Evaluation Estimators Under Essential Heterogeneity

Advanced Microeconometrics
ITAM

Instructor: Cristián Sánchez

Spring 2022

Background Papers

- Heckman and Navarro (2003) (Matching)
- Card (2001) (IV)
- Imbens and Angrist (1994) (LATE)
- Heckman and Robb (1985) (Control Function)
- Heckman and Vytlacil (2007) (Treatment Parameters)
- Heckman, Urzua, and Vytlacil (2006) (Essential Heterogeneity)

A Prototypical Policy Evaluation Problem

- To motivate the discussion, consider the following prototypical policy problem.
- Suppose a policy is proposed for adoption in a country.
- It has been tried in other countries and we know the outcomes there. We also know outcomes in countries where it was not adopted.
- From the historical record, what can we conclude about the likely effectiveness of the policy in countries that have not implemented it?

- To answer questions of this sort, economists build models of counterfactuals.
- Consider the following model.
- Let Y_0 be the outcome of a country (e.g. GDP) under a no-policy regime. Y_1 is the outcome if the policy is implemented.
- $Y_1 - Y_0$ is the "treatment effect" of the policy. It may vary among countries.
- We observe characteristics X of various countries (e.g. level of democracy, level of population literacy, etc.).

- It is convenient to decompose Y_1 into its mean given X, $\mu_1(X)$ and deviation from mean U_1.

- One can make a similar decomposition for Y_0:

$$Y_1 = \mu_1(X) + U_1 \tag{1}$$
$$Y_0 = \mu_0(X) + U_0.$$

- Decomposition not necessarily structural.

- Additive separability simplifies the exposition.

- It may happen that controlling for the X , $Y_1 - Y_0$ is the same for all countries.
- This is the case of homogeneous treatment effects given X.
- More likely, countries vary in their responses to the policy even after controlling for X.

Roy Model

- The Roy model writes

$$D = 1 \text{ if sector 1 chosen; } D = 0 \text{ otherwise}$$

$$D = 1[Y_1 - Y_0 \geq 0]$$
$$Y = DY_1 + (1 - D)Y_0$$

- The generalized Roy model writes

$$D = 1[Y_1 - Y_0 - C \geq 0]$$

  - C is the direct cost of choosing 1.
  - Y_0 is the opportunity cost of choosing 1.
  - C varies among people.

- The extended Roy model treats C as a constant.

- Recall $D = 1$ if a country adopts a policy; $D = 0$ if it does not.

- The observed outcome is

$$Y = DY_1 + (1 - D)Y_0. \tag{2}$$

- Substituting the previous expressions into this expression, and keeping all X implicit, one obtains

$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \\ &= \mu_0 + (\mu_1 - \mu_0 + U_1 - U_0)D + U_0. \end{aligned}$$

- Using conventional regression notation,

$$Y = \alpha + \beta D + \varepsilon,$$

  where $\alpha = \mu_0$

- $\beta = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$.

- $\varepsilon = U_0$.

- I will also use the notation that $\upsilon = U_1 - U_0$.

- Let $\bar{\beta} = \mu_1 - \mu_0$ and $\beta = \bar{\beta} + \upsilon$.

- The coefficient on D is the "treatment effect".
- The case where β is the same for every country is the case conventionally assumed. (Homogeneity)
- More elaborate versions assume that β depends on X ($\beta(X)$) and estimate interactions of D with X.
- The case where β varies even after accounting for X is called the "random coefficient" or "heterogenous treatment effect" case.
- The case where $\upsilon = U_1 - U_0$ depends on D is the case of essential heterogeneity analyzed by Heckman, Urzua, and Vytlacil (2006).

Treatment Parameters

- Treatment parameters are various conditional means of the distribution of $Y_1 - Y_0$.
- Common treatment parameters:

$$\text{ATE} = E(Y_1 - Y_0)$$
(Average Treatment Effect)

$$\text{TT} = E(Y_1 - Y_0 \mid D = 1)$$
(Treatment on the Treated)

$$\text{MTE} = E(Y_1 - Y_0 \mid Y_1 - Y_0 - C = 0)$$
(Marginal Treatment Effect)

Example: Extended Roy Economy

Outcomes	Choice Model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$
$Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D^* = Y_1 - Y_0 - C$
General Case	
$(U_1 - U_0) \not\perp D$	
ATE $\neq$ TT $\neq$ TUT	

Example: Extended Roy Economy (cont.)

