Instrumental Variable Methods in Program Evaluation

Manuel Arellano
CEMFI
March 4, 2009

1. Preliminaries

1.1 Structural and treatment effect approaches

- The classic approach to quantitative policy evaluation in economics has been the *structural approach*.
- Its goals are to specify a class of theory-based models of individual choice, choose the one within the class that best fits the data, and use it for ex-post or ex-ante policy simulation.
- During the last 20 years the *treatment effect approach* has established itself as a formidable competitor that has introduced a different language, different priorities, techniques and practices in applied work.
- Not only that, it has also changed the perception of evidence-based economics among economists, public opinion, and policy makers.
- The ambition in a structural exercise is to use data from a particular context to identify, with the help of theory, deep rules of behavior that can be extrapolated to other contexts.

- A treatment effect (TE) exercise is context-specific and addresses less ambitious policy questions.
- The goal is to evaluate the impact of an *existing* policy by comparing the distribution of a chosen outcome variable for individuals affected by the policy (treatment group) with the distribution of unaffected individuals (control group).
- The aim is to choose the control and treatment groups in such a way that membership of one or the other, either results from randomization or can be regarded as if they were the result of randomization.
- In this way one hopes to achieve the standards of empirical credibility on causal evidence that are typical of experimental biomedical studies.

1.2 Potential outcomes and causality

- Association and causation have always been known to be different, but a mathematical framework for an unambiguous characterization of *statistical causal effects* is surprisingly recent (Rubin, 1974; despite precedents in statistics and economics, Neyman, 1923; Roy, 1951).
- Think of a population of individuals that are susceptible of treatment. Let Y_1 be the outcome for an individual if exposed to treatment and let Y_0 be the outcome for the same individual if not exposed. The treatment effect for that individual is $Y_1 - Y_0$.
- In general, individuals differ in how much they gain from treatment, so that we can imagine a distribution of gains over the population with mean
$$\alpha_{ATE} = E\left(Y_1 - Y_0\right).$$
- The average treatment effect so defined is a standard measure of the causal effect of treatment 1 relative to treatment 0 on the chosen outcome.
- Suppose that treatment has been administered to a fraction of the population, and we observe whether an individual has been treated or not ($D = 1$ or 0) and the person's outcome Y. Thus, we are observing Y_1 for the treated and Y_0 for the rest:
$$Y = (1 - D)\, Y_0 + DY_1.$$

- Because Y_1 and Y_0 can never be observed for the same individual, the distribution of gains lacks empirical entity. It is just a conceptual device that can be related to observables.
- This notion of causality is statistical because it is not interested in finding out causal effects for specific individuals. Causality is defined in an average sense.

Connection with regression

- A standard measure of association between Y and D is:

$$\beta = E(Y \mid D = 1) - E(Y \mid D = 0)$$
$$= E(Y_1 - Y_0 \mid D = 1) + \{E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)\}$$

- The second expression makes it clear that in general β differs from the *average gain for the treated* (another standard measure of causality, that we call α_{TT}).
- The reason is that treated and nontreated units may have different average outcomes in the absence of treatment.
- For example, this will be the case if treatment status is the result of individual decisions, and those with low Y_0 choose treatment more frequently than those with high Y_0.

- From a structural model of D and Y one could obtain the implied average treatment effects, but here α_{ATE} or α_{TT} have been directly defined with respect to the distribution of potential outcomes, so that relative to a structure they are reduced form causal effects.
- Econometrics has conventionally distinguished between reduced form effects (uninterpretable but useful for prediction) and structural effects (associated with rules of behavior).
- The TE literature emphasizes "reduced form causal effects" as an intermediate category between predictive and structural effects.

Social feedback

- The potential outcome representation is predicated on the assumption that the effect of treatment is independent of how many individuals receive treatment, so that the possibility of different outcomes depending on the treatment received by other units is ruled out.
- This excludes general equilibrium or feedback effects, as well as strategic interactions among agents.
- So the framework is not well suited to the evaluation of system-wide reforms which are intended to have substantial equilibrium effects.

1.3 Social experiments

- In the TE approach, a randomized field trial is regarded as the ideal research design.
- Observational studies seen as "more speculative" attempts to generate the force of evidence of experiments.
- In a controlled experiment, treatment status is randomly assigned by the researcher, which by construction ensures:

$$(Y_0, Y_1) \perp D$$

  In such a case, $F(Y_1 \mid D = 1) = F(Y_1)$ and $F(Y_0 \mid D = 0) = F(Y_0)$. The implication is $\alpha_{ATE} = \alpha_{TT} = \beta$.
- Analysis of data takes a simple form: An unbiased estimate of α_{ATE} is the difference between the average outcomes for treatments and controls:

$$\widehat{\alpha}_{ATE} = \overline{Y}_T - \overline{Y}_C$$

- In a randomized setting, there is no need to "control" for covariates, rendering multiple regression unnecessary, except if interested in effects for specific groups.

