MTE*.

11.5.1 Policy-relevant treatment effects vs. policy effects

What is the average gross gain from treatment following policy intervention? This is a common question posed in the study of accounting. Given uniformity (one way flows into or away from participation in response to a change in instrument) and policy invariance, *IV* can identify the average treatment effect for policy a compared with policy a', that is, a policy-relevant treatment effect (*PRTE*). Policy invariance means the policy impacts the likelihood of treatment but not the potential outcomes (that is, the distributions of $\{y_{1a}, y_{0a}, V_{Da} \mid X_a = x\}$ and $\{y_{1a'}, y_{0a'}, V_{Da'} \mid X_{a'} = x\}$ are equal).

The policy-relevant treatment effect is

$$\begin{aligned} PRTE &= E[Y \mid X = x, a] - E[Y \mid X = x, a'] \\ &= \int_0^1 MTE(x, u_D)\left[F_{P(a')|X}(u_D \mid x) - F_{P(a)|X}(u_D \mid x)\right] du_D \end{aligned}$$

where $F_{P(a)|X}(u_D \mid x)$ is the distribution of P, the probability of treatment conditional on $X = x$, and the weight function is $h_{PRTE}(x, u_D)$.[6]

$$h_{PRTE}(x, u_D) = \left[F_{P(a')|X}(u_D \mid x) - F_{P(a)|X}(u_D \mid x)\right]$$

Intuition for the above connection can be seen as follows, where conditioning on X is implicit.

$$\begin{aligned} E[Y \mid a] &= \int_0^1 E[Y \mid P(Z) = p]\, dF_{P(a)}(p) \\ &= \int_0^1 \left(\begin{array}{c} \int_0^1 \Im_{[0,p]}(u_D)\, E(Y_1 \mid U = u_D) \\ +\Im_{(p,1]}(u_D)\, E(Y_0 \mid U = u)\, du_D \end{array} \right) dF_{P(a)}(p) \\ &= \int_0^1 \left(\begin{array}{c} \left[1 - F_{P(a)}(u_D)\right] E[Y_1 \mid U = u_D] \\ +F_{P(a)}(u_D)\, E[Y_0 \mid U = u_D] \end{array} \right) du_D \end{aligned}$$

[6]Heckman and Vytlacil [2005] also identify the per capita weight for policy-relevant treatment as

$$\frac{\Pr\left(P\left(\tilde{W}\right) \leq u_D \mid X = x, a'\right) - \Pr\left(P\left(\tilde{W}\right) \leq u_D \mid X = x, a\right)}{\int_0^1 \Pr\left(P\left(\tilde{W}\right) \leq u_D \mid X = x, a'\right) du_d - \int_0^1 \Pr\left(P\left(\tilde{W}\right) \leq u_D \mid X = x, a\right) du_d}$$

where $\Im_A(u_D)$ is an indicator function for the event $u_D \in A$. Hence, comparing policy a to a', we have

$$\begin{aligned}
& E[Y \mid X = x, a] - E[Y \mid X = x, a'] \\
= & \int_0^1 \left(\begin{array}{c} \left[1 - F_{P(a)}(u_D)\right] E[Y_1 \mid U = u_D] \\ +F_{P(a)}(u_D) E[Y_0 \mid U = u_D] \end{array} \right) du_D \\
& - \int_0^1 \left(\begin{array}{c} \left[1 - F_{P(a')}(u_D)\right] E[Y_1 \mid U = u_D] \\ +F_{P(a')}(u_D) E[Y_0 \mid U = u_D] \end{array} \right) du_D \\
= & \int_0^1 \left[F_{P(a')}(u_D) - F_{P(a)}(u_D)\right] E[Y_1 - Y_0 \mid U = u_D] \, du_D \\
= & \int_0^1 \left[F_{P(a')}(u_D) - F_{P(a)}(u_D)\right] MTE(U = u_D) \, du_D
\end{aligned}$$

On the other hand, we might be interested in the policy effect or net effect of a policy change rather than the treatment effect. In which case it is perfectly sensible to estimate the net impact with some individuals leaving and some entering, this is a policy effect not a treatment effect. The policy effect parameter is $E[Y \mid Z_{a'} = z'] - E[Y \mid Z_a = z]$

$$\begin{aligned}
= \; & E[Y_1 - Y_0 \mid D(z') > D(z)] \Pr(D(z') > D(z)) \\
& - E[Y_1 - Y_0 \mid D(z') \leq D(z)] \Pr(D(z') \leq D(z))
\end{aligned}$$

Notice the net impact may be positive, negative, or zero as two way flows are allowed (see Heckman and Vytlacil [2006]).

11.5.2 Linear IV weights

As mentioned earlier, *HV* argue that linear *IV* produces a complex weighting of effects that can be difficult to interpret and depends on the instruments chosen. This argument is summarized by their linear *IV* weight distribution. Let $J(Z)$ be any function of Z such that $Cov[J(Z), D] \neq 0$. The population analog of the *IV* estimator is $\frac{Cov[J(Z),Y]}{Cov[J(Z),D]}$. Consider the numerator.

$$\begin{aligned}
Cov[J(Z), Y] &= E[(J(Z) - E[J(Z)]) Y] \\
&= E[(J(Z) - E[J(Z)]) (Y_0 + D(Y_1 - Y_0))] \\
&= E[(J(Z) - E[J(Z)]) D(Y_1 - Y_0)]
\end{aligned}$$

