Recitation Notes 5

Konrad Menzel

October 13, 2006

1 Instrumental Variables (continued)

1.1 Omitted Variables and the Wald Estimator

Consider a Wald estimator for the Angrist (1991) approach to estimating the intertemporal elasticity of substitution of labor supply for the regression

$$h_{it} = \alpha + \beta w_{it} + \varepsilon_{it}$$

and assume that there are actually aggregate shocks to wages, in a way such that changes in the workers' marginal utility of wealth generate a linear trend in labor supply, $\gamma_0 + \gamma_1 t$. Therefore, the "right" regression we should actually run is

$$h_{it} = \gamma_0 + \gamma_1 t + \beta w_{it} + \varepsilon_{it}$$

Then, using some - arbitary - aggregate binary variable Z_{1t} as an instrumental variable gives - by the 2SLS formula from last week - the Wald estimand

$$\begin{aligned} \text{plim}_N \hat{\beta}_1 &= \frac{\mathbb{E}[h_{it}|Z_{1t}=1] - \mathbb{E}[h_{it}|Z_{1t}=0]}{\mathbb{E}[w_{it}|Z_{1t}=1] - \mathbb{E}[w_{it}|Z_{1t}=0]} \\ &= \frac{\left(\mathbb{E}[w_{it}|Z_{1t}=1] - \mathbb{E}[w_{it}|Z_{1t}=0]\right)\beta + \mathbb{E}[t\gamma + \varepsilon_{it}|Z_{1t}=1] - \mathbb{E}[t\gamma + \varepsilon_{it}|Z_{1t}=0]}{\mathbb{E}[w_{it}|Z_{1t}=1] - \mathbb{E}[w_{it}|Z_{1t}=0]} \end{aligned}$$

Assuming that $\text{Cov}(Z_{1t}, \varepsilon_{it}) = 0$, this simplifies to

$$\text{plim}_N \hat{\beta}_1 = \beta + \frac{\mathbb{E}[t|Z_{1t}=1] - \mathbb{E}[t|Z_{1t}=0]}{\mathbb{E}[w_{it}|Z_{1t}=1] - \mathbb{E}[w_{it}|Z_{1t}=0]}\gamma_1 =: \beta + \omega_1\gamma_1$$

If e.g. we are looking at a balanced panel for the years 1969 to 1979 and $Z_{1t} = \mathbb{1}\{t > 1974\}$, we can calculate that

$$\mathbb{E}[t|Z_{1t}=1] - \mathbb{E}[t|Z_{1t}=0] = \frac{75+76+77+78+79}{5} - \frac{69+70+71+72+73+74}{6} = 77 - 71.5 = 5.5$$

So if there is an increase in wages over time which is accompanied by a negative time trend in the marginal utility of wealth, $\gamma_1 < 0$, our IV derived from the year dummies gives downward biased estimates.
Now. if we had a second binary instrument Z_2, we could similarly obtain

$$\text{plim}_N \hat{\beta}_2 = \beta + \frac{\mathbb{E}[t|Z_{2t}=1] - \mathbb{E}[t|Z_{2t}=0]}{\mathbb{E}[w_{it}|Z_{2t}=1] - \mathbb{E}[w_{it}|Z_{2t}=0]}\gamma_1 =: \beta + \omega_2\gamma_1$$

You should notice that all information we need to compute ω_2 is observed in the data, and the only reason why we can't compute the bias right away is that we don't know γ_1.
But since we have two different instruments, it is possible to subtract out the biases by computing

$$\tilde{\beta} = \frac{\omega_2\hat{\beta}_1 - \omega_1\beta_2}{\omega_2 - \omega_1} \longrightarrow \frac{\omega_2(\beta + \omega_1\gamma_1) - \omega_1(\beta + \omega_2\gamma_1)}{\omega_2 - \omega_1} = \beta$$

Note that for this we need two instruments that generate different ωs, and that this only works under the presumption that we know what the conditional averages of the omitted variable are (which in this case is trivial since we actually know the values of t). If we don't know enough about the omitted variable (which is almost always the case), there is in general not much we can do to fix a "contaminated" instrumental variable.

