Instrumental Variable Methods in Program Evaluation

Class Notes

Manuel Arellano

January 21, 2009

1 Introduction and examples

- Basic ideas when gains are heterogeneous:
  - Availability of IVs by itself does not guarantee identification of average treatment effects.
  - Different instruments lead to different parameters even under instrument validity (counter to standard GMM thinking).

1.1 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.

1.2 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, 177–178).

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, *AER*, 1990, 313–336).
- Here we have:
  - Instrumental variable: draft lottery eligibility.
  - Treatment variable: Veteran status.
  - Outcome variable: Log earnings.

- 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. The value of that ceiling varied from 95 to 195 depending on the year.
- Abadie (2002, *JASA*) 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".
- The focus in Angrist (1990) was estimating "the average treatment effect for compliers" (the LATE parameter, see below). His data:
  - Data: $N = 11637$ white men born 1950–1953.
  - March Population Surveys of 1979 nand 1981–1985.
- There was a strong selection process in the military during the Vietnam period. Some volunteered, while others avoided enrollment using student or job deferments. Others disqualified for health problems or criminal records.
- Presumably, enrollment was influenced by variables associated with future potential earnings.
- Result: $\widehat{\alpha}_{LATE} = -1278$ dollars, i.e. a negative impact on earnings for compliers (but not significantly different from zero).
- See more examples of sources of IVs in Angrist and Krueger (*Journal of Economic Perspectives*, Fall 2001, Table 1, p. 82).

2 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 the fact that the reasoning can be conditional on X.

2.1 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 general

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i$$

and in the homogeneous case

$$Y_i = Y_{0i} + \alpha D_i.$$

- 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

$$E(D_i \mid Z_i = 1) \neq E(D_i \mid Z_i = 0).$$

- Intuitively, the effect of D on Y can be measured through the effect of Z because we have assumed that Z only affects Y through D.

2.2 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 ($D = 0, 1$, eg. high school and college) with associated potential wages (Y_0, Y_1), so that individual returns are $Y_1 - Y_0$ and observed wages are $Y = Y_0 + (Y_1 - Y_0) D$. Also consider an exogenous determinant of schooling ($Z = 0, 1$ e.g. low or high tuition fee) with associated potential schooling levels (D_0, D_1), so that $D = D_0 + (D_1 - D_0) Z$. The instrumental variable Z is exogenous in the sense that it is independent of (Y_0, Y_1, D_0, D_1).
- Another example of Z in the same context is proximity to college:
  - $Z = 0$ college far away
  - $Z = 1$ college nearby
  - $D = 1$ if going to college
  - 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

$$E(Y \mid Z=1) - E(Y \mid Z=0) = E[(Y_1 - Y_0)(D_1 - D_0)]$$

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

- $E(Y \mid Z=1) - E(Y \mid Z=0)$ 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:

$$E(Y \mid Z=1) = E(Y_0) + E[(Y_1 - Y_0) D \mid Z=1] = E(Y_0) + E(Y_1 - Y_0 \mid D=1, Z=1) E(D \mid Z=1)$$

  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.

In the next section we consider an alternative assumption that leads to the identification of average treatment effects for certain subpopulations.

3 Local average treatment effects (LATE)

3.1 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). This point is further discussed below.

- What happens if there are no compliers? Suppose there are no compliers, i.e.

$$\Pr(D_1 - D_0 = 1) = 0.$$

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, in the absence of defiers, lack of compliers implies lack of instrument relevance, hence underidentification.

3.2 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. One can see the parallel with the IV case presenting this situation in a regression format as follows.

- 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.

3.3 Conditional estimation with instrumental variables

- So far we have abstracted from the fact that the validity of the instrument may only be conditional on a covariate vector X. That is, it may be that $(Y_0, Y_1) \perp Z$ does not hold, but the following does:

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

- For example, in the analysis of wage 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 assumption of validity of Z may be more credible given family background variables X (e.g. Card (1995) analyzes this situation).

- In a linear version of the problem:
  - First stage: Regress D on Z and $X \to$ 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, that is, if for an individual with a given value of 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:

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

- Calculation of β_C would appear to be problematic because $F(X \mid compliers)$ is not observable. However, note that

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

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 two (matching) non-parametric imputation estimators (c.f. Marcus Frölich, 2003; Heckman–Vytlacil (2005), p. 686, footnote 25).