Define $\tilde{J}(Z) = J(Z) - E[J(Z)]$. Then, $Cov[J(Z), Y]$

$$\begin{aligned}
&= E\left[\tilde{J}(Z) D (Y_1 - Y_0)\right] \\
&= E\left[\tilde{J}(Z) I[U_D \leq P(Z)] (Y_1 - Y_0)\right] \\
&= E\left[\tilde{J}(Z) I[U_D \leq P(Z)] E[(Y_1 - Y_0) \mid Z, V_D]\right] \\
&= E\left[\tilde{J}(Z) I[U_D \leq P(Z)] E[(Y_1 - Y_0) \mid V_D]\right] \\
&= E_{V_D}\left[E_Z\left[\tilde{J}(Z) I[U_D \leq P(Z)] \mid U_D\right] E[(Y_1 - Y_0) \mid U_D]\right] \\
&= \int_0^1 E\left[\tilde{J}(Z) \mid P(Z) \geq u_D\right] \Pr(P(Z) \geq u_D) \\
&\quad \times E[(Y_1 - Y_0) \mid U_D = u_D] \, du_D \\
&= \int_0^1 \Delta_{MTE}(x, u_D) E\left[\tilde{J}(Z) \mid P(Z) \geq u_D\right] \Pr(P(Z) \geq u_D) \, du_D
\end{aligned}$$

where $P(Z)$ is propensity score utilized as an instrument.

For the denominator we have, by iterated expectations,

$$Cov[J(Z), D] = Cov[J(Z), P(Z)]$$

Hence,

$$h_{IV}(x, u_D) = \frac{E\left[\tilde{J}(Z) \mid P(Z) \geq u_D\right] \Pr(P(Z) \geq u_D)}{Cov[J(Z), P(Z)]}$$

where $Cov[J(Z), P(Z)] \neq 0$. Heckman, Urzua, and Vytlacil [2006] illustrate the sensitivity of treatment effects identified via linear *IV* to choice of instruments.

11.5.3 OLS weights

It's instructive to identify the effect exogenous dummy variable *OLS* estimates as a function of *MTE*. While not a true weighted distribution (as the weights can be negative and don't necessarily sum to one), for consistency we'll write $h_{OLS}(x, u_D) =$

$$\begin{matrix}
1 + \frac{E[Y_1|x,u_D] h_{ATT}(x,u_D) - E[v_0|x,u_D] h_{ATUT}(x,u_D)}{MTE(x,u_D)} & MTE(x, u_D) \neq 0 \\
0 & otherwise
\end{matrix}$$

Table 11.1: Comparison of identification conditions for common econometric strategies (adapted from Heckman and Navarro-Lozano's [2004] table 3)

Method	Exclusion required?	Separability of observables and unobservables in outcome equations?
Matching	no	no
Control function	yes, for nonparametric identification	conventional, but not required
IV (linear)	yes	yes
LIV	yes	no

Method	Functional form required?	Marginal = Average (given X, Z)?
Matching	no	yes
Control function	conventional, but not required	no
IV (linear)	no	no (yes, in standard case)
LIV	no	no

Method	Key identification conditions for means (assuming separability)
Matching	$E[U_1 \mid X, D=1, Z] = E[U_1 \mid X, Z]$ $E[U_0 \mid X, D=1, Z] = E[U_0 \mid X, Z]$
Control function	$E[U_0 \mid X, D=1, Z]$ and $E[U_1 \mid X, D=1, Z]$ can be varied independently of $\mu_0(X)$ and $\mu_1(X)$, respectively, and intercepts can be identified through limit arguments (identification at infinity), or symmetry assumptions
IV (linear)	$E[U_0 + D(U_1 - U_0) \mid X, Z] = E[U_0 + D(U_1 - U_0) \mid X]$ (*ATE*) $E[U_0 + D(U_1 - U_0) - E[U_0 + D(U_1 - U_0) \mid X] \mid P(W), X]$ $= E[U_0 + D(U_1 - U_0) - E[U_0 + D(U_1 - U_0) \mid X] \mid X]$ (*ATT*)
LIV	(U_0, U_1, U_D) independent of $Z \mid X$

Method	Key identification conditions for propensity score
Matching	$0 < \Pr(D=1 \mid Z, X) < 1$
Control function	$0 \leq \Pr(D=1 \mid Z, X) \leq 1$ is a nontrivial function of Z for each X
IV (linear)	not needed
LIV	$0 < \Pr(D=1 \mid X) < 1$ $0 \leq \Pr(D=1 \mid Z, X) \leq 1$ is a nontrivial function of Z for each X

11.6 Comparison of identification strategies

Following Heckman and Navarro-Lozano [2004], we compare and report in table 11.1 treatment effect identification strategies for four common econometric approaches: matching (especially, propensity score matching), control functions (selection models), conventional (linear) instrumental variables (*IV*), and local instrumental variables (*LIV*).

All methods define treatment parameters on common support — the intersection of the supports of X given $D = 1$ and X given $D = 0$, that is,

$$Support\left(X \mid D = 1\right) \cap Support\left(X \mid D = 0\right)$$

LIV employs common support of the propensity score — overlaps in $P\left(X, Z\right)$ for $D = 0$ and $D = 1$. Matching breaks down if there exists an explanatory variable that serves as a perfect classifier. On the other hand, control functions exploit limit arguments for identification,[7] hence, avoiding the perfect classifier problem. That is, identification is secured when $P\left(X, Z\right) = 1$ for some $Z = z$ but there exists $P\left(X, Z\right) < 1$ for some $Z = z'$. Similarly, when $P\left(W\right) = 0$, where $W = \left(X, Z\right)$, for some $Z = z$ there exists $P\left(X, Z\right) > 0$ for some $Z = z''$.