The Researcher Observes (Y, D, C)

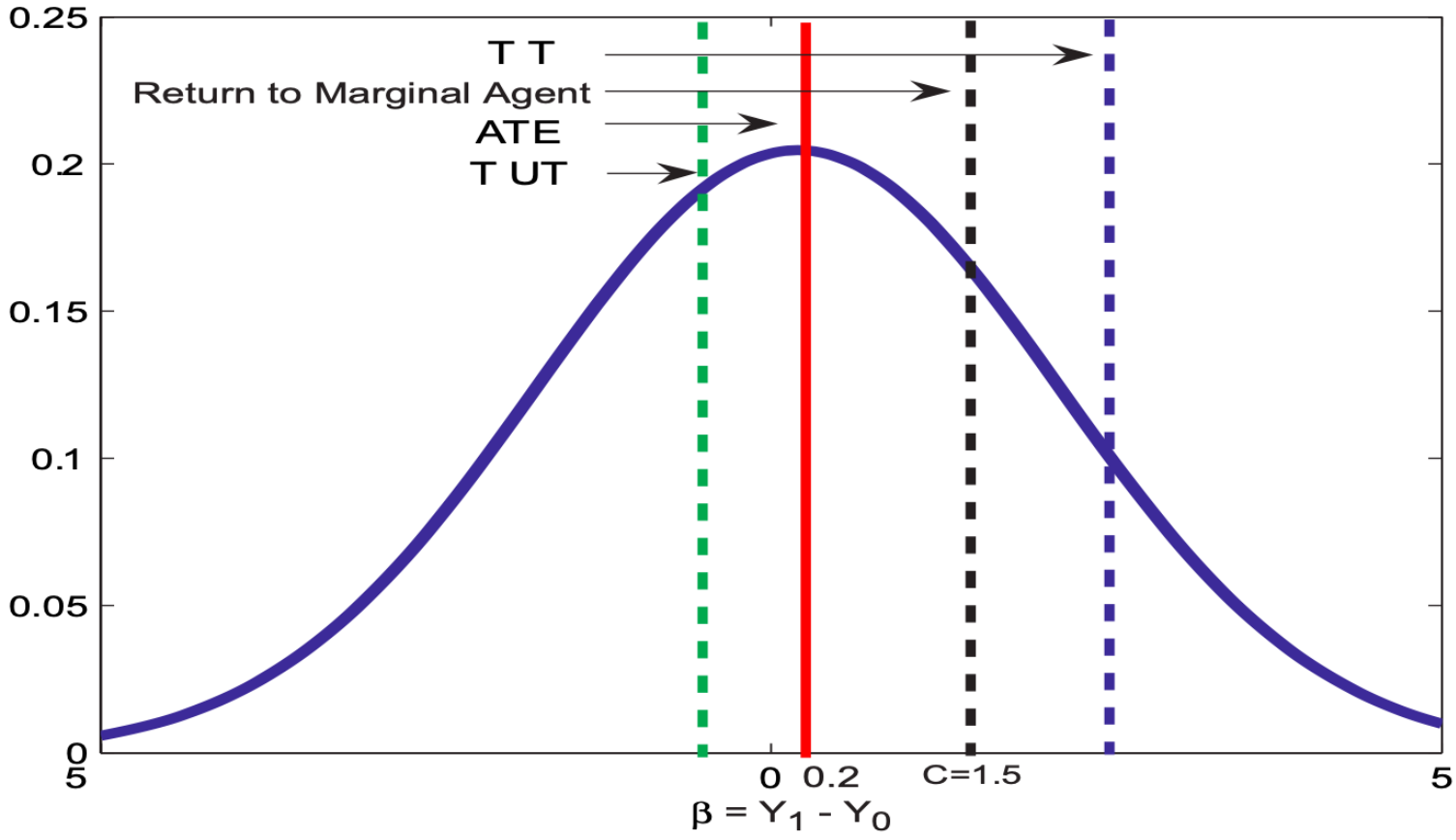
$$Y = \alpha + \beta D + U_0 \text{ where } \beta = Y_1 - Y_0$$

