The Principles Underlying Evaluation Estimators

James J. Heckman
University of Chicago

Econ 312, Spring 2022

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

- Two potential outcomes (Y_0, Y_1).
- $D = 1$ if Y_1 is observed.
- $D = 0$ corresponds to Y_0 being observed.
- Observed outcome:

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

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

- Not possible to identify the individual level treatment effect $Y_1 - Y_0$ for any person.
- ***Question:*** *Suppose $Y_1 - Y_0$ is a random variable that depends on X. Can you identify individual-level treatment effects?*
- Typical solution: reformulate the problem at the population level rather than at the individual level.
- Identify certain mean outcomes or quantile outcomes or various distributions of outcomes. See, e.g., Heckman and Vytlacil (2007).

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

- 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, $j \in \{0, 1\}$.
- 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:

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

- 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 $\mathcal{I}$.

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

- 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

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

- Matching assumes a version of (R-1) conditional on matching variables X: $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$.
- Z affects costs and affects $C(Z)$ and hence D, but is not in X.
- ***Question:*** *In a Generalized Roy model in which agents have as much information as the observing economist, and both use the information in making decisions and forming estimates, show that conditional on* (X, Z) *(the assumed information set) (R-1) is satisfied.*

- 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 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 for people right at the margin of indifference:

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

are all the same (i.e., MTE for $C = Y_1 - Y_0$).

- 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 ($A = 1$) or untreated ($A = 0$) 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.

- 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 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 $X = x$.
- Since $\Delta(x)$ can be identified from $E(Y_1 \mid A = 1, X = x) - E(Y_0 \mid A = 0, X = x)$ if 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 $\Delta(X)$ (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.

- Consider a model for (Y_0, Y_1)

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

- (μ_1, μ_0) are *structural*
- What does randomization of assignment with full compliance identify?

What Does Full Compliance Random Assignment Identify?

- We get

$$E(Y_1|X) = \mu_1(X) + E(U_1|X)$$
$$E(Y_0|X) = \mu_0(X) + E(U_0|X)$$

- Identifies

$$E(Y_1|X) - E(Y_0|X) = \overbrace{\mu_1(X) - \mu_0(X) + \underbrace{E(U_1|X) - E(U_0|X)}_{\substack{\text{constructed over the} \\ \text{whole population}}}}^{\text{ATE}}$$

- Does not induce $X \perp\!\!\!\perp (U_0, U_1)$.
- Randomization with respect to X would.

- Suppose

$$\begin{aligned} U_1 = U_0 = U & \qquad \text{(common coefficient model)} \\ Y = DY_1(1) + (1-D)Y_0 &= Y_0 + D(Y_1 - Y_0) \\ Y = \mu_0(X) + D(\mu_1(X) - \mu_0(X)) + U & \end{aligned}$$

- Suppose treatment assignment is randomized with perfect compliance

$$E(Y|D,X) = \mu_0(X) + D(\mu_1(X) - \mu_0(X)) + E[U|D,X]$$

- $D = 1$ means agent wants to go into program; $D = 0$ otherwise.
- But $E(U|D,X) = E(U|X)$
- $\therefore \mu_1(X) - \mu_0(X)$ is identified.
- In this case, randomization **balances the bias**
- In this case, $E(U_1|D,X) = E(U_0|D,X) = E(U|D,X)$.

- 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.
- Purposive choice

Imperfect Compliance

- 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 $D = 1$.
- **Question:** What causal parameter, if any, can be identified from an experiment with imperfect compliance?