- Note that similarly to the case of matching, here it is also possible to consider versions of the estimator based on the propensity score.

- Again, the common support condition can be particularly relevant in practice.

4 Relating LATE to parametric models of the potential outcomes

4.1 The endogenous dummy explanatory variable probit model

- The model as usually written in terms of observables is

$$\begin{aligned}
Y &= \mathbf{1}(\alpha + \beta D + U \geq 0) \\
D &= \mathbf{1}(\pi_0 + \pi_1 Z + V \geq 0)
\end{aligned}$$

$$\begin{pmatrix} U \\ V \end{pmatrix} \mid Z \sim \mathcal{N}\left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

- 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:

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

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

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

- In less parametric specifications $E(Y_1 - Y_0)$ 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:

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

- 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 $\Phi(-\pi_0 - \pi_1) = 1 - \Phi(\pi_0 + \pi_1)$.
- 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 $\Phi(-\pi_0) - \Phi(-\pi_0 - \pi_1) = \Phi(\pi_0 + \pi_1) - \Phi(\pi_0)$.
- Always-takers: Units with $V \geq -\pi_0$. They have $D_1 = 1$ and $D_0 = 1$. Their mass is $1 - \Phi(-\pi_0) = \Phi(\pi_0)$.

LATE under joint probit assumptions

- Let us obtain the average treatment effect for the subpopulation of compliers:

$$\theta_{LATE} = E(Y_1 - Y_0 \mid D_1 - D_0 = 1) \equiv E(Y_1 - Y_0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0).$$

- We have

$$E\left(Y_1 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) = \Pr\left(\alpha + \beta + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right)$$
$$= 1 - \frac{\Pr\left(U \leq -\alpha - \beta, V \leq -\pi_0\right) - \Pr\left(U \leq -\alpha - \beta, V \leq -\pi_0 - \pi_1\right)}{\Pr\left(V \leq -\pi_0\right) - \Pr\left(V \leq -\pi_0 - \pi_1\right)}$$

  and similarly

$$E\left(Y_0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right) = \Pr\left(\alpha + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0\right)$$
$$= 1 - \frac{\Pr\left(U \leq -\alpha, V \leq -\pi_0\right) - \Pr\left(U \leq -\alpha, V \leq -\pi_0 - \pi_1\right)}{\Pr\left(V \leq -\pi_0\right) - \Pr\left(V \leq -\pi_0 - \pi_1\right)}.$$

- Finally,

$$\begin{aligned}\theta_{LATE} &= \frac{1}{\Phi\left(-\pi_0\right) - \Phi\left(-\pi_0 - \pi_1\right)}\left[\Phi_2\left(-\alpha, -\pi_0; \rho\right) - \Phi_2\left(-\alpha, -\pi_0 - \pi_1; \rho\right)\right. \\ &\quad \left. -\Phi_2\left(-\alpha - \beta, -\pi_0; \rho\right) + \Phi_2\left(-\alpha - \beta, -\pi_0 - \pi_1; \rho\right)\right].\end{aligned}$$

  where we are using $\Phi_2\left(r, s; \rho\right) = \Pr\left(U \leq r, V \leq s\right)$ as the notation for standard normal bivariate probabilities.

Remarks

- 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 the index model assumption for Y_1 and Y_0. So we do not need monotonicity in the relationship between Y and D.
- We have

$$\begin{aligned}\theta_{ATE} &= \theta_{LATE} \Pr\left(\text{compliers}\right) + E\left(Y_1 - Y_0 \mid \text{never-takers}\right) \Pr\left(\text{never-takers}\right) \\ &\quad + E\left(Y_1 - Y_0 \mid \text{always-takers}\right) \Pr\left(\text{always-takers}\right).\end{aligned}$$

4.2 Models with additive errors: switching regressions

The switching regression model with endogenous switch

- The model is as follows:

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

- 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.

- In the current model β_i may also 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 (both notations are useful):

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

  where $\alpha = \mu_0$ and $U_i = V_{0i}$.

- 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)[1]

- 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:

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

- 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.