Parametrization

$$\alpha = 0.67 \qquad (U_1, U_0) \sim N(0, \Sigma)$$

$$\bar{\beta} = 0.2 \qquad \Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \qquad C = 1.5$$

Figure: Distribution of Gains, Extended Roy Economy, $U_1 - U_0 \not\perp D$



TT = 2.666, TUT = −0.632, Return to Marginal Agent = C = 1.5
ATE = $\mu_1 - \mu_0 = \bar{\beta} = 0.2$.
Source: Heckman, Urzua and Vytlacil (2006)

The Basic Principles Underlying the Identification of the Leading Econometric Evaluation Estimators

- Assume two potential outcomes (Y_0, Y_1).
- $D = 1$ if Y_1 is observed, and $D = 0$ corresponds to Y_0 being observed.
- The observed outcome is

$$Y = DY_1 + (1 - D)Y_0. \tag{3}$$

- The *evaluation problem* arises because for each person we observe either Y_0 or Y_1, but not both.

- In general it is not possible to identify the individual level treatment effect $Y_1 - Y_0$ for any person.

- The typical solution to this problem is to reformulate the problem at the population level rather than at the individual level and to identify certain mean outcomes or quantile outcomes or various distributions of outcomes as described in Heckman and Vytlacil (2007).

- For example, a commonly used approach focuses attention on average treatment effects, such as

$$\text{ATE} = E(Y_1 - Y_0).$$

- If treatment is assigned or chosen on the basis of potential outcomes, so that

$$(Y_0, Y_1) \not\perp\!\!\!\perp D,$$

where $\not\perp\!\!\!\perp$ denotes "is not independent" and $\perp\!\!\!\perp$ denotes independence, we encounter the problem of selection bias.

- Suppose that we observe people in each treatment state $D = 0$ and $D = 1$.

- If $Y_j \not\perp\!\!\!\perp D$, then the observed Y_j will be selectively different from randomly assigned Y_j, $j = 0, 1$.

- Then $E(Y_0 \mid D = 0) \neq E(Y_0)$ and

$$E(Y_1 \mid D = 1) \neq E(Y_1).$$

- Using unadjusted data to construct $E(Y_1 - Y_0)$ will produce one source of evaluation bias:

$$E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0) \neq E(Y_1 - Y_0).$$

- The selection problem underlies the evaluation problem.

- Many methods have been proposed to solve both problems.

Randomization

- The method with the greatest intuitive appeal, which is sometimes called the "gold standard" in policy evaluation analysis, is the method of random assignment.

- Nonexperimental methods can be organized by how they attempt to approximate what can be obtained by an ideal random assignment.

- If treatment is chosen at random with respect to (Y_0, Y_1), or if treatments are randomly assigned and there is full compliance with the treatment assignment,

$$(Y_0, Y_1) \perp\!\!\!\perp D. \tag{R-1}$$

- It is useful to distinguish several cases where (R-1) will be satisfied.
- The first is that agents (decision makers whose choices are being analyzed) pick outcomes that are random with respect to (Y_0, Y_1).
- Thus agents may not know (Y_0, Y_1) at the time they make their choices to participate in treatment or at least do not act on (Y_0, Y_1), so that $\Pr(D = 1 \mid X, Y_0, Y_1) = \Pr(D = 1 \mid X)$ for all X.

- Thus consider a Roy model where the agent information set is $\mathcal{I}$.

$$D = \mathbf{1}[E(Y_1 - Y_0 \mid \mathcal{I}) \geq 0]$$

- If agents do not know (Y_1, Y_0) at the time they make their decision or if they only know X (but not U_0, U_1), then

$$\Pr(D = 1 \mid Y_1, Y_0, X) = \Pr(D = 1 \mid X)$$

- A second case arises when individuals are randomly assigned to treatment status even if they would choose to self select into no-treatment status, and they comply with the randomization protocols.

- Let ξ be randomized assignment status.

- With full compliance, $\xi = 1$ implies that Y_1 is observed and $\xi = 0$ implies that Y_0 is observed.

