# Applied Econometrics

# Lecture 12: Treatment Effects Part II.

Måns Söderbom*

29 September 2009

*Department of Economics, University of Gothenburg. Email: mans.soderbom@economics.gu.se. Web: www.economics.gu.se/soderbom, www.soderbom.net.

# 1. Introduction

In the notes for Lecture 11 we discussed how the Difference-in-Differences estimator can be used to identify the average treatment effect even if there is selection on unobservables, provided that the selection mechanism is time invariant. We now discuss the instrumental variable (IV) approach for solving the problem posed by selection on unobservables.

References for this lecture are as follows:

Angrist and Pischke (2009), Chapter 4.4 (read carefully) and 4.5 (focus on the intuition).

Chapter 18.4 in Wooldridge (2002) "Cross Section and Panel Data".

# 2. Selection on Unobservables (continued)

## 2.1. Estimating ATE using IV

- Consider the following two equations for potential outcomes:

$$\begin{aligned} y_0 &= \mu_0 + v_0, \\ y_1 &= \mu_1 + v_1, \end{aligned}$$

  with $E(v_0) = E(v_1) = 0$. Hence, $(\mu_1 - \mu_0)$ is the average treatment effect.

- Observed outcome can be written as a 'switching regression':

$$y = \mu_0 + (\mu_1 - \mu_0) w + v_0 + w (v_1 - v_0),$$

  thus the coefficient on $w$ is the ATE.

- Now consider the possibility that $v_1, v_0$ are not mean independent of $w$, even if you control for the observables, i.e. $x$. That is, ignorability of treatment (or CIA) fails.

- Suppose we have a vector of instruments $z$, and consider the following assumptions:

**Assumption ATE.2** (Wooldridge, 2002, p. 621):

(a) $v_1 = v_0$

(b) The linear projection of $v_0$ on $x$ and $z$ instrument is identical to the linear projection of $v_0$ on $x$ only. Think: $z$ is uncorrelated with $v_0$, conditional on $x$ - i.e. instrument validity.

(c) The linear projection of $w$ on $x$ and $z$ instrument is not identical to the linear projection of $z$ on $x$ only. Think: $z$ is correlated with $w$, conditional on $x$ - i.e. instrument is informative.

Part (a) implies that the interaction term $w(v_1 - v_0)$ disappears:

$$y = \mu_0 + (\mu_1 - \mu_0)\, w + v_0.$$

Furthermore, write $v_0$ in terms of observables:

$$v_0 = x\beta_0 + u_0,$$

and our estimable equation becomes

$$y = \mu_0 + (\mu_1 - \mu_0)\, w + x\beta_0 + u_0.$$

Under ATE.2 we can estimate the average treatment effect $(\mu_1 - \mu_0)$ by means of IV. Because the only endogenous explanatory variable in the equation is a binary, this model is often called a dummy endogenous variable model.

- One variant on the above is to use the **propensity score** as instrument**.**This will give a more efficient IV estimator than the one just discussed, but we also need some additional assumptions - see Assumption ATE.2' if you are interested (Wooldridge, 2002, p. 623). Mechanically, this involves fitting a probit or logit modelling the likelihood of treatment as a function of all exogenous variables and the instrument $z$, calculating the propensity score, and then using the propensity score as an instrument. Notice that standard errors in the second stage will have to be modified to take into account the fact that the propensity score is itself an **estimate**. The simplest way of doing this is through bootstrapping.

- This all sounds fine and straightforward - and very similar to "conventional" instrumental variables estimation (see lecture 2). But the validity of the procedure clearly hinges on the validity of assumption ATE.2.

- In general, if in the equations

$$
\begin{aligned}
y_0 &= \mu_0 + v_0, \\
y_1 &= \mu_1 + v_1,
\end{aligned}
$$

we have $v_1 \neq v_0$, the IV estimator does **not** generally consistently estimate $ATE$ (or $ATE_1$). Why? Go back to the general switching regression:

$$
y = \mu_0 + (\mu_1 - \mu_0)\, w + \{v_0 + w\,(v_1 - v_0)\}\,,
$$

and think of the term inside { } as a residual.

