Generalized Roy Model and Cost-Benefit Analysis of Social Programs[1]

James J. Heckman
The University of Chicago
University College Dublin

Philipp Eisenhauer
University of Mannheim

Edward Vytlacil
Columbia University

This Draft, March 15, 2011

[1] This research was supported by NIH R01-HD32058-3, NSF SES-024158, and NSF SES-05-51089.

1 Introduction

In a series of papers, Heckman and Vytlacil (1999, 2005a) use the marginal benefit of treatment B^{MTE} introduced in Björklund and Moffitt (1987) to unify the benefit of treatment parameters and to consider their identification. These correspond to average changes in the outcome due to treatment, and do not reflect the cost of the treatment as viewed by the agents. We extend their work by imposing more structure, the nonparametric generalized Roy model, and exploit the additional structure to consider the subjective cost and surplus, or net benefit, from the program. We define parameters for the generalized Roy model corresponding to the average cost and surplus of participating in a program; these parameters are parallel to the treatment effect parameters considered by Heckman and Vytlacil (1999, 2005a). The central feature of the generalized Roy model is that the agent chooses treatment if the benefit exceeds the subjective cost as perceived by the agent. This creates a simple relationship between the cost and benefit parameters that we exploit for identifying the cost and surplus parameters. Our main result is that cost parameters and surplus parameters in the generalized Roy model can be identified without direct information on the costs of treatment. Our analysis complements and extends the analysis of Björklund and Moffitt (1987) who first noted the duality between cost and benefit parameters in the generalized Roy model.

Our identification analysis highlights the role of having regressors that shift both, the cost and benefit from the program. The greater the variation in costs induced by variation in the observed cost shifters, the richer the information on the benefits of the program that is identified. The greater the variation in benefits induced by variation in the regressors that enter the outcome equation, the richer the information on the costs of the program that is identified. Both forms of variation play a central role in identification of the surplus from the program.

The plan of this paper is as follows. Section 2 introduces our generalized Roy model in which agents select into treatment if the surplus from doing so is positive. Section 3 reviews the average benefit of treatment parameters from Heckman and Vytlacil (1999, 2005a), and develops and analyzes the dual cost parameters to match the benefit parameters. Section 4 presents an identification analysis of the cost and surplus parameters, based on first identifying the corresponding marginal benefits of treatment through local instrumental variables and then integrating them appropriately. To illustrate the empirical content of our approach, we apply the results to a example from educational choice in Section 5. Section 6 concludes.

2 The Nonparametric Generalized Roy Model

Suppose there are two potential outcomes (Y_0, Y_1), and a choice indicator D with $D = 1$ if the agent selects into treatment so that Y_1 is observed and $D = 0$ if the agent does not select into treatment so that Y_0 is observed. The observed outcome Y can be written in switching regression form (Quandt, 1958, 1973)

$$Y = DY_1 + (1 - D)Y_0, \tag{2.1}$$

where we impose

$$Y_j = \mu_j(X) + U_j \tag{2.2}$$

for $j = 0, 1$. X is a vector of regressors observed by the econometrician while (U_1, U_0) are not. Combining equations (2.1) and (2.2),

$$Y = \mu_0(X) + \{[\mu_1(X) - \mu_0(X)] + U_1 - U_0\}D + U_0. \tag{2.3}$$

The individual benefit of treatment associated with moving an otherwise identical person from "0" to "1" is $B = Y_1 - Y_0$ and is defined as the causal effect on Y of a *ceteris paribus* move from "0" to "1". We denote the subjective cost of choosing treatment as perceived by the agent as C, determined by

$$C = \mu_C(Z) + U_C, \tag{2.4}$$

where Z is an observed random vector of cost shifters and U_C is an unobserved random variable. The decision rule for program participation is determined according to the generalized Roy model, i.e. individuals choose treatment if the benefit from treatment is greater than the subjective cost:

$$D = 1 \quad \text{if} \quad S \geq 0\,; \qquad D = 0 \quad \text{otherwise}, \tag{2.5}$$

where S is the surplus, i.e. net gain, from treatment,

$$\begin{aligned} S &= (Y_1 - Y_0) - C \\ &= \{[\mu_1(X) - \mu_0(X)] - \mu_C(Z)\} - [U_C - (U_1 - U_0)] \\ &= \mu_S(X, Z) - V \end{aligned}$$

with $\mu_S(X, Z) = [\mu_1(X) - \mu_0(X)] - \mu_C(Z)$ and $V = U_C - (U_1 - U_0)$. We do not assume any particular functional form for the functions μ_0, μ_1 and μ_C, and we do not assume that the distribution of U_0, U_1, or U_C is known. We maintain equations (2.1) – (2.5) throughout this paper.

The original Roy (1951) model assumes that there are no observed X regressors, that the cost of treatment is identically zero (i.e. $\mu_C = 0, U_C = 0$), and that $(U_0, U_1) \sim N(0, \Sigma)$. Heckman and Honoré (1990) develop a nonparametric version of the Roy model with X regressors and no parametric assumption on the distribution of (U_0, U_1). Their version of the Roy model also imposes that the cost of treatment is identically zero. In contrast, we allow non-zero cost of treatment, and for our identification analysis we require nondegenerate cost of treatment and observed cost-shifters.

From the point of view of the econometrician (X, Z) is observed and (U_1, U_0, U_C) is unobserved. This model supposes that agents know the true benefit, $B = Y_1 - Y_0$, of the treatment. As shown in Appendix (A), our results extend to a broader class of models in which the agents participate in the program if the expected benefits given the information available to them is greater than their cost of treatment. This model also supposes that there is no other aspect of the benefit of the treatment other than $Y_1 - Y_0$. Implicitly, any subjective benefits of the program are being incorporated into the costs of treatment, i.e. the cost function includes the subjective benefits of the treatment. For example, if training allows the individual to work in a job with preferred amenities, this is being modeled as a (negative) contribution to the subjective cost of treatment. We will suppose that Z and X do not contain any common elements. This supposition is purely for ease of exposition, all of the analysis of this paper can be seen as implicitly conditioning on all common elements of X and Z.

We invoke the following assumptions:

(A-1) (U_0, U_1, U_C) *are independent of* (X, Z).

(A-2) *The distribution of* $\mu_C(Z)$ *conditional on* X *is absolutely continuous with respect to Lebesgue measure.*

(A-3) *The distribution of* $V = U_C - (U_1 - U_0)$ *is absolutely continuous with respect to Lebesgue measure and has a cumulative distribution function that is strictly increasing.*

(A-4) *The values of* $E|Y_1|$, $E|Y_0|$ *and* $E|C|$ *are finite.*

(A-5) $0 < \Pr(D = 1 \mid X, Z) < 1$ *w.p. 1.*

(A-1) assumes that (U_0, U_1, U_C) is independent of (X, Z). Thus, D is endogenous but other regressors in both the treatment equation and the outcome equation are exogenous. Recall that we are implicitly conditioning on any regressors that enter both the outcome equations and the cost equation, so that this condition should be interpreted as an independence assumption of the error terms from the unique elements of X and Z conditional on the regressors that enter both equations. (A-2) requires that there exist at least one continuous component of Z conditional on X. This assumption will only be required for our identification analysis, and is not needed for our definitions or analysis of the cost and surplus parameters. (A-3) is a regularity condition. It allows for the possibility that U_C is degenerate (costs don't vary conditional on Z) or that $U_1 - U_0$ is degenerate (treatment effects don't vary conditional on X), though not both. Assumption (A-4) is needed to satisfy standard integration conditions. It also guarantees that the mean benefit and cost parameters are well defined. Assumption (A-5) is the assumption in the population of both a treatment and a control group for (a.e.) (X, Z). Note that this assumption still allows the support of $\Pr(D = 1|X, Z)$ to be the full unit interval.

