New Developments in Econometrics **Cemmap, UCL, June 2009**

Lecture 5, Tuesday June 16th , 15.15-16.15

Instrumental Variables with Treatment Effect Heterogeneity: Local Average Treatment Effects

1. Introduction

In this lecture we discuss the interpretation of instrumental variables estimators allowing for general heterogeneity in the effect of the endogenous regressor. We shall see that instrumental variables estimators generally estimate average treatment effects, with the specific average depending on the choice of instruments. Initially we focus on the case where both the instrument and the endogenous regressor are binary. The example we will use is based on one of the best known examples of instrumental variables, the paper by Joshua Angrist on estimating the effect of veteran status on earnings (Angrist, 1990). We also discuss the case where the instrument and or the endogenous variable take on multiple values, and incorporate the presence of covariates.

The general theme of this lecture is that with heterogenous treatment effects, endogeneity creates severe problems for identification of population averages. Population average causal effects are only estimable under very strong assumptions on the effect of the instrument on the endogenous regressor (sometimes referred to as "identification at infinity", Chamberlain, 1986), or under the constant treatment effect assumptions. Without such assumptions we can only identify average effects for subpopulations that are induced by the instrument to change the value of the endogenous regressors. Following Angrist, Imbens and Rubin (1996), we refer to such subpopulations as *compliers*, and we refer to the average treatment effect that is point identifed as the *local average treatment effect* (Imbens and Angrist, 1994). The "complier" terminology stems from the canonical example of a randomized experiment with noncompliance. In this example a random subpopulation is assigned to the treatment, but some of the individuals do not comply with their assigned treatment.

These complier subpopulations are not necessarily the subpopulations that are *ex ante* the most interesting subpopulations. The reason to nevertheless focus on these subpopulations is that the data are generally not informative about average effects for other subpopulations without extrapolation, similar to the way in which a randomized experiment conducted on men is not informative about average effects for women without extrapolation. The set up here allows the researcher to sharply separate the extrapolation to the (sub-)population of interest, from exploration of the information in the data about the causal effect of interest. The latter analysis relies primarily on relatively interpretable, and substantively meaningful assumptions, and it avoids functional form or distributional assumptions. Subsequently, given estimates for the compliers, one can these estimates in combination with the data to assess the plausibility of extrapolating the local average treatment effect to other subpopulations, using the information on outcomes given one of the two treatment levels and covariates, or construct bounds on the average effects for the primary population of interest using the bounds approach from Manski (e.g., Manski, 2008).

With multiple instruments, and/or with covariates, one can assess the evidence for heterogeneity, and therefore investigate the plausibility of extrapolation to the full population more extensively.

2. Linear Instrumental Variables with Constant Coefficients

First let us briefly review standard textbook linear instrumental variables methods (e.g., Wooldridge, 2000). In the example from Angrist (1990) we use to illustrate the concepts discussed in this lecture we are interested in the causal effect of military service on earnings, using eligibility for the draft as the instrument. Let Y_i be the outcome of interest for unit i (log earnings in the example), W_i the binary endogenous regressor (an indicator for veteran status), and Z_i the binary instrument (a binary indicator for draft eligibility). The standard set up is as follows. A linear model is postulated for the relation between the outcome and the endogenous regressor:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \varepsilon_i. \tag{1}$$

This is a structural, behavioral, or causal relationship (we use the terms interchangeably).

The concern is that the regressor W_i is endogenous, that is, that W_i is correlated with the unobserved component of the outcome, ε_i. Suppose that we are confident that a second observed covariate, the instrument Z_i, is both uncorrelated with the unobserved component ε_i and correlated with the endogenous regressor W_i. The solution is to use Z_i as an instrument for W_i. There are a couple of ways to implement this.

In Two-Stage-Least-Squares (TSLS) we first estimate a linear regression of the endogenous regressor on the instrument by least squares. Let the estimated regression function be

$$\hat{W}_i = \hat{\pi}_0 + \hat{\pi}_1 \cdot Z_i.$$

Then we regress the outcome on the predicted value of the endogenousr regressor, using least squares:

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau}^{\text{tsls}} \cdot \hat{W}_i.$$

Alternatively, with a single instrument we can estimate the two reduced form regressions

$$\hat{Y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot Z_i, \qquad \text{and} \quad \hat{W}_i = \hat{\pi}_0 + \hat{\pi}_1 \cdot Z_i,$$

by least squares and estimate β_1 through Indirect Least Squares (ILS) as the ratio

$$\hat{\tau}^{\text{ils}} = \hat{\gamma}_1 / \hat{\pi}_1,$$

irrespective of the validity of the behavioral model.

In the case with a single instrument and single endogenous regressor, we end up in both cases with the ratio of the sample covariance of Y_i and Z_i to the sample covariance of W_i and Z_i.

$$\hat{\tau}^{\text{iv}} = \hat{\tau}^{\text{ils}} = \hat{\tau}^{\text{tsls}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y}) \cdot (Z_i - \bar{Z})}{\frac{1}{N}\sum_{i=1}^{N}(W_i - \bar{W}) \cdot (Z_i - \bar{Z})}.$$

This estimator is consistent for

$$\tau^{\text{iv}} = \frac{\mathbb{E}\left[(Y_i - \mathbb{E}[Y_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]}{\mathbb{E}\left[(W_i - \mathbb{E}[W_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]}. \tag{2}$$

Using a central limit theorem for all the moments and the delta method we can infer the large sample distribution without additional assumptions:

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{iv}} - \tau^{\text{iv}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\varepsilon_i^2 \cdot (Z_i - \mathbb{E}[Z_i])^2\right]}{\left(\mathbb{E}\left[(W_i - \mathbb{E}[W_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]\right)^2}\right),$$

where $\varepsilon_i = Y_i - \mathbb{E}[Y_i] - \tau^{\text{iv}} \cdot (W_i - \mathbb{E}[W_i])$. Under independence between the residual ε_i and the instrument Z_i, the asymptotic distribution further simplifies to:

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{iv}} - \tau^{\text{iv}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\varepsilon_i^2\right] \cdot \mathbb{E}\left[(Z_i - \mathbb{E}[Z_i])^2\right]}{\left(\mathbb{E}\left[(W_i - \mathbb{E}[W_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]\right)^2}\right),$$