1.4 Matching

- There are many situations where experiments are too expensive, unfeasible, or unethical. A classical example is the analysis of the effects of smoking on mortality rates.
- Experiments guarantee the independence condition
$$(Y_1, Y_0) \perp D$$
but with observational data it is not very plausible.
- A less demanding condition for nonexperimental data is:
$$(Y_1, Y_0) \perp D \mid X.$$
- Conditional independence implies
$$\begin{aligned} E(Y_1 \mid X) &= E(Y_1 \mid D = 1, X) = E(Y \mid D = 1, X) \\ E(Y_0 \mid X) &= E(Y_0 \mid D = 0, X) = E(Y \mid D = 0, X). \end{aligned}$$
Therefore, for α_{ATE} we can calculate (and similarly for α_{TT}):
$$\begin{aligned} \alpha_{ATE} &= E(Y_1 - Y_0) = \int E(Y_1 - Y_0 \mid X)\, dF(X) \\ &= \int [E(Y \mid D = 1, X) - E(Y \mid D = 0, X)]\, dF(X). \end{aligned}$$
- The following is a matching expression for $\alpha_{TT} = E(Y_1 - Y_0 \mid D = 1)$:
$$E[Y - E(Y_0 \mid D = 1, X) \mid D = 1] = E[Y - \mu_0(X) \mid D = 1]$$
where $\mu_0(X) = E(Y \mid D = 0, X)$ is used as an imputation for Y_0.

2. Instrumental variable assumptions

- Suppose we have non-experimental data with covariates, but cannot assume conditional independence as in matching:

$$(Y_1, Y_0) \perp D \mid X.$$

- Suppose, however, that we have a variable Z that is an "exogenous source of variation in D" in the sense that it satisfies the *independence assumption*:

$$(Y_1, Y_0) \perp Z \mid X$$

and the *relevance assumption*:

$$Z \not\perp D \mid X.$$

- Matching can be regarded as a special case of IV in which $Z = D$, i.e. all variation in D is exogenous given X.

Examples

Example 1: Non-compliance in randomized trials

- In a classic example, Z indicates assignment to treatment in an experimental design. Therefore, $(Y_1, Y_0) \perp Z$.
- However, "actual treatment" D differs from Z because some individuals in the treatment group decide not to treat (non-compliers). Z and D will be correlated in general.
- Assignment to treatment is not a valid instrument in the presence of externalities that benefit members of the treatment group even if they are not treated themselves. In such case the exclusion restriction fails to hold.
- An example of this situation arises in a study of the effect of deworming on school participation in Kenya using school-level randomization (Miguel and Kremer, *Econometrica*, 2004).

Example 2: Ethnic enclaves and immigrant outcomes

- Interest in the effect of leaving in a highly concentrated ethnic area on labor success. In Sweden 11% of the population was born abroad. Of those, more than 40% live in an ethnic enclave (Edin, Fredriksson and Åslund, *QJE*, 2003).
- The causal effect is ambiguous. Residential segregation lowers the acquisition rate of local skills, preventing access to good jobs. But enclaves act as opportunity-increasing networks by disseminating information to new immigrants.
- Immigrants in ethnic enclaves have 5% lower earnings, after controlling for age, education, gender, family background, country of origin, and year of immigration.
- But this association may not be causal if the decision to live in an enclave depends on expected opportunities.
- Swedish governments of 1985-1991assigned initial areas of residence to refugee immigrants. Motivated by the belief that dispersing immigrants promotes integration.
- Let Z indicate initial assignment (8 years before measuring ethnic enclave indicator D). Edin et al. assumed that Z is independent of potential earnings Y_0 and Y_1.
- IV estimates implied a 13% gain for low-skill immigrants associated with one std. deviation increase in ethnic concentration. For high-skill immigrants there was no effect.

Example 3: Vietnam veterans and civilian earnings

- Did military service in Vietnam have a negative effect on earnings? (Angrist, 1990).
- Here we have:
  - Instrumental variable: draft lottery eligibility.
  - Treatment variable: Veteran status.
  - Outcome variable: Log earnings.
  - Data: $N = 11637$ white men born 1950–1953.
  - March Population Surveys of 1979 and 1981–1985.
- This lottery was conducted annually during 1970-1974. It assigned numbers (from 1 to 365) to dates of birth in the cohorts being drafted. Men with lowest numbers were called to serve up to a ceiling determined every year by the Department of Defense.
- Abadie (2002) uses as instrument an indicator for lottery numbers lower than 100.
- The fact that draft eligibility affected the probability of enrollment along with its random nature makes this variable a good candidate to instrument "veteran status".
- There was a strong selection process in the military during the Vietnam period. Some volunteered, while others avoided enrollment using student or job deferments.
- Presumably, enrollment was influenced by future potential earnings.

3. Identification of causal effects in IV settings

- The question is whether the availability of an instrumental variable identifies causal effects. To answer it, I consider a binary Z, and abstract from conditioning.

Homogeneous effects

- If the causal effect is the same for every individual
$$Y_{1i} - Y_{0i} = \alpha$$
the availability of an IV allows us to identify α. This is the traditional situation in econometric models with endogenous explanatory variables.