- Then, under randomized assignment,

$$(Y_0, Y_1) \perp\!\!\!\perp \xi, \tag{R-2}$$

  even if in a regime of self-selection, $(Y_0, Y_1) \not\perp\!\!\!\perp D$.

- If randomization is performed conditional on X, we obtain

$$(Y_0, Y_1) \perp\!\!\!\perp \xi \mid X.$$

- Let A denote actual treatment status.

- If the randomization has full compliance among participants, $\xi = 1 \Rightarrow A = 1$ and $\xi = 0 \Rightarrow A = 0$.

- This is entirely consistent with a regime in which a person would choose $D = 1$ in the absence of randomization, but would have no treatment $(A = 0)$ if suitably randomized, even though the agent might desire treatment.

- If treatment status is chosen by self-selection,

$$D = 1 \Rightarrow A = 1 \text{ and } D = 0 \Rightarrow A = 0.$$

- If there is imperfect compliance with randomization,

$$\xi = 1 \nRightarrow A = 1$$

  because of agent choices.

- In general, $A = \xi D$, so that $A = 1$ only if $\xi = 1$ and $D = 1$.

- If treatment status is randomly assigned, either through randomization or randomized self-selection,

$$(Y_0, Y_1) \perp\!\!\!\perp A. \tag{R-3}$$

- This version of randomization can also be defined conditional on X.

If $(Y_0, Y_1) \perp\!\!\!\perp D$, keeping X implicit, the parameters treatment on the treated

$$\text{TT} = E(Y_1 - Y_0 \mid D = 1)$$

and treatment on the untreated

$$\text{TUT} = E(Y_1 - Y_0 \mid D = 0)$$

and the average treatment effect

$$\text{ATE} = E(Y_1 - Y_0)$$

and the marginal treatment effect

$$\text{MTE} = E(Y_1 - Y_0 \mid Y_1 - Y_0 - C = 0)$$

are all the same (i.e., MTE for $C = 0$).

- These parameters can be identified from population means:

$$\text{TT} = \text{TUT} = \text{ATE} = E(Y_1 - Y_0) = E(Y_1) - E(Y_0).$$

- Forming averages over populations of persons who are treated ($D = 1$) or untreated ($D = 0$) suffices to identify this parameter.

- If there are conditioning variables X, we can define the mean treatment parameters for all X where (R-1), (R-2), or (R-3) hold.

- Full compliance randomization when $(Y_0, Y_1) \not\perp D$, identifies $E(Y_1 - Y_0 \mid X)$, not the other parameter.

- Observe that even with random assignment of treatment status and full compliance, one cannot, in general, identify the distribution of the treatment effects $(Y_1 - Y_0)$.

- One can, however, identify the marginal distributions

$$F_1(Y_1 \mid A = 1, X = x) = F_1(Y_1 \mid X = x)$$

and

$$F_0(Y_0 \mid A = 0, X = x) = F_0(Y_0 \mid X = x).$$

- One special assumption, common in conventional econometrics, is that $Y_1 - Y_0 = \Delta(x)$, a constant given x.

- Since $\Delta(x)$ can be identified from $E(Y_1 \mid A = 1, X = x) - E(Y_0 \mid A = 0, X = x)$ because A is allocated by randomization, in this special case the analyst can identify the joint distribution of (Y_0, Y_1).

- This approach assumes that (Y_0, Y_1) have the same distribution up to a parameter Δ (Y_0 and Y_1 are perfectly dependent).

- One can make other assumptions about the dependence across ranks from perfect positive or negative ranking to independence.

- The joint distribution of (Y_0, Y_1) or of $(Y_1 - Y_0)$ is not identified unless the analyst can pin down the dependence across (Y_0, Y_1).

- Thus, even with data from a randomized trial one cannot, without further assumptions, identify the proportion of people who benefit from treatment in the sense of gross gain ($\Pr(Y_1 \geq Y_0)$).

- This problem plagues all evaluation methods.

- Assumption (R-1) is very strong.

- In many cases, it is thought that there is *selection bias* with respect to Y_0, Y_1, so persons who select into status 1 or 0 are selectively different from randomly sampled persons in the population.

Method of Matching

- One assumption commonly made to circumvent problems with (R-1) is that even though D is not random with respect to potential outcomes, the analyst has access to variables X that effectively produce a randomization of D with respect to (Y_0, Y_1) given X.