- The source of the problem is the interaction term $w\,(v_1 - v_0)$. Remember our task is to find an instrument $z$ that is correlated with $w$ while at the same time uncorrelated with the residual - but since the latter contains an interaction term depending on $w$ you see how it will be hard to get both conditions fulfilled unless you bring in more assumptions. Clearly you can attempt to proxy $v_0$ and $v_1$ with observable $x$-variables as above:

$$
\begin{aligned}
v_0 &= \eta_0 + x\beta_0 + e_0 \\
v_1 &= \eta_1 + x\beta_1 + e_1,
\end{aligned}
$$

where $\eta_0, \beta_0, \eta_1, \beta_1$ denote coefficients (vectors) and $e_0$ and $e_1$ are unobservable random terms. Still, unless $e_0 = e_1$, you are stuck with an interaction term between the treatment variable $w$ and an unobserved term $(e_1 - e_0)$ which accordingly will go into the residual. It is true that you can still recover the $ATE$ by making further assumptions, but from an applied point of view I am not sure how interesting this is - see pp. 625-628 in Wooldridge for details, if you feel like it.

- The main point I think we can take away from this is that the assumptions you need for it to be possible to identify $ATE$ by means of IV, while at the same time allowing for individual heterogeneity in treatment effects, are potentially quite strong.
- We will now discuss how we can give meaningful (and potentially interesting) interpretation to the IV estimator under weaker assumptions than you need in order to identify $ATE$

### 2.2. Estimating LATE using IV

Reference: Angrist-Pischke, Chapter 4.4.

#### 2.2.1. LATE: Setting the scene

- Common features of the type of environment for which we may want to, and be able to, estimate the Local Average Treatment Effect, LATE:
  - The treatment status, from now on denoted $D_i$, depends on an underlying instrument $Z_i$.
  - The effect of $Z_i$ on treatment is heterogeneous.
  - The effect of treatment $D_i$ on the outcome variable of interest $Y_i$ is also heterogeneous.
- Thus, the causal chain is as follows:

$$Z_i \rightarrow D_i \rightarrow Y_i,$$

  and we are primarily interested in the effect of treatment on outcomes; i.e. $D_i \rightarrow Y_i$.
- Define $Y_i(d, z)$ as the potential outcome of individual i, were this individual to have treatment status $D_i = d$ and instrument value $Z_i = z$. We focus on the case where both $d$ and $z$ can take two values, 0 or 1. That is, $D_i$ and $Z_i$ are dummy variables.

- Following Angrist-Pischke, we relate the exposition to a specific application, namely Angrist (1990), who looks at the effect of veteran status on earnings in the US. The instrument is defined as follows:

$$\begin{aligned} Z_i &= 1 \text{ if lottery implied individual } i \text{ would be draft eligible,} \\ Z_i &= 0 \text{ if lottery implied individual } i \text{ would not be draft eligible.} \end{aligned}$$

- The instrument affects treatment, which is this application amounts to entering the military service. The econometrician observes treatment status as follows:

$$\begin{aligned} D_i &= 1 \text{ if individual } i \text{ served in the Vietnam war (veteran),} \\ D_i &= 0 \text{ if individual } i \text{ did not serve in the Vietnam war (not veteran);} \end{aligned}$$

- Now define potential outcomes for $D_i$ as $D_{0i}$ and $D_{1i}$, respectively, where $D_{0i}$ is the treatment status when $Z_i = 0$ and $D_{1i}$ is the treatment status when $Z_i = 1$. We thus have:

$$\begin{aligned} D_{0i} &= 0 \text{ if individual } i \text{ would not serve in the military if not draft eligible} \\ D_{0i} &= 1 \text{ if individual } i \text{ would serve in the military even though not draft eligible} \\ D_{1i} &= 0 \text{ if individual } i \text{ would not serve in the military even though draft eligible} \\ D_{1i} &= 1 \text{ if individual } i \text{ would serve in the military if draft eligible.} \end{aligned}$$