Let $P(X, Z)$ denote the probability of selecting treatment given (X, Z), or the "propensity score" $P(X, Z) \equiv \Pr(D = 1 \mid X, Z) = F_V(\mu_S(X, Z))$, where $F_V(\cdot)$ denotes the distribution of V.[1] We sometimes denote $P(X, Z)$ by P, suppressing the (X, Z) argument. We also work with U_S, a uniform random variable ($U_S \sim \text{Unif}[0, 1]$) defined by $U_S = F_V(V)$. Thus different values of u_S denote different quantiles of V. Given our previous assumptions, F_V is strictly increasing, and $P(X, Z)$ is continuous random variable conditional on X.

The generalized Roy model of this paper is a special case of the model of Heckman and Vytlacil (1999, 2005a). Our model of equations (2.1) – (2.5) and our assumptions (A-1) – (A-5) imply the model and assumptions of Heckman and Vytlacil (1999, 2005a), and thus, by the result of Vytlacil (2002), imply the Local Average Treatment Effect (LATE) model of Imbens and Angrist (1994). We are imposing more restrictions here. In particular, we are imposing the generalized Roy model and the corresponding assumptions that will allow us to exploit the generalized Roy model for identification of subjective cost parameters. As is conventional for the Roy model, we are imposing additive separability in the outcome equations (2.2). This additive separability is not imposed in Heckman and Vytlacil (1999, 2005a), but is required by our analysis to make additive separability in the latent index equation (2.5) consistent with the generalized Roy model. Recall again that we are implicitly conditioning on all common elements of (X, Z), so that these need not be additively separable from the error term. We are imposing conditions on X that are not

[1] We will refer to the cumulative distribution function of a random vector A by $F_A(\cdot)$ and to the cumulative distribution function of a random vector A conditional on random vector B by $F_{A|B}(\cdot)$. We will write the cumulative distribution function of A conditional on $B = b$ by $F_{A|B}(\cdot \mid b)$.

required by Heckman and Vytlacil (1999, 2005a). In their analysis, they fully condition on X, and thus do not need to assume that X is independent of the error vector. In contrast, to exploit the generalized Roy model to recover subjective cost parameters, we require that the unique elements X are independent of the error vector.[2] We are implicitly fully conditioning on any common elements of X and Z, and no independence condition is required on the common elements.

[2] In this respect, our analysis is broadly analogous to the identification strategies and conditions of Vytlacil and Yildiz (2007) and Shaikh and Vytlacil (2005), who also require that there be exogenous regressors in the outcome equation and exploit variation in such regressors for identification .

3 Benefit, Cost, and Surplus Parameters

In this section, we define and analyze the benefit, cost, and surplus parameters. We maintain the model of equations (2.1) – (2.5), and impose assumptions (A-1) and (A-3) – (A-5). We do not require assumption (A-2) for the definition or analysis of the parameters, but will use this assumption in the next section for the identification analysis.

Standard treatment effect analysis seeks averaged parameters of the benefit of treatment, $B = Y_1 - Y_0$. The most commonly invoked treatment effect parameter is the average benefit of treatment $B^{ATE}(x) \equiv E(Y_1 - Y_0 \mid X = x) = \mu_1(x) - \mu_0(x)$. This is the effect of assigning treatment randomly to everyone of type $X = x$ assuming full compliance, and ignoring general equilibrium effects. Another commonly invoked parameter is the average benefit of treatment on persons who actually take the treatment, referred to as the benefit of treatment on the treated $B^{TT}(x) \equiv E(Y_1 - Y_0 \mid X = x, D = 1) = \mu_1(x) - \mu_0(x) + E(U_1 - U_0|X = x, D = 1)$. Heckman and Vytlacil (1999, 2005a) unify a broad class of treatment effect parameters including the $B^{ATE}(x)$ and $B^{TT}(x)$ through the marginal benefit of treatment, defined as $B^{MTE}(x, u_S) \equiv E(Y_1 - Y_0|X = x, U_S = u_S) = \mu_1(x) - \mu_0(x) + E(U_1 - U_0|U_S = u_S)$. B^{MTE} is the treatment effect parameter that conditions on the first stage error term, i.e. conditions on the unobserved desire to select into treatment.

The conventional treatment analysis does not define, seek to identify, or estimate any aspect of the cost of the treatment. We define a set of cost parameters parallel to the benefit of treatment parameters, where cost is the subjective cost as perceived by the agent. Thus, we define the average cost of treatment, the average cost of treatment on those treated, and the marginal cost of treatment:

$$\begin{aligned} C^{ATE}(z) &= E(C|Z = z) &&= \mu_C(z) \\ C^{TT}(z) &= E(C|Z = z, D = 1) &&= \mu_C(z) + E(U_C|Z = z, D = 1) \\ C^{MTE}(z, u_S) &= E(C|Z = z, U_S = u_S) &&= \mu_C(z) + E(U_C|U_S = u_S). \end{aligned}$$

Recalling that $S = B - C = \mu_S(X, Z) - V$, where $\mu_S(X, Z) = [\mu_1(X) - \mu_0(X)] - \mu_C(Z)$ and $V = U_C - (U_1 - U_0)$, we now define the corresponding surplus parameters,

$$\begin{aligned} S^{ATE}(x, z) &= E(S|X = x, Z = z) &&= \mu_S(x, z) \\ S^{TT}(x, z) &= E(S|X = x, Z = z, D = 1) &&= \mu_S(x, z) - E(V|X = x, Z = z, D = 1) \\ S^{MTE}(x, z, u_S) &= E(S|X = x, Z = z, U_S = u_S) &&= \mu_S(x, z) - E(V|U_S = u_S). \end{aligned}$$

With these parameters, we can now ask questions not only about the outcome change from the treatment but also the subjective cost of the treatment and the net surplus as well.

Following Heckman and Vytlacil (1999, 2005a), we can represent the average treatment effects and treatment on the treated as averaged versions of the marginal effects of treatment:

$$\begin{array}{lcl} B^{ATE}(x) & = & \int_0^1 B^{MTE}(x,u_S)du_S \\ B^{TT}(x) & = & \int_0^1 B^{MTE}(x,u_S)\frac{1-F_{P|X}(u_S|x)}{\int_0^1(1-F_{P|X}(t|x))dt}du_S. \end{array} \tag{3.1}$$

Following the same arguments as Heckman and Vytlacil (1999, 2005a), we have

$$\begin{array}{lcl} C^{ATE}(z) & = & \int_0^1 C^{MTE}(z,u_S)du_S \\ C^{TT}(z) & = & \int_0^1 C^{MTE}(z,u_S)\frac{1-F_{P|Z}(u_S|z)}{\int_0^1(1-F_{P|Z}(t|z))dt}du_S, \end{array} \tag{3.2}$$

and

$$\begin{array}{lcl} S^{ATE}(x,z) & = & \int_0^1 S^{MTE}(x,z,u_S)du_S \\ S^{TT}(x,z) & = & \frac{1}{P(x,z)}\int_0^{P(x,z)} S^{MTE}(x,z,u_S)du_S. \end{array} \tag{3.3}$$

Next we point to some relationships between the marginal effects of treatment. At first, consider the marginal surplus parameter. Using that $U_S = F_V(V)$ with F_V strictly increasing, we have that $U_S = u_S$ is equivalent to $V = F_V^{-1}(u_S)$, and thus

$$S^{MTE}(x,z,u_S) = \mu_S(x,z) - E\left(V|V = F_V^{-1}(u_S)\right) = \mu_S(x,z) - F_V^{-1}(u_S).$$