[1] Willis, R. and S. Rosen (1979): "Education and Self -Selection," *JPE*, 87, S7-S36.

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

$$\overline{\beta} = \frac{Cov(Z, Y)}{Cov(Z, D)}.$$

- Otherwise, $\overline{\beta}$ does not coincide with the IV estimand. Note that 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. Let $\varphi(Z_i) = E(V_{1i} - V_{0i} \mid D_i = 1, Z_i)$, then the model can be written as

$$Y_i = \alpha + \overline{\beta} D_i + \varphi(Z_i) D_i + \zeta_i$$

with $E(\zeta_i \mid Z_i) = 0$.

- When we do ordinary IV estimation we are not taking into account the variable $\varphi(Z_i) D_i$.

- The quantity $\varphi(z)$ is the average excess return for college-educated people with $Z_i = z$. In the distance to college example ($Z = 1$ college near, $Z = 0$ college far), we would expect $\varphi(1) \leq \varphi(0)$.

- The average treatment effect on the treated is

$$\alpha_{TT} = E(Y_{1i} - Y_{0i} \mid D_i = 1) = \overline{\beta} + E(V_{1i} - V_{0i} \mid D_i = 1),$$

whereas the LATE is

$$\alpha_{LATE} = E(Y_{1i} - Y_{0i} \mid D_{1i} - D_{0i} = 1) = \overline{\beta} + E(V_{1i} - V_{0i} \mid -\gamma_0 - \gamma_1 \leq \varepsilon_i < -\gamma_0).$$

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 efficiently 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:

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

where $\Phi_i = \Phi\left(\gamma_0 + \gamma_1 Z_i\right)$, $\phi_i = \phi\left(\gamma_0 + \gamma_1 Z_i\right)$, and $\lambda_i = \lambda\left(\gamma_0 + \gamma_1 Z_i\right)$.

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 (discussed in another class note) applies.
- Namely, with a continuous exclusion restriction $E\left(Y_{1i} \mid X_i\right)$ and $E\left(Y_{0i} \mid X_i\right)$ will be identified up to a constant (here X_i denotes controls that so far we omitted for simplicity).
- However, the constants are important in this case because they determine the average treatment effect of D on Y. Unfortunately, identification of the constants require an identification at infinity argument.

5 Marginal treatment effects

Introduction

- When the support of Z is not binary, there is a multiplicity of causal effects that can be calculated.
- In any event, the question arises: 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 of values (z, z'):

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

The multiplicity is of an even higher dimension when there is more than one instrument available.

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

  $$D_{zi} \geq D_{z'i}$$

  for all units in the population, or $D_{zi} \leq D_{z'i}$.

Latent index representation

- Alternatively we can postulate an index model for D_z:

  $$D_z = 1(\mu(z) - U > 0) \quad \text{and } U \perp Z,$$

  which can be an economically meaningful way of organizing different LATEs. This is the approach adopted in Heckman and Vytlacil (2005).
- Note that the observed D is $D = D_Z$.
- Vytlacil (2002) showed that the monotonicity and index model assumptions are equivalent.
- Showing that the latent index assumption implies independence and monotonicity is immediate. What Vytlacil proves is that independence and monotonicity imply a latent index representation.
- This result is nice because it connects LATE thinking with the kind of thinking used in econometric selection models.
- Without loss of generality we can set $\mu(z) = P(z)$ and take U as uniformly distributed in the $(0, 1)$ interval. To see this note that

  $$1(\mu(z) > U) = 1\{F_U[\mu(z)] > F_U(U)\} = 1\left(P(z) > \widetilde{U}\right)$$

  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 $P(0)$ and $P(1)$. Suppose that $P(0) < P(1)$. Always-takers have $U < P(0)$, compliers have a value of U between $P(0)$ and $P(1)$, and never-takers have $U > P(1)$. 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

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

  or

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

- 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 called "local IV" or MTE by Heckman and Vytlacil:

$$MTE(P(z)) = \frac{\partial E(Y \mid P(Z) = P(z))}{\partial P(z)}.$$

- $\alpha_{LATE}(P(z), P(z'))$ gives the ATE for individuals who would change schooling status from changing $P(Z)$ from $P(z')$ to $P(z)$:

$$\alpha_{LATE}\left(P(z), P\left(z'\right)\right) = E\left[Y_1 - Y_0 \mid P\left(z'\right) < U < P(z)\right]$$