- In view of this, the following way of categorizing types of individuals is useful (why will be clear later):

| | |
|---|---|
| Compliers: | $D_{1i} = 1, D_{0i} = 0$ |
| Never-takers: | $D_{1i} = 0, D_{0i} = 0$ |
| Always-takers: | $D_{1i} = 1, D_{0i} = 1$ |
| Defiers: | $D_{1i} = 0, D_{0i} = 1$ |

Note that "defiers" are very odd cases - as we shall see, the basic LATE estimator assumes there are no defiers. In the present context, at least, it's hard to see why there might be defiers.

- The outcome variable of interest is earnings, and the main research question is whether veteran status causes earnings. The causal effect of veteran status, conditional on draft eligibility status, is defined as

$$Y_i(1, Z_i) - Y_i(0, Z_i).$$

- As usual, we can't identify individual treatment effects, because we don't observe all potential outcomes. We will not even try to estimate $ATE$ or $ATE_1$ because, as we have seen, we would need pretty strong assumptions. Before discussing assumptions and interpretation further, let's remind ourselves of what the OLS and IV estimators would look like in the present context.

#### 2.2.2. Estimation by regression: OLS and IV

- If I use **OLS** to estimate a model of the following kind:

$$Y_i = \alpha + \theta D_i + \varepsilon_i,$$

where $\alpha$ is a constant and $\varepsilon_i$ a zero–mean residual, we know from last time (lecture 11) that $\theta^{OLS}$ is interpretable as an estimate of $ATE$ and $ATE_1$, provided potential outcomes are independent of actual treatment status. That is, provided treatment is (as good as) randomly assigned. If that doesn't hold, OLS does not identify $ATE$ or $ATE_1$. In the present context, it seems likely there are lots of unobservables correlated with veteran status, so the OLS estimator is hard to justify here.

- Suppose I were to use an **IV** estimator instead, with $Z_i$ as a single instrument:

$$\begin{aligned} D_i &= \gamma + \phi Z_i + u_i \\ Y_i &= \alpha + \theta D_i + \varepsilon_i, \end{aligned}$$

For now, don't worry about my reasons for doing this - just think about what I "would get" if I were to do this. An old result in econometrics has it that in the **special case** where $Z_i$ is a dummy variable, the IV estimator can be written simply as:

$$\theta^{IV} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}.$$

See Appendix 1 for a (somewhat laborious) proof. This is known as the **Wald estimator** in the IV literature. Inevitably, this is what I will get if I use IV to model earnings as a function of veteran status, while using draft eligibility status as an instrument for veteran status. But how should this quantity be interpreted? Does it estimate an average treatment effect?

- Yes, potentially. As ever, your IV estimator can only be interpreted in the light of the assumptions you are making. As we have seen, under assumptions ATE2 or ATE2' above (from Wooldridge's exposition), the IV estimator identifies the $ATE$. But those assumptions are potentially strong ones.

#### 2.2.3. LATE: A distinct evaluation parameter

- One common "evaluation parameter" estimated by means of instrumental variable techniques is the **Local Average Treatment Effect** (LATE).

- Suppose we are concerned that OLS doesn't identify $ATE$ because there are unobserved differences between veterans and nonveterans (the standard endogeneity concern). We propose to use the draft lottery outcome as an instrument for veteran status. Suppose we are prepared to make **four assumptions** as follows:

**Assumption A1: Independence** between the potential outcomes $[Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}]$ and the instrument $Z_i$. That is, the instrument is as good as randomly assigned. This is what we mean by "exogenous" in the present context.

**Assumption A2: Exclusion restriction.** The potential outcomes $Y_i(d, z)$ is **only** a function of $d$; they are only affected by the instrument $Z_i$ through the treatment variable $D_i$. Implies $Y_i(d, 0) = Y_i(d, 1)$.

**Assumption A3: Relevance, first stage**. $E[D_{1i} - D_{0i} \neq 0]$. The average causal effect of the instrument on veteran status is not zero.