11.7 *LIV* estimation

We've laid the groundwork for the potential of marginal treatment effects to address various treatment effects in the face of unobserved heterogeneity, it's time to discuss estimation. Earlier, we claimed *LIV* can estimate *MTE*

$$\frac{\partial E[Y \mid X=x, P(Z)=p]}{\partial p} \Big|_{p=u_D} = E\left[Y_1 - Y_0 \mid X = x, U_D = u_D\right]$$

For the linear separable model we have

$$Y_1 = \delta + \alpha + X\beta_1 + V_1$$

and

$$Y_0 = \delta + X\beta_0 + V_0$$

Then,

$$E\left[Y \mid X = x, P\left(Z\right) = p\right] = X\beta_0 + X\left(\beta_1 - \beta_0\right)\Pr\left(Z\right) + \kappa\left(p\right)$$

where

$$\kappa\left(p\right) = \alpha \Pr\left(Z\right) + E\left[v_0 \mid \Pr\left(Z\right) = p\right] + E\left[v_1 - v_0 \mid D = 1, \Pr\left(Z\right) = p\right]\Pr\left(Z\right)$$

Now, *LIV* simplifies to

$$LIV = X\left(\beta_1 - \beta_0\right) + \frac{\partial \kappa(p)}{\partial p} \Big|_{p=u_D}$$

[7] This is often called "identification at infinity."

Since *MTE* is based on the partial derivative of expected outcome with respect to p

$$\frac{\partial}{\partial p}E\left[Y \mid X=x, P\left(Z\right)=p\right]=X\left(\beta_1-\beta_0\right)+\frac{\partial\kappa\left(p\right)}{\partial p},$$

the objective is to estimate $(\beta_1 - \beta_0)$ and the derivative of $\kappa(p)$. Heckman, Urzua, and Vytlacil's [2006] local *IV* estimation strategy employs a relaxed distributional assignment based on the data and accommodates unobservable heterogeneity. *LIV* employs nonparametric (local linear kernel density; see chapter 6) regression methods.

LIV Estimation proceeds as follows:

Step 1: Estimate the propensity score, $P(Z)$, via probit, nonparametric discrete choice, etc.

Step 2: Estimate β_0 and $(\beta_1 - \beta_0)$ by employing a nonparametric version of *FWL* (double residual regression). This involves a local linear regression (*LLR*) of each regressor in X and $X * P(Z)$ onto $P(Z)$. *LLR* for X_k (the kth regressor) is $\{\tau_{0k}(p), \tau_{1k}(p)\} =$

$$\underset{\{\tau_0(p),\tau_1(p)\}}{\arg\min}\left\{\sum_{j=1}^{n}\left(X_k\left(j\right)-\tau_0-\tau_1\left(P\left(Z_j\right)-p\right)\right)^2 K\left(\frac{P\left(Z_j\right)-p}{h}\right)\right\}$$

where $K(W)$ is a (Gaussian, biweight, or Epanechnikov) kernel evaluated at W. The bandwidth h is estimated by leave-one out generalized cross-validation based on the nonparametric regression of $X_k(j)$ onto $(\tau_{0k} + \tau_{1k}p)$.

For each regressor in X and $X * P(Z)$ and for the response variable y estimate the residuals from *LLR*. Denote the matrix of residuals from the regressors (ordered with X followed by $X * P(Z)$) as e_X and the residuals from Y, e_Y.

Step 3: Estimate $[\beta_0, \beta_1 - \beta_0]$ from a no-intercept linear regression of e_Y onto e_X. That is, $\left[\widehat{\beta_0}, \widehat{\beta_1-\beta_0}\right]^T = \left[e_X^T e_X\right]^{-1} e_X^T e_Y$.

Step 4: For $E[Y \mid X = x, P(Z) = p]$, we've effectively estimated $\beta_0 X_i + (\beta_1 - \beta_0) X_i * P(Z_i)$. What remains is to estimate the derivative of $\kappa(p)$. We complete nonparametric *FWL* by defining the restricted response as follows.

$$\tilde{Y}_i = Y_i - \widehat{\beta_0}X_i - \left(\widehat{\beta_1-\beta_0}\right)X_i * P\left(Z_i\right)$$

The intuition for utilizing the restricted response is as follows. In the textbook linear model case

$$Y = X\beta + Z\gamma + \varepsilon$$

FWL produces

$$E\left[Y \mid X, Z\right] = P_Z Y + \left(I - P_Z\right) X b$$

where b is the *OLS* estimator for β and P_Z is the projection matrix $Z(Z^T Z)^{-1} Z^T$. Rewriting we can identify the estimator for γ, g, from

$$E\left[Y \mid X, Z\right] = Xb + P_Z\left(Y - Xb\right) = Xb + Zg$$

Hence, $g = \left(Z^T Z\right)^{-1} Z^T (Y - Xb)$. That is, g is estimated from a regression of the restricted response $(Y - Xb)$ onto the regressor Z. LIV employs the non-parametric analog.

Step 5: Estimate $\tau_1(p) = \frac{\partial \kappa(p)}{\partial p}$ by *LLR* of $Y_i - \widehat{\beta_0} X_i - \left(\widehat{\beta_1 - \beta_0}\right) X_i * P(Z_i)$ onto $P(Z_i)$ for each observation i in the set of overlaps. The set of overlaps is the region for which *MTE* is identified — the subset of common support of $P(Z)$ for $D = 1$ and $D = 0$.

Step 6: The *LIV* estimator of *MTE*(x, u_D) is $\left(\widehat{\beta_1 - \beta_0}\right) X + \widehat{\tau_1(p)}$.

MTE depends on the propensity score p as well as X. In the homogeneous response setting, MTE is constant and $MTE = ATE = ATT = ATUT$. While in the heterogeneous response setting, *MTE* is nonlinear in p.

11.8 Discrete outcomes

Aakvik, Heckman, and Vytlacil [2005] (*AHV*) describe an analogous *MTE* approach for the discrete outcomes case. The setup is analogous to the continuous case discussed above except the following modifications are made to the potential outcomes model.

$$\begin{aligned} Y_1 &= \mu_1(X, U_1) \\ Y_0 &= \mu_0(X, U_0) \end{aligned}$$

A linear latent index is assumed to generate discrete outcomes

$$\mu_j(X, U_j) = I\left[X\beta_j \geq U_j\right]$$

AHV describe the following identifying conditions.