3. Potential Outcomes

First we set up the problem in a slightly different way, using Rubin's (1974) potential outcomes approach to causality. This set up of the instrumental variables problem originates with Imbens and Angrist (1994). Let $Y_i(0)$ and $Y_i(1)$ be two potential outcomes for unit i, one for each value of the endogenous regressor or treatment. The first potential outcome $Y_i(0)$ measures the outcome if person i were not to serve in the military, irrespective of whether this person served or not. The second potential outcome, $Y_i(1)$, measures the outcome given military service, again irrespective of whether the person served or not. We are interested in the causal effect of military service, $Y_i(1) - Y_i(0)$. We cannot directly observe this since we can only observe either $Y_i(0)$ or $Y_i(1)$, never both. Let W_i be the realized value of the endogenous regressor, equal to zero or one. We observe W_i and

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 1 \\ Y_i(0) & \text{if } W_i = 0. \end{cases}$$

So far the set up is identical to that in the analysis under unconfoundedness in Lecture XX. Now we introduce additional notation by defining similar potential outcomes for the treatment. Initially we focus on the case with a binary instrument Z_i. In the Angrist example, Z_i is a binary indicator for having a draft number below the cutoff value that implied a potential recruit would get called up for military service, and thus an indicator for being draft eligible. Define two potential outcomes $W_i(0)$ and $W_i(1)$, representing the value of the endogenous regressor given the two values for the instrument. The actual or realized (and observed) value of the endogenous variable is

$$W_i = Y_i(Z_i) = \begin{cases} W_i(1) & \text{if } Z_i = 1 \\ W_i(0) & \text{if } Z_i = 0. \end{cases}$$

In summary, we observe the triple (Z_i, W_i, Y_i), where $W_i = W_i(Z_i)$ and $Y_i = Y_i(W_i(Z_i))$.

4. Local Average Treatment Effects

In this section we interpret the estimand (2) under weaker assumptions than the linear additive model set up in (1).

4.1. Assumptions

The key instrumental variables assumption is

Assumption 1 (Independence)

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1), W_i(0), W_i(1)).$$

This assumption requires that the instrument is as good as randomly assigned, and that it does not directly affect the outcome. The assumption is formulated in a nonparametric way, without definitions of residuals that are tied to functional forms.

It is important to note that this assumption is *not* implied by random assignment of Z_i. To see this, an alternative formulation of the assumption, slightly generalizing the notation, is useful. First we postulate the existence of four potential outcomes, $Y_i(z, w)$, corresponding to the outcome that would be observed if the instrument was exogenously set to $Z_i = z$ and the treatment was exogenously set to $W_i = w$. Then the independence assumption is the combination of two assumptions.

Assumption 2 (Random Assignment)

$$Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1)).$$

Assumption 3 (Exclusion Restriction)

$$Y_i(z,w) = Y_i(z',w), \qquad \text{for all } z, z', w.$$

The first of these two assumptions is implied by random assignment of Z_i. It can be weakened in the presence of covariates to unconfoundedness. The second assumption is substantive, and randomization has no bearing on it. It corresponds to the notion that there is no direct effect of the instrument on the outcome other than through the treatment. In the model-based version of this, (1), it is captured by the absence of Z_i in the behavioral equation. This assumption has to be argued on a case-by-case basis.

It is useful for our approach to think about the compliance behavior of the different individuals or units, that is how they respond in terms of the treatment received to different values of the instrument. Table 1 gives the four possible pairs of values $(W_i(0), W_i(1))$, given the binary nature of the treatment and instrument and their labels. The labels refer to the canonical example of a randomized experiment with imperfect compliance.

Table 1: Compliance Types

		$W_i(0)$	
		0	1
$W_i(1)$	0	never-taker	defier
	1	complier	always-taker

We cannot directly establish the type of an individual based on what we observe for them (the triple $Z_i, W_i, Y_i)$) since we only see the pair (Z_i, W_i), not the pair $(W_i(0), W_i(1))$

(typically observing Y_i is immaterial for this argument). Nevertheless, we can rule out some possibilities. Table 2 summarizes the information about compliance behavior from observed treatment status and instrument. For each pair of (Z_i, W_i) values there are two possible types, with the two others ruled out.

Table 2: Compliance Type by Treatment and Instrument

		Z_i	
		0	1
W_i	0	complier/never-taker	never-taker/defier
	1	always-taker/defier	complier/always-taker

To make additional progress we we consider a *monotonicity* assumption, also known as the *no-defiers* assumption, introduced by Imbens and Angrist (1994):

Assumption 4 (Monotonicity/No-Defiers)

$$W_i(1) \geq W_i(0).$$

This monotonicity assumption is very apealling in many applications. It is implied directly by many (constant coefficient) latent index models of the type:

$$W_i(z) = 1\{\pi_0 + \pi_1 \cdot z + \varepsilon_i > 0\}, \tag{3}$$

which would imply $W_i(1) \geq W_i(0)$ if $\pi_1 \geq 0$ and $W_i(1) \leq W_i(0)$ otherwise. In the canonical example of a randomized experiment with non-compliance this assumption is very plausible: if Z_i is assignment to a treatment, and W_i is an indicator for receipt of treatment, it makes sense that there are few, if any, individuals who always to the exact opposite of what their assignment is.

4.2. The Local Average Treatment Effect

Given monotonicity we can infer more about an individual's compliance behavior, as summarized in Table 3. For individuals with (Z_i, W_i) equal to $(0, 1)$ or $(1, 0)$ we can now determine their type. For individuals with (Z_i, W_i) equal to $(0, 0)$ or $(1, 1)$ there are still multiple types consistent with the observed behavior. Nevertheless, we can stochastically infer the compliance types.