- This is the method of matching, which is based on the conditional independence assumption

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X. \qquad \text{(M-1)}$$

- Conditioning on X randomizes D with respect to (Y_0, Y_1).

- (M-1) assumes that any selective sampling of (Y_0, Y_1) can be adjusted by conditioning on observed variables.

- (R-1) and (M-1) are different assumptions and neither implies the other.

- In a linear equations model, assumption (M-1) that D is independent from (Y_0, Y_1) given X justifies application of least squares on D to eliminate selection bias in mean outcome parameters.

- For means, matching is just nonparametric regression.

- In order to be able to compare X-comparable people in the treatment regime one must assume

$$0 < \Pr(D = 1 \mid X = x) < 1. \qquad \text{(M-2)}$$

- Assumptions (M-1) and (M-2) justify matching.
- Assumption (M-2) is required for *any* evaluation estimator that compares treated and untreated persons.
- It is produced by random assignment if the randomization is conducted for all $X = x$ and there is full compliance.

- Observe that from (M-1) and (M-2), it is possible to identify $F_1(Y_1 \mid X = x)$ from the observed data $F_1(Y_1 \mid D = 1, X = x)$, since we observe the left hand side of

$$\begin{aligned} F_1(Y_1 \mid D = 1, X = x) &= F_1(Y_1 \mid X = x) \\ &= F_1(Y_1 \mid D = 0, X = x). \end{aligned}$$

- The first equality is a consequence of conditional independence assumption (M-1).

- The second equality comes from (M-1) and (M-2).

- By a similar argument, we observe the left hand side of

$$F_0(Y_0 \mid D = 0, X = x) = F_0(Y_0 \mid X = x)$$
$$= F_0(Y_0 \mid D = 1, X = x).$$

- The equalities are a consequence of (M-1) and (M-2).

- Since the pair of outcomes (Y_0, Y_1) is not identified for anyone, as in the case of data from randomized trials, the joint distributions of (Y_0, Y_1) given X or of $Y_1 - Y_0$ given X are not identified without further information.

- This is a problem that plagues all selection estimators.

- From the data on Y_1 given X and $D = 1$ and the data on Y_0 given X and $D = 0$ it follows that

$$\begin{aligned} E(Y_1 \mid D = 1, X = x) &= E(Y_1 \mid X = x) \\ &= E(Y_1 \mid D = 0, X = x) \end{aligned}$$

and

$$\begin{aligned} E(Y_0 \mid D = 0, X = x) &= E(Y_0 \mid X = x) \\ &= E(Y_0 \mid D = 1, X = x). \end{aligned}$$

- Thus,

$$\begin{aligned} E(Y_1 - Y_0 \mid X = x) &= E(Y_1 - Y_0 \mid D = 1, X = x) \\ &= E(Y_1 - Y_0 \mid D = 0, X = x). \end{aligned}$$

- Effectively, we have a randomization for the subset of the support of X satisfying (M-2).

- At values of X that fail to satisfy (M-2), there is no variation in D given X. One can define the residual variation in D not accounted for by X as

$$\mathcal{E}(x) = D - E(D \mid X = x) = D - \Pr(D = 1 \mid X = x).$$

- If the variance of $\mathcal{E}(x)$ is zero, it is not possible to construct contrasts in outcomes by treatment status for those X values and (M-2) is violated.

- To see the consequences of this violation in a regression setting, use $Y = Y_0 + D(Y_1 - Y_0)$ and take conditional expectations, under (M-1), to obtain

$$E(Y \mid X, D) = E(Y_0 \mid X) + D[E(Y_1 - Y_0 \mid X)].$$

- If $\text{Var}(\mathcal{E}(x)) > 0$ for all x in the support of X, one can use nonparametric least squares to identify

$$E(Y_1 - Y_0 \mid X = x) = \text{ATE}(x)$$

by regressing Y on D and X.

- The function identified from the coefficient on D is the average treatment effect.

- If $\text{Var}(\mathcal{E}(x)) = 0$, $\text{ATE}(x)$ is not identified at that x value because there is no variation in D that is not fully explained by X.