- In the homogeneous case
$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i = Y_{0i} + \alpha D_i.$$

- Also, taking into account that $Y_{0i} \perp Z_i$
$$\begin{aligned} E(Y_i \mid Z_i = 1) &= E(Y_{0i}) + \alpha E(D_i \mid Z_i = 1) \\ E(Y_i \mid Z_i = 0) &= E(Y_{0i}) + \alpha E(D_i \mid Z_i = 0). \end{aligned}$$

- Subtracting both equations we obtain
$$\alpha = \frac{E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)}{E(D_i \mid Z_i = 1) - E(D_i \mid Z_i = 0)}$$
which determines α as long as
$$E(D_i \mid Z_i = 1) \neq E(D_i \mid Z_i = 0).$$

- Get the effect of D on Y through the effect of Z because Z only affects Y through D.

Heterogeneous effects

Summary

- In the heterogeneous case the availability of IVs is not sufficient to identify a causal effect.
- An additional assumption that helps identification of causal effects is the following "monotonicity" condition: Any person that was willing to treat if assigned to the control group, would also be prepared to treat if assigned to the treatment group.
- The plausibility of this assumption depends on the context of application.
- Under monotonicity, the IV coefficient coincides with the average treatment effect for those whose value of D would change when changing the value of Z (local average treatment effect or LATE).

Indicator of potential treatment status

- In preparation for the discussion below let us introduce the following notation:

$$D = \begin{cases} D_0 \text{ if } Z = 0 \\ D_1 \text{ if } Z = 1 \end{cases}$$

- Given data on (Y, D) there are 4 observable groups but 8 underlying groups, which can be classified as never-takers, compliers, defiers, and always-takers.

Example

- Consider two levels of schooling ($D = 0, 1$, high school and college) with associated potential wages (Y_0, Y_1), so that individual returns are $Y_1 - Y_0$. Also consider an exogenous determinant of schooling Z with associated potential schooling levels (D_0, D_1). The IV Z is exogenous in the sense that it is independent of (Y_0, Y_1, D_0, D_1).
- An example of Z is proximity to college:
  - $Z = 0$ college far away
  - $Z = 1$ college nearby
  - Defier with $D = 1, Z = 0$ (ie. $D_1 = 0$): Person who goes to college when is far but would not go if it was near.
  - Defier with $D = 0, Z = 1$ (ie. $D_0 = 1$): Person does not go to college when it is near but would go if it was far.

Table 1
Observable and Latent Types

	Z	D	D_0	D_1		
Type 1	0	0	0	0	Type 1A	Never-taker
				1	Type 1B	Complier
Type 2	0	1	1	0	Type 2A	Defier
				1	Type 2B	Always-taker
Type 3	1	0	0	0	Type 3A	Never-taker
			1		Type 3B	Defier
Type 4	1	1	0	1	Type 4A	Complier
			1		Type 4B	Always-taker

Availability of IV is not sufficient by itself to identify causal effects

- Note that since

$$\begin{aligned} E(Y \mid Z=1) &= E(Y_0) + E[(Y_1 - Y_0) D_1] \\ E(Y \mid Z=0) &= E(Y_0) + E[(Y_1 - Y_0) D_0] \end{aligned}$$

  we have

$$\begin{aligned} E(Y \mid Z=1) - E(Y \mid Z=0) &= E[(Y_1 - Y_0)(D_1 - D_0)] \\ &= E(Y_1 - Y_0 \mid D_1 - D_0 = 1) \Pr(D_1 - D_0 = 1) \\ &\quad - E(Y_1 - Y_0 \mid D_1 - D_0 = -1) \Pr(D_1 - D_0 = -1) \end{aligned}$$

- $E(Y \mid Z=1) - E(Y \mid Z=0)$ could be negative and yet the causal effect be positive for everyone, as long as the probability of defiers is sufficiently large.

Additional assumption: Eligibility rules

- An additional assumption that helps to identify α_{TT} is an eligibility rule of the form:
$$\Pr(D = 1 \mid Z = 0) = 0$$
i.e. individuals with $Z = 0$ are denied treatment.

- In this situation:
$$\begin{aligned} E(Y \mid Z = 1) &= E(Y_0) + E[(Y_1 - Y_0) D \mid Z = 1] \\ &= E(Y_0) + E(Y_1 - Y_0 \mid D = 1, Z = 1) E(D \mid Z = 1) \end{aligned}$$
and since $E(D \mid Z = 0) = 0$
$$E(Y \mid Z = 0) = E(Y_0) + E(Y_1 - Y_0 \mid D = 1, Z = 0) E(D \mid Z = 0) = E(Y_0)$$

- Therefore,
$$\text{Wald parameter} \equiv \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1)} = E(Y_1 - Y_0 \mid D = 1, Z = 1).$$

- Moreover,
$$\alpha_{TT} \equiv E(Y_1 - Y_0 \mid D = 1) = E(Y_1 - Y_0 \mid D = 1, Z = 1).$$
This is so because $\Pr(Z = 1 \mid D = 1) = 1$. That is,
$$\begin{aligned} E(Y_1 - Y_0 \mid D = 1) &= E(Y_1 - Y_0 \mid D = 1, Z = 1) \Pr(Z = 1 \mid D = 1) \\ &\quad + E(Y_1 - Y_0 \mid D = 1, Z = 0) [1 - \Pr(Z = 1 \mid D = 1)]. \end{aligned}$$

- Thus, if $\Pr(D = 1 \mid Z = 0) = 0$ the IV coefficient coincides with the average treatment effect on the treated.