1.2 2SLS as a Grouped Data IV

Now assume that we have k discrete instrumental variables Z, and we can think of them as dividing your sample into r cells. Without loss of generality, we can therefore assume that the instruments are indicator functions for the R *mutually exclusive* groups, i.e.

$$Z_{ri} := \begin{cases} 1 & \text{if observation } i \text{ falls into cell } r \\ 0 & \text{otherwise} \end{cases}$$

In vector notation, we can therefore write the matrix of all instruments as

$$\mathbf{Z} = [\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_R}] = \begin{bmatrix} \iota_{N_1} & 0 & 0 & \ldots & 0 \\ 0 & \iota_{N_2} & 0 & \ldots & 0 \\ 0 & 0 & \iota_{N_3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \iota_{N_R} \end{bmatrix}$$

possibly after reordering the observations, where ι_n is an n-dimensional column vector of ones, and N_r is the number of observations that falls into the rth cell.
For 2SLS, the first stage consists of fitting the endogenous right-hand side variable X onto the Zs in order to obtain (again in matrix notation)

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{P_Z X} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \\ &= \begin{bmatrix} \iota_{N_1} & 0 & 0 & \ldots & 0 \\ 0 & \iota_{N_2} & 0 & \ldots & 0 \\ 0 & 0 & \iota_{N_3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \iota_{N_R} \end{bmatrix} \begin{bmatrix} N_1 & 0 & 0 & \ldots & 0 \\ 0 & N_2 & 0 & \ldots & 0 \\ 0 & 0 & N_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & N_R \end{bmatrix}^{-1} \begin{bmatrix} \iota_{N_1} & 0 & 0 & \ldots & 0 \\ 0 & \iota_{N_2} & 0 & \ldots & 0 \\ 0 & 0 & \iota_{N_3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \iota_{N_R} \end{bmatrix}' \mathbf{X} \\ &= \frac{1}{N_1}\mathbf{Z_1}\mathbf{Z_1}'\mathbf{X} + \frac{1}{N_2}\mathbf{Z_2}\mathbf{Z_2}'\mathbf{X} + \ldots + \frac{1}{N_R}\mathbf{Z_R}\mathbf{Z_R}'\mathbf{X} = \begin{bmatrix} \iota_{N_1}\bar{X}_1 \\ \iota_{N_2}\bar{X}_2 \\ \vdots \\ \iota_{N_R}\bar{X}_R \end{bmatrix} = \Big[\mathbb{E}_N[X_i|Z_i] \Big]_{i=1}^{N} \end{aligned}$$

since inside the inverse, $\iota_n'\iota_n = \sum_{j=1}^{n} 1 = n$. Notice that the whole trick is that groups are mutually exclusive, so that the inner product of the instrument matrix with itself is diagonal. The interpretation

of what is going on here is also straightforward: fitting to/projecting onto the group dummies is exactly the same thing as replacing the variable with group-wise means.
In the second stage, we do plain OLS of the left-hand-side endogenous variable Y on the fitted right-hand-side endogenous variable, $\hat{X}$, which gives us

$$\hat{\beta}_{2SLS} = \frac{\text{Cov}(\hat{X}, Y)}{\text{Var}(\hat{X})} = \frac{\text{Cov}(\mathbf{P_Z}X, Y)}{\text{Var}(\mathbf{P_Z}X)} = \frac{\text{Cov}(\mathbb{E}_N[X|Z], \mathbb{E}_N[Y|Z])}{\text{Var}(\mathbb{E}_N[X|Z])}$$

For the pairwise Wald IV between groups j and k, we basically do the same thing, only that we ignore all other groups, which is equivalent to doing 2SLS using instruments $\mathbf{Z_{jk}} = [\mathbf{Z_j}, \mathbf{Z_k}]$, but since now we only have one binary instrument, the estimator simplifies significantly to

$$\hat{\beta}_{jk} = \frac{\mathbb{E}_N[Y|Z_j = 1] - \mathbb{E}_N[Y|Z_k = 1]}{\mathbb{E}_N[X|Z_j = 1] - \mathbb{E}_N[X|Z_k = 1]} = \frac{\bar{Y}_j - \bar{Y}_k}{\bar{X}_j - \bar{X}_k}$$

The asymptotic variance of this estimator under homoskedasticity of ε_i, i.e. $\text{Var}(\varepsilon_i|Z_i) = \text{Var}(\varepsilon_i) = \sigma^2$ can be computed as

$$\text{Var}(\hat{\beta}_{jk}) = \frac{\text{Var}(\bar{Y}_j - \bar{Y}_k)}{(\bar{X}_j - \bar{X}_k)^2} = \frac{\sigma^2\left(\frac{1}{N_j} + \frac{1}{N_k}\right)}{(\bar{X}_j - \bar{X}_k)^2} = \frac{\sigma^2(N_j + N_k)}{N_j N_k (\bar{X}_j - \bar{X}_k)^2}$$