- A special case of matching is linear least squares where one can write

$$Y_0 = X\alpha + U \qquad Y_1 = X\alpha + \beta + U.$$

- $U_0 = U_1 = U$, and hence under (M-1)

$$E(Y \mid X, D) = X\alpha + \beta D + E(U \mid X).$$

- If D is perfectly predictable by X, one cannot identify β because of a multicollinearity problem.
- (M-2) rules out perfect collinearity.
- Matching is a nonparametric version of least squares that does not impose functional form assumptions on outcome equations, and that imposes support condition (M-2).
- It identifies β but not necessarily α (look at the term $E(U \mid X)$).

- Conventional econometric choice models make a distinction between variables that appear in outcome equations (X) and variables that appear in choice equations (Z).

- The same variables may be in (X) and (Z), but more typically there are some variables not in common.

- For example, the instrumental variable estimator we discuss next is based on variables that are not in X but that are in Z.

- Matching makes no distinction between the X and the Z.
- It does not rely on exclusion restrictions.
- The conditioning variables used to achieve conditional independence can in principle be a set of variables Q distinct from the X variables (covariates for outcomes) or the Z variables (covariates for choices).
- I use X solely to simplify the notation.

- The key identifying assumption is the assumed existence of a random variable X with the properties satisfying (M-1) and (M-2).

- Conditioning on a larger vector (X augmented with additional variables) or a smaller vector (X with some components removed) may or may not produce suitably modified versions of (M-1) and (M-2).

- Without invoking further assumptions there is no objective principle for determining what conditioning variables produce (M-1).

- Assumption (M-1) is strong.
- Many economists do not have enough faith in their data to invoke it.
- Assumption (M-2) is testable and requires no act of faith.
- To justify (M-1), it is necessary to appeal to the quality of the data.

- Using economic theory can help guide the choice of an evaluation estimator.
- A crucial distinction is the one between the information available to the analyst and the information available to the agent whose outcomes are being studied.
- Assumptions made about these information sets drive the properties of all econometric estimators.
- Analysts using matching make strong informational assumptions in terms of the data available to them.

- In fact, all econometric estimators make assumptions about the presence or absence of informational asymmetries.

- The possibility of asymmetry in information between the agent making participation decisions and the observing economist creates the potential for a major identification problem that is ruled out by assumption (M-1).

- The methods of control functions and instrumental variables estimators (and closely related regression discontinuity design methods) address this problem but in different ways.

- Accounting for this possibility is a more conservative approach to the selection problem than the one taken by advocates of least squares, or its nonparametric counterpart, matching.

- Those advocates assume that they know the X that produces a relevant information set.
- Conditional independence condition (M-1) cannot be tested without maintaining other assumptions.
- Choice of the appropriate conditioning variables is a problem that plagues *all* econometric estimators.

- The methods of control functions, replacement functions, proxy variables, and instrumental variables (to be defined next) all recognize the possibility of asymmetry in information between the agent being studied and the econometrician.

- They recognize that even after conditioning on X (variables in the outcome equation) and Z (variables affecting treatment choices, which may include the X), analysts may fail to satisfy conditional independence condition (M-1).

- Agents generally know more than econometricians about their choices and act on this information.

- These methods postulate the existence of some unobservables θ, which may be vector valued, with the property that

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta, \tag{U-1}$$

but allow for the possibility that

$$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z. \tag{U-2}$$

- In the event (U-2) holds, these approaches model the relationships of the unobservable θ with Y_1, Y_0, and D in various ways.

- The content in the control function principle is to specify the exact nature of the dependence of the relationship between observables and unobservables in a nontrivial fashion that is consistent with economic theory.

- The early literature focused on mean outcomes conditional on covariates.

- The normal Roy selection model makes distributional assumptions and identifies the joint distribution of outcomes.

- Replacement functions (Heckman and Robb, 1985) are methods that proxy θ. They substitute out for θ using observables.

- Aakvik, Heckman, and Vytlacil (1999, 2005), Carneiro, Hansen, and Heckman (2001, 2003), Cunha, Heckman, and Navarro (2005), and Heckman and Urzua (2008) develop methods that integrate out θ from the model, assuming $\theta \perp\!\!\!\perp (X, Z)$, or invoking weaker mean independence assumptions, and assuming access to proxy measurements for θ.