Table 3: Compliance Type by Treatment and Instrument given Monotonicity

		Z_i	
		0	1
W_i	0	complier/never-taker	never-taker
	1	always-taker	complier/always-taker

Now we proceed to identifying the marginal distribution of types and conditional potential outcome distributions. Let π_c, π_n, and π_a be the population proportions of compliers, never-takers and always-takers respectively. We can identify those from the population distribution of treatment and instrument status:

$$\mathbb{E}[W_i|Z_i = 0] = \pi_a, \qquad \mathbb{E}[W_i|Z_i = 1] = \pi_a + \pi_c,$$

which we can invert to infer the population shares of the different types:

$$\pi_a = \mathbb{E}[W_i|Z_i = 0], \qquad \pi_c = \mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0],$$

and

$$\pi_n = 1 - \pi_a - \pi_c = 1 - \mathbb{E}[W_i|Z_i = 1].$$

Now consider average outcomes by instrument and treatment status. In the (Z_i, W_i) equal to $(0,1)$ or $(1,0)$ subpopulations these expectations have a simple interpretation:

$$\mathbb{E}[Y_i|W_i = 0, Z_i = 1] = \mathbb{E}[Y_i(0)|\text{never} - \text{taker}], \tag{4}$$

and

$$\mathbb{E}[Y_i|W_i = 1, Z_i = 0] = \mathbb{E}[Y_i(1)|\text{always} - \text{taker}]. \tag{5}$$

In the (Z_i, W_i) equal to $(0,0)$ or $(1,1)$ the conditional outcome expectations are mixtures of expected values for compliers and nevertakers and compliers and alwaystakers respectively:

$$\mathbb{E}[Y_i|W_i = 0, Z_i = 0] = \frac{\pi_c}{\pi_c + \pi_n} \cdot \mathbb{E}[Y_i(0)|\text{complier}] + \frac{\pi_n}{\pi_c + \pi_n} \cdot \mathbb{E}[Y_i(0)|\text{never} - \text{taker}], \tag{6}$$

and

$$\mathbb{E}[Y_i|W_i = 1, Z_i = 1] = \frac{\pi_c}{\pi_c + \pi_a} \cdot \mathbb{E}[Y_i(1)|\text{complier}] + \frac{\pi_a}{\pi_c + \pi_a} \cdot \mathbb{E}[Y_i(1)|\text{always} - \text{taker}]. \tag{7}$$

From these relationships we can infer the average outcome by treatment status for compliers, first by combining (4) and (6),

$$\mathbb{E}[Y_i(0)|\text{complier}] = \frac{\pi_c + \pi_n}{\pi_n} \cdot \mathbb{E}[Y_i|W_i = 0, Z_i = 0] - \frac{\pi_c}{\pi_n} \cdot \mathbb{E}[Y_i|W_i = 0, Z_i = 1],$$

and then by combining (5) and (7)

$$\mathbb{E}[Y_i(1)|\text{complier}] = \frac{\pi_c + \pi_a}{\pi_a} \cdot \mathbb{E}[Y_i|W_i = 1, Z_i = 1] - \frac{\pi_c}{\pi_a} \cdot \mathbb{E}[Y_i|W_i = 1, Z_i = 0].$$

Thus we can infer the average effect for compliers, $\mathbb{E}[Y(1) - Y_i(0)|\text{complier}] = \mathbb{E}[Y_i(1)|\text{complier}] - \mathbb{E}[Y_i(0)|\text{complier}]$.

It turns out this is equal to the instrumental variables estimand (2). Consider the least squares regression of Y_i on a constant and Z_i. The slope coefficient in that regression estimates

$$\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0].$$

The two terms are equal to:

$$\mathbb{E}[Y_i|Z_i = 1] = \mathbb{E}[Y_i(1)|\text{complier}] \cdot \pi_c + \mathbb{E}[Y_i(0)|\text{never} - \text{taker}] \cdot \pi_0 + \mathbb{E}[Y_i(1)|\text{always} - \text{taker}] \cdot \pi_a.$$

and

$$\mathbb{E}[Y_i|Z_i = 0] = \mathbb{E}[Y_i(0)|\text{complier}] \cdot \pi_c + \mathbb{E}[Y_i(0)|\text{never} - \text{taker}] \cdot \pi_0 + \mathbb{E}[Y_i(1)|\text{always} - \text{taker}] \cdot \pi_a.$$

Hence the difference is

$$\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] = \mathbb{E}[Y_i(1) - Y_i(0)|\text{complier}] \cdot \pi_c.$$

The same argument can be used to show that the slope coefficient in the regression of W_i on Z_i is

$$\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0] = \pi_c.$$

Hence the instrumental variables estimand, the ratio of these two reduced form estimands, is equal to the local average treatment effect

$$\beta^{\text{iv}} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]} = \mathbb{E}[Y_i(1) - Y_i(0)|\text{complier}]. \tag{8}$$

The key insight is that the data are informative only about the average effect for compliers only. Put differently, the data are not informative about the average effect for nevertakers because nevertakers are never seen receiving the treatment, and they are not informative about the average effect for alwaystakers because alwaystakers are never seen without the treatment. A similar insight in a parametric settings is discussed in Björklund and Moffitt (1987). (These results do not take away from the fact that one can construct informative bounds about the average effect for nevertakers or alwaystakers based on the outcomes we do observe for such individuals, in the spirit of the work by Manski, 2008.)

A special case of considerable interest is that with one-side non-compliance. Suppose that $W_i(0)=0$, so that those assigned to the control group cannot receive the active treatment (but those assigned to the active treatment can choose to receive it or not, so that $W_i(1) \in \{0,1\}$). In that case only two compliance types remain, compliers and always-takers. Monotonicity is automatically satisfied, and the average effect for compliers is now equal to the average effect for the treated, since any one receiving the treatment is by definition a complier. This case was first studied in Bloom (1984). It also has a useful connection to Chamberlain's notion of "identification at infinity," (see also Heckman, 1990). Suppose that we have a selection model with a participation equation as in (3), with $\pi_1 > 0$. If Z_i is a continuous instrument, then in order to idenfify the average effect for the treated we need Z_i to have unbounded support. Within this specific selection model this is, as Chamberlain (1987) in a different context, an unattractive identification condition. However, in many application it is plausible that there is some value of the instrument such that individuals do not have access to the treatment, implying identification of the average effect for the treated.