- Specifically, compute the $\boldsymbol{ITT}$ reported in many journal articles (especially in *QJE*) for persons who would have participated in the program in absence of randomization (i.e., $D = 1$).

$$\begin{aligned} R &= 1 \text{ :(Randomized in)} \\ R &= 0 \text{ :(Randomized out)} \end{aligned}$$

<u>For Two Outcome Model</u>

$$\begin{aligned} D &= 1 \text{ :(You want 1)} \\ D &= 0 \text{ :(You want 0)} \end{aligned}$$

- You cannot compel people to participate

$$
\begin{aligned}
&E(Y|R=1) - E(Y|R=0) \\
=&\{E(Y_1|D=1,R=1) \quad Pr(D=1|R=1) \\
+&E(Y_0|D=0,R=1) \quad Pr(D=0|R=1)\} \\
\\
-&\{E(Y_1|D=1,R=0) \quad Pr(D=1|R=0)(\tau) \\
+&E(Y_0|D=0,R=0) \quad Pr(D=0|R=0) \\
+&E(Y_0|D=1,R=0) \quad Pr(D=1|R=0)(1-\tau)\}.
\end{aligned}
$$

where τ is the proportion of people who want to go in among $R = 0$ people who do not comply with the assignment and, in fact, get into treatment anyway

- This is a mix of people with different preferences for and access to the program
- **What interesting economic question does this estimate?**

- With “perfect compliance”

$$\begin{aligned} Pr(D=1|R=1)&=1 \quad Pr(D=0|R=1)=0 \\ Pr(D=1|R=0)&=0 \quad Pr(D=0|R=0)=1 \\ E(Y|R=1)-E(Y|R=0)&=E(Y_1-Y_0|D=1) \end{aligned}$$

- **Question: Is full compliance credible? What is the assumed decision problem?**

Method of Matching

- One assumption commonly made to circumvent problems with satisfying (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.

Method of Matching

- $$(Y_0, Y_1) \perp\!\!\!\perp D \mid X. \tag{M-1}$$
- Conditioning on X randomizes D with respect to (Y_0, Y_1).
- (M-1) assumes that any selective sampling of (Y_0, Y_1) with respect to D can be adjusted by conditioning on observed variables.
- (R-1) and (M-1) are different assumptions and neither implies the other.

- In order to be able to compare X-comparable people in the treatment regime a sufficient condition is

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

- Assumptions (M-1) and (M-2) justify matching.
- Assumption (M-2) is required for *any* evaluation estimator that compares treated and untreated persons.
- Clearly we can invoke a restricted version (common support for $D = 1$ and $D = 0$).
- 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).
- X eliminates differences.

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

$$\begin{aligned} F_0(Y_0 \mid D = 0, X = x) &= F_0(Y_0 \mid X = x) \\ &= F_0(Y_0 \mid D = 1, X = x). \end{aligned}$$

- 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.
- Problem 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).

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

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

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.
- Thus cannot make counterfactual comparisons.

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

$$Y_0 = X\alpha + U_0 \qquad\qquad Y_1 = X\alpha + \beta + U_1.$$

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

$$E(Y \mid X, D) = \varphi(X) + \beta D,$$

where $\varphi(X) = X\alpha + E(U \mid X)$.

- If D is perfectly predictable by X, one cannot identify β.
- 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 $E(U \mid X)$).

- Observe that we do not need $E(U \mid X) = 0$ to identify β.

- 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 (to be discussed) 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.
- Crucial distinction:
  - **The information available to the analyst.**
  - **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.

Implicit Information Assumptions

- All econometric estimators make assumptions about the presence or absence of informational asymmetries.

Five Distinct Information Sets

- To analyze the informational assumptions invoked in matching, and other econometric evaluation strategies, it is helpful to introduce **five distinct information sets** and establish some relationships among them.

  1. An information set $\sigma(I_{R^*})$ with an associated random variable that satisfies conditional independence (M-1) is defined as a *relevant* information set.
  2. The minimal information set $\sigma(I_R)$ with associated random variable needed to satisfy conditional independence (M-1) is defined as the *minimal relevant* information set.
  3. The information set $\sigma(I_A)$ available to the agent at the time decisions to participate are made. Here A means agent, not assignment.

4. The information available to the economist, $\sigma(I_{E^*})$.
5. The information $\sigma(I_E)$ used by the economist in conducting an empirical analysis.

- Denote the random variables generated by these sets as I_{R^*}, I_R, I_A, I_{E^*}, and I_E, respectively.