Condition 11.8 (U_0, V_D) *and* (U_1, V_D) *are independent of* (Z, X) *(conditional independence),*

Condition 11.9 $\mu_D(Z)$ *is a nondegenerate random variable conditional on* X *(rank condition),*

Condition 11.10 (V_0, V_D) *and* (V_1, V_D) *are continuous,*

Condition 11.11 *the values of* $E[|Y_0|]$ *and* $E[|Y_1|]$ *are finite (finite means is trivially satisfied for discrete outcomes),*

Condition 11.12 $0 < Pr(D = 1 \mid X) < 1$.

Mean treatment parameters for dichotomous outcomes are

$$\begin{aligned}
MTE\left(x,u\right) &= \Pr\left(Y_1 = 1 \mid X = x, U_D = u\right) \\
&\quad - \Pr\left(Y_0 = 1 \mid X = x, U_D = u\right) \\
ATE\left(x\right) &= \Pr\left(Y_1 = 1 \mid X = x\right) - \Pr\left(Y_0 = 1 \mid X = x\right) \\
ATT\left(x, D = 1\right) &= \Pr\left(Y_1 = 1 \mid X = x, D = 1\right) \\
&\quad - \Pr\left(Y_0 = 1 \mid X = x, D = 1\right) \\
ATUT\left(x, D = 0\right) &= \Pr\left(Y_1 = 1 \mid X = x, D = 0\right) \\
&\quad - \Pr\left(Y_0 = 1 \mid X = x, D = 0\right)
\end{aligned}$$

AHV also discuss and empirically estimate treatment effect distributions utilizing a (single) factor-structure strategy for model unobservables.[8]

11.8.1 Multilevel discrete and continuous endogenous treatment

To this point, our treatment effects discussion has been limited to binary treatment. In this section, we'll briefly discuss extensions to the multilevel discrete (ordered and unordered) case (Heckman and Vytlacil [2007b]) and continuous treatment case (Florens, Heckman, Meghir, and Vytlacil [2003] and Heckman and Vytlacil [2007b]). Identification conditions are similar for all cases of multinomial treatment.

FHMV and *HV* discuss conditions under which control function, *IV*, and *LIV* equivalently identify *ATE* via the partial derivative of the outcome equation with respect to (continuous) treatment. This is essentially the homogeneous response case. In the heterogenous response case, *ATE* can be identified by a control function or *LIV* but under different conditions. *LIV* allows relaxation of the standard single index (uniformity) assumption. Refer to *FHMV* for details. Next, we return to *HV*'s *MTE* framework and briefly discuss how it applies to ordered choice, unordered choice, and continuous treatment.

Ordered choice

Consider an ordered choice model where there are S choices. Potential outcomes are

$$Y_s = \mu_s\left(X, U_s\right) \quad \text{for } s = 1, \ldots, S$$

Observed choices are

$$D_s = 1\left[C_{s-1}\left(W_{s-1}\right) < \mu_D\left(Z\right) - V_D < C_s\left(W_s\right)\right]$$

for latent index $U = \mu_D\left(Z\right) - V_D$ and cutoffs $C_s\left(W_s\right)$ where Z shift the index generally and W_s affect s-specific transitions. Intuitively, one needs an instrument

[8]Carneiro, Hansen, and Heckman [2003] extend this by analyzing panel data, allowing for multiple factors, and more general choice processes.

(or source of variation) for each transition. Identifying conditions are similar to those above.

Condition 11.13 *(U_s, V_D) are independent of (Z, W) conditional on X for $s = 1, \ldots, S$ (conditional independence),*

Condition 11.14 *$\mu_D(Z)$ is a nondegenerate random variable conditional on (X, W) (rank condition),*

Condition 11.15 *the distribution of V_D is continuous,*

Condition 11.16 *the values of $E[|Y_s|]$ are finite for $s = 1, \ldots, S$ (finite means),*

Condition 11.17 *$0 < Pr(D_s = 1 \mid X) < 1$ for $s = 1, \ldots, S$ (in large samples, there are some individuals in each treatment state).*

Condition 11.18 *For $s = 1, \ldots, S-1$, the distribution of $C_s(W_s)$ conditional on (X, Z) and the other $C_j(W_j)$, $j = 1, \ldots S$, $j \neq s$, is nondegenerate and continuous.*

The transition-specific *MTE* for the transition from s to $s+1$ is

$$\Delta^{MTE}_{s,s+1}(x, v) = E[Y_{s+1} - Y_s \mid X = x, V_D = v] \quad \text{for } s = 1, \ldots, S-1$$

Unordered choice

The parallel conditions for evaluating causal effects in multilevel unordered discrete treatment models are:

Condition 11.19 *(U_s, V_D) are independent of Z conditional on X for $s = 1, \ldots, S$ (conditional independence),*

Condition 11.20 *for each Z_j there exists at least one element $Z^{[j]}$ that is not an element of Z_k, $j \neq k$, and such that the distribution of $\mu_D(Z)$ conditional on $\left(X, Z^{[-j]}\right)$ is not degenerate,*

or

Condition 11.21 *for each Z_j there exists at least one element $Z^{[j]}$ that is not an element of Z_k, $j \neq k$, and such that the distribution of $\mu_D(Z)$ conditional on $\left(X, Z^{[-j]}\right)$ is continuous.*

Condition 11.22 *the distribution of V_D is continuous,*

Condition 11.23 *the values of $E[|Y_s|]$ are finite for $s = 1, \ldots, S$ (finite means),*

Condition 11.24 *$0 < Pr(D_s = 1 \mid X) < 1$ for $s = 1, \ldots, S$ (in large samples, there are some individuals in each treatment state).*

The treatment effect is $Y_j - Y_k$ where $j \neq k$. And regime j can be compared with the best alternative, say k, or other variations.