4.3 Extrapolating to the Full Population

Although we cannot consistently estimate the average effect of the treatment for always-takers and never-takers, we do have some information about the potential outcomes for these subpopulations that can aid in assessing the plausibility of extrapolating to average effects for the full population. They key insight is that we can infer the average outcome for never-takers and always-takers in one of the two treatment arms. Specifically, we can estimate

$$\mathbb{E}\left[Y_i(0)|\mathrm{never-taker}\right], \qquad \text{and} \;\; \mathbb{E}\left[Y_i(1)|\mathrm{always-taker}\right], \tag{9}$$

but not

$$\mathbb{E}\left[Y_i(1)|\mathrm{never-taker}\right], \qquad \text{and} \;\; \mathbb{E}\left[Y_i(0)|\mathrm{always-taker}\right],$$

We can learn from the expectations in (9) whether there is any evidence of heterogeneity in outcomes by compliance status, by comparing the pair of average outcomes of $Y_i(0)$;

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \qquad \text{and} \;\; \mathbb{E}\left[Y_i(0)|\text{complier}\right],$$

and the pair of average outcomes of $Y_i(1)$:

$$\mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right], \qquad \text{and} \;\; \mathbb{E}\left[Y_i(1)|\text{complier}\right].$$

If compliers, never-takers and always-takers are found to be substantially different in levels, based on evidence of substantial difference between $\mathbb{E}[Y_i(0)|\text{never}-\text{taker}]$ and $\mathbb{E}[Y_i(0)|\text{complier}]$, and or/between $\mathbb{E}[Y_i(1)|\text{always}-\text{taker}]$, and $\mathbb{E}[Y_i(1)|\text{complier}]$, then it appears much less plausible that the average effect for compliers is indicative of average effects for other compliance types. On the other hand, if one finds that outcomes given the control treatment for never-takers and compliers are similar, and outcomes given the treatment are similar for compliers and always-takers (and especially if this holds within various subpopulations defined by observed covariates), then it appears to be more plausible that average treatment effects for these groups are also comparable.

4.4 Covariates

The local average treatment effect result in (8) implies in general that one cannot consistently estimate average effects for subpopulations other than compliers. This still holds in cases where we observe covariates. One can incorporate the covariates into the analysis in a number of different ways. Traditionally the TSLS or ILS set up is used with the covariates entering in the structural outcome equation linearly and additively, as

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2' X_i + \varepsilon_i,$$

with the covariates added to the set of instruments. Given the potential outcome set up with general heterogeneity in the effects of the treatment, one may also wish to allow for more heterogeneity in the correlations with the covariates. Here we describe a general way of doing so. Unlike TSLS-type approaches, this involves modelling both the dependence of the outcome and the treatment on the covariates. Although there is often a reluctance to model the relation between the treatment, there is no apparent reason that economic theory is more informative about the relation between covariates and outcomes than about the relation between covariates and the choices that lead to the treatment.

A full model can be decomposed into two parts, a model for the compliance type given covariates, and a model for the potential outcomes given covariates for each compliance type. A traditional parametric model with a dummy endogenous variables might have the form (translated to the potential outcome set up used here):

$$W_i(z) = 1\{\pi_0 + \pi_1 \cdot z + \pi_2' X_i + \eta_i \geq 0\}, \tag{10}$$

$$Y_i(w) = \beta_0 + \beta_1 \cdot w + \beta_2' X_i + \varepsilon_i, \tag{11}$$

with (η_i, ε_i) jointly normally distributed and independent of the instruments(e.g., Heckman, 1978). A more general model would allow for separate outcome equations by treatment status:

$$Y_i(0) = \beta_{00} + \beta_{20}' X_i + \varepsilon_{0i}, \tag{12}$$

$$Y_i(1) = \beta_{01} + \beta_{21}' X_i + \varepsilon_{1i}, \tag{13}$$

in combination with (10), (e.g., Björklund and Moffitt, 1987). Such models can be viewed as imposing various restrictions on the relation between compliance types, covariates and outcomes. For example, in the model characterized by equations (10) and (11), if $\pi_1 > 0$, compliance type depends on η_i:

$$\text{unit } i \text{ is a } \begin{cases} \text{never} - \text{taker} & \text{if } \ \eta_i < -\pi_0 - \pi_1 - \pi_2' X_i \\ \text{complier} & \text{if } \ -\pi_0 - \pi_1 - \pi_2' X_i \leq \eta_i < -\pi_0 - \pi_1 - \pi_2' X_i \\ \text{always} - \text{taker} & \text{if } \ -\pi_0 - \pi_2' X_i \leq \eta_i. \end{cases}$$

Not only does this impose monotonicity, by ruling out the presence of defiers, it also implies strong restrictions on the relationship between type and outcomes. Specifically, the selection equation implies that compliers correspond to intermediate values of η_i, implying that conditional expectations of $Y_i(0)$ and $Y_i(1)$ for compliers are in between those for never-takers and always-takers.

An alternative approach to the conventional selection model that exploits the identification results more directly, is to model the potential outcome $Y_i(w)$ for units with compliance type t given covariates X_i through a common functional form with type and treatment specific parameters:

$$f_{Y(w)|X,T}(y(w)|x,t) = f(y|x;\theta_{wt}),$$

for $(w,t) = (0,n),(0,c),(1,c),(1,a)$. For example, using a normal model,

$$Y_i(w)|T_i = t, X_i = x \sim \mathcal{N}\left(x'\beta_{wt}, \sigma^2_{wt}\right), \tag{14}$$

for $(w,t) = (0,n),(0,c),(1,c),(1,a)$.

A natural model for the distribution of type is a trinomial logit model:

$$\mathrm{pr}(T_i = \mathrm{complier}|X_i) = \frac{1}{1+\exp(\pi_n'X_i)+\exp(\pi_a'X_i)},$$

$$\mathrm{pr}(T_i = \mathrm{never-taker}|X_i) = \frac{\exp(\pi_n'X_i)}{1+\exp(\pi_n'X_i)+\exp(\pi_a'X_i)},$$

and

$$\mathrm{pr}(T_i = \mathrm{always-taker}|X_i) = 1 - \mathrm{Pr}(T_i = \mathrm{complier}|X_i) - \mathrm{Pr}(T_i = \mathrm{never-taker}|X_i).$$