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

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

- Moreover,

$$\alpha_{ATE} = \int_0^1 MTE(u)\, du,$$

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

Policy-relevant treatment effects

- Constructing suitably integrated $MTE(u)\,s$ 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

  $$E(Y \mid P(Z) = P(z), X = x)$$

  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 $Y = Y_0 + (Y_1 - Y_0)\,D$ and write

  $$\begin{aligned} E(Y \mid P(Z)) &= E(Y_0 \mid P(Z)) + E((Y_1 - Y_0)\,D \mid P(Z)) \\ &= E(Y_0) + E[Y_1 - Y_0 \mid D = 1, P(Z)]\,P(Z) \end{aligned}$$

  The quantity $E[Y_1 - Y_0 \mid D = 1, P(Z)]$ is constant under homogeneity, so that $E(Y \mid P(Z))$ is linear in $P(Z)$.
- Moffitt (2008) Outcomes are a nonlinear function of participation probabilities. The degree of this nonlinearity, and hence the shape of the marginal response curve, can be estimated with series methods such as power series or splines. An illustration is provided for the returns to higher education in the U.K, indicating that marginal returns to higher education fall as the proportion of the population with higher education rises, thus providing evidence of heterogeneity in returns.

Concluding remarks

- Two remarks about the importance of unobserved heterogeneity in IV settings:
  - 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 (the "selection problem").

References

[1] Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models", *Journal of the American Statistical Association*, 97, 284-292.

[2] Angrist, J., G. Imbens, and K. Graddy (2000): "The Interpretation of Instrumental Variable Estimators in Simultaneous Equations Models with an Application to the Demand for Fish", *Review of Economic Studies*, 67, 499-528.

[3] Imbens, G. W. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 467-475.

[4] Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variable Models", *Review of Economic Studies*, 64, 555-574.

[5] Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation", *Econometrica*, 73, 669-738.

[6] Moffitt, R. (2008): "Estimating Marginal Treatment Effects in Heterogeneous Populations", unpublished.

[7] Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Results" *Econometrica*, 70, 331-341.

A Local Linear Regression

- Let us consider estimating the regression function $g(x) = E(Y \mid X = x)$ from given observations $\{Y_i, X_i\}_{i=1}^n$.
- A linear approximation to $g(x)$ at a fixed point r is

$$g(x) \approx a(r) + b(r)'(x - r)$$

where $a(r) = g(r)$ and $b(r) = \partial g(r)/\partial r$ for x in a neighborhood of r.

- Thus, locally, the problem of finding $g(r)$ is equivalent to finding the intercept of the approximating regression line.
- The local neighborhood may be determined by a kernel function K and a smoothing parameter γ_n, which suggests using the least squares criterion

$$\sum_{i=1}^{n} K\left(\frac{X_i - r}{\gamma_n}\right)\left[Y_i - a - b'(X_i - r)\right]^2.$$

- Minimization with respect to a and b gives an estimate $\left[\widehat{a}(r), \widehat{b}(r)\right]$ of $g(r)$ and $\partial g(r)/\partial r$.
- Letting $K_i(r) = K\left(\frac{X_i - r}{\gamma_n}\right)$ and

$$\begin{aligned}
\overline{Y}(r) &= \frac{1}{\sum_{i=1}^n K_i(r)} \sum_{i=1}^{n} K_i(r) Y_i \\
\overline{D}_X(r) &= \frac{1}{\sum_{i=1}^n K_i(r)} \sum_{i=1}^{n} K_i(r)(X_i - r) \\
\widetilde{Y}_i(r) &= Y_i - \overline{Y}(r) \\
\widetilde{D}_{Xi}(r) &= (X_i - r) - \overline{D}_X(r),
\end{aligned}$$

the estimates are

$$\begin{aligned}
\widehat{b}(r) &= \left(\sum_{i=1} K_i(r) \widetilde{D}_{Xi}(r) \widetilde{D}_{Xi}(r)'\right)^{-1} \sum_{i=1} K_i(r) \widetilde{D}_{Xi}(r) \widetilde{Y}_i(r) \\
\widehat{a}(r) &= \overline{Y}(r) - \overline{D}_X(r)' \widehat{b}(r).
\end{aligned}$$