4. Local average treatment effects (LATE)

Monotonicity and LATEs

- If we rule out defiers i.e. $\Pr(D_1 - D_0 = -1) = 0$, we have
  $E(Y \mid Z = 1) - E(Y \mid Z = 0) = E(Y_1 - Y_0 \mid D_1 - D_0 = 1) \Pr(D_1 - D_0 = 1)$
  and
  $$E(D \mid Z = 1) - E(D \mid Z = 0) = E(D_1) - E(D_0) = \Pr(D_1 - D_0 = 1).$$
- Therefore,
  $$E(Y_1 - Y_0 \mid D_1 - D_0 = 1) = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)}$$
- Imbens and Angrist called this parameter "local average treatment effects" (LATE).
- Different IV's lead to different parameters, even under instrument validity, which is counter to standard GMM thinking.
- Policy relevance of a LATE parameter depends on the subpopulation of compliers defined by the instrument. Most relevant LATE's are those based on instruments that are policy variables (eg college fee policies or college creation).
- What happens if there are no compliers? In the absence of defiers, the probability of compliers satisfies
  $$\Pr(D_1 - D_0 = 1) = E(D \mid Z = 1) - E(D \mid Z = 0).$$
  So, lack of compliers implies lack of instrument relevance, hence underidentification.

Distributions of potential wages for compliers

- Imbens and Rubin (1997) showed that under monotonicity not only the average treatment effect for compliers is identified but also the entire marginal distributions of Y_0 and Y_1 for compliers.
- Abadie (2002) gives a simple proof that suggests a Wald calculation. For any function $h(.)$ let us consider

$$W = h(Y) D = \begin{cases} W_1 = h(Y_1) & \text{if } D = 1 \\ W_0 = 0 & \text{if } D = 0 \end{cases}.$$

  Because (W_1, W_0, D_1, D_0) are independent of Z, we can apply the LATE formula to W and get

$$E(W_1 - W_0 \mid D_1 - D_0 = 1) = \frac{E(W \mid Z = 1) - E(W \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)},$$

  or substituting

$$E(h(Y_1) \mid D_1 - D_0 = 1) = \frac{E(h(Y) D \mid Z = 1) - E(h(Y) D \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)}.$$

- If we choose $h(Y) = 1(Y \leq r)$, the previous formula gives as an expression for the *cdf* of Y_1 for the compliers.

- Similarly, if we consider

$$V = h(Y)(1-D) = \begin{cases} V_1 = h(Y_0) & \text{if } 1-D=1 \\ V_0 = 0 & \text{if } 1-D=0 \end{cases}$$

then

$$E(V_1 - V_0 \mid D_1 - D_0 = 1) = \frac{E(V \mid Z=1) - E(V \mid Z=0)}{E(1-D \mid Z=1) - E(1-D \mid Z=0)}$$

or

$$E(h(Y_0) \mid D_1 - D_0 = 1) = \frac{E(h(Y)(1-D) \mid Z=1) - E(h(Y)(1-D) \mid Z=0)}{E(1-D \mid Z=1) - E(1-D \mid Z=0)}$$

from which we can get the *cdf* of Y_0 for the compliers, again setting $h(Y) = 1(Y \leq r)$.

- To see the intuition, suppose that D is exogenous (i.e. $Z = D$), then the *cdf* of $Y \mid D=0$ coincides with the *cdf* of Y_0, and the *cdf* of $Y \mid D=1$ coincides with the *cdf* of Y_1.

- If we regress $h(Y)D$ on D, the OLS regression coefficient is

$$E[h(Y)D \mid D=1] - E[h(Y)D \mid D=0] = E[h(Y_1)]$$

which for $h(Y) = 1(Y \leq r)$ gives us the *cdf* of Y_1.

- Similarly, if we regress $h(Y)(1-D)$ on $(1-D)$, the regression coefficient is

$$E[h(Y)(1-D) \mid 1-D=1] - E[h(Y)(1-D) \mid 1-D=0] = E[h(Y_0)].$$

- In the IV case, we are running similar IV (instead of OLS) regressions using Z as instrument and getting expected $h(Y_1)$ and $h(Y_0)$ for compliers.

Conditional estimation with instrumental variables

- So far we abstracted from the fact that the validity of the instrument may only be conditional on X: It may be that $(Y_0, Y_1) \perp Z$ does not hold, but the following does:

$$(Y_0, Y_1) \perp Z \mid X \quad \text{(conditional independence)}$$
$$Z \not\perp D \mid X \quad \text{(conditional relevance)}$$

- For example, in the analysis of returns to college where Z is an indicator of proximity to college. The problem is that Z is not randomly assigned but chosen by parents, and this choice may depend on characteristics that subsequently affect wages. The validity of Z may be more credible given family background variables X.