- The normal selection model produces partial identification of a generalized Roy model and full identification of a Roy model under separability and normality.

- The normal selection framework models the conditional expectation of U_0 and U_1 given X, Z, D.

- In terms of (U-1), it models the conditional mean dependence of Y_0 and Y_1 on D and θ given X and Z.

- Central to both the selection approach and the instrumental variable approach for a model with heterogenous responses is the probability of selection.

- Let Z denote variables in the choice equation. Fixing Z at different values (denoted z), define D(z) as an indicator function that is "1" when treatment is selected at the fixed value of z and that is "0" otherwise.

- In terms of a separable index mode $U_D = \mu_D(Z) - V$, for a fixed value of z,

$$D(z) = \mathbf{1}[\mu_D(z) \geq V],$$

where $Z \perp\!\!\!\perp V \mid X$.

- Thus fixing $Z = z$, values of z do not affect the realizations of V for any value of X.

- An alternative way of representing the independence between Z and V given X due to Imbens and Angrist (1994) writes that $D(z) \perp\!\!\!\perp Z$ for all $z \in \mathcal{Z}$, where $\mathcal{Z}$ is the support of Z.

- The Imbens-Angrist independence condition for IV is

$$\{D(z)\}_{z \in \mathcal{Z}} \perp\!\!\!\perp Z \mid X \Leftrightarrow V \perp\!\!\!\perp Z \mid X$$

  (See Vytlacil, 2002)

- Thus the probabilities that $D(z) = 1$, $z \in \mathcal{Z}$ are not affected by the occurrence of Z.

The Method of Instrumental Variables

- The method of instrumental variables (IV) postulates that

$$(Y_0, Y_1, \{D(z)\}_{z \in \mathcal{Z}}) \perp\!\!\!\perp Z \mid X. \quad \textbf{(Independence)} \tag{IV-1}$$

- One consequence of this assumption is that $E(D \mid Z) = P(Z)$, the propensity score, is random with respect to potential outcomes.

- Thus $(Y_0, Y_1) \perp\!\!\!\perp P(Z) \mid X$. So are all other functions of Z given X.

- The method of instrumental variables also assumes that

$$E(D \mid X, Z) = P(X, Z) \tag{IV-2}$$

  is a nondegenerate function of Z given X. **(Rank Condition)**

- Comparing (IV-1) to (M-1) in the method of instrumental variables, Z is independent of (Y_0, Y_1) given X, whereas in matching D is independent of (Y_0, Y_1) given X.

- So in (IV-1), Z plays the role of D in matching condition (M-1).

- Comparing (IV-2) with (M-2), in the method of IV the choice probability $\Pr(D = 1 \mid X, Z)$ is assumed to vary with Z conditional on X, whereas in matching, D varies conditional on X.

- No explicit model of the relationship between D and (Y_0, Y_1) is required in applying IV.

- (IV-2) is a rank condition and can be empirically verified.
- (IV-1) is not testable as it involves assumptions about counterfactuals.
- In a conventional common coefficient regression model

$$Y = \alpha + \beta D + U,$$

where β is a constant and where it is possible that $\text{Cov}(D, U) \neq 0$, (IV-1) and (IV-2) identify β.

- When β varies in the population and is correlated with D, additional assumptions must be invoked for IV to identify interpretable parameters.

- Heckman, Urzua, and Vytlacil (2006) and Heckman and Vytlacil (2007) discuss these conditions.

- Assumptions (IV-1) and (IV-2), with additional assumptions in the case where β varies in the population, can be used to identify mean treatment parameters.

- In matching, the variation in D that arises after conditioning on X provides the source of randomness that switches people across treatment status.

- Nature is assumed to provide an experimental manipulation conditional on X that replaces the randomization assumed in (R-1)-(R-3).

- When D is perfectly predictable by X, there is no variation in it conditional on X, and the randomization by nature breaks down.

- Heuristically, matching assumes a residual $\mathcal{E}(X) = D - E(D \mid X)$ that is nondegenerate and is one manifestation of the randomness that causes persons to switch status.