If we write out the 2SLS estimator, we get

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \frac{\text{Cov}(\mathbb{E}_N[X|Z], \mathbb{E}_N[Y|Z])}{\text{Var}(\mathbb{E}_N[X|Z])} \\
&= \frac{\sum_{r=1}^R N_r \bar{X}_r (\bar{Y}_r - \bar{Y})}{\sum_{r=1}^R N_r \bar{X}_r (\bar{X}_r - \bar{X})} \\
&= \frac{\sum_{r=1}^R N_r \bar{X}_r \sum_{s=1}^R \frac{N_s}{N} (\bar{Y}_r - \bar{Y}_s)}{\sum_{r=1}^R N_r \bar{X}_r (\bar{X}_r - \bar{X})} \\
&= \frac{\sum_{r=1}^R \sum_{s>r} N_r N_s (\bar{X}_r - \bar{X}_s)(\bar{Y}_r - \bar{Y}_s)}{N \sum_{r=1}^R N_r \bar{X}_r (\bar{X}_r - \bar{X})} \\
&= \sum_{r=1}^R \sum_{s>r} \frac{N_r N_s (\bar{X}_r - \bar{X}_s)^2}{N \sum_{t=1}^R N_t \bar{X}_t (\bar{X}_t - \bar{X})} \hat{\beta}_{rs} =: \sum_{r=1}^R \sum_{s=r+1}^R w_{rs} \hat{\beta}_{rs}
\end{aligned}$$

Notice that under homoskedasticity of ε_i, the weights on the pairwise Wald estimates $\hat{\beta}_{rs}$ in the 2SLS estimator are

$$w_{rs} = \frac{N_r N_s (\bar{X}_r - \bar{X}_s)^2}{N \sum_{t=1}^R N_t \bar{X}_t (\bar{X}_t - \bar{X})} \propto \frac{N_r + N_s}{\text{Var}(\hat{\beta}_{rs})}$$

where "$\propto$" stands for "proportional to". This is exactly what the statement in the Angrist (1991) paper means that 2SLS is an "efficient GLS-combination" of the pairwise Wald-IV estimators.

1.3 Heterogeneous Treatment Effects (not covered in recitation)

Assume you have data about a training program for unemployed workers, and you want to tell a policy maker whether the program was successful so that the government should continue to finance it. Say, your main outcome of interest is earnings six months after the training program, y_i, and you know whether a

particular person participated ($D_i = 1$) or not ($D_i = 0$).
Now, so far we have only looked at regressions of the type

$$Y_i = \alpha + D_i\beta + \varepsilon_i$$

This implies that across the whole population, everyone who participates earns exactly β dollars more than if he hadn't received training. This doesn't make much sense in a real-world application. For example we'd think that, say, a basic literacy program wouldn't have much of an effect on more educated individuals, or that the causal effect of the number of children in a household on the mother's labor supply differs a lot with the mother's age, education and other characteristics.
So we'd rather like to write the model as

$$Y_i = \alpha + D_i\beta_i + \varepsilon_i$$

which allows individuals not only to have different initial earnings levels, but each person could also benefit from the program to a different degree.
So if people participated in the program regardless of their initial earnings level (say by random assignment), $D_i \perp \varepsilon_i$, and we could run OLS

$$\begin{aligned}\hat{\beta}_{LS} \quad &\longrightarrow \quad \frac{\mathbb{E}[D_iY_i] - \mathbb{E}[Y_i]\mathbb{E}[D_i]}{\mathbb{E}[D_i] - \mathbb{E}[D_i]^2} = \frac{\mathbb{E}[D_i\beta_i] - \mathbb{E}[D_i\beta_i]\mathbb{E}[D_i]}{\mathbb{E}[D_i] - \mathbb{E}[D_i]^2} \\ &= \quad \frac{\left(\mathbb{E}[D_i] - \mathbb{E}[D_i]^2\right)\mathbb{E}[\beta_i|D_i = 1]}{\mathbb{E}[D_i] - \mathbb{E}[D_i]^2} = \mathbb{E}[\beta_i|D_i = 1]\end{aligned}$$

by the law of iterated expectations. Therefore, OLS estimates the *treatment effect on the treated* individuals for our training program, which need in general not be the same as the *average treatment effect*