F_V^{-1} is strictly increasing, and thus $S^{MTE}(x,z,u_S)$ is strictly decreasing in u_S. Individuals with low U_S most want to enter the program and are those with the highest surplus from the program, while individuals with high U_S least want to enter the program and have the smallest surplus from the program. Again using that F_V is strictly increasing and that $P(X,Z) = F_V(\mu_S(X,Z))$, we have that conditioning on $U_S = P(x,z)$ is equivalent to conditioning on $V = \mu_S(x,z)$, and thus

$$S^{MTE}(x,z,P(x,z)) = \mu_S(x,z) - E\left(V|V = \mu_S(x,z)\right) = 0.$$

An individual with $u_S = P(x,z)$ is an individual who is indifferent between treatment or not if they are assigned $X = x, Z = z$. Since $S^{MTE}(x,z,u_S)$ is strictly decreasing in u_S, we have $S^{MTE}(x,z,u_S)$ is positive for $u_S < P(x,z)$, $= 0$ at $u = P(x,z)$, and is negative for $u_S > P(x,z)$. If we instead consider holding the u_S evaluation point fixed and consider how $S^{MTE}(x,z,u_S)$ varies with (x,z), we have that $S^{MTE}(x,z,u_S)$ will be positive for all (x,z) such that $P(x,z) > u_S$ and will be negative for all (x,z) such that $P(x,z) < u_S$.

We have thus far discussed only the marginal surplus function. Using that $S^{MTE}(x,z,u_S) = B^{MTE}(x,u) - C^{MTE}(z,u_S)$, we can translate these statements into relative statements about the marginal benefit and marginal cost functions. We thus have that

$$\begin{array}{ccccc} B^{MTE}(x,u_S) & > & C^{MTE}(z,u_S) & \forall & (x,z,u_S) \text{ s.t. } P(x,z) < u_S \\ B^{MTE}(x,u_S) & = & C^{MTE}(z,u_S) & \forall & (x,z,u_S) \text{ s.t. } P(x,z) = u_S \\ B^{MTE}(x,u_S) & < & C^{MTE}(z,u_S) & \forall & (x,z,u_S) \text{ s.t. } P(x,z) > u_S \end{array}$$

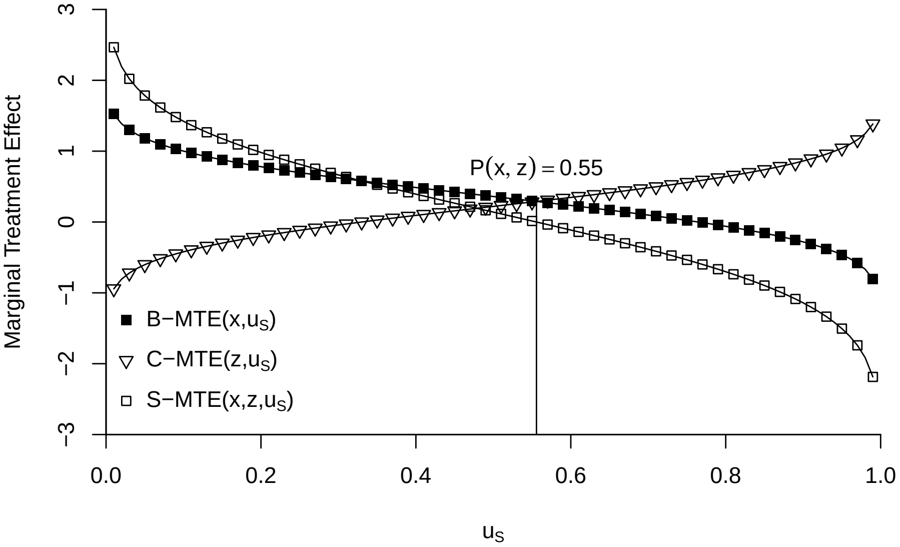
The two parameters coincide when evaluated at $u_S = P(x,z)$, because at this margin marginal cost equal the marginal benefit. The equality between marginal benefit and marginal cost for people at the margin will be used in the next session to allow identification of cost parameters.

To fix ideas, we show the full set of marginal effects of treatment for a numerical example in Figure (1).[3] Evaluated at fixed values of the observables (X,Z), agents select their treatment status based on gains unobservable by the econometrican. The marginal benefit of treatment $B^{MTE}(x,u_S)$ decreases when moving along the quantiles of the first-stage unobservable V. Agents with high values of u_S are unlikely to take up treatment. The opposite is true for the marginal cost of treatment $C^{MTE}(z,u_S)$. Agents wich are likely to select treatment not only have higher benefits, but also lower cost. For very low values of u_S, the cost of treatment is actually negative. Individuals just indifferent to treatment, $P(x,z) = 0.55$, the marginal surplus of treatment $S^{MTE}(x,z,u_S)$ is zero. The benefit of treatment are still positive, but so are cost. For the marginal agent, the benefits of treatment are just offset by the subjective cost.

Remark 3.1. Consider some special cases of the analysis. If benefits do not vary across individuals conditional on X, i.e. if $U_1 - U_0$ is degenerate, than $B^{MTE}(x,u_S) = B^{ATE}(x) = B^{TT}(x)$. In

[3] More details on the example are provided in the Web Appendix.

Figure 1: Marginal Effects of Treatment



addition, $U_1 - U_0$ degenerate implies that $V = U_C$ and $U_S = F_{U_C}(U_C)$ so that

$$\begin{aligned} C^{MTE}(z, u_S) &= \mu_C(z) + E(U_C|U_S = u_S) \\ &= \mu_C(z) + E(U_C|U_C = F_{U_C}^{-1}(u_S)) = \mu_C(z) + F_{U_C}^{-1}(u_S) \end{aligned}$$

which is increasing in u_S. In this case, variation in unobserved costs drives selection conditional on (X, Z) and those who most want to enter the program (have lowest U_S) have the least costs of treatment. Using equation (3.2) and that $C^{MTE}(z, u_S)$ is increasing in u_S, we also have $C^{TT}(z) < C^{ATE}(z)$ so that, conditional on Z, those who chose treatment have lower cost of treatment than those who did not select into treatment. Symmetrically, if costs do not vary across individuals conditional on Z, i.e. if U_C is degenerate, then heterogeneity in the benefits of treatment drive selection and (1) $C^{MTE}(x, u_S) = C^{ATE}(x) = C^{TT}(x)$; (2) $B^{MTE}(x, u_S)$ is decreasing in u_S; and (3) $B^{TT}(x) > B^{ATE}(x)$.

We have shown that the marginal surplus parameter is highest for those who most want to participate in the program. Adding equation (3.3) we thus have that the average surplus among the treated is higher than the unconditional average surplus of treatment. As discussed in Remark (3.1), degeneracy of either $U_1 - U_0$ or of U_C implies that treatment parameters and cost parameters will have intuitive properties, such as highest benefit or lowest cost for those who most want to participate in the program. We now show a more general set of conditions under which these properties of the treatment effect parameters and cost parameters will hold.