- In a linear version of the problem:
  - First stage: Regress D on Z and $X \rightarrow$ get $\widehat{D}$.
  - Second stage: Regress Y on $\widehat{D}$ and X.

- In general we now have conditional LATE given X:

$$\gamma(X) = E(Y_1 - Y_0 \mid D_1 \neq D_0, X).$$

- On the other hand, we have conditional IV estimands:

$$\beta(X) = \frac{E(Y \mid Z = 1, X) - E(Y \mid Z = 0, X)}{E(D \mid Z = 1, X) - E(D \mid Z = 0, X)}$$

- What is the relevant aggregate effect? If the treatment effect is homogeneous given X

$$Y_1 - Y_0 = \beta(X),$$

then a parameter of interest is:

$$E[\beta(X)] = \int \beta(X)\, dF(X).$$

- However, in the case of heterogeneous effects, it makes sense to consider an average treatment effect for the overall subpopulation of compliers:

$$\beta_C = \int \beta(X)\, dF(X \mid compliers).$$

- Calculating β_C appears problematic because $F(X \mid compliers)$ is unobservable, but

$$\begin{aligned}\beta_C &= \int \beta(X) \frac{\Pr(compliers \mid X)}{\Pr(compliers)} dF(X) \\ &= \int [E(Y \mid Z=1, X) - E(Y \mid Z=0, X)] \frac{1}{\Pr(compliers)} dF(X)\end{aligned}$$

where

$$\Pr(compliers) = \int [E(D \mid Z=1, X) - E(D \mid Z=0, X)]\, dF(X).$$

- Therefore,

$$\beta_C = \frac{\int [E(Y \mid Z=1, X) - E(Y \mid Z=0, X)]\, dF(X)}{\int [E(D \mid Z=1, X) - E(D \mid Z=0, X)]\, dF(X)},$$

which can be estimated as a ratio of matching estimators (Frölich, 2003).

5. Relating LATE to parametric models of the potential outcomes

5.1 The endogenous dummy explanatory variable probit model

- The model as usually written in terms of observables is

$$\begin{aligned} Y &= \mathbf{1}\left(\alpha + \beta D + U \geq 0\right) \\ D &= \mathbf{1}\left(\pi_0 + \pi_1 Z + V \geq 0\right) \\ \begin{pmatrix} U \\ V \end{pmatrix} &\mid Z \sim \mathcal{N}\left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]. \end{aligned}$$

- In this model D is an endogenous explanatory variable as long as $\rho \neq 0$. D is exogenous if $\rho = 0$.

- In this model there are only two potential outcomes:

$$\begin{aligned} Y_1 &= \mathbf{1}\left(\alpha + \beta + U \geq 0\right) \\ Y_0 &= \mathbf{1}\left(\alpha + U \geq 0\right) \end{aligned}$$

- The average probability effect of interest (ATE) is given by

$$\theta = E\left(Y_1 - Y_0\right) = \Phi\left(\alpha + \beta\right) - \Phi\left(\alpha\right).$$

- In less parametric specifications $E\left(Y_1 - Y_0\right)$ may not be point identified, but we may still be able to estimate LATE.

Monotonicity is equivalent to the index model assumption for D

- The equivalence between monotonicity and index models provides a link with economic assumptions.
- Consider the case where Z is a scalar 0–1 instrument, so that there are only two potential values of D:

$$\begin{aligned} D_1 &= \mathbf{1}\left(\pi_0 + \pi_1 + V \geq 0\right) \\ D_0 &= \mathbf{1}\left(\pi_0 + V \geq 0\right). \end{aligned}$$

- Suppose without lack of generality that $\pi_1 \geq 0$. Then we can distinguish three subpopulations depending on an individual's value of V:
- Never-takers: Units with $V < -\pi_0 - \pi_1$. They have $D_1 = 0$ and $D_0 = 0$. Their mass is $1 - \Phi\left(\pi_0 + \pi_1\right)$.
- Compliers: Units with $V \geq -\pi_0 - \pi_1$ but $V < -\pi_0$. They have $D_1 = 1$ and $D_0 = 0$. Their mass is $\Phi\left(\pi_0 + \pi_1\right) - \Phi\left(\pi_0\right)$.
- Always-takers: Units with $V \geq -\pi_0$. They have $D_1 = 1$ and $D_0 = 1$. Their mass is $\Phi\left(\pi_0\right)$.