Continuous treatment

Continue with our common setup except assume outcome Y_d is continuous in d. This implies that for d and d' close so are Y_d and $Y_{d'}$. The average treatment effect can be defined as

$$ATE_d\left(x\right) = E\left[\frac{\partial}{\partial d} Y_d \mid X = x\right]$$

The average treatment effect on treated is

$$ATT_d\left(x\right) = E\left[\frac{\partial}{\partial d_1} Y_{d_1} \mid D = d_2, X = x\right] \Bigg|_{d=d_1=d_2}$$

And the marginal treatment effect is

$$MTE_d\left(x, u\right) = E\left[\frac{\partial}{\partial d} Y_d \mid X = x, U_D = u\right]$$

See Florens, Heckman, Meghir, and Vytlacil [2003] and Heckman and Vytlacil [2007b, pp.5021-5026] for additional details regarding semiparametric identification of treatment effects.

11.9 Distributions of treatment effects

A limitation of the discussion to this juncture is we have focused on population means of treatment effects. This prohibits discussion of potentially important properties such as the proportion of individuals who benefit or who suffer from treatment.

Abbring and Heckman [2007] discuss utilization of factor models to identify the joint distribution of outcomes (including counterfactual distributions) and accordingly the distribution of treatment effects $Y_1 - Y_0$. Factor models are a type of replacement function (Heckman and Robb [1986]) where conditional on the factors, outcomes and choice equations are independent. That is, we rely on a type of conditional independence for identification. A simple one-factor model illustrates. Let θ be a scalar factor that produces dependence amongst the unobservables (unobservables are assumed to be independent of (X, Z)). Let M be a proxy measure for θ where $M = \mu_M\left(X\right) + \alpha_M \theta + \varepsilon_M$

$$\begin{aligned}
V_0 &= \alpha_0 \theta + \varepsilon_0 \\
V_1 &= \alpha_1 \theta + \varepsilon_1 \\
V_D &= \alpha_D \theta + \varepsilon_D
\end{aligned}$$

$\varepsilon_0, \varepsilon_1, \varepsilon_D, \varepsilon_M$ are mutually independent and independent of θ, all with mean zero. To fix the scale of the unobserved factor, normalize one coefficient (loading) to, say, $\alpha_M = 1$. The key is to exploit the notion that all of the dependence arises from θ.

$$\begin{aligned}
Cov\left[Y_0, M \mid X, Z\right] &= \alpha_0 \alpha_M \sigma_\theta^2 \\
Cov\left[Y_1, M \mid X, Z\right] &= \alpha_1 \alpha_M \sigma_\theta^2 \\
Cov\left[Y_0, D^* \mid X, Z\right] &= \alpha_0 \frac{\alpha_D}{\sigma_{U_D}} \sigma_\theta^2 \\
Cov\left[Y_1, D^* \mid X, Z\right] &= \alpha_1 \frac{\alpha_D}{\sigma_{U_D}} \sigma_\theta^2 \\
Cov\left[D^*, M \mid X, Z\right] &= \frac{\alpha_D}{\sigma_{U_D}} \alpha_M \sigma_\theta^2
\end{aligned}$$

From the ratio of $Cov\left[Y_1, D^* \mid X, Z\right]$ to $Cov\left[D^*, M \mid X, Z\right]$, we find α_1 ($\alpha_M = 1$ by normalization). From $\frac{Cov[Y_1, D^* \mid X, Z]}{Cov[Y_0, D^* \mid X, Z]} = \frac{\alpha_1}{\alpha_0}$, we determine α_0. Finally, from either $Cov\left[Y_0, M \mid X, Z\right]$ or $Cov\left[Y_1, M \mid X, Z\right]$ we determine scale σ_θ^2. Since $Cov\left[Y_0, Y_1 \mid X, Z\right] = \alpha_0 \alpha_1 \sigma_\theta^2$, the joint distribution of objective outcomes is identified.

See Abbring and Heckman [2007] for additional details, including use of proxies, panel data and multiple factors for identification of joint distributions of subjective outcomes, and references.

11.10 Dynamic timing of treatment

The foregoing discussion highlights one time (now or never) static analysis of the choice of treatment. In some settings it's important to consider the impact of acquisition of information on the option value of treatment. It is important to distinguish what information is available to decision makers and when and what information is available to the analyst. Distinctions between ex ante and ex post impact and subjective versus objective gains to treatment are brought to the fore.

Policy invariance (P-1 through P-4) as well as the distinction between the evaluation problem and the selection problem lay the foundation for identification. The evaluation problem is one where we observe the individual in one treatment state but wish to determine the individual's outcome in another state. The selection problem is one where the distribution of outcomes for an individual we observe in a given state is not the same as the marginal outcome distribution we would observe if the individual is randomly assigned to the state. Policy invariance simplifies the dynamic evaluation problem to (a) identifying the dynamic assignment of treatments under the policy, and (b) identifying dynamic treatment effects on individual outcomes.

Dynamic treatment effect analysis typically takes the form of a duration model (or time to treatment model; see Heckman and Singer [1986] for an early and extensive review of the problem). A variety of conditional independence, matching, or dynamic panel data analyses supply identification conditions. Discrete-time and continuous-time as well as reduced form and structural approaches have been proposed. Abbring and Heckman [2007] summarize this work, and provide additional details and references.

11.11 General equilibrium effects

Policy invariance pervades the previous discussion. Sometimes policies or programs to be evaluated are so far reaching to invalidate policy invariance. Interactions among individuals mediated by markets can be an important behavioral consideration that invalidates the partial equilibrium restrictions discussed above and mandates general equilibrium considerations (for example, changing prices and/or supply of inputs as a result of policy intervention). As an example, Heckman, Lochner, and Tabor [1998a, 1998b, 1998c] report that static treatment effects overstate the impact of college tuition subsidy on future wages by ten times compared to their general equilibrium analysis. See Abbring and Heckman [2007] for a review of the analysis of general equilibrium effects.