Theorem 1. *Assume that equations (2.1) – (2.5) and (2.5) and our assumptions (A-1) – (A-5) hold.*

1. $S^{TT}(x) > S^{ATE}(x)$*, and* $S^{MTE}(x, u_S)$ *is monotonically decreasing in* u_S*.*
2. *Suppose* $U_C \perp\!\!\!\perp U_1 - U_0$*. Then* $C^{TT}(z) \leq C^{ATE}(z)$*,* $B^{TT}(x) \geq B^{ATE}(x)$*.*
3. *Suppose* $U_C \perp\!\!\!\perp U_1 - U_0$*, and that* U_C *and* $U_1 - U_0$ *have log concave densities. Then* $C^{MTE}(z, u_S)$ *is monotonically increasing in* u_S *and* $B^{MTE}(x, u_S)$ *is monotonically decreasing in* u_S*.*

Proof. Assertion (1) was proven in the preceding text. For assertion (2), first consider the cost parameters. $C^{ATE}(z) - C^{TT}(z) = E(U_C) - E(U_C|Z = z, D = 1)$, and $E(U_C|Z = z, D = 1) = \int E(U_C|Z = z, X = x, U_S \leq P(x, z))dF_{X|Z}(x|z) = \int E(U_C|U_S \leq P(x, z))dF_{X|Z}(x|z)$ using $(X, Z) \perp\!\!\!\perp (U_C, U_S)$. Thus, using that $U_S = F_V(V)$, it will be sufficient to show that $E(U_C|V \leq t) \leq E(U_C)$ for all t, and thus sufficient to show that $\Pr[U_C \leq s|U_C - (U_1 - U_0) \leq t] \geq \Pr[(U_C \leq s]$. Using Bayes' rule, this is equivalent to $\Pr[U_C - (U_1 - U_0) \leq t|U_C \leq s] \geq \Pr[U_C - (U_1 - U_0) \leq t]$, and this last assertion can now easily be shown using $U_C \perp\!\!\!\perp (U_1 - U_0)$. We can thus conclude that $C^{ATE}(z) - C^{TT}(z) \geq 0$. The same argument *mutatis mutandis* shows that $B^{ATE}(x) - B^{TT}(x) \leq 0$. Now consider assertion (3). The densities of U_C and $U_1 - U_0$ being log concave is equivalent to their densities being Polya frequency functions of order 2 (PF2) (Klein (1968)). Using that $U_1 - U_0 \perp\!\!\!\perp U_C$, one can now easily verify that $(U_C, U_C - (U_1 - U_0))$ and $(-(U_1 - U_0), U_C - (U_1 - U_0))$ have joint densities that are totally positive of order 2 (TP2). By Joe (1968) (Theorems 2.2, 2.3), $(U_C, U_C - (U_1 - U_0))$ and $(-(U_1 - U_0), U_C - (U_1 - U_0))$ having TP2 densities implies that U_C and $-(U_1 - U_0)$ are stochastically increasing in $U_C - (U_1 - U_0)$ and thus stochastically increasing in U_S using that U_S is a strictly monotonic function of $U_C - (U_1 - U_0)$. Thus $E(U_C|U_S = u_S)$ is increasing in u_S while $E(U_1 - U_0|U_S = u_S)$ is decreasing in u, establishing the assertion. □

The theorem provides intuitive results. If the unobservables related to the cost and benefit are independent, then the average benefit among those who select into treatment is larger than the unconditional average benefit. At the same time, the average cost among those who select into treatment is lower than the unconditional average cost. In other words, under the independence of the unobservables related to benefits and costs, it is the high benefit and low cost individuals who select into treatment in the generalized Roy model. Part (2) and (3) of the theorem state that, under a regularity condition, the expected gain is decreasing while the expected cost is increasing in U_S. Note that the normal density as well as many other standard densities are log concave.[4]

As the numerical example previously introduced in Figure (1) is based on unobservables drawn from a normal distribution with uobservable benefits $(U_1 - U_0)$ independent of the uobservable component U_C of cost, the marginal effects of treatment exhibit the shape predicted by Theorem (1), Assertion (3). The $B^{MTE}(x, u_S)$ is decreasing in U_S, while the opposite is true for $C^{MTE}(z, u_S)$.

[4] Heckman and Honoré (1990) exploit the restriction of log-concave density functions for the disturbance terms in a Roy model with zero costs. See Bagnoli and Bergstrom (2005) for a review of log concave densities and economic applications.

4 Identification Analysis

Following Heckman and Vytlacil (1999, 2005a), we have that local instrumental variables (LIV) identifies the marginal benefit of treatment:

$$\frac{\partial}{\partial p} E(Y|X = x, P = p) = B^{MTE}(x, p). \tag{4.1}$$

We identify $E(Y|X = x, P = p)$ and its derivative for all $(x, p) \in \text{Supp}(X, P)$, where $\text{Supp}(X, P)$ denotes the support of $(X, P(X, Z))$.[5] We thus have identification of $B^{MTE}(x, u_S)$ for all values of $(x, u_S) \in \text{Supp}(X, P)$. For a fixed x, we identify $B^{MTE}(x, u_S)$ for $u_S \in \text{Supp}(P|X = x)$. The more variation in propensity scores conditional on $X = x$, the larger the set of evaluation points u_S for which we identify $B^{MTE}(x, u_S)$. Variation in propensity scores conditional on X is driven by variation in Z, the cost shifters. Thus, if we observe regressors that cause large variations in costs, we will be able to identify $B^{MTE}(x, u_S)$ at a larger set.

We can identify $B^{ATE}(x)$ and $B^{TT}(x)$ by identifying $B^{MTE}(x, u_S)$ over the appropriate support and then integrating the latter with the appropriate weights. By equation (3.1), we identify $B^{ATE}(x)$ if $\text{Supp}(P|X = x) = [0, 1]$. This requires, for fixed $X = x$, there to be enough variation in the cost shifters Z to drive the probabilities $P(x, Z)$ all the way to zero and to one. In other words, holding fixed the regressors that enter the outcome equation, we must observe costs shifters such that conditional on some values of those cost shifters, the cost to the agent is so low that the agent will select into treatment with probability one; and conditional on other values of the cost shifters, the cost to the agent is so high that the agent will select into treatment with probability zero. Likewise, we identify $B^{TT}(x)$ if $\text{Supp}(P|X = x) = [0, p_x^{max}]$ where p_x^{max} is the supremum of $\text{Supp}(P|X = x)$. This support requirement in turn requires that, for fixed $X = x$, for there to be enough variation in the cost shifters Z to drive the selection probability to zero.[6]

Using equation (4.1) and that $B^{MTE}(x, P(x, z)) = C^{MTE}(z, P(x, z))$, we have

$$\frac{\partial}{\partial p} E(Y|X = x, P = p)\Big|_{p=P(x,z)} = C^{MTE}(z, P(x, z)).$$

[5] For any random vectors A and B, we will write the support of the distribution of A as $\text{Supp}(A)$, and the support of distribution of A conditional on $B = b$ as $\text{Supp}(A|B = b)$.

[6] As shown by Heckman and Vytlacil (2001), we can identify $B^{ATE}(x)$ and $B^{TT}(x)$ under weaker conditions than those required to follow this strategy of first identifying $B^{MTE}(x, u)$ over the appropriate support.

We thus identify $C^{MTE}(z, u_S)$ for all values of $(z, u_S) \in \text{Supp}(Z, P)$. One can thus identify the marginal cost parameter without direct information on the cost of treatment by using the structure of the Roy model and by identifying the marginal benefit of treatment for individuals at the margin of participation. For a fixed z, we identify $C^{MTE}(z, u_S)$ for $u_S \in \text{Supp}(P|Z = z)$. The more variation in propensity scores conditional on $Z = z$, the larger the set of evaluation points for which we identify $C^{MTE}(z, u_S)$. Variation in propensity scores conditional on $Z = z$ is driven by variation in X, the regressors that drive the outcome. Thus, if we observe regressors that cause large variations in benefits, we will be able to identify $C^{MTE}(z, u_S)$ at a larger set of u_S evaluation points. In contrast, if there are no X regressors, then P only depends on Z and we can only identify $C^{MTE}(z, u_S)$ for $u_S = P(z)$.