- The Nadaraya-Watson (NW) estimate of $g(r)$ is $\overline{Y}(r)$.
- If the distribution of the X's in a neighborhood of r is symmetric around r, then $\overline{D}_X(r) \approx 0$ and $\widehat{a}(r) \approx \overline{Y}(r)$ (i.e. the NW and local linear regression estimates of $g(r)$ will be close to each other).

- However, if the X's in a neighborhood of r are mostly below (above) r then $\overline{D}_X(r)$ will be negative (positive). In such case the local linear regression estimate applies a first-order correction to $\overline{Y}(r)$ using the local slope estimate $\widehat{b}(r)$.

- Thus, NW can be regarded as a local regression approximation to $g(r)$ of order zero, whereas $\widehat{a}(r)$ is a similar approximation of order one.

- Note that in the case where X_i is discrete and $K\left(\frac{X_i - r}{\gamma_n}\right) = 1(X_i = r)$, the criterion boils down to $\sum_{X_i=r}(Y_i - a)^2$ which is minimized by the sample mean of Y_i for the observations with $X_i = r$.

- Jianqing Fan (*JASA*, 1992) showed that local linear regression avoids the drawbacks of other types of kernel estimators such as NW.

- Local linear regression adapts to various types of designs (random, highly clustered, nearly uniform) and reduces boundary effects.

B Heterogeneity of gains vs. heterogeneity of treatments

- Hamermesh and Donald (2008)[2] look at the relationship between college major and earnings using a new dataset. The nice aspects of the sample they use are its high homogeneity and the wealth of background controls available. Yet the authors find significant differences in returns to different majors.

- The Hamermesh–Donald results (and other similar results on differences in returns to majors) have implications for the interpretation of the heterogeneity in returns to college found in the literature. We elaborate on some of these implications in what follows.

- Suppose for the sake of the argument that there are two majors with indicators (D_A, D_B) and potential outcomes (Y_0, Y_{1A}, Y_{1B}) corresponding to high-school, college with major A, and college with major B. Observed earnings are

  $$Y = Y_0 + (Y_{1A} - Y_0) D_A + (Y_{1B} - Y_0) D_B$$

  and the college–high school indicator is

  $$D = D_A + D_B$$

- Suppose we observe Y and D. Under exogeneity $(Y_0, Y_{1A}, Y_{1B}) \perp (D_A, D_B)$, the OLS impact is a linear combination of the to average returns:

  $$\beta = E(Y \mid D = 1) - E(Y \mid D = 0) = \pi E(Y_{1A} - Y_0) + (1 - \pi) E(Y_{1B} - Y_0)$$

  where $\pi = \Pr(D_A = 1 \mid D_A + D_B = 1)$. Since π is the result of choice, β does not measure any meaningful causal effect.

  – To see this, note that

  $$E(Y \mid D = 0) = E(Y_0 \mid D = 0) = E(Y_0)$$

  and

  $$\begin{aligned} E(Y \mid D = 1) &= E(Y_{1A} D_A + Y_{1B} D_B \mid D = 1) \\ &= E(Y_{1A} \mid D_A = 1) \pi + E(Y_{1B} \mid D_B = 1)(1 - \pi) \\ &= E(Y_{1A}) \pi + E(Y_{1B})(1 - \pi). \end{aligned}$$

[2] Daniel Hamermesh and Stephen Donald (2008): "The effect of college curriculum on earnings: An affinity identifier for non-ignorable non-response bias", *Journal of Econometrics*, 144(2), 479-491.

- Note that also

$$Y = Y_0 + (Y_{1B} - Y_0) D + (Y_{1A} - Y_{1B}) D_A$$

so that under exogeneity, if we use data for graduates only, OLS of Y on D_A gives $E(Y_{1A} - Y_{1B})$.

- Now suppose non-exogeneity but that the IV assumption holds $(Y_0, Y_{1A}, Y_{1B}) \perp Z$, with Z binary.