The log likelihood function is then, factored in terms of the contribution by observed (W_i, Z_i) values, using the normal model for the conditional outcomes in (14):

$$\mathcal{L}(\pi_n,\pi_a,\beta_{0n},\beta_{0c},\beta_{1c},\beta_{1a},\sigma_{0n},\sigma_{0c},\sigma_{1c},\sigma_{1a}) =$$

$$\times \prod_{i|W_i=0,Z_i=1} \frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{0n}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{0n}}{\sigma_{0n}}\right)$$

$$\times \prod_{i|W_i=0,Z_i=0} \left(\frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)} \cdot \frac{1}{\sigma_{0n}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{0n}}{\sigma_{0n}}\right) + \frac{1}{1+\exp(\pi_n' X_i)} \cdot \frac{1}{\sigma_{0c}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{0c}}{\sigma_{cn}}\right)\right)$$

$$\times \prod_{i|W_i=1,Z_i=1} \left(\frac{\exp(\pi_a' X_i)}{1+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{1a}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{1a}}{\sigma_{1a}}\right) + \frac{1}{1+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{1c}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{1c}}{\sigma_{1c}}\right)\right)$$

$$\times \prod_{i|W_i=1,Z_i=0} \frac{\exp(\pi_a' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{1a}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{1a}}{\sigma_{1a}}\right).$$

For example, the second factor consists of the contributions of individuals with $Z_i = 0$, $W_i = 0$, who are known to be either compliers or never-takers. Maximizing a likelihood function with this mixture structure is straightforward using the EM algorithm (Dempster, Laird, and Rubin, 1977). For an empirical example of this approach see Hirano, Imbens, Rubin and Zhou (2000), and Imbens and Rubin (1997).

In small samples one may wish to incorporate restrictions on the effects of the covariates, and for example assume that the effect of covariates on the potential outcome is the same irrespective of compliance type, or even irrespective of the treatment status. An advantage of this approach is that it can easily be generalized. The type probabilities are nonparametricaly identified as functions of the covariates, and the similarly the outcome distributions are nonparametrically identified, by type as a function of the covariates,.

5. Effects of Military Service on Earnings

In a classic application of instrumental variables methods Angrist (1989) was interested in estimating the effect of serving in the military on earnings. He was concerned about the possibility that those choosing to serve in the military are different from those who do not in ways that affects their subsequent earnings irrespective of serving in the military. To avoid biases in simple comparisons of veterans and non-veterans, he exploited the Vietnam era draft lottery. Specifically he uses the binary indicator whether or not someone’s draft lottery number made him eligible to be drafted as an instrument. The lottery number was tied to an individual's day of birth, so more or less random. Even so, that in itself does not make it valid as an instrument as we shall discuss below. As the outcome of interest Angrist uses total earnings for a particular year.

The simple ols regression leads to:

$$\begin{array}{rcccl} \widehat{\log(\text{earnings})}_i & = & 5.4364 & - & 0.0205 \cdot \widehat{\text{veteran}}_i \\ & & (0079) & & (0.0167) \end{array}$$

In Table 4 we present population sizes of the four treatment/instrument subsamples. For example, with a low lottery number 5,948 individuals do not, and 1,372 individuals do serve in the military.

Table 4: Treatment Status by Assignment

		Z_i 0	Z_i 1
W_i	0	5,948	1,915
	1	1,372	865

Using these data we get the following proportions of the various compliance types, given in Table 5, under the no-defiers or monotonicity assumption. For example, the proportion of nevertakers is estimated as the conditional probability of $W_i = 0$ given $Z_i = 1$:

$$\text{pr(nevertaker)} = \frac{1915}{1915 + 865} = xxx.$$

Table 6 gives the average outcomes for the four groups, by treatment and instrument status.

Table 5: Compliance Types: Estimated Proportions

		$W_i(0)$ 0	1
$W_i(1)$	0	never-taker (0.6888)	defier (0)
	1	complier (0.1237)	always-taker (0.1874)

Table 6: Estimated Average Outcomes by Treatment and Instrument

		Z_i 0	1
W_i	0	$\widehat{\mathbb{E}[Y]} = 5.4472$	$\widehat{\mathbb{E}[Y]} = 5.4028$
	1	$\widehat{\mathbb{E}[Y]} = 5.4076,$	$\widehat{\mathbb{E}[Y]} = 5.4289$

Table 7 gives the estimated averages for the four compliance types, under the exclusion restriction. This restriction is the key assumption here. There are a number of reasons why it may be violated in this application. For example, never-takers may need to taking active action to avoid military service if draft eligible, for example by continuing their formal education, or by moving to Canada. Always-takers may be affected their lottery number if draftees were treated differently in the military compared to volunteers. The local average treatment effect is -0.2336, a 23% drop in earnings as a result of serving in the military.

Simply doing IV or TSLS would give you the same numerical results:

$$\begin{array}{ccccc} \widehat{\log(\text{earnings})}_i & = & 5.4836 & - & 0.2336 \cdot \widehat{\text{veteran}}_i \\ & & (0.0289) & & (0.1266) \end{array}$$

It is interesting in this application to inspect the average outcome for different compli-

Table 7: Compliance Types: Estimated Average Outcomes

		$W_i(0)$	
		0	1
$W_i(1)$	0	never-taker: $\widehat{\mathbb{E}[Y_i(0)}] = 5.4028$	defier (NA)
	1	complier: $\widehat{\mathbb{E}[Y_i(0)}] = 5.6948$, $\widehat{\mathbb{E}[Y_i(1)}] = 5.4612$	always-taker: $\widehat{\mathbb{E}[Y_i(1)}] = 5.4076$

ance groups. Average log earnings for never-takers are 5.40, lower by 29% than average earnings for compliers who do not serve in the military. This suggests that never-takers are substantially different than compliers, and that the average effect of 23% for compliers need not be informative never-takers. In contrast, average log earnings for always-takers are only 6% lower than those for compliers who serve, suggesting that the differences between always-takers and compliers are considerably smaller. Note that compliers have better outcomes without the treatment than never-takers and better outcomes than always-takers given the treatment. This is inconsistent with the simple normal selection model in(10)-(11).

6. Multivalued Instruments

For any two values of the instrument z_0 and z_1 satisfying the local average treatment effect assumptions we can define the corresponding local average treatment effect:

$$\tau_{z_1,z_0} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i(z_1) = 1, W_i(z_0) = 0].$$

Note that these local average treatment effects need not be the same for different pairs of instrument values. Comparisons of estimates based on different instruments underlies tests of overidentifying restrictions in TSLS settings. An alternative interpretation of rejections in such testing procedures is therefore the presence of heterogeneity in causal effects, rather than that some of the instruments are invalid. Without restrictions on the heterogeneity of the causal effects there are no tests in general for the validity of the instruments.