$$\beta_{ATE} := \mathbb{E}[\beta_i]$$

for all individuals regardless of their actual treatment status. We might be tempted to think that it is a disadvantage of OLS that it doesn't pick up the treatment effect for the whole population, but for practical purposes, this typically isn't so. Often individuals which aren't reached by our program aren't too relevant for our evaluation question either - e.g. we wouldn't observe unemployed economics PhDs participating in a basic literacy program, but we wouldn't be too interested either in the effect of the literacy program on them because we'd never choose to send them to that program anyway on apriori grounds. The treatment effect on the treated answers the question about how much better off we are by running the program (ignoring the cost of running it) compared to a world in which we shut it down entirely and for everyone.
In many situations, there is always some degree of self-selection into, or imperfect compliance with a particular treatment, but sometimes we have a good instrument Z for participation, e.g. a randomized encouragement to participate, some exogenous eligibility rule, or some factor that shifts exogenously individuals' cost of taking up the program.
This assignment Z makes only some people switch from control to treatment (*compliers*), and from treatment to control (*defiers*). But there are also individuals which participate no matter what (*always-takers*) or don't in any case (*never-takers*). In order to formalize that, we denote the treatment a person would receive if $Z_i = 1$ with D_{1i}, and for $Z_i = 0$, we would observe D_{0i}. Now let's also assume

1. Independence: $(D_i, \beta_i, \varepsilon_i) \perp\!\!\!\perp Z_i$
2. Monotonicity: $D_{1i} \geq D_{0i}$ for all individuals

Under these assumptions

$$\begin{aligned}
\hat{\beta}_{Wald} &= \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]} = \frac{\mathbb{E}[D_i\beta_i|Z_i=1]-\mathbb{E}[D_i\beta_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]} \\
&= \frac{\mathbb{E}[D_{1i}\beta_i|Z_i=1]-\mathbb{E}[D_{0i}\beta_i|Z_i=0]}{\mathbb{E}[D_{1i}|Z_i=1]-\mathbb{E}[D_{0i}|Z_i=0]} \overset{indep.}{=} \frac{\mathbb{E}[(D_{1i}-D_{0i})\beta_i]}{\mathbb{E}[D_{1i}-D_{0i}]} \\
&\overset{monot.}{=} \frac{\mathbb{E}[D_{1i}-D_{0i}]\mathbb{E}[\beta_i|D_{1i}=1,D_{0i}=0]}{\mathbb{E}[D_{1i}-D_{0i}]} \\
&= \mathbb{E}[\beta_i|D_{1i}>D_{0i}] =: \beta_{LATE}
\end{aligned}$$

That is, Wald (and thereby all instrumental variables estimators) estimate the average effect of the treatment on the subpopulation of compliers with a particular instrument. The important point to take away from this is that each instruments has a different set of compliers, so e.g. in the Angrist and Evans paper on the twin births and same-sex instruments we'd expect the LATE for the effect of the third child on mothers of twins to be different from the LATE on mothers whose first two children were of the same sex - the group of mothers that have a third child in order to balance the sex composition of their offspring is different from those mothers who have twins at their second birth (and therefore automatically have a third child). This is again not a weakness of the estimator, but we just have to be aware that each instrument defines its own Wald *estimand*. If Z is our policy intervention (e.g. offering a training program for which participation is voluntary), the instrumental variables estimator gives us directly the answer about the effect on those individuals which were affected by that policy, excluding people who would have received treatment in any case. And that's the actual question we'd typically want to ask if we evaluate the policy corresponding to Z: we want to know by how much e.g. offering that particular training program makes everyone better off given that on the one hand in a world without that intervention there could still be close substitutes available, and that on the other hand, many people wouldn't want to take part in the program either way.

1.4 A Few Hints About Calculations with Scalars in Stata

After running a regression in Stata, e.g.

```
ivreg lnh year (lnw=z)
```

you can retrieve the estimated coefficient on lnw by typing

```
local beta=_b[lnw]
```

where "local" simply means that you define a scalar (formally a local macro), e.g.

```
local abc=sin(345)+5
```

which you can then manipulate or display by typing, e.g.

```
local def=‘abc’*456
display ‘def’+1.2345
```

The ‘’ marks around the name of the local name tell Stata to evaluate it - i.e. to put the number stored under the name `abc` into whichever expression the `‘abc’` part appears in. `_b[var1]` refers to the coefficient on the variable `var1` in the last regression of any kind you ran, and in a similar fashion you can retrieve the standard error of that coefficient in order to run a simple t-test:

```
local beta=_b[lnw]
local se=_se[lnw]
display ‘beta’/‘se’
```