In any social setting, policy invariance conditions PI-2 and PI-4 are very strong. They effectively claim that untreated individuals are unaffected by who does receive treatment. Relaxation of invariance conditions or entertainment of general equilibrium effects is troublesome for standard approaches like difference - in - difference estimators as the "control group" is affected by policy interventions but a difference-in-difference estimator fails to identify the impact. Further, in stark contrast to conventional uniformity conditions of microeconometric treatment effect analysis, general equilibrium analysis must accommodate two way flows.

11.12 Regulated report precision example

LIV estimation of marginal treatment effects is illustrated for the regulated report precision example from chapter 10. We don't repeat the setup here but rather refer the reader to chapters 2 and 10. Bayesian data augmentation and analysis of marginal treatment effects are discussed and illustrated for regulated report precision in chapter 12.

11.12.1 Apparent nonnormality and MTE

We explore the impact of apparent nonnormality on the analysis of report precision treatment effects. In our simulation, α_d is observed by the owner prior to selecting report precision, α_d^L is drawn from an exponential distribution with rate $\frac{1}{0.02}$ (reciprocal of the mean), α_d^H is drawn from an exponential distribution with rate $\frac{1}{0.04}$, α is drawn from an exponential distribution with rate $\frac{1}{0.03}$ and γ is drawn from an exponential distribution with rate $\frac{1}{5}$.[9] This means the unobservable (by the analyst) portion of the choice equation is apparently nonnormal. Setting parameters are summarized below.

Stochastic parameters
$\alpha_d^L \sim \exp\left(\frac{1}{0.02}\right)$
$\alpha_d^H \sim \exp\left(\frac{1}{0.04}\right)$
$\alpha \sim \exp\left(\frac{1}{0.03}\right)$
$\gamma \sim \exp\left(\frac{1}{5}\right)$
$\beta^L \sim N(7, 1)$
$\beta^H \sim N(7, 1)$

First, we report benchmark *OLS* results and results from *IV* strategies developed in chapter 10. Then, we apply *LIV* to identify *MTE*-estimated average treatment effects.

OLS results

Benchmark *OLS* simulation results are reported in table 11.2 and sample statistics for average treatment effects in table 11.3. Although there is little difference between *ATE* and *OLS*, *OLS* estimates of other average treatment effects are poor, as expected. Further, *OLS* cannot detect outcome heterogeneity. *IV* strategies may be more effective.

Ordinate *IV* control model

The ordinate control function regression is

$$E[Y \mid s, D, \phi] = \beta_0 + \beta_1 (s - \overline{s}) + \beta_2 D (s - \overline{s}) + \beta_3 \phi(Z\theta) + \beta_4 D$$

and is estimated via two stage *IV* where instruments

$$\{\iota, (s - \overline{s}), m(s - \overline{s}), \phi(Z\theta), m\}$$

are employed and

$$m = \Pr\left(D = 1 \mid Z = \begin{bmatrix} \iota & w_1 & w_2 \end{bmatrix}\right)$$

is estimated via probit. The coefficient on D, β_4, estimates *ATE*. Simulation results are reported in table 11.4. Although, on average, the rank ordering of *ATT*

[9] Probability as logic implies that if we only know the mean and support is nonnegative, then we conclude α_d has an exponential distribution. Similar reasoning implies knowledge of the variance leads to a Gaussian distribution (see Jaynes [2003] and chapter 13).

Table 11.2: Continuous report precision but observed binary OLS parameter estimates for apparently nonnormal DGP

$statistic$	β_0	β_1	β_2
$mean$	635.0	0.523	$-.006$
$median$	635.0	0.526	-0.066
$std.dev.$	1.672	0.105	0.148
$minimum$	630.1	0.226	-0.469
$maximum$	639.6	0.744	0.406
$statistic$	$\beta_3\,(estATE)$	$estATT$	$estATUT$
$mean$	4.217	4.244	4.192
$median$	4.009	4.020	4.034
$std.dev.$	2.184	2.183	2.187
$minimum$	-1.905	-1.887	-1.952
$maximum$	10.25	10.37	10.13
$E[Y \mid s, D] = \beta_0 + \beta_1 (s - \overline{s}) + \beta_2 D (s - \overline{s}) + \beta_3 D$			

Table 11.3: Continuous report precision but observed binary average treatment effect sample statistics for apparently nonnormal DGP

$statistic$	ATE	ATT	$ATUT$
$mean$	-1.053	62.04	-60.43
$median$	-1.012	62.12	-60.44
$std.dev.$	1.800	1.678	1.519
$minimum$	-6.007	58.16	-64.54
$maximum$	3.787	65.53	-56.94

Table 11.4: Continuous report precision but observed binary ordinate control IV parameter estimates for apparently nonnormal DGP

$statistic$	β_0	β_1	β_2	β_3
$mean$	805.7	-2.879	5.845	54.71
$median$	765.9	-2.889	5.780	153.3
$std.dev.$	469.8	1.100	1.918	1373
$minimum$	-482.7	-5.282	0.104	-3864
$maximum$	2135	0.537	10.25	3772
$statistic$	$\beta_4\,(estATE)$	$estATT$	$estATUT$	
$mean$	-391.4	-369.6	-411.7	
$median$	-397.9	-336.5	-430.7	
$std.dev.$	164.5	390.4	671.2	
$minimum$	-787.4	-1456	-2190	
$maximum$	130.9	716.0	1554	
$E[Y \mid s, D, \phi] = \beta_0 + \beta_1 (s - \overline{s}) + \beta_2 D (s - \overline{s}) + \beta_3 \phi (Z\theta) + \beta_4 D$				

Table 11.5: Continuous report precision but observed binary inverse Mills IV parameter estimates for apparently nonnormal DGP

$statistic$	β_0	β_1	β_2	β_3	β_4
$mean$	636.7	0.525	0.468	2.074	0.273
$median$	636.1	0.533	0.467	0.610	−4.938
$std.dev.$	30.61	0.114	0.114	39.74	41.53
$minimum$	549.2	0.182	0.108	−113.5	−118.4
$maximum$	724.4	0.809	0.761	116.0	121.4
$statistic$		$\beta_5\,(estATE)$	$estATT$	$estATUT$	
$mean$		2.168	0.687	3.555	
$median$		5.056	0.439	12.26	
$std.dev.$		48.44	63.22	66.16	
$minimum$		−173.4	−181.4	−192.9	
$maximum$		117.8	182.6	190.5	

$$E[Y \mid s, D, \lambda] = \beta_0 + \beta_1 (1-D)(s-\overline{s}) + \beta_2 D (s-\overline{s}) + \beta_3 (1-D)\lambda^H + \beta_4 D \lambda^L + \beta_5 D$$

and *ATUT* is consistent with the sample statistics,.the ordinate control function treatment effect estimates are inconsistent (biased downward) and extremely variable, In other words, the evidence suggests nonnormality renders the utility of a normality-based ordinate control function approach suspect.