- Let the potential indicators of major choice be

$$D_A = \begin{cases} D_{1A} & \text{if } Z = 1 \\ D_{0A} & \text{if } Z = 0 \end{cases} \qquad D_B = \begin{cases} D_{1B} & \text{if } Z = 1 \\ D_{0B} & \text{if } Z = 0 \end{cases}$$

and assume absence of defiers in both groups.

- It turns out that the IV impact is a linear combination of average treatment effects for two different groups of compliers:

$$\frac{Cov(Z, Y)}{Cov(Z, D)} = E(Y_{1A} - Y_0 \mid D_{1A} - D_{0A} = 1)(1 - \lambda) + E(Y_{1B} - Y_0 \mid D_{1B} - D_{0B} = 1)\lambda$$

where λ is given by the odds of compliers in one college major relative to the other. Specifically, $\lambda = 1/(1 + \varphi)$ and $\varphi = \Pr(D_{1A} - D_{0A} = 1) / \Pr(D_{1B} - D_{0B} = 1)$.

  - A detailed derivation is as follows. We have

$$\begin{aligned} E(Y \mid Z = 1) &= E[Y_0 + (Y_{1A} - Y_0) D_A + (Y_{1B} - Y_0) D_B \mid Z = 1] \\ &= E(Y_0) + E[(Y_{1A} - Y_0) D_{1A}] + E[(Y_{1B} - Y_0) D_{1B}] \end{aligned}$$

Similarly,

$$E(Y \mid Z = 0) = E(Y_0) + E[(Y_{1A} - Y_0) D_{0A}] + E[(Y_{1B} - Y_0) D_{0B}]$$

and subtracting both expressions

$$E(Y \mid Z = 1) - E(Y \mid Z = 0) = E[(Y_{1A} - Y_0)(D_{1A} - D_{0A})] + E[(Y_{1B} - Y_0)(D_{1B} - D_{0B})]$$

$$\begin{aligned} &= E(Y_{1A} - Y_0 \mid D_{1A} - D_{0A} = 1) \Pr(D_{1A} - D_{0A} = 1) \\ &\quad + E(Y_{1B} - Y_0 \mid D_{1B} - D_{0B} = 1) \Pr(D_{1B} - D_{0B} = 1) \end{aligned}$$

Moreover,

$$\Pr(D_{1A} - D_{0A} = 1) = E(D_{1A} - D_{0A}) = E(D_A \mid Z = 1) - E(D_A \mid Z = 0)$$

and similarly for $\Pr(D_{1B} - D_{0B} = 1)$. Next, we have

$$E(D \mid Z = 1) = E(D_A + D_B \mid Z = 1) = E(D_A \mid Z = 1) + E(D_B \mid Z = 1)$$

so that

$$\begin{aligned} E(D \mid Z=1) - E(D \mid Z=0) &= [E(D_A \mid Z=1) - E(D_A \mid Z=0)] \\ &\quad + [E(D_B \mid Z=1) - E(D_B \mid Z=0)] \end{aligned}$$

Finally,

$$\begin{aligned} \alpha_{IV} &= \frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(D \mid Z=1) - E(D \mid Z=0)} \\ &= E(Y_{1A} - Y_0 \mid D_{1A} - D_{0A} = 1)(1-\lambda) + E(Y_{1B} - Y_0 \mid D_{1B} - D_{0B} = 1)\lambda \end{aligned}$$

where

$$\lambda = \frac{E(D_B \mid Z=1) - E(D_B \mid Z=0)}{[E(D_A \mid Z=1) - E(D_A \mid Z=0)] + [E(D_B \mid Z=1) - E(D_B \mid Z=0)]}.$$

- If $Y_{1A} - Y_0$ and $Y_{1B} - Y_0$ are constant, the IV impact is just a weighted combination of the two returns:

$$\alpha_{IV} = (Y_{1A} - Y_0)\lambda + (Y_{1B} - Y_0)(1-\lambda).$$

- In neither case, the IV parameter provides a meaningful causal effect.
- It is not difficult to generalize these expressions to more than two majors.