**Assumption A4: Monotonicity.** $D_{1i} - D_{0i} \geq 0$ for all individuals (or vice versa). That is, no defiers.

- A1 states that the instrument is **exogenous**. Draft eligibility was determined by a lottery, thus exogeneity holds by design in this case.

- A2 says that the instrument can have no direct effect on the outcome variable (earnings). May or may not hold in this case.

- A3 says that the instrument impacts on treatment - easy to check in practice.

- A4 says that any man who would serve if not draft eligible, would also serve if draft eligible. A reasonable assumption in this case it would seem.

- Under these assumptions the parameter you're estimating in the second stage of your IV procedure (the coefficient on veteran status, $D_i$) is interpretable as measuring the average effect of military service on earnings for *men who served because they were draft eligible, but who would not have served had they not been draft eligible.* That is, the average affect for the group of men whose treatment status can be changed by the instrument - the "compliers". Note that this group of people does not include volunteers (always-takers) or men who were exempted from service (never takers).

- The average effect for the compliers is a parameter called the LATE. Mathematically, we define the LATE as

$$LATE = E[Y_{1i} - Y_{0i} | D_{1i} - D_{0i} > 0],$$

where $Y_{1i} - Y_{0i}$ denotes the difference in outcomes due to treatment, $D_{1i}$ is the potential treatment status when the instrument $Z_i = 1$ and $D_{0i}$ is the potential treatment status when $Z_i = 0$. Clearly $D_{1i} - D_{0i} > 0$ only applies for compliers.

- Under assumptions A1-A4 we can show that the Wald estimator coincides with the expression for $LATE$. In other words, IV identifies $LATE$, in this case.

**2.2.4. Analysis: Why Wald = LATE?**

- Relate observed treatment status to potential treatment outcomes:

$$\begin{aligned} D_i &= D_{0i} + (D_{1i} - D_{0i}) Z_i, \\ D_i &= \pi_0 + \pi_{1i} Z_i + \xi_i, \end{aligned}$$

where $\pi_0 = E(D_{0i})$ and $\pi_{1i} = (D_{1i} - D_{0i})$ is the (note) heterogeneous causal effect of the instrument on $D_i$. Assumption A1 (independence) implies $\pi_{1i}$ is interpretable as the causal effect of $Z_i$ on treatment (compare this to the case where treatment is randomized; see lecture 11). Assumption A4 (monotonicity) implies that $\pi_{1i} \geq 0$ for all $i$ or $\pi_{1i} \leq 0$ for all $i$.

- Recall from above that the potential outcome of our main "dependent variable" is defined $Y_i(d, z)$. Assumption A2 (exclusion restriction) implies $Y_i(d, 0) = Y_i(d, 1)$, hence we can write the observed outcome:

$$\begin{aligned} Y_i &= Y_i(0, Z_i) + [Y_i(1, Z_i) - Y_i(0, Z_i)] D_i, \\ Y_i &= Y_{0i} + [Y_{1i} - Y_{0i}] D_i \\ Y_i &= \alpha_0 + \rho_i D_i + \eta_i, \end{aligned}$$

where $\alpha_0 = E(Y_{0i})$, $\rho_i = Y_{1i} - Y_{0i}$ is a random coefficient and $\eta_i$ measures the discrepancy between $E(Y_{0i})$ and $Y_{0i}$. the heterogeneous causal effect of treatment (e.g. veteran status) on your outcome variable of interest (e.g. earnings).