By equation (3.2), we thus identify $C^{ATE}(x)$ if $\text{Supp}(P|Z = z) = [0, 1]$. This requires, for fixed $Z = z$, for there to be enough variation in the outcome shifters X to drive the probabilities $P(X, Z)$ all the way to zero and to one. In other words, holding fixed the regressors that enter the cost equation, we must observe outcome shifters such that conditional on some values of those outcome shifters, the benefit to the agent is so high that the agent will select into treatment with probability one; and conditional on other values of the outcome shifters, the benefit to the agent is so high that the agent will select into treatment with probability zero. Likewise, we thus identify $C^{TT}(x)$ if $\text{Supp}(P|Z = z) = [0, p_z^{max}]$ where p_z^{max} is the supremum of $\text{Supp}(P|Z = z)$. This support requirement in turn requires that, for fixed $Z = z$, for there to be enough variation in the outcome shifters X to drive the probabilities to zero.

Finally, consider identification of the surplus parameters. Using that $S^{MTE}(x, z, u_S) = B^{MTE}(x, u_S) - C^{MTE}(z, u_S)$, we identify the marginal surplus parameter at (x, z, u_S) such that $(x, u_S) \in \text{Supp}(X, P)$ and $(z, u_S) \in \text{Supp}(Z, P)$. By equation (3.3), we can now integrate $S^{MTE}(x, z, u_S)$ using the appropriate weights to identify $S^{ATE}(x, z)$ and $S^{TT}(x, z)$ under the appropriate support conditions. For example, we identify $S^{ATE}(x, z)$ if $\text{Supp}(P|X = x) = [0, 1]$ and $\text{Supp}(P|Z = z) = [0, 1]$.

Thus, for identification of the treatment parameters we need sufficient variation in cost shifters conditional on the outcome shifters; for identification of the cost parameters we need sufficient variation in the outcome shifters conditional on the cost shifters; and for identification of the surplus parameters we need sufficient variation in both sets of regressors. We can thus identify the marginal cost, the average cost, and the cost of treatment on the treated parameters and the corresponding surplus parameters without direct information on the cost of treatment. Our ability to do so is directly related to the extend of variation in observed regressors that shift the benefit of the treatment.

We summarize this discussion in the form of a theorem:

Theorem 2. *Assume that equations (2.1) – (2.5) and (2.5) and our assumptions (A-1) – (A-5) hold.*

1. $B^{MTE}(x, u_S)$ *is identified for* $(x, u_S) \in$ *Supp*(X, P); $C^{MTE}(z, u_S)$ *is identified for* $(z, u_S) \in$ *Supp*(Z, P); *and* $S^{MTE}(x, z, u_S)$ *is identified for* (x, z, u_S) *such that* $(x, u_S) \in$ *Supp*(X, P) *and* $(z, u_S) \in$ *Supp*(Z, P).

2. $B^{ATE}(x)$ *is identified if Supp*$(P|X = x) = [0, 1]$; $C^{ATE}(z)$ *is identified if Supp*$(P|Z = z) = [0, 1]$; $S^{ATE}(x, z)$ *is identified if Supp*$(P|X = x) = [0, 1]$ *and Supp*$(P|Z = z) = [0, 1]$.

3. $B^{TT}(x)$ *is identified if Supp*$(P|X = x) = [0, p_x^{max}]$; $C^{TT}(z)$ *is identified if if Supp*$(P|Z = z) = [0, p_z^{max}]$; $S^{TT}(x, z)$ *is identified if Supp*$(P|X = x) = [0, p_x^{max}]$ *and Supp*$(P|Z = z) = [0, p_z^{max}]$.

Remark 4.1. As discussed in Remark (3.1), if there is no unobserved heterogeneity in costs of treatment, $U_C = 0$, then $C^{MTE}(z, u_S) = C^{TT}(z) = C^{ATE}(z)$. Thus, in the case of no unobserved heterogeneity in costs of treatment, we immediately identify the cost of treatment on the treated and average cost parameters without the additional support conditions. Likewise, if there is no unobserved heterogeneity in the treatment effects, $U_1 - U_0 = 0$, we have $B^{MTE}(z, u_S) = B^{TT}(z) = B^{ATE}(z)$ and thus identify all of the treatment effect parameters without additional support conditions.

We have thus far considered identification of $B^{ATE}(x) = \mu_1(X) - \mu_0(X)$, and of $C^{ATE}(z) = \mu_C(z)$. We can identify $\mu_1(X) - \mu_0(X)$ and $\mu_C(z)$ up to a location shift under weaker conditions than those required to full identify the functions. From the analysis of the previous section, we identify $B^{MTE}(x, p) = \mu_1(x) - \mu_0(x) + E(U_1 - U_0|U_S = p)$. By varying x holding p constant, we trace out $\mu_1(x) - \mu_0(x)$ up to an additive constant. Likewise, consider $C^{MTE}(z, p) = \mu_C(z) + E(U_C|U_S = p)$. Varying z holding p fixed for the marginal cost parameter, we identify $\mu_C(z)$ up to an additive constant. Given our previous identification analysis, we can identify $B^{MTE}(x, p)$ over $x \in \text{Supp}(X|P = p)$ and $C^{MTE}(z, p)$ over $z \in \text{Supp}(Z|P = p)$, but not over the unconditional supports of X and Z. Thus, we immediately have identification of shifts in $\mu_1(x) - \mu_0(x)$ for $x \in \text{Supp}(X|P = p)$ and of $\mu_C(z)$ for $z \in \text{Supp}(Z|P = p)$ for some $p \in \text{Supp}(P)$, but not immediately, e.g., of $\mu_C(z_0) - \mu_C(z_1)$ if there does not exist a p such that $z_0, z_1 \in \text{Supp}(Z|P = p)$. However, given a rank condition, we can combine information across different values of p to identify $\mu_C(z)$ and $\mu_1(x) - \mu_0(x)$ up to an additive constant for all z and x in their unconditional supports. In particular, consider the following rank assumption.

(A-6) *X and $P(X, Z)$ are measurably separated, i.e., any function of X that almost surely equals a function of $P(X, Z)$ must be almost surely equal to a constant.*

Theorem 3. *Assume that equations (2.1) – (2.5) and our assumptions (A-1) – (A-5) hold. Additionally, suppose (A-6) holds. Then $\mu_C(\cdot)$ is identified over the support of Z up to an additive constant, and $\mu_1(\cdot) - \mu_0(\cdot)$ is identified over the support of X up to an additive constant.*

Proof. Let $\mu_{10}(\cdot) = \mu_1(\cdot) - \mu_0(\cdot)$. From our previous analysis, we have

$$\frac{\partial}{\partial p} E(Y|X = x, P = p) = \mu_{10}(x) + \Upsilon(p) \quad \text{a.e. } (x, p) \tag{4.2}$$

where $\Upsilon(p) = E(U_1 - U_0 \mid U_S = p)$. Let μ_{10}^1, Υ^1 and μ_{10}^0, Υ^0 denote two candidate functions that both satisfy equation (4.2) for a.e. (x, p). We then have $\mu_{10}^1(x) - \mu_{10}^0(x) = \Upsilon^0(p) - \Upsilon^1(p)$ for a.e. (x, p). By the rank condition (A-6), we have $\mu_{10}^1(x) - \mu_{10}^0(x)$ equals a constant for a.e. x, so that μ_{10} is identified up to an additive constant. The same argument *mutatis mutandis* shows identification of μ_C up to an additive constant. □

Measurable separability between X and P is a rank condition. As discussed by Florens et al. (2006), measurable separability is a relatively weak regularity condition in this context. See their paper for more discussion of this condition, including sufficient conditions for measurable separability.

Finally, consider testable restrictions on $E(Y|X = x, P = p)$ as a function of p that result from additional restrictions including those considered in Theorem (1).