LATE under joint probit assumptions

- Let us obtain the average treatment effect for the subpopulation of compliers:
$$\theta_{LATE} = E\left(Y_1 - Y_0 \mid D_1 - D_0 = 1\right) \equiv E\left(Y_1 - Y_0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right).$$
- We have
$$\begin{aligned} E\left(Y_1 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) &= \Pr\left(\alpha + \beta + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) \\ &= 1 - \frac{\Pr\left(U \leq -\alpha - \beta, V \leq -\pi_0\right) - \Pr\left(U \leq -\alpha - \beta, V \leq -\pi_0 - \pi_1\right)}{\Pr\left(V \leq -\pi_0\right) - \Pr\left(V \leq -\pi_0 - \pi_1\right)} \end{aligned}$$
and similarly
$$\begin{aligned} E\left(Y_0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) &= \Pr\left(\alpha + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) \\ &= 1 - \frac{\Pr\left(U \leq -\alpha, V \leq -\pi_0\right) - \Pr\left(U \leq -\alpha, V \leq -\pi_0 - \pi_1\right)}{\Pr\left(V \leq -\pi_0\right) - \Pr\left(V \leq -\pi_0 - \pi_1\right)}. \end{aligned}$$
- Finally,
$$\begin{aligned} \theta_{LATE} = \frac{1}{\Phi\left(-\pi_0\right) - \Phi\left(-\pi_0 - \pi_1\right)} &\left[\Phi_2\left(-\alpha, -\pi_0; \rho\right) - \Phi_2\left(-\alpha, -\pi_0 - \pi_1; \rho\right)\right. \\ &\left. -\Phi_2\left(-\alpha - \beta, -\pi_0; \rho\right) + \Phi_2\left(-\alpha - \beta, -\pi_0 - \pi_1; \rho\right)\right]. \end{aligned}$$
where $\Phi_2\left(r, s; \rho\right) = \Pr\left(U \leq r, V \leq s\right)$ is a standard normal bivariate probability.
- The nice thing about θ_{LATE} is that it is identified from the Wald formula in the absence of joint normality.
- In fact, it does not even require monotonicity in the relationship between Y and D.

5.2 Models with additive errors: switching regressions

The switching regression model with endogenous switch

- The model is as follows:

$$\begin{aligned} Y_i &= \alpha + \beta_i D_i + U_i \\ D_i &= 1\left(\gamma_0 + \gamma_1 Z_i + \varepsilon_i \geq 0\right) \end{aligned} \tag{1}$$

- The potential outcomes are

$$\begin{aligned} Y_{1i} &= \alpha + \beta_i + U_i \equiv \mu_1 + V_{1i} \\ Y_{0i} &= \alpha + U_i \equiv \mu_0 + V_{0i} \end{aligned}$$

  so that the treatment effect $\beta_i = Y_{1i} - Y_{0i}$ is heterogeneous.

- Traditional models assume that β_i is constant or that it varies only with observable characteristics. In these models D may be exogenous (independent of U) or endogenous (correlated with U) but in either case $Y_1 - Y_0$ is constant, at least given controls.

- β_i may depend on unobservables and D_i may be correlated with both U_i and β_i.

- We assume the exclusion restriction holds in the sense that $(V_{1i}, V_{0i}, \varepsilon_i)$ or $(U_i, \beta_i, \varepsilon_i)$ are independent of Z_i.

- In terms of the alternative notation (letting $\alpha = \mu_0$ and $U_i = V_{0i}$):

$$Y_i = \mu_0 + (Y_{1i} - Y_{0i})\, D_i + V_{0i} = \mu_0 + (\mu_1 - \mu_0)\, D_i + \left[V_{0i} + (V_{1i} - V_{0i})\, D_i\right].$$

- Let us write the ATE as $\overline{\beta} = \mu_1 - \mu_0$ and $\xi_i = V_{1i} - V_{0i}$ so that $\beta_i = \overline{\beta} + \xi_i$.

Example: Rosen and Willis (1979)

- Consider the effect of education on earnings and the decision to become educated. We are interested in the decision of college education ($D = 1$) vs. high school ($D = 0$).
- The model consists of potential earnings with or without college education (Y_1, Y_0) and a schooling decision rule:

$$D = 1\left(Y_1 - Y_0 > C\right).$$

- There are determinants of costs (C) like distance to college, tuition fees, availability of scholarships, opportunity costs or borrowing constraints, which are potential instruments. $Y_1 - Y_0$ is the return to college education for a particular individual. Equation (1) can be regarded as a reduced form version of the schooling decision rule.
- In the Rosen & Willis model $Y_1 - Y_0$ may also depend on unobservables because they think of multiple abilities and comparative advantage. Moreover, the model suggests that D_i may be correlated with both U_i and β_i.