- Now consider the formula for the Wald estimator:

$$\frac{E\left[Y_i|Z_i=1\right]-E\left[Y_i|Z_i=0\right]}{E\left[D_i|Z_i=1\right]-E\left[D_i|Z_i=0\right]}.$$

The exclusion restriction (A2) and independence (A1) assumptions imply

$$\begin{aligned} E\left[Y_i|Z_i=1\right] &= E\left[Y_{0i}+\left[Y_{1i}-Y_{0i}\right]D_{1i}|Z_i=1\right], \\ E\left[Y_i|Z_i=1\right] &= E\left[Y_{0i}+\left[Y_{1i}-Y_{0i}\right]D_{1i}\right]. \end{aligned}$$

By the same principles,

$$E\left[Y_i|Z_i=0\right]=E\left[Y_{0i}+\left[Y_{1i}-Y_{0i}\right]D_{0i}\right],$$

and so the numerator in the Wald estimator can be written

$$\begin{aligned} \text{Wald-numerator} &= E\left[Y_i|Z_i=1\right]-E\left[Y_i|Z_i=0\right] \\ &= E\left[Y_{0i}+\left[Y_{1i}-Y_{0i}\right]D_{1i}\right]-E\left[Y_{0i}+\left[Y_{1i}-Y_{0i}\right]D_{0i}\right] \\ &= E\left[\left(Y_{1i}-Y_{0i}\right)\left(D_{1i}-D_{0i}\right)\right]. \end{aligned}$$

Now, monotonicity implies $(D_{1i}-D_{0i})$ is either equal to 1 or 0; hence

$$\begin{aligned} \text{Wald-numerator} &= E\left[\left(Y_{1i}-Y_{0i}\right)\left(D_{1i}-D_{0i}\right)\right] \\ &= P\left(D_{1i}-D_{0i}>0\right)E\left[\left(Y_{1i}-Y_{0i}\right)|D_{1i}-D_{0i}>0\right]. \end{aligned}$$

The denominator of the Wald formula is

$$\text{Wald-denominator}=E\left[D_i|Z_i=1\right]-E\left[D_i|Z_i=0\right].$$

We can use exactly the same principles as for the numerator, and arrive at

$$\begin{aligned}
\text{Wald-denominator} &= P\left(D_{1i}-D_{0i}>0\right) E\left[\left(D_{1i}-D_{0i}\right)|D_{1i}-D_{0i}>0\right] \\
&= P\left(D_{1i}-D_{0i}>0\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{E\left[Y_i|Z_i=1\right]-E\left[Y_i|Z_i=0\right]}{E\left[D_i|Z_i=1\right]-E\left[D_i|Z_i=0\right]} &= E\left[\left(Y_{1i}-Y_{0i}\right)|D_{1i}-D_{0i}>0\right] \\
&= E\left[\rho_i|\pi_{1i}>0\right]
\end{aligned}$$

i.e. the IV estimator identifies $LATE$.

[Example: Simulating LATE. Appendix 2]

### 2.2.5. Discussion: How useful is LATE?

Summing up, we have seen how we can identify LATE - i.e. the average effect of treatment for the subpopulation of compliers. Compliers in the specific application referred to above are individuals who were induced by the draft lottery to serve in the military. Never-takers who would not serve irrespective of their lottery number, and always-takers, who would volunteer irrespective of their lottery number, clearly do not belong to this group.

Is LATE an economically interesting quantity? It may be. Suppose policy makers want to compensate those who were involuntarily made to serve in the war - in such a case, policy makers need to know how much earnings the group of compliers lost as a consequence of being compliers.

The local average treatment effect is not necessarily the main effect of interest. However, it may be that we cannot identify average treatment effects for the population because we cannot identify the average causal effect of treatment amongst the never-takers or the always-takers, if our instrument has no effect on individuals belonging to these groups.

- Extensions (see Section 4.5 in Angrist-Pischke for details):
  - $LATE$ with multiple instruments. The LATE is always closely connected to the underlying instrument, since whether someone is a complier likely depends on what the instrument is..Different instruments will therefore identify different LATE:s. If we have, say, two instruments with distinct complier groups and thus distinct LATEs, using 2SLS with both instruments simultaneously produces a linear combination of the instrument-specific LATEs. Whether or not that is interesting clearly depends on the context.
  - Covariates - where did the $x$-variables go? As soon as we have instruments that are randomly assigned, we don't really need to control for $x$-variables (as these will be orthogonal to the instrument anyway). However, the instrument may in fact covary with $x$-variables that also impact potential outcomes, in which case we should control for $x$-variables.
  - Variable treatment intensity. Suppose that our treatment variable, rather than being binary,

can take on more than two values. Years of schooling is a popular example.