Theorem 4. *Assume that equations (2.1) – (2.5) and (2.5) and our assumptions (A-1)–(A-5) hold.*

1. *Suppose that $U_1 - U_0$ is degenerate. Then $E(Y|X = x, P = p)$ is linear in p.*

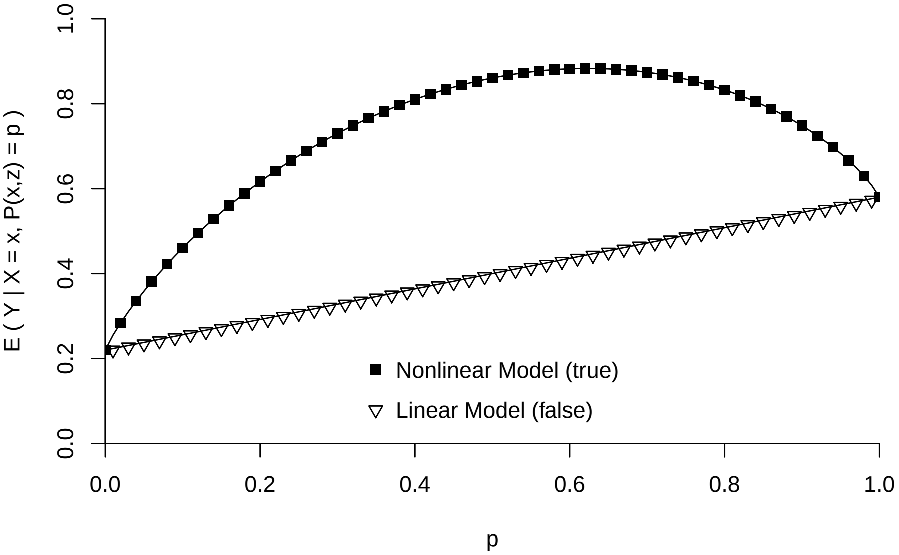
2. *Suppose $U_1 - U_0 \perp\!\!\!\perp U_C$. For a fixed x, consider a line $a + bp$, where $a = E(Y|X = x, P(X, Z) = 0)$ and $b = E(Y|X = x, P(X, Z) = 1) - E(Y|X = x, P(X, Z) = 0)$. Then $E(Y|X = x, P(X, Z) = p) \geq a + bp$ for all $p \in$Supp$(P|X = x)$.*

3. *Suppose $U_1 - U_0 \perp\!\!\!\perp U_C$, and suppose $U_1 - U_0$ and U_C have log concave densities. Then $E(Y|X = x, P(X, Z) = p)$ is a concave function of p.*

Proof. Assertion (1) follows from equation (4.1) and $B^{MTE}(x, u_S) = \mu_1(x) - \mu_0(x)$ if $U_1 - U_0$ is degenerate. Assertion (2) follows from $E(Y|X = x, P(X, Z) = 1) - E(Y|X = x, P(X, Z) = 0) = B^{ATE}(x)$, $[E(Y|X = x, P(X, Z) = p) - E(Y|X = x, P(X, Z) = 0)]/p = E(B|X = x, P(X, Z) = p, D = 1)$, and that $B^{ATE}(x) \leq E(B|X = x, P(X, Z) = p, D = 1)$ by the arguments used to prove assertion (2) of Theorem (1). Assertion (3) follows from equation (4.1) and Assertion (3) of Theorem (1). □

Recalling, that the unobservables in the numerical example are all normal, but independently, distributed, Figure (2) depicts the corresponding $E\,(\,Y_1 - Y_0\,|\,X = x, P(X, Z) = p\,)$. As perdicted by Theorem (4), it is concave. This is a direct consequence of the fact that those agents with a high propensity of treatment (low values of u_S) have the highest gains even after conditioning on observables. As the p increases, the share of inviduals participating increases constantly, but at the same time the gain for agents at the margin decreases. Individuals with high values of V, which enter treatment only for high values of p, have the least to gain from treatment.

Figure 2: Testable Implication



5 Application

Following Carneiro et al. (2011), we estimate the marginal effects of treatment for a sample of white males from the National Longitudinal Survey of Youth of 1979 (NLSY) imposing full independence between the observables (X, Z) and unobservables (U_1, U_0, U_C) of the model. We group individuals in two groups: persons with a high school education or below who do not go to college (D = 0) and persons with some college or above (D = 1). The outcome variable is the log of the average of non-missing values of the hourly wage between 1989 and 1993, which we interpret as an estimate of the log hourly wage in 1991. Schooling is measured in 1991 when individuals are between 28 and 34 years of age. As described in Heckman et al. (2006), we estimate the B^{MTE} using a semiparametric version of local instrumental variables relying on a probit estimate for the selection probability. We present annualized returns, obtained my dividing the marginal effects of treatment by four (corresponding to four years of college) and assume a linear-in-parameter version of the generalized Roy model presented, thus $Y_1 = X\beta_1 + U_1$ and $Y_0 = X\beta_0 + U_0$.

In our specification, we include regressors that affect benefits as well as cost. Thus, X denotes all variables in the outcome equation, whereas Z indicates all variables that affect choice. The central requirement to be able to identify the full set of marginal effects of treatment is the availability of two types of exclusion restrictions. First, to identify the B^{MTE} we require cost-shifters in Z, that do not affect the benefits of treatment. While conditioning on permanent local labor market conditions, we use short-run fluctuations at the time of the treatment decision in labor market conditions that only affect the cost of treatment, but leave benefits unchanged. Second, for identification of C^{MTE}, we require in addition, access to benefit-shifters in X, that do not affect the cost of treatment. For this, we use long-run labor market conditions in adolescence.[7]

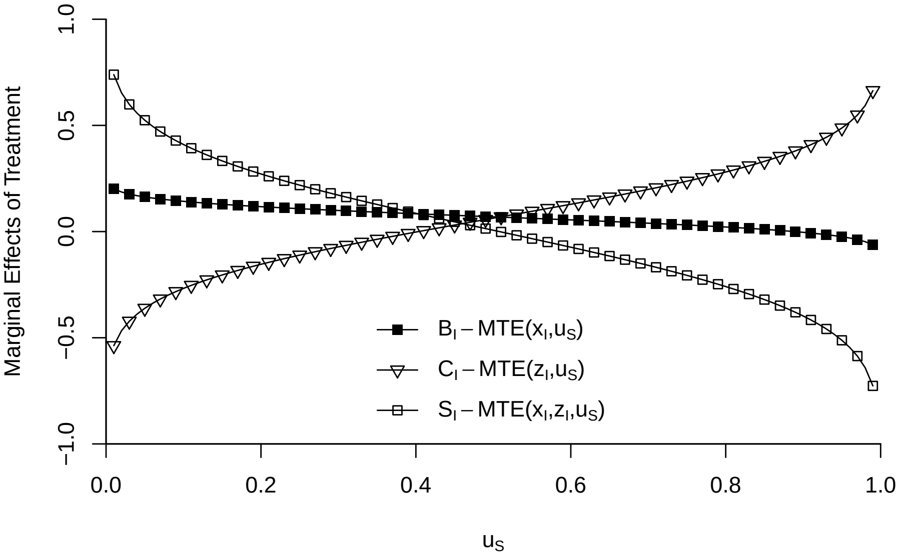
An additional complication arises, as some determinants of benefits are unknown to agent at the time when making their treatment decision. In our application, this refers to the local labor market environment in 1991. However, as shown in Appendix (A), the analysis extends easily to this case if agents form rational expectations based only on observables known to them at time of treatment. A central aspect of this type of analysis is the fact that LIV identifies ex-post benefits, not the benefits as perceived by the agent at time of treatment. In an intermediate step, the B^{MTE} needs to be projected on the information set available at that time. The resulting perceived benefits are then used to identify perceived cost and surplus. To contrast these to the ex-post

[7] See our Web Appendix for more details on the dataset and exact specification.

realized counterparts, we add the subscript I.