The presence of multi-valued, or similarly, multiple, instruments, does, however, provide an opportunity to assess variation in treatment effects, as well as an opportunity to obtain average effects for subpopulations closer to the one of ultimate interest. Suppose that we have an instrument Z_i with support $z_0, z_1, \ldots, z_K$. Suppose also that the monotonicity assumption holds for all pairs z and z', and suppose that the instruments are ordered in such a way that

$$p(z_{k-1}) \leq p(z_k), \qquad \text{where } \; p(z) = \mathbb{E}[W_i|Z_i = z].$$

Also suppose that the instrument is relevant, so that for some function $g(Z)$,

$$\mathbb{E}[g(Z_i) \cdot (W_i - \mathbb{E}[W_i])] \neq 0.$$

Then the instrumental variables estimator based on using $g(Z)$ as an instrument for W estimates a weighted average of the local average treatment effects $\tau_{z_k,z_{k-1}}$:

$$\tau_g = \frac{\text{Cov}(Y_i, g(Z_i))}{\text{Cov}(W_i, g(Z_i))} = \sum_{k=1}^{K} \lambda_k \cdot \tau_{z_k,z_{k-1}},$$

where the weights λ_k are non-negative and satisfy

$$\lambda_k = \frac{(p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]}{\sum_{k=1}^{K} p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]},$$

for

$$\pi_k = \text{pr}(Z_i = z_k),$$

implying that $\sum_{k=1}^{K} \lambda_k = 1$.

Choosing the function $g(z)$ corresponds to choosing the weight function. There are obviously limits to the weight functions that can be choosen. One can only estimate a weighted average of the local average treatment effects defined for all pairs of instrument values in the support of the instrument. If $p(z_0) = 0$ for some z_0 in the support of Z, one can estimate the average effect on the treated as τ_{z_K,z_0}.

If the instrument Z has a continuous distribution, and the probability of receiving the treatment given the instrument, $p(z)$, is continuous in z, we can define the limit of the local average treatment effects

$$\tau_z = \lim_{z' \downarrow z, z'' \uparrow z} \tau_{z',z''}.$$

If the monotonicity assumption holds for all pairs z and z', we can use the implied structure on the compliance behavior by modelling $W_i(z)$ as a threshold crossing process,

$$W_i(z) = 1\{h(z) + \eta_i \geq 0\}, \tag{15}$$

with the scalar unobserved component η_i independent of the instrument Z_i. This type of latent index model is used extensively in work by Heckman (Heckman and Robb, 1985; Heckman,1990; Heckman and Vytlacil, 2005), as well as in Vytlacil (2000). Vytlacil shows that if the earlier three assumptions (independence, the exclusion restriction and monotonicity) hold for all pairs z and z', than there is a function $h(\cdot)$ such that this latent index structure is consistent with the joint distribution of the observables. The latent index structure implies that individuals can be ranked in terms of an unobserved component η_i such that if for two individuals i and j we have $\eta_i > \eta_j$, than $W_i(z) \geq W_j(z)$ for all z.

Given this assumption, we can define the marginal treatment effect $\tau(\eta)$ as

$$\tau(\eta) = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| \, \eta_i = \eta \right].$$

In a parametric setting this was introduced by Björklund and Moffitt (1987). In the continuous Z case this marginal treatment effect relates directly to the limit of the local average treatment effects:

$$\tau(\eta) = \tau_z, \qquad \text{with } \; \eta = -h(z)).$$

Note that we can only define $\tau(\eta)$ for values of η for which there is a z such that $\tau = -h(z)$. Normalizing the marginal distribution of η to be uniform on $[0, 1]$ (Vytlacil, 2002), this restricts η to be in the interval $[\inf_z p(z), \sup_z p(z)]$, where $p(z) = \text{pr}(W_i = 1|Z_i = z)$. Heckman and Vytlacil (2005) characterize various average treatment effects (e.g., the population average treatment effect, the average treatment effect for the treated, the local average treatment effect) in terms of this marginal treatment effect. For example, the population average treatment effect is simply the average of the marginal treatment effect over the marginal distribution of η:

$$\tau = \int_\eta \tau(\eta) dF_\eta(\eta).$$

In practice the same limits remain on the identification of average effects. A necessary condition for identification of the population average effect is that the instrument moves the probability of participation from zero to one. Note that identification of the population average treatment effect does not require identification of $\tau(\eta)$ at every value of η. The latter is sufficient, but not necessary. For example, in a randomized experiment (corresponding to a binary instrument with the treatment indicator equal to the instrument) the population average treatment effect is obviously identified, but the marginal treatment effect is not identified for any value of η.

7. Multivalued Endogenous Variables

Now suppose that the endogenous variable W_i takes on values $0, 1, \ldots, J$. We still assume that the instrument Z_i is binary. We study the interpretation of the instrumental variables estimand

$$\tau^{\text{iv}} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(W_i, Z_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]}.$$

We make the exclusion assumption that for all z in the support of Z_i,

$$Y_i(w), W_i(z) \perp\!\!\!\perp Z_i,$$

and a version of the monotonicity assumption,

$$W_i(1) \geq W_i(0).$$

Then we can write the instrumental variables estimand as

$$\tau^{\rm iv} = \sum_{j=1}^{J} \lambda_j \cdot \mathbb{E}[Y_i(j) - Y_i(j-1) | W_i(1) \geq j > W_i(0)], \tag{16}$$

where

$$\lambda_j = \frac{{\rm pr}(W_i(1) \geq j > W_i(0)}{\sum_{i=1}^{J} {\rm pr}(W_i(1) \geq i > W_i(0)}. \tag{17}$$

The weights are non-negative and add up to one.

Note that we can estimate the weights λ_j because

$${\rm pr}(W_i(1) \geq j > W_i(0) = {\rm pr}(W_i(1) \geq j) - {\rm pr}(W_i(0) \geq j)$$

$$= {\rm pr}(W_i(1) \geq j | Z_i = 1) - {\rm pr}(W_i(0) \geq j | Z_i = 0)$$

$$= {\rm pr}(W_i \geq j | Z_i = 1) - {\rm pr}(W_i \geq j | Z_i = 0),$$

using the monotonicity assumption.