## Appendix 1: Deriving the Wald estimator.

Consider the regression model

$$y_i = b_0 + b_1 x_i + u,$$

and let $z$ be a binary instrumental variable for $x$. In this case, the IV estimator can be written as

$$b_1^{IV} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0},$$

where $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of $y_i$ and $x_i$ over the part of the sample with $z_i = 0$; and $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of $y_i$ and $x_i$ over the part of the sample with $z_i = 1$

**Proof**: The basic definition:

$$b_1^{IV} = \frac{\sum_{i=1}^{N} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N} (z_i - \bar{z})(x_i - \bar{x})},$$

where $z_i$ is the instrument, $y_i$ is the dependent variable in the second stage, $x_i$ is the endogenous explanatory variable and $\bar{z}, \bar{y}, \bar{x}$ are sample means. Expanding this expression gives

$$b_1^{IV} = \frac{\sum_{i=1}^{N} (z_i y_i - \bar{z} y_i - z_i \bar{y} + \bar{z}\bar{y})}{\sum_{i=1}^{N} (z_i x_i - \bar{z} x_i - z_i \bar{x} + \bar{z}\bar{x})}.$$

Since in the present application $z_i$ is a dummy variable, we have:

$$\begin{aligned}
\sum_{i=1}^{N} z_i y_i &= N_1 \bar{y}_1, \\
\sum_{i=1}^{N} \bar{z} y_i &= \frac{N_1}{N} \sum_{i=1}^{N} y_i = \frac{N_1}{N} N \bar{y} = \bar{y} N_1, \\
\sum_{i=1}^{N} z_i \bar{y} &= \bar{y} \sum_{i=1}^{N} z_i = \bar{y} N_1, \\
\sum_{i=1}^{N} \bar{z}\bar{y} &= \bar{z}\bar{y} \sum_{i=1}^{N} 1 = \frac{N_1}{N} \bar{y} N = \bar{y} N_1,
\end{aligned}$$

where $N_1$ is the number of observations for which $z_i = 1$ and $\bar{y}_1$ is the sample average of $y_i$ over the part of the sample with $z_i = 1$. Analogously,

$$\begin{aligned}\sum_{i=1}^{N} z_i x_i &= N_1 \bar{x}_1, \\ \sum_{i=1}^{N} \bar{z} x_i &= \bar{x} N_1, \\ \sum_{i=1}^{N} z_i \bar{x} &= \bar{x} N_1, \\ \sum_{i=1}^{N} \bar{z}\bar{x} &= \frac{N_1}{N}\bar{x}N = \bar{x}N_1,\end{aligned}$$

where $\bar{x}_1$ is the average of $x_i$ over the part of the sample with $z_i = 1$. Thus,

$$\begin{aligned}b_1^{IV} &= \frac{N_1\bar{y}_1 - \bar{y}N_1 - \bar{y}N_1 + \bar{y}N_1}{N_1\bar{x}_1 - \bar{x}N_1 - \bar{x}N_1 + \bar{x}N_1}, \\ b_1^{IV} &= \frac{N_1\bar{y}_1 - \bar{y}N_1}{N_1\bar{x}_1 - \bar{x}N_1}, \\ b_1^{IV} &= \frac{\bar{y}_1 - \bar{y}}{\bar{x}_1 - \bar{x}}. \qquad (2.1)\end{aligned}$$