Definition 1

Define $\sigma(I_{R^*})$ as a **relevant information set** if the information set is generated by the random variable I_{R^*}, possibly vector valued, and satisfies condition (M-1), so

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

Definition 2

Define $\sigma(I_R)$ as a **minimal relevant information set** if it is the intersection of all sets $\sigma(I_{R^*})$ and satisfies $(Y_0, Y_1) \perp\!\!\!\perp D \mid I_R$. The associated random variable I_R is a minimum amount of information that guarantees that condition (M-1) is satisfied. **There may be no such set. But in most cases, there is.**

- The intersection of all sets $\sigma(I_{R^*})$ may be empty and hence may not be characterized by a (possibly vector valued) random variable I_R that guarantees $(Y_1, Y_0) \perp\!\!\!\perp D \mid I_R$.
- If the information sets that produce conditional independence are nested, then the intersection of all sets $\sigma(I_{R^*})$ producing conditional independence is well defined and has an associated random variable I_R with the required property, although it may not be unique.
- E.g., strictly monotonic measure-preserving transformations and affine transformations of I_R also preserve the property.

- In the more general case of non-nested information sets with the required property, it is possible that no uniquely defined minimal relevant set exists.
- Among collections of nested sets that possess the required property, there is a minimal set defined by intersection but there may be multiple minimal sets corresponding to each collection.

- If one defines the relevant information set as one that produces conditional independence, it may not be unique.
- If the set $\sigma(I_{R^*})$ satisfies the conditional independence condition, then the set $\sigma(I_{R^*}, Q)$ such that $Q \perp\!\!\!\perp (Y_0, Y_1) \mid I_{R^*}$ would also guarantee conditional independence.
- For this reason, when it is possible to do so I define the relevant information set to be minimal, that is, to be the intersection of all relevant sets that still produce conditional independence between (Y_0, Y_1) and D.
- However, no minimal set may exist.

Definition 3

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

- By the agent I mean the person making the treatment decision, not necessarily the person whose outcomes are being studied (e.g., the agent may be the parent, the person being studied may be a child).

Definition 4

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

Definition 5

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

- For the case where a unique minimal relevant information set exists, only three restrictions are implied by the structure of these sets:

$$\sigma(I_R) \subseteq \sigma(I_{R^*}), \quad \sigma(I_A) \subseteq \sigma(I_R), \text{ and } \quad \sigma(I_E) \subseteq \sigma(I_{E^*}).$$

- First restriction previously discussed.
- Second restriction requires that the minimal relevant information set must include the information the agent uses when deciding which treatment to take or assign.
- It is the information in $\sigma(I_A)$ that gives rise to the selection problem which in turn gives rise to the evaluation problem.

- The third restriction requires that the information used by the econometrician must be part of the information that he/she observes.
- Aside from these orderings, the econometrician's information set may be different from the agent's or the relevant information set.
- The econometrician may know something the agent doesn't know, for typically he is observing events after the decision is made.
- At the same time, there may be private information known to the agent but not the econometrician.

- Matching assumption (M-1) implies that $\sigma(I_R) \subseteq \sigma(I_E)$, so that the econometrician uses at least the minimal relevant information set, but of course he or she may use more.
- However, using more information is not guaranteed to produce a model with conditional independence property (M-1) satisfied for the augmented model.
- Thus an analyst can “overdo” it.

- 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**.

Control Functions, Replacement Functions, Proxy Variables, IV and Panel Approaches

- The methods of control functions, replacement functions, proxy variables, and instrumental variables 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

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

- If (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.

- **Replacement functions:** (Heckman and Robb, 1985) proxy θ. They substitute out for θ using observables. Olley & Pakes (1993) is an application.
- Aakvik, Heckman, and Vytlacil (1999, 2005), Carneiro, Hansen, and Heckman (2001, 2003), Cunha, Heckman, and Navarro (2005), and Cunha, Heckman, and Schennach (2006a,b) 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.

- 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 model $U_D = \mu_D(Z) - V$, for a fixed value of z,

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

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

The Method of Instrumental Variables

- The method of instrumental variables (IV) postulates that

$$\left(Y_0, Y_1, \{D(z)\}_{z\in\mathcal{Z}}\right) \perp\!\!\!\perp Z \mid X \text{ (Independence)} \qquad \text{(IV-1)}$$

- $E(D \mid X, Z) = P(X, Z)$ is random with respect to potential outcomes.
- Thus $(Y_0, Y_1) \perp\!\!\!\perp P(X, Z) \mid X$.
- So are all other functions of Z given X.