8. Instrumental Variables Estimates of the Returns to Education Using Quarter of Birth as an Instrument

Here we use a subset of the data used by Angrist and Krueger in their 1991 study of the returns to education. Angrist and Krueger were concerned with the endogeneity of education, with the standard argument that individuals with higher ability are likely to command higher wages at any level of education, as well as be more likely to choose high levels of education. In that case simple least squares estimates would over estimate the returns to education. Angrist and Krueger realized that individuals born in different parts of the year are subject to slightly different compulsory schooling laws. If you are born before a fixed cutoff date you enter school at a younger age than if you are born after that cutoff date, and given that you are allowed to leave school when you turn sixteen, those individuals born before the cutoff date are required to completely more years of schooling. The instrument can therefore be thought of as the tightness of the compulsory schooling laws, with the tightness being measured by the individual's quarter of birth.

Angrist and Krueger implement this using census data with quarter of birth indicators as the instrument. Table 8 gives average years of education and sample sizes by quarter of birth.

Table 8: Average Level of Education by Quarter of Birth

quarter	1	2	3	4
average level of education	12.69	12.74	12.81	12.84
standard error	0.01	0.01	0.01	0.01
number of observations	81,671	80,138	86,856	80,844

In the illustrations below we just use a single instrument, an indicator for being born in the first quarter. First let us look at the reduced form regressions of log earnings and years of education on the first quarter of birth dummy:

$$\begin{array}{rcccc} \widehat{\text{educ}_i} & = & 12.797 & - & 0.109 \cdot \text{qob}_i \\ & & (0.006) & & (0.013) \end{array}$$

and

$$\begin{array}{rcccc} \log(\widehat{\text{earnings}})_i & = & 5.903 & - & 0.011 \cdot \text{qob}_i \\ & & (0.001) & & (0.003) \end{array}$$

The instrumental variables estimate is the ratio of the reduced form coefficients,

$$\hat{\beta}^{\text{iv}} = \frac{-0.1019}{-0.011} = 0.1020.$$

Now let us interpret this estimate in the context of heterogeneous returns to education, using (16) and (17).. This estimate is an average of returns to education, consisting of two types of averaging. The first averaging is over different levels of education. That is, it is a weighted average of the return to the tenth year of education, to the elevent year of education, and so on.. In addition, for any level, e.g., to moving from nine to ten years of education, it is an average effect where the averaging is over those people whose schooling would have been at least ten years of education if more restrictive compulsory schooling laws had been in effect for them, and who would have had less than ten years of education had they been subject to the looser compulsory schooling laws.

Furthermore, we can estimate how large a fraction of the population is in these categories. First we estimate the

$$\gamma_j = \text{pr}(W_i(1) \geq j > W_i(0) = \text{pr}(W_i \geq j | Z_i = 1) - \text{pr}(W_i \geq j | Z_i = 0)$$

as

$$\hat{\gamma}_j = \frac{1}{N_1} \sum_{i|Z_i=1} 1\{W_i \geq j\} - \frac{1}{N_0} \sum_{i|Z_i=0} 1\{W_i \geq j\}.$$

This gives the unnormalized weight function. We then normalize the weights so they add up to one, $\hat{\lambda}_j = \hat{\gamma}_j / \sum_i \hat{\gamma}_i$.

Figure 1-4 present some of the relevant evidence here. First, Figure 1 gives the distribution of years of education for the Angrist-Krueger data. Figure 2 gives the normalized and Figure 3 gives the unnormalized weight functions. Figure 4 gives the distribution functions of years of education by the two values of the instrument. The most striking feature of these figures (not entirely unanticipated) is that the proportion of individuals in the "complier" subpopulations is extremely small, never more than 2% of the population. This implies that these instrumental variables estimates are averaged only over a very small subpopulation, and that there is little reason to believe that they generalize to the general population. (Nevertheless, this may well be a very interesting subpopulation for some purposes.) The nature of the instrument also suggests that most of the weight would be just around the number of years that would be required under the compulsory schooling laws. The weight function is actually surprisingly flat, putting weight even on fourteen to fifteen years of education.