Final comment

- A problem of aggregating educational categories is that returns are less meaningful. In fact, it is conceivable that one would find less unobserved heterogeneity in returns if using a finer classification of educational achievement.
- Sometimes the choice of aggregation of educational categories into just two categories has a methodological motivation (possibly because the techniques under consideration are only well developed or feasible for binary endogenous explanatory variables). Thus, a methodological emphasis may offer new opportunities but also impose constraints on the analysis.

C Potential outcome distributions and local instrumental variables

- Carneiro and Lee (2007)[3] extend the Heckman–Vytlacil method of local instrumental variables to the estimation of distributions of potential outcomes. They apply the marginal treatment effect methodology to estimating distributions of potential wages for each schooling category.

- Recall that a LATE defined for two points of the propensity score $P(Z) = \Pr(D = 1 \mid Z)$ is:

$$\alpha_{LATE}\left(P(z), P\left(z'\right)\right) = \frac{E[Y \mid P(Z) = P(z)] - E[Y \mid P(Z) = P(z')]}{P(z) - P(z')}. \tag{2}$$

- Taking limits as $z \to z'$, we get Heckman–Vytlacil's "local IV" or MTE:

$$MTE(P(z)) = \frac{\partial E[Y \mid P(Z) = P(z)]}{\partial P(z)},$$

  which satisfies

$$MTE(P(z)) = E[Y_1 - Y_0 \mid U = P(z)]$$

  where U is the uniformly distributed error term in the index model.

- Using $P(Z)$ as instrument, we can write an expression equivalent to (2) for Abadie (2002)'s formulation of the *cdf*'s of potential outcomes in the subpopulation of compliers (section 3.2).

- Letting $h(.)$ be an arbitrary function, for Y_1 we have

$$E\left[h(Y_1) \mid P\left(z'\right) < U < P(z)\right] = \frac{E[h(Y)D \mid P(Z) = P(z)] - E[h(Y)D \mid P(Z) = P(z')]}{P(z) - P(z')}.$$

- Taking limits as $z \to z'$, we get the local IV version:

$$E[h(Y_1) \mid U = P(z)] = \frac{\partial E[h(Y)D \mid P(Z) = P(z)]}{\partial P(z)} \tag{3}$$

- Note that $E[h(Y)D \mid P(Z) = P(z)] = E[h(Y) \mid D = 1, P(Z) = P(z)]P(z)$, so that also

$$E[h(Y_1) \mid U = P(z)] = E[h(Y) \mid D = 1, P(Z) = P(z)] + P(z)\frac{\partial E(h(Y) \mid D = 1, P(Z) = P(z))}{\partial P(z)}. \tag{4}$$

- For $h(Y) = 1(Y \leq r)$, (3) or (4) gives the *cdf* of potential wages as college graduates for individuals that are indifferent between high school and college when $P(Z) = P(z)$.

[3] Pedro Carneiro and Sokbae Lee (2007): "Changes in College Enrollment and Wage Inequality: Distinguishing Price and Composition Effects", unpublished.

- Similarly, we have

$$E\left[h\left(Y_0\right) \mid U = P\left(z\right)\right] = \frac{\partial E\left[h\left(Y\right)\left(1-D\right) \mid P\left(Z\right) = P\left(z\right)\right]}{\partial\left[1-P\left(z\right)\right]} = -\frac{\partial E\left[h\left(Y\right)\left(1-D\right) \mid P\left(Z\right) = P\left(z\right)\right]}{\partial P\left(z\right)}$$

- Moreover, due to

$$E\left[h\left(Y\right)\left(1-D\right) \mid P\left(Z\right) = P\left(z\right)\right] = E\left[h\left(Y\right) \mid D = 0, P\left(Z\right) = P\left(z\right)\right]\left[1-P\left(z\right)\right],$$

  we can also write

$$E\left[h\left(Y_0\right) \mid U = P\left(z\right)\right] = E\left[h\left(Y\right) \mid D = 0, P\left(Z\right) = P\left(z\right)\right] - \left[1-P\left(z\right)\right]\frac{\partial E\left[h\left(Y\right) \mid D = 0, P\left(Z\right) = P\left(z\right)\right]}{\partial P\left(z\right)} \tag{5}$$

- Expressions (4) and (5) are the objects analyzed in Carneiro and Lee (2007).