- The method of instrumental variables also assumes that

$$E(D \mid X, Z) = P(X, Z) \text{ is a nondegenerate function of } Z \text{ given } X. \text{ (Rank Condition)} \tag{IV-2}$$

- Alternatively, one can write that

$$\operatorname{Var}\left(E\left(D \mid X, Z\right)\right) \neq \operatorname{Var}\left(E\left(D \mid X\right)\right).$$

Comparing Instrumental Variables and Matching

$$(Y_0, Y_1) \perp\!\!\!\perp Z|X \text{ } \textbf{IV}$$
$$(Y_0, Y_1) \perp\!\!\!\perp D|X \text{ } \textbf{Matching}$$

- 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 $\Pr(D = 1 \mid X, Z)$ varies with Z conditional on X.
- In matching, D varies conditional on X. This is the source of identifying information in this method.
- No *explicit* model of the relationship between D and (Y_0, Y_1) is required in applying IV.
- An explicit model is required to interpret what IV estimates.

- (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

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

- β is a constant.
- If $\mathrm{Cov}(D, U) \neq 0$, (IV-1) and (IV-2) identify β.

Opposite Roles for $D - P(X, Z)$

- 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 IV**, the choice probability $E(D \mid X, Z) = P(X, Z)$ is random with respect to (Y_0, Y_1), conditional on X.

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

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

- Components of variation in D not predictable by (X, Z) do not produce the required independence.
- They are assumed to be the **source** of the problem.
- The predicted component provides the required independence.
- Just the opposite in matching where they are the source of identification.

Control and 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.
- Missing data problem.

- 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:

$$\theta = \tau(X, Z, Q). \tag{2}$$

- This approach based on a perfect proxy is called the **method of replacement functions** (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).
- A version later used by Olley and Pakes (1996).

- If θ is ability and τ 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 $\tau = \alpha_0 + \alpha_1 X + \alpha_2 Q + \alpha_3 Z + \theta$, one can write $\theta = \tau - \alpha_0 - \alpha_1 X - \alpha_2 Q - \alpha_3 Z$, and use this as the proxy function.
- Controlling for τ, 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 τ, X, Q, Z.

Factor Models

- The method of replacement functions assumes that (2) is a perfect proxy.
- In many applications, θ is measured with error.
- This produces a **factor model or measurement error** model.

- One can represent the factor model in a general way by a system of equations:

$$Y_j = g_j\left(X, Z, Q, \theta, \varepsilon_j\right) , \qquad j = 0, 1. \tag{3}$$

- A linear factor model separable in the unobservables writes

$$Y_j = g_j\left(X, Z, Q\right) + \alpha_j \theta + \varepsilon_j , \qquad j = 0, 1, \tag{4}$$

where

$$(X, Z) \perp\!\!\!\perp (\theta, \varepsilon_j), \varepsilon_j \perp\!\!\!\perp \theta , \qquad j = 0, 1, \tag{5}$$

and the ε_j are mutually independent.

- Observe that under (3) and (4), Y_j controlling for X, Z, only imperfectly proxies θ because of the presence of ε_j.
- θ is called a factor, α_j factor loadings, and the ε_j "uniquenesses".

- The key to identification is multiple, but imperfect (because of ε_j), measurements on θ from the Y_j, $j = 0, 1$, and X, Z, Q, and possibly other measurement systems that depend on θ.
- Carneiro, Hansen, and Heckman (2003), Cunha, Heckman, and Navarro (2005, 2006), and Cunha and Heckman (2006a,b) apply and develop these methods.
- Under assumption (5), they show how to nonparametrically identify the econometric model and the distributions of the unobservables $F_{\Theta}(\theta)$ and $F_{\xi_j}(\varepsilon_j)$.
- See notes on Factor Models.