The perceived marginal effects of treatment are depicted in Figure (3) evaluated at the mean of the observable characteristics (X_I, Z_I). Agents select their treatment based on gains unobservable to the econometrician. Individuals with a high propensity for treatment have higher gains. These range from 40% for individuals with very low values of u_S to -20% at the upper tail of the distribution of V. The opposite shape is true for cost, these are initially negative for a small subset of agents, but strictly increasing when moving along the quantiles of V. Given the point of evaluation, the individuals at the margin are $u_S = P(\bar{x}, \bar{z}) = 0.45$. Individuals at this quantile are just indifferent towards treatment. Their surplus is zero as their benefit are just offset by their cost from participation.

Figure 3: Marginal Effects of Treatment

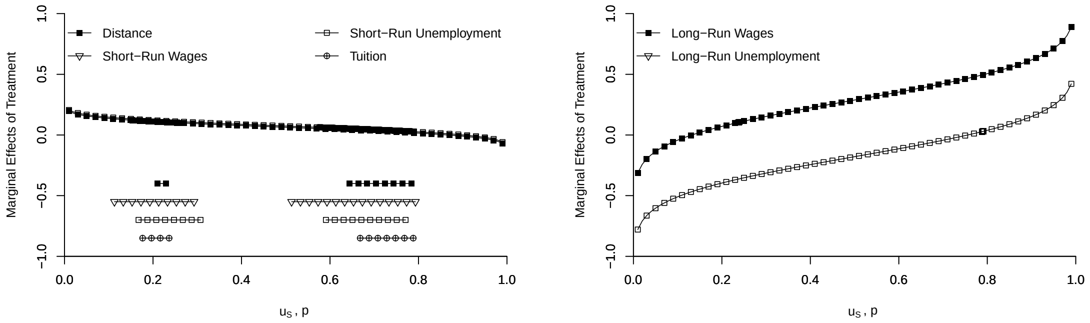


Assuming independence between all observables and unobservables of the model is stronger assumption than required for identification. The identification analysis in Section (4) was performed conditioning implicitly on all common elements that affect benefits as well as cost of treatment. Independence from the unobservable V was only crucial for the exclusions for those observables that shift cost independent of benefits and those that shift benefits independent of cost. When imposing full independence, it is the unconditional support of P that determines the range of identification for the marginal effects of treatment. However, if this is not the case, the range over which we can identify the marginal effects of treatment is stated in Theorem (2). It is the variation in P conditional on the X (and all common elements) that matters for the B^{MTE} and the variation in P conditional on Z (and all common elements) for the C^{MTE}

To emphasize this issue, we graph the area of local identification for the marginal effects of treatment in Figure (4). In each figure, we plot the marginal benefit or marginal cost of treatment evaluated at the 25^{th} and 75^{th} percentile of their index functions $(\mu_1(X) - \mu_0(X))$ and $\mu_C(Z)$. Then we show, which part of the marginal effect is identified by what instrument. We do so, by fixing all instruments at their mean value at the analyzed quantile of the index functions and then allowing each one to vary separately.[8] The continuously plotted part of the marginal cost and benefits of treatment that can be identified if all the relevant exclusions are allowed to vary. Figure

(4) makes clear, which exclusions aid in the identification of the marginal effects of treatment and for which portion. Using multiple instruments at once allows to increases the area of identification. Concerning marginal cost, the variation in long-run wages aids most in the identification. Recalling, that marginal surplus of treatment is identified only in the area of overlap of the local benefits and cost function, this highlights the importance of having both strong cost shifters and powerful regressors that shift the benefit of the program for a comprehensive policy analysis.

Figure 4: Local Identification



(a) $B_I^{MTE}(x_I, u_S)$

(b) $C_I^{MTE}(z_I, u_S)$

[8] See our Web Appendix for details on the implementation.

6 Conclusion

This pape extends the analysis of Heckman and Vytlacil (1999, 2005b, 2007) by using the marginal benefit of treatment B^{MTE} to identify the subjective cost and surplus of treatment. An empirical application from educational choice illustrates the empirical feasability of this analysis.

References

Bagnoli, M. and T. Bergstrom (2005). Log-concave probability and its applications. *Economic Theory 26*, 445–469.

Björklund, A. and R. Moffitt (1987, February). The estimation of wage gains and welfare gains in self-selection. *Review of Economics and Statistics 69*(1), 42–49.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating Marginal and Average Returns to Education. *American Economic Review*, forthcoming.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. J. Vytlacil (2006). Control functions for nonparametric models without large support. Unpublished manuscript, University of Chicago.

Heckman, J. J. and B. E. Honoré (1990, September). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*, 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2001). Instrumental variables, selection models, and tight bounds on the average treatment effect. In M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluation of Labour Market Policies*, pp. 1–15. New York: Center for European Economic Research.

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

Heckman, J. J. and E. J. Vytlacil (2005b). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica 73*(3), 669–739.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast their Effects in New Environments. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144. Amsterdam, NL: Elsevier Science.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Joe, H. (1968). *Multivariate Models and Dependence Concepts.* Stanford: Chapman and Hall.

Klein, S. (1968). *Total Positivity: Volume 1.* Stanford: Stanford University Press.

Quandt, R. E. (1958). The Estimatin of the Parameters of a Linear Regression System Obeying two Separate Regimes. *Journal of the American Statistical Association 53*(284), 873–880.

Quandt, R. E. (1973). A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association 67*(338).

Roy, A. (1951, June). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Shaikh, A. M. and E. J. Vytlacil (2005). Threshold crossing models and bounds on treatment effects: A nonparametric analaysis. Unpublished manuscript, Columbia University and University of Chicago.

Vytlacil, E. J. (2002, January). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Vytlacil, E. J. and N. Yildiz (2007). Dummy endogenous variables in weakly separable models. *Econometrica*. forthcoming.

A Extension to Limited Information by Agent

Our analysis thus far has imposed equation (2.5), that is $D = \mathbf{1}[S \geq 0]$ where $S = (Y_1 - Y_0) - C$. Hence assumed that agents have perfect foresight of their individual benefit of treatment $B = Y_1 - Y_0$ as well as cost C. We now relax the model of equation (2.5) to allow limited information on the part of the agents, while maintaining the model on latent outcomes Y_0, Y_1 and cost C of equations (2.2) and (2.4). We impose that agents form valid expectations of their outcomes and costs given the information that they have at the time of treatment choice and that they select into treatment if the expected surplus is positive. We allow agents to know only some elements of (X, Z), and to possibly have incomplete knowledge of (U_0, U_1, U_C) and thus of their own idiosyncratic benefit and cost of treatment. We now show that the analysis essentially goes through with minor modifications, though it is now important to distinguish conditioning sets that condition on what is known to the agent at the time of treatment choice (which might include some information not known to the econometrician), what is known to the econometrician (which might include some information not known to the agent at the time of treatment choice), and what is realized ex post. The essential change in the procedure under incomplete information is that the marginal benefit of treatment identified by LIV must be projected onto the information set when selecting treatment to form the expected marginal benefit of treatment conditional on the information available to the agent. It is this coarsened version of B^{MTE} used to identify the marginal cost parameter. In addition, only components of X which are known to the agent at the time of treatment choice can aid in identification of the cost parameters, so that the exclusion restriction for identification of the cost parameter are variables in X that are not in Z and which were known to the agent at the time of treatment selection.

Let (X_I, Z_I) denote components of (X, Z) that are observed by the agent when choosing whether to select into treatment. Suppose that the agent's information set equals (X_I, Z_I, U_I).[9] U_I is the private information of the agent relevant to their own benefits and cost of treatment, and is not observed by the econometrician.