Endogeneity and self-selection

- Write
$$E(Y_i \mid Z_i) = \mu_0 + (\mu_1 - \mu_0) E(D_i \mid Z_i) + E(V_{1i} - V_{0i} \mid D_i = 1, Z_i) E(D_i \mid Z_i).$$
- If β_i is mean independent of D_i
$$E(Y_i \mid Z_i) = \mu_0 + (\mu_1 - \mu_0) E(D_i \mid Z_i).$$
so that $\overline{\beta} = Cov(Z, Y) / Cov(Z, D)$.
- Otherwise, $\overline{\beta}$ does not coincide with the IV estimand. A special case of mean independence of β_i with respect to D_i occurs when β_i is constant.
- The failure of IV can be seen as the result of a missing variable. The model can be written as
$$Y_i = \alpha + \overline{\beta} D_i + \varphi(Z_i) D_i + \zeta_i$$
where $\varphi(Z_i) = E(V_{1i} - V_{0i} \mid D_i = 1, Z_i)$. Note that $E(\zeta_i \mid Z_i) = 0$.
- When we do ordinary IV estimation we are not taking into account the variable $\varphi(Z_i) D_i$.
- $\varphi(z)$ is the average excess return for college-educated people with $Z_i = z$. In the distance to college example ($Z = 1$ if college near), we would expect $\varphi(1) \leq \varphi(0)$.
- The average treatment effect on the treated and the LATE are, respectively,
$$\alpha_{TT} = E(Y_{1i} - Y_{0i} \mid D_i = 1) = \overline{\beta} + E(V_{1i} - V_{0i} \mid D_i = 1),$$
$$\alpha_{LATE} = E(Y_{1i} - Y_{0i} \mid D_{1i} - D_{0i} = 1) = \overline{\beta} + E(V_{1i} - V_{0i} \mid -\gamma_0 - \gamma_1 \leq \varepsilon_i < -\gamma_0).$$

The Gaussian model

- The model is completed with the assumption

$$\begin{pmatrix} V_{1i} \\ V_{0i} \\ \varepsilon_i \end{pmatrix} \mid Z_i \sim \mathcal{N}\left[0, \begin{pmatrix} \sigma_1^2 & \sigma_{10} & \sigma_{1\varepsilon} \\ & \sigma_0^2 & \sigma_{0\varepsilon} \\ & & 1 \end{pmatrix}\right].$$

- In this case we have a parametric likelihood model that can be estimated by ML.
- We can also consider a variety of two-step methods. Note that

$$E\left(V_{1i} - V_{0i} \mid D_i = 1, Z_i\right) = \left(\sigma_{1\varepsilon} - \sigma_{0\varepsilon}\right) \lambda\left(\gamma_0 + \gamma_1 Z_i\right),$$

  so that we can do IV estimation in

$$Y_i = \alpha + \overline{\beta} D_i + \left(\sigma_{1\varepsilon} - \sigma_{0\varepsilon}\right) \lambda_i D_i + \zeta_i,$$

  or OLS estimation in:

$$Y_i = \alpha + \overline{\beta} \Phi_i + \left(\sigma_{1\varepsilon} - \sigma_{0\varepsilon}\right) \phi_i + \zeta_i^*.$$

Identification without parametric distributional assumptions

- The current model can be regarded as the combination of two generalized selection models. So the identification result for that model applies.
- Namely, with a continuous exclusion restriction $E\left(Y_{1i} \mid X_i\right)$ and $E\left(Y_{0i} \mid X_i\right)$ are identified up to a constant (X_i denotes controls that so far we omitted for simplicity).
- However, the constants are important because they determine the average treatment effect of D on Y. Unfortunately, they require an identification at infinity argument.

6. Marginal treatment effects

Introduction

- When the support of Z is not binary, there is a multiplicity of causal effects.
- What causal effects are relevant for evaluating a given policy?
- The natural experiment literature has been satisfied with identifying "causal effects", without paying much attention to their relevance.
- If Z is continuous we can define a different LATE parameter for every pair (z, z'):

$$\alpha_{LATE}(z, z') = \frac{E(Y \mid Z = z) - E(Y \mid Z = z')}{E(D \mid Z = z) - E(D \mid Z = z')}.$$

  The multiplicity is even higher when there is more than one instrument.

IV assumptions and monotonicity

- For a general instrument vector Z, there are as many potential treatment status indicators D_z as possible values z of the instrument. The IV assumptions become:
  - Independence: $(Y_1, Y_0, D_z) \perp Z$.
  - Relevance: $\Pr(D = 1 \mid Z = z) = P(z)$ is a nontrivial function of z.
- The monotonicity assumption for general Z can be expressed as follows. For any pair of values (z, z') either

$$D_{zi} \geq D_{z'i} \quad \text{or} \quad D_{zi} \leq D_{z'i}$$

  for all units in the population.

Latent index representation

- Alternatively we can postulate an index model for D_z:
$$D_z = 1\left(\mu\left(z\right) - U > 0\right) \quad \text{and } U \perp Z,$$
which can be a useful way of organizing different LATEs (Heckman & Vytlacil, 2005).
- Note that the observed D is $D = D_Z$.
- Monotonicity and index model assumptions are equivalent (Vytlacil, 2002).
- This result connects LATE thinking with econometric selection models.
- Without loss of generality we can set $\mu\left(z\right) = P\left(z\right)$ and take U as uniformly distributed in the $(0, 1)$ interval. To see this note that
$$1\left(\mu\left(z\right) > U\right) = 1\left\{F_U\left[\mu\left(z\right)\right] > F_U\left(U\right)\right\} = 1\left(P\left(z\right) > \widetilde{U}\right)$$
where $\widetilde{U}$ is uniformly distributed.
- To connect with the earlier discussion, if Z is a 0–1 scalar instrument there are only two values of the propensity score $P\left(0\right)$ and $P\left(1\right)$. Suppose that $P\left(0\right) < P\left(1\right)$. Always-takers have $U < P\left(0\right)$, compliers have a value of U between $P\left(0\right)$ and $P\left(1\right)$, and never-takers have $U > P\left(1\right)$. A similar argument can be made for any pair (z, z') in the case of a general Z.
- So under monotonicity we can always invoke and index equation and imagine each member of the population as having a particular value of the unobserved variable U.