Inverse-Mills *IV* model

Heckman's inverse-Mills ratio regression is

$$\begin{aligned} E[Y \mid s, D, \lambda] &= \beta_0 + \beta_1 (1-D)(s-\overline{s}) + \beta_2 D (s-\overline{s}) \\ &\quad + \beta_3 (1-D)\lambda^H + \beta_4 D \lambda^L + \beta_5 D \end{aligned}$$

where $\overline{s}$ is the sample average of s, $\lambda^H = -\frac{\phi(Z\theta)}{1-\Phi(Z\theta)}$, $\lambda^L = \frac{\phi(Z\theta)}{\Phi(Z\theta)}$, and θ is the estimated parameters from a probit regression of precision choice D on $Z = \begin{bmatrix} \iota & w_1 & w_2 \end{bmatrix}$ (ι is a vector of ones). The coefficient on D, β_5, is the estimate of the average treatment effect, *ATE*. Simulation results including estimated average treatment effects on treated (*estATT*) and untreated (*estATUT*) are reported in table 11.5. The inverse-Mills estimates of the treatment effects are inconsistent and sufficiently variable that we may not detect nonzero treatment effects — though estimated treated effects are not as variable as those estimated by the ordinate control *IV* model. Further, the inverse-Mills results suggest greater homogeneity (all treatment effects are negative, on average) which suggests we likely would be unable to identify outcome heterogeneity based on this control function strategy.

MTE estimates via *LIV*

Next, we employ Heckman's *MTE* approach for estimating the treatment effects via a semi-parametric local instrumental variable estimator (*LIV*). Our *LIV* semi-

Table 11.6: Continuous report precision but observed binary LIV parameter estimates for apparently nonnormal DGP

$statistic$	β_1	β_2	$estATE$	$estATT$	$estATUT$
$mean$	1.178	−1.390	17.98	14.73	25.79
$std.dev.$	0.496	1.009	23.54	26.11	38.08
$minimum$	0.271	−3.517	−27.63	−32.86	−55.07
$maximum$	2.213	0.439	64.67	69.51	94.19
$E[Y \mid s, D, \tau_1(p)] = \beta_1(s-\overline{s}) + \beta_2 D(s-\overline{s}) + \tau_1(p)$					

parametric approach only allows us to recover estimates from the outcome equations for β_1 and β_2 where the reference regression is

$$E[Y \mid s, D, \tau_1(p)] = \beta_1(s-\overline{s}) + \beta_2 D(s-\overline{s}) + \tau_1(p)$$

We employ semi-parametric methods to estimate the outcome equation. Estimated parameters and treatment effects based on bootstrapped semi-parametric weighted *MTE* are in table 11.6.[10] While the *MTE* results may more closely approximate the sample statistics than their parametric counterpart *IV* estimators, their high variance and apparent bias compromises their utility. Could we reliably detect endogeneity or heterogeneity? Perhaps — however the ordering of the estimated treatment effects doesn't correspond well with sample statistics for the average treatment effects.

Are these results due to nonnormality of the unobservable features of the selection equation? Perhaps, but a closer look suggests that our original thinking applied to this *DGP* is misguided. While expected utility associated with low (or high) inverse report precision equilibrium strategies are distinctly nonnormal, selection involves their relative ranking or, in other words, the unobservable of interest comes from the difference in unobservables. Remarkably, their difference (V_D) is not distinguishable from Gaussian draws (based on descriptive statistics, plots, etc.).

Then, what is the explanation? It is partially explained by the analyst observing binary choice when there is a multiplicity of inverse report precision choices. However, we observed this in an earlier case (see chapter 10) with a lesser impact than demonstrated here. Rather, the feature that stands out is the quality of the instruments. The same instruments are employed in this "nonnormal" case as previously employed but, apparently, are much weaker instruments in this allegedly nonnormal setting. In table 11.7 we report the analogous sample correlations to those reported in chapter 10 for Gaussian draws. Correlations between the instruments, w_1 and w_2, and treatment, D, are decidedly smaller than the examples reported in chapter 10. Further, α and γ offer little help.

[10]Unlike other simulations which are developed within R, these results are produced using Heckman, Urzua, and Vytlacil's *MTE* program. Reported results employ a probit selection equation. Similar results obtain when either a linear probability or nonparametric regression selection equation is employed.