By definition:

$$\bar{y} = \frac{N_1}{N}\bar{y}_1 + \frac{N - N_1}{N}\bar{y}_0,$$

and

$$\bar{x} = \frac{N_1}{N}\bar{x}_1 + \frac{N - N_1}{N}\bar{x}_0,$$

where $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of $y_i$ and $x_i$ over the part of the sample with $z_i = 0$. Use these expressions in (2.1):

$$
\begin{aligned}
b_1^{IV} &= \frac{\bar{y}_1 - \left(\frac{N_1}{N}\bar{y}_1 + \frac{N-N_1}{N}\bar{y}_0\right)}{\bar{x}_1 - \left(\frac{N_1}{N}\bar{x}_1 + \frac{N-N_1}{N}\bar{x}_0\right)}, \\
b_1^{IV} &= \frac{\frac{N\bar{y}_1 - (N_1\bar{y}_1 + (N-N_1)\bar{y}_0)}{N}}{\frac{N\bar{x}_1 - (N_1\bar{x}_1 + (N-N_1)\bar{x}_0)}{N}}, \\
b_1^{IV} &= \frac{N\bar{y}_1 - N_1\bar{y}_1 - (N-N_1)\,\bar{y}_0}{N\bar{x}_1 - N_1\bar{x}_1 - (N-N_1)\,\bar{x}_0}, \\
b_1^{IV} &= \frac{(N-N_1)}{(N-N_1)}\frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}, \\
b_1^{IV} &= \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}, \qquad (2.2)
\end{aligned}
$$

**PhD Programme: Applied Econometrics**
**Department of Economics, University of Gothenburg**
**Appendix 2: Lecture 12**
Måns Söderbom

## The Local Average Treatment Effect

## 1. Simulating LATE in Stata

Stata code:

```
clear
set seed 54687
set obs 20000

/* first, randomly assign the instrument  - say half-half */
ge z = uniform()>.5

/* then, generate never-takers (d00), always-takers (d11) and compliers
(d01), independent of z */

ge d00=(_n<=5000)
ge d11=(_n>5000 & _n<=10000)
ge d01=(_n>10000)

/* observed outcomes: always zero for never-takers, always one for
always-takers, depends on the IV for compliers */
ge D=d11+z*d01

/* now give the three groups different LATE. Without loss of
generality, assume within group homogeneity. */

ge late=-1 if d00==1
replace late=0 if d11==1
replace late=1 if d01==1

/* next generate potential outcomes y0,y1 */

ge y0=0.25*invnorm(uniform())
ge y1=y0+late

/* actual outcome depends on treatment status */
ge y = D*y1+(1-D)*y0

/* the average treatment effect is simply the sample mean of late */
sum late

/* OLS doesn't give you ATE or LATE */
reg y D

/* IV gives you the LATE for the compliers */
ivreg y (D=z)

exit
```

Results:

```
. /* the average treatment effect is simply the sample mean of late */
. sum late

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        late |     20000         .25    .8291769         -1          1

.
. /* OLS doesn't give you ATE or LATE */
. reg y D

      Source |       SS       df       MS              Number of obs =   20000
-------------+------------------------------           F(  1, 19998) = 6631.34
       Model |  1246.68658     1  1246.68658           Prob > F      =  0.0000
    Residual |  3759.60651 19998  .187999125           R-squared     =  0.2490
-------------+------------------------------           Adj R-squared =  0.2490
       Total |  5006.29309 19999  .250327171           Root MSE      =  .43359

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   .4993491    .006132    81.43   0.000     .4873298    .5113684
       _cons |   .0005316   .0043511     0.12   0.903     -.007997    .0090602
------------------------------------------------------------------------------

.
. /* IV gives you the LATE for the compliers */
. ivreg y (D=z)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =   20000
-------------+------------------------------           F(  1, 19998) = 5048.40
       Model | -86.6843722     1 -86.6843722           Prob > F      =  0.0000
    Residual |  5092.97746 19998   .25467434           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |  5006.29309 19999  .250327171           Root MSE      =  .50465

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |   1.015767   .0142961    71.05   0.000     .9877453    1.043788
       _cons |  -.2594847   .0080341   -32.30   0.000    -.2752322   -.2437373
------------------------------------------------------------------------------
Instrumented:  D
Instruments:   z
------------------------------------------------------------------------------
```

Recall: Treatment effect is 1.0 for the compliers, 0.0 for the always-takers and -1.0 for the never-takers.