Marginal Treatment Effect

- Using the propensity score $P(Z) = \Pr(D = 1 \mid Z)$ as instrument, LATE becomes
$$\alpha_{LATE}(P(z), P(z')) = \frac{E(Y \mid P(Z) = P(z)) - E(Y \mid P(Z) = P(z'))}{P(z) - P(z')}.$$
- If Z is binary this is equivalent to what we had in the first place, but if Z is continuous, taking limits as $z \to z'$, we get a limiting form of LATE or MTE:
$$MTE(P(z)) = \frac{\partial E(Y \mid P(Z) = P(z))}{\partial P(z)}.$$
- $\alpha_{LATE}(P(z), P(z'))$ gives the ATE for individuals who would change schooling status from changing $P(Z)$ from $P(z')$ to $P(z)$:
$$\alpha_{LATE}(P(z), P(z')) = E[Y_1 - Y_0 \mid P(z') < U < P(z)]$$
- Similarly $MTE(P(z))$ gives the ATE for individuals who would change schooling status following a marginal change in $P(z)$ or, in other words, who are indifferent between schooling choices at $P(Z) = P(z)$.
- Using the error term in the index model, we can say that
$$MTE(P(z)) = E(Y_1 - Y_0 \mid U = P(z))$$

- Integrating $MTE\left(P\left(z\right)\right)$ over different ranges of U we can get other ATE measures. For example,

$$\alpha_{LATE}\left(P\left(z\right),P\left(z'\right)\right)=\frac{\int_{P(z')}^{P(z)}MTE\left(u\right)du}{P\left(z\right)-P\left(z'\right)}$$

- Moreover,

$$\alpha_{ATE}=\int_{0}^{1}MTE\left(u\right)du,$$

which makes it clear that to be able to identify α_{ATE} we need identification of $MTE\left(u\right)$ over the entire $\left(0,1\right)$ range.

Policy-relevant treatment effects

- Constructing suitably integrated $MTE\left(u\right)s$ it may be possible to identify policy relevant treatment effects.
- LATE gives the per capita effect of the policy in those induced to change by the policy when the instrument is precisely an indicator of the policy change.
- For example, policies that change college fees or distance to school, under the assumption that the policy change affects the probability of participation but not the gain itself.

Estimation: Local IV method

- Heckman and Vytlacil suggest to estimate MTE by estimating the derivative of the conditional mean
$$E\left(Y \mid P\left(Z\right) = P\left(z\right), X = x\right)$$
using kernel-based local linear regression techniques.
- Note that in this context the propensity score plays a very different role to matching.
- *Testing for homogeneity (or absence of self-selection)* : A test of linearity on the propensity score (conditional on X) is a test of homogeneity of treatment effects.
- To see this use $Y = Y_0 + \left(Y_1 - Y_0\right) D$ and write
$$\begin{aligned} E\left(Y \mid P\left(Z\right)\right) &= E\left(Y_0 \mid P\left(Z\right)\right) + E\left(\left(Y_1 - Y_0\right) D \mid P\left(Z\right)\right) \\ &= E\left(Y_0\right) + E\left[Y_1 - Y_0 \mid D = 1, P\left(Z\right)\right] P\left(Z\right) \end{aligned}$$
- The quantity $E\left[Y_1 - Y_0 \mid D = 1, P\left(Z\right)\right]$ is constant under homogeneity, so that the conditional mean $E\left(Y \mid P\left(Z\right)\right)$ is linear in $P\left(Z\right)$.

Concluding remarks about unobserved heterogeneity

- How important is it?
  - The balance between observed and unobserved heterogeneity depends on how detailed information on agents is available (an empirical issue).
  - The worry for IV-based identification of treatment effects is not heterogeneity *per se*, but the fact that heterogeneous gains may affect program participation.
- Warnings:
  - In the absence of an economic model or a clear notional experiment, it is often difficult to interpret what IV estimates estimate.
  - Knowing that IV estimates can be interpreted as averages of heterogeneous effects is not very useful if understanding the heterogeneity itself is first order (Deaton, 2009).
- Heterogeneity of gains vs. heterogeneity of treatments
  - Heterogeneity of treatments may be more important. For example, the literature has found significant differences in returns to different college majors.
  - A problem of aggregating educational categories is that returns are less meaningful.
  - Sometimes education outcomes are aggregated into just two categories because some techniques are only well developed for binary explanatory variables.
  - A methodological emphasis may offer new opportunities but also impose constraints.