Restate assumption (A-1) as

$$(U_0, U_1, U_C, U_I) \perp\!\!\!\perp (X, Z),$$

[9] In other words, the information set of the agent equals $\sigma(X, Z, U_I)$, the sigma-algebra generated by (X, Z, U_I).

so that the private information of the agent is independent of the observed regressors. Note that, under this independence assumption,

$$E(V|X,Z,U_I)=E(V|X_I,Z_I,U_I)=E(V|U_I)$$

using that $V=U_C-(U_1-U_0)$. Furthermore, given this assumption, $(X,Z)\perp\!\!\!\perp U_I|(X_I,Z_I)$, so that U_I does not help the agent predict elements of (X,Z) that are not contained in (X_I,Z_I). Thus, we are allowing the agents to have private information about their own idiosyncratic benefits (U_1-U_0) and costs U_C, though we are imposing the restriction that the agent's information relevant to (X,Z) is only that they know some components (X_I,Z_I).

Restate assumption (A-3) as the distribution of $\tilde{V}=E(V|U_I)$ is absolutely continuous with respect to Lebesgue measure, and the cumulative distribution function of $\tilde{V}$ is strictly increasing. We are thus requiring that the agent has some nontrivial information on their own cost or benefit from treatment, we are not allowing $E(V|U_I)$ to be a degenerate random variable. Maintain assumptions (A-2), (A-4), and (A-5) as before.

Define $\mu_j^I(X_I)=E(Y_j|X_I)$ for $j=0,1$, and $\mu_C^I(Z_I)=E(C|Z_I)$, and note that given our independence assumptions and the law of iterated expectations we have $\mu_j^I(X_I)=E(\mu_j(X)|X_I)$, $\mu_C^I(Z_I)=E(\mu_C(Z)|Z_I)$. Define $\mu_S^I(X_I,Z_I)=E(S|X_I,Z_I)$. Given these assumptions we have

$$E(S|X_I,Z_I,U_I)=\mu_S^I(X_I,Z_I)-\tilde{V}=\mu_1^I(X_I)-\mu_0^I(X_I)-\mu_C^I(Z_I)-\tilde{V}.$$

The previous decision rule, equation (2.5), under perfect certainty is now replaced with

$$D=1 \quad \text{if} \quad E(S|X_I,Z_I,U_I)\geq 0\,; \qquad D=0 \quad \text{otherwise}, \tag{A.1}$$

where $E(S|X_I,Z_I,U_I)$ is the expected surplus, i.e. net gain, from treatment, with the expectation conditional on the agents information set. We thus have

$$D=\mathbf{1}[\mu_S^I(X_I,Z_I)-\tilde{V}\geq 0]$$

where our independence assumptions imply $\tilde{V}\perp\!\!\!\perp(X_I,Z_I)$, and thus the selection model is of the same form as Heckman and Vytlacil (1999), which allows to use LIV to identify B^{MTE}. Redefining

$U_S = F_{\tilde{V}}(\tilde{V})$ and $P(X_I, Z_I) = \Pr[D = 1|X_I, Z_I] = F_{\tilde{V}}(\mu_S^I(X_I, Z_I))$, we have

$$D = \mathbf{1}[P(X_I, Z_I) - U_S \geq 0]$$

with U_S distributed unit uniform and independent of (X, Z) and thus independent of (X_I, Z_I).

Define $B_I^{MTE}(x_I, u_S) \equiv E(Y_1 - U_0|X_I = x_I, U_S = u_S)$, $C_I^{MTE}(z_I, u_S) \equiv E(C|Z_I = z_I, U_S = u_S)$, and $S_I^{MTE}(x_I, z_I, u_S) \equiv B_I^{MTE}(x_I, u_S) - C_I^{MTE}(z_I, u_S)$, the marginal benefit, cost, and net surplus of treatment conditional on the agent's information set, where again by the law of iterated expectations and our independence assumptions

$$\begin{array}{rcccl} B_I^{MTE}(x_I, u_S) & = & E(B^{MTE}(X, u_S)|X_I = x_I, U_S = u_S) & = & E(B^{MTE}(X, u_S)|X_I = x_I) \\ C_I^{MTE}(z_I, u_S) & = & E(C^{MTE}(Z, u_S)|Z_I = z_I, U_S = u_S) & = & E(C^{MTE}(Z, u_S)|Z_I = z_I) \end{array}$$

Evaluating $S_I^{MTE}(x_I, z_I, u_S)$ at $u_S = P(x_I, z_I)$, we have

$$\begin{array}{rcl} S_I^{MTE}(x_I, z_I, P(x_I, z_I)) & = & \mu_S^I(x_I, z_I) - E(V|U_S = P(x_I, z_I)) \\ & = & \mu_S^I(x_I, z_I) - E(V|\tilde{V} = \mu_S^I(x_I, z_I)) \\ & = & \mu_S^I(x_I, z_I) - E(V|E(V|U_I) = \mu_S^I(x_I, z_I)) \\ & = & \mu_S^I(x_I, z_I) - E(E(V|U_I)|E(V|U_I) = \mu_S^I(x_I, z_I)) \\ & = & \mu_S^I(x_I, z_I) - \mu_S^I(x_I, z_I) \\ & = & 0 \end{array}$$

where the second equality is plugging in the definition of U_S, the third equality is plugging in the definition of $\tilde{V}$, and the fourth equality is using law of iterated expectations and that $E(V|U_I)$ is degenerate given U_I. Since $S_I^{MTE}(x_I, z_I, u_S) = B_I^{MTE}(x_I, u_S) - C_I^{MTE}(z_I, u_S)$, we have

$$B_I^{MTE}(x_I, u_S) = C_I^{MTE}(z_I, u_S) \text{ for } u_S \text{ such that } u_S = P(x_I, z_I)$$

thus, identification of $B_I^{MTE}(x_I, P(x_I, z_I))$ provides identification of $C_I^{MTE}(z_I, P(x_I, z_I))$.

Since our model is a special case of Heckman and Vytlacil (1999) and we can follow them in using LIV to identify $B^{MTE}(x, u_S)$ for (x, U_S) in the support of $(X, P(X_I, Z_I))$. It is important to note that LIV does not identify the B^{MTE} that is relevant to the agent's decision problem, LIV identifies $B^{MTE}(x, u_S) = E(Y_1 - Y_0|X = x, U_S = u_S)$, not $B_I^{MTE}(x_I, u_S) = E(Y_1 - Y_0|X_I = x_I, U_S = u_S)$.

However, we can project the B^{MTE} identified by LIV on the information known to the agent at the time of treatment selection, coarsen B^{MTE}, to identify the B_I^{MTE} relevant to the agent's decision problem. It is the latter, that is relevant for identifying cost. By the law of iterated expectations (and using that X_I is degenerate given X and that U_I is independent of X), we have

$$B_I^{MTE}(x_I, u_S) = E(B^{MTE}(X, u_S)|X_I = x_I)$$

for (x_I, U_S) in the support of $(X_I, P(X_I, Z_I))$. Using that $B_I^{MTE}(x_I, P(x_I, z_I)) = C_I^{MTE}(z_I, P(x_I, z_I))$, we identify $C_I^{MTE}(z_I, u_S)$ for (z_I, u) in the support of $(Z_I, P(X_I, Z_I))$. We have thus identified the marginal cost parameter, and can integrate it to obtain other cost parameters, and combine it with the benefit parameters to identify net surplus parameters as before. The only additional change is that the only elements of X that are useful to identification of the cost parameter are those elements that are in X, not in Z, and which are known to the agent at the time of selection into treatment (are in X_I).