Figure 1: histogram estimate of density of years of education

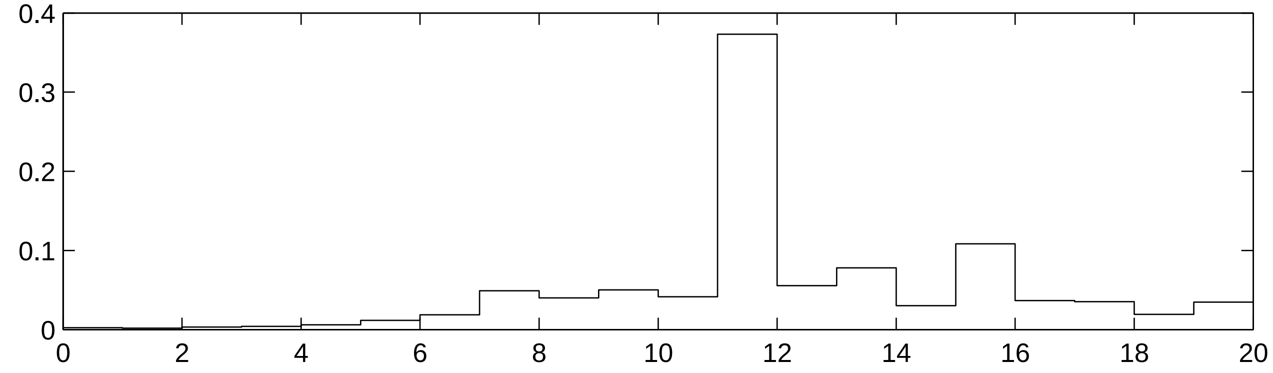


Figure 2: Normalized Weight Function for Instrumental Variables Estimand

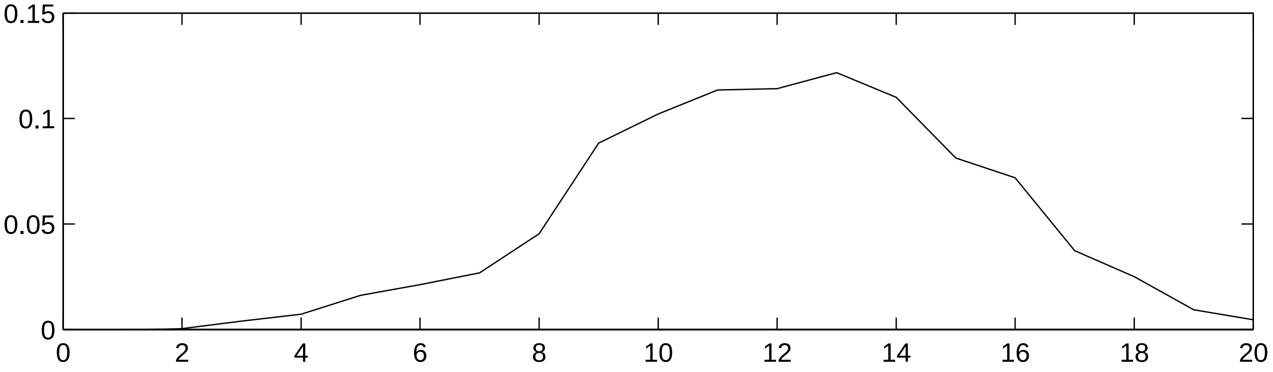


Figure 3: Unnormalized Weight Function for Instrumental Variables Estimand

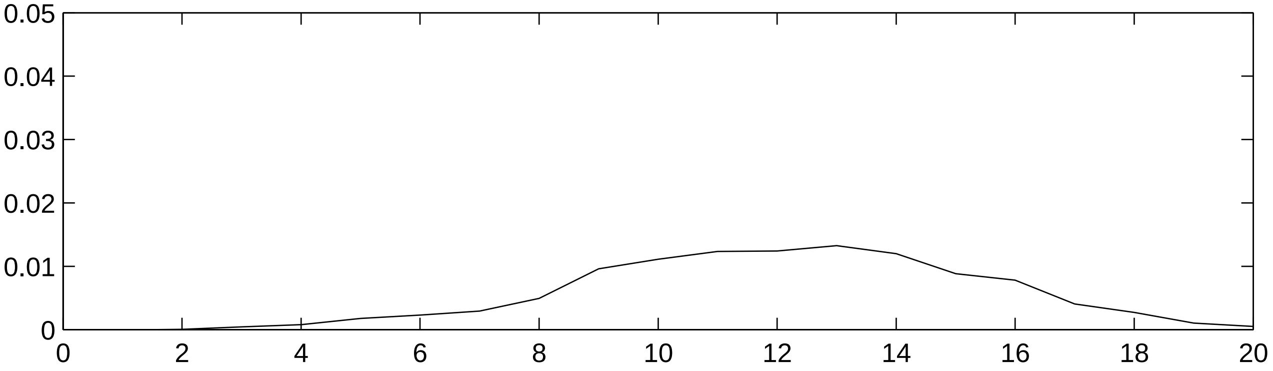
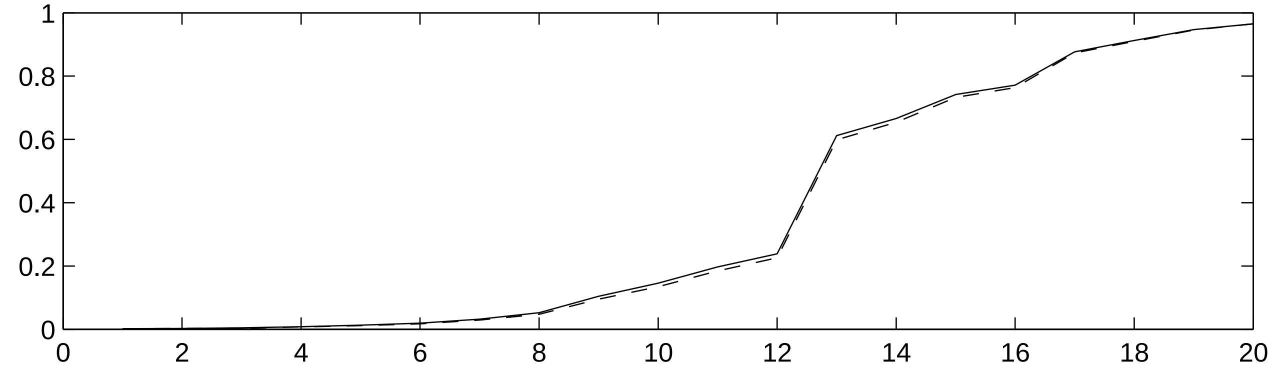


Figure 3: Education Distribution Function by Quarter



References

Abadie, A., (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, Vol 97, No. 457, 284-292.

Abadie, A., (2003), "Semiparametric Instrumental Variable Estimation of Treatment Reponse Models," *Journal of Econometrics*, Vol 113, 231-263.

Angrist, J. D., and G. W. Imbens, (1995), "Two–Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, Vol 90, No. 430, 431-442.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," (with discussion) *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

Angrist, J. and A. Krueger, (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association* 87, June.

Björklund, A. and R. Moffitt, (1987), "The Estimation of Wage Gains and Welfare Gains in Self–Selection Models", *Review of Economics and Statistics*, Vol. LXIX, 42–49.

Bloom, H., (1984), "Accounting for No–shows in Experimental Evaluation Designs,"*Evaluation Review*, 8(2) 225–246.

Chamberlain, G., (1986), "Asymptotic Efficiency in Semi-parametric Models with Censoring," *Journal of Econometrics*, Vol 32, 189-218.

Dempster, A., N. Laird, and D. Rubin (1977), "Maximum Likelihood Estimation from Incomplete Data Using the EM Algorithm (with discussion)," *Journal of the Royal*

Statistical Society, Series B, 39, 1-38.

Heckman, J. (1990), "Varieties of Selection Bias," *American Economic Review* Vol 80, Papers and Proceedings, 313-318.

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Heckman, J., and E. Vytlacil, (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, Vol. 73(3), 669-738.

Hirano, K., G. Imbens, D. Rubin, and X. Zhou (2000), "Identification and Estimation of Local Average Treatment Effects," *Biostatistics*, Vol. 1(1), 69-88.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G. W., and D. B. Rubin, (1997), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance,"*Annals of Statistics*, Vol. 25, No. 1, 305–327.

textscManski, C., (2008), *Identification for Prediction and Decision*, Harvard University Press, Cambridge, MA.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Vytlacil, E., (2002), "Independence, Monotonicity, and Latent Index Models: Am Equivalence Result," *Econometrica*, Vol. 70(1), 331-341.

Wooldridge, J. (2001) *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.