- In the IV method, it is the choice probability $E(D \mid X, Z) = P(X, Z)$ that is random with respect to (Y_0, Y_1), not components of D not predictable by (X, Z).

- Variation in Z for a fixed X provides the required variation in D that switches treatment status and still produces the required conditional independence:

$$(Y_0, Y_1) \perp\!\!\!\perp P(X, Z) \mid X.$$

- Variation in $P(X, Z)$ produces variations in D that switch treatment status.

Replacement Functions

- Versions of the method of control functions use measurements to proxy θ in (U-1) and (U-2) and remove spurious dependence that gives rise to selection problems.

- These are called "replacement functions" or "control variates".

- The methods of replacement functions and proxy variables all start from characterizations (U-1) and (U-2).

- θ is not observed and (Y_0, Y_1) are not observed directly, but Y is observed:

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

- Missing variables θ produce selection bias which creates a problem with using observational data to evaluate social programs.

- From (U-1), if one conditions on θ, condition (M-1) for matching would be satisfied, and hence one could identify the parameters and distributions that can be identified if the conditions required for matching are satisfied.

- The most direct approach to controlling for θ is to assume access to a function $\tau(X, Z, Q)$ that perfectly proxies θ:

$$\theta = \tau(X, Z, Q). \tag{4}$$

- This approach based on a perfect proxy is called the method of replacement functions by Heckman and Robb (1985).

- In (U-1), one can substitute for θ in terms of observables (X, Z, Q).
- Then
$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, Q.$$
- This is a version of matching.
- It is possible to condition nonparametrically on (X, Z, Q) and without having to know the exact functional form of τ.
- θ can be a vector and τ can be a vector of functions.

- This method has been used in the economics of education for decades (see the references in Heckman and Robb, 1985).

- If θ is ability and T is a test score, it is sometimes assumed that the test score is a perfect proxy (or replacement function) for θ and that one can enter it into the regressions of earnings on schooling to escape the problem of ability bias.

- Thus if $T = \alpha_0 + \alpha_1 X + \alpha_2 Q + \alpha_3 Z + \theta$, one can write $\theta = T - \alpha_0 - \alpha_1 X - \alpha_2 Q - \alpha_3 Z$, and use this as the proxy function.

- Controlling for T, X, Q, Z controls for θ.

- Notice that one does not need to know the coefficients $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ to implement the method. One can condition on T, X, Q, Z.

Control Functions

- The recent econometric literature applies in special cases the idea of the control function principle introduced in Heckman and Robb (1985).
- This principle, versions of which can be traced back to Telser (1964), partitions θ in (U-1) into two or more components, $\theta = (\theta_1, \theta_2)$, where only one component of θ is the source of bias.
- Thus it is assumed that (U-1) is true, and (U-1') is also true:

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta_1. \qquad \text{(U-1')}$$

- Thus (U-2) holds.

- For example, in a normal selection model with additive separability, one can break U_1, the error term associated with Y_1, into two components,

$$U_1 = E(U_1 \mid V) + \varepsilon,$$

where V plays the role of θ_1 and is associated with the choice equation.

- Further,

$$E(U_1 \mid V) = \frac{\text{Cov}(U_1, V)}{\text{Var}(V)} V, \tag{5}$$

assuming $E(U_1) = 0$ and $E(V) = 0$.

- Under normality, ε is independent of $E(U_1 \mid V)$.

- Heckman and Robb (1985) show how to construct a control function in the context of the choice model

$$D = \mathbf{1}[\mu_D(Z) > V].$$

- Controlling for V controls for the component of θ_1 in (U-1') that gives rise to the spurious dependence.

- As developed in Heckman and Robb (1985) and Heckman and Vytlacil (2007), under additive separability for the outcome equation for Y_1, one can write

$$E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + \underbrace{E(U_1 \mid \mu_D(Z) > V)}_{\text{control function}},$$

- The analyst "expects out" rather than solves out the effect of the component of V on U_1, and thus controls for selection bias under the maintained assumptions.

- In terms of the propensity score, under the conditions specified in Heckman and Vytlacil (2007), one may write the preceding expression in terms of $P(Z)$:

$$E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + K_1(P(Z)),$$

where

$$K_1(P(Z)) = E(U_1 \mid X, Z, D = 1).$$