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:

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

- Thus (U-2) holds, conditional on θ_1.

- 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,

  $$E(U_1 \mid V) = \frac{\mathrm{Cov}(U_1, V)}{\mathrm{Var}(V)} V, \tag{6}$$

  assuming $E(U_1) = 0$ and $E(V) = 0$.
- Under normality, $\varepsilon \perp\!\!\!\perp E(U_1 \mid V)$.

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

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

- 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 (2007a,b), 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 $P(Z)$:

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

where

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

- The most commonly used panel data method is **difference-in-differences** as discussed in Heckman and Robb (1985), Blundell, Duncan, and Meghir (1998), Heckman, LaLonde, and Smith (1999), and Bertrand, Duflo, and Mullainathan (2004).
- All of the estimators can be adapted to a panel data setting.

- Heckman, Ichimura, Smith, and Todd (1998): difference -in -differences matching estimators.
- Abadie (2002) extends this work.
- Separability between errors and observables is a key feature of the panel data approach in its standard application.
- Altonji and Matzkin (2005) and Matzkin (2003) present analyses of nonseparable panel data methods.
- Regression discontinuity estimators, which are versions of IV estimators, are discussed by Heckman and Vytlacil (2007b).

- Table 1 summarizes some of the main lessons of this lecture. The stated conditions are necessary. There are many versions of the IV and control functions principle and extensions of these ideas which refine these basic postulates.

Table 1: Identifying Assumptions Under Commonly Used Methods

(Y_0, Y_1) are potential outcomes that depend on X.

$$D = \begin{cases} 1 \text{ if assigned (or chose) status 1} \\ 0 \text{ otherwise.} \end{cases}$$

Z are determinants of D, θ is a vector of unobservables.

For random assignments, A is a vector of actual treatment status.

$A = 1$ if treated; $A = 0$ if not.

$\xi = 1$ if a person is randomized to treatment status; $\xi = 0$ otherwise.

	Identifying Assumptions	Identifies marginal distributions?	Exclusion condition needed?
Random Assignment	$(Y_0, Y_1) \perp\!\!\!\perp \xi$, $\xi = 1 \implies A = 1$, $\xi = 0 \implies A = 0$ (full compliance) Alternatively, if self-selection is random with respect to outcomes, $(Y_0, Y_1) \perp\!\!\!\perp D$. Assignment can be conditional on X.	Yes	No
Matching	$(Y_0, Y_1) \not\perp\!\!\!\perp D$, but $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$, $0 < \Pr(D = 1 \mid X) < 1$ for all X. D conditional on X is a nondegenerate random variable	Yes	No

Table 1: Identifying Assumptions Under Commonly Used Methods, Cont.

(Y_0, Y_1) are potential outcomes that depend on X

$$D = \begin{cases} 1 \text{ if assigned (or choose) status 1} \\ 0 \text{ otherwise} \end{cases}$$

Z are determinants of D, θ is a vector of unobservables

For random assignments, A is a vector of actual treatment status. $A = 1$ if treated; $A = 0$ if not.

$\xi = 1$ if a person is randomized to treatment status; $\xi = 0$ otherwise.

	Identifying Assumptions	Identifies marginal distributions?	Exclusion condition needed?
Control Functions and Extensions	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp D \mid X, Z, \theta$. The method models dependence induced by θ or else proxies θ (replacement function). Version (i) Replacement functions (substitute out θ by observables) (Blundell and Powell, 2003; Heckman and Robb, 1985; Olley and Pakes, 1994). Factor models (Carneiro, Hansen and Heckman, 2003) allow for measurement error in the proxies. Version (ii) Integrate out θ assuming $\theta \perp\!\!\!\perp (X, Z)$ (Aakvik, Heckman, and Vytlacil, 2005; Carneiro, Hansen, and Heckman, 2003) Version (iii) For separable models for mean response expect θ conditional on X, Z, D as in standard selection models (control functions in the same sense of Heckman and Robb).	Yes	Yes (for semiparametric models)
IV	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp Z \mid X$, $\Pr(D = 1 \mid Z)$ is a nondegenerate function of Z.	Yes	Yes