Table 11.7: Continuous report precision but observed binary sample correlations for apparently nonnormal DGP

$statistic$	$r(\alpha, U^L)$	$r(\alpha, U^H)$	$r(\gamma, U^L)$	$r(\gamma, U^H)$
$mean$	-0.004	0.000	0.005	-0.007
$median$	-0.005	-0.001	0.007	-0.006
$std.dev.$	0.022	0.024	0.023	0.022
$minimum$	-0.081	-0.056	-0.048	-0.085
$maximum$	0.054	0.064	0.066	0.039
$statistic$	$r(\alpha, D)$	$r(\gamma, D)$	$r(w_1, D)$	$r(w_2, D)$
$mean$	0.013	-0.046	-0.114	0.025
$median$	0.013	-0.046	-0.113	0.024
$std.dev.$	0.022	0.021	0.012	0.014
$minimum$	-0.042	-0.106	-0.155	-0.011
$maximum$	0.082	0.017	-0.080	0.063

Stronger instruments

To further explore this explanation, we create a third and stronger instrument, w_3, and utilize it along with w_1 in the selection equation where $W = \begin{bmatrix} w_1 & w_3 \end{bmatrix}$. This third instrument is the residuals of a binary variable

$$\Im\left(EU\left(\sigma_2^L, \overline{\sigma}_2^L\right) > EU\left(\sigma_2^H, \overline{\sigma}_2^L\right)\right)$$

regressed onto U^L and U^H where $\Im(\cdot)$ is an indicator function. Below we report in table 11.8 ordinate control function results. Average treatment effect sample statistics for this simulation including the *OLS* effect are reported in table 11.9. Although the average treatment effects are attenuated a bit toward zero, these results are a marked improvement of the previous, wildly erratic results. Inverse-Mills results are reported in table 11.10. These results correspond quite well with treatment effect sample statistics. Hence, we're reminded (once again) the value of strong instruments for logically consistent analysis cannot be over-estimated.

Finally, we report in table 11.11 *LIV*-estimated average treatment effects derived from *MTE* with this stronger instrument, w_3. Again, the results are improved relative to those with the weaker instruments but as before the average treatment effects are attenuated.[11] Average treatment on the untreated along with the average treatment effect correspond best with their sample statistics. Not surprisingly, the results are noisier than the parametric results. For this setting, we conclude that strong instruments are more important than relaxed distributional assignment (based on the data) for identifying and estimating various average treatment effects.

[11]Reported results employ a probit regression for the selection equations (as is the case for the foregoing parametric analyses). Results based on a nonparametric regression for the treatment equation are qualitatively unchanged.

Table 11.8: Continuous report precision but observed binary stronger ordinate control IV parameter estimates for apparently nonnormal DGP

$statistic$	β_0	β_1	β_2	β_3
$mean$	596.8	0.423	0.024	137.9
$median$	597.0	0.414	0.025	138.2
$std.dev.$	4.168	0.140	0.238	14.87
$minimum$	586.8	−0.012	−0.717	90.56
$maximum$	609.8	0.829	0.728	179.2
$statistic$	$\beta_4\ (estATE)$	$estATT$	$estATUT$	
$mean$	−2.494	40.35	−43.77	
$median$	−2.449	40.07	−43.58	
$std.dev.$	2.343	−4.371	5.598	
$minimum$	−8.850	28.50	−58.91	
$maximum$	4.162	52.40	−26.60	
$E[Y \mid s, D, \phi] = \beta_0 + \beta_1 (s - \overline{s}) + \beta_2 D (s - \overline{s}) + \beta_3 \phi (W\theta) + \beta_4 D$				

Table 11.9: Continuous report precision but observed binary average treatment effect sample statistics for apparently nonnormal DGP

$statistic$	ATE	ATT	$ATUT$	OLS
$mean$	−0.266	64.08	−62.26	0.578
$median$	−0.203	64.16	−62.30	0.764
$std.dev.$	1.596	1.448	1.584	2.100
$minimum$	−5.015	60.32	−66.64	−4.980
$maximum$	3.746	67.48	−57.38	6.077

Table 11.10: Continuous report precision but observed binary stronger inverse Mills IV parameter estimates for apparently nonnormal DGP

$statistic$	β_0	β_1	β_2	β_3	β_4
$mean$	608.9	0.432	0.435	−48.27	61.66
$median$	608.9	0.435	0.438	−48.55	61.60
$std.dev.$	1.730	0.099	0.086	2.743	3.949
$minimum$	603.8	0.159	0.238	−54.85	51.27
$maximum$	613.3	0.716	0.652	−40.70	72.70
$statistic$		$\beta_5\ (estATE)$	$estATT$	$estATUT$	
$mean$		−8.565	57.61	−72.28	
$median$		−8.353	57.44	−72.28	
$std.dev.$		2.282	3.294	4.628	
$minimum$		−15.51	48.44	−85.37	
$maximum$		−2.814	67.11	−60.39	
$E[Y \mid s, D, \lambda] = \beta_0 + \beta_1 (1 - D)(s - \overline{s}) + \beta_2 D (s - \overline{s})$ $+\beta_3 (1 - D) \lambda^H + \beta_4 D \lambda^L + \beta_5 D$					

Table 11.11: Continuous report precision but observed binary stronger LIV parameter estimates for apparently nonnormal DGP

$statistic$	β_1	β_2	$estATE$	$estATT$	$estATUT$
$mean$	0.389	0.220	−7.798	9.385	−24.68
$std.dev.$	0.159	0.268	9.805	14.17	16.38
$minimum$	0.107	−0.330	−26.85	−17.69	−57.14
$maximum$	0.729	0.718	11.58	37.87	−26.85
$statistic$		OLS	ATE	ATT	$ATUT$
$mean$		3.609	1.593	63.76	−61.75
$median$		3.592	1.642	63.91	−61.70
$std.dev.$		2.484	1.894	1.546	1.668
$minimum$		−3.057	−4.313	59.58	−66.87
$maximum$		11.28	5.821	67.12	−58.11
$E[Y \mid s, D, \tau_1] = \beta_1 (s - \overline{s}) + \beta_2 D (s - \overline{s}) + \tau_1 (p)$					

11.13 Additional reading

There are numerous contributions to this literature. We suggest beginning with Heckman's [2001] Nobel lecture, Heckman and Vytlacil [2005, 2007a, 2007b], and Abbring and Heckman [2007]. These papers provide extensive discussions and voluminous references. This chapter has provided at most a thumbnail sketch of this extensive and important work. A FORTRAN program and documentation for estimating Heckman, Urzua, and Vytlacil's [2006] marginal treatment effect can be found at URL: http://jenni.uchicago.edu/underiv/.