14.661: Recitation 4

Chris Walters

October 7, 2010

1 IV and Grouped Regression

Suppose that we want to estimate

$$y_i = \alpha + \beta x_i + \epsilon_i$$

and we have instruments $z_1 ... z_K$, which are mutually exclusive and exhaustive dummy variables. This is the situation in Angrist (1991), where the z's are year dummies, y is log hours, and x is log wages. Suppose that n_k observations have $z_{ik} = 1$. As we saw in lecture, 2SLS in this setup will turn out to be algebraically equivalent to GLS on means of y and x, where the groups are observations with different z's switched on.

We want to use these instruments to run 2SLS. Recall the 2SLS procedure:

1. Regress x_i on the z's and get the fitted values, $\hat{x}_i$
2. Regress y_i on $\hat{x}_i$

The first stage is

$$x_i = z_i' \pi + u_i$$

where z_i is now a vector including all of the z's. Using the multivariate OLS formula, the first stage coefficient vector will be

$$\hat{\pi} = \left(\sum_i z_i z_i' \right)^{-1} \left(\sum_i z_i x_i \right)$$

It is worth thinking about what this object is. We know that z_i looks like

$$z_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

That is, for individuals in group k, it will have a 1 in the k-th position and zeros everywhere else. Then the matrix $z_i z_i'$ is

$$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}_{K\times 1} \cdot [0,\ldots,1,\ldots,0]_{1\times K} = \begin{bmatrix} 0 & & \cdots & & 0 \\ & \ddots & & \ddots & \\ \vdots & & 1 & & \vdots \\ & \ddots & & \ddots & \\ 0 & & \cdots & & 0 \end{bmatrix}_{K\times K}$$

For individuals in group k, this $K \times K$ matrix has a 1 on the k-th position on the diagonal, and it has zeros everywhere else. The sum over all observations is

$$\sum_i z_i z_i' = \begin{bmatrix} n_1 & & \cdots & & 0 \\ & \ddots & & \ddots & \\ \vdots & & n_k & & \vdots \\ & \ddots & & \ddots & \\ 0 & & \cdots & & n_K \end{bmatrix}_{K\times K}$$

and the inverse is

$$\left(\sum_i z_i z_i'\right)^{-1} = \begin{bmatrix} \frac{1}{n_1} & & \cdots & & 0 \\ & \ddots & & \ddots & \\ \vdots & & \frac{1}{n_k} & & \vdots \\ & \ddots & & \ddots & \\ 0 & & \cdots & & \frac{1}{n_K} \end{bmatrix}_{K\times K}$$

Similarly, we have

$$z_i x_i = \begin{bmatrix} 0 \\ \vdots \\ x_i \\ \vdots \\ 0 \end{bmatrix}$$

for an individual in group k. Summing these up,

$$\sum_i z_i x_i = \begin{bmatrix} \sum_{i\in 1} x_i \\ \vdots \\ \sum_{i\in k} x_i \\ \vdots \\ \sum_{i\in K} x_i \end{bmatrix}$$

So the first-stage coefficient vector is

$$\hat{\pi} = \left(\sum_i z_i z_i'\right)^{-1} \left(\sum_i z_i x_i\right) = \begin{bmatrix} \frac{1}{n_1} & & \cdots & & 0 \\ & \ddots & & \ddots & \\ \vdots & & \frac{1}{n_k} & & \vdots \\ & \ddots & & \ddots & \\ 0 & & \cdots & & \frac{1}{n_K} \end{bmatrix}_{K \times K} \cdot \begin{bmatrix} \sum_{i \in 1} x_i \\ \vdots \\ \sum_{i \in k} x_i \\ \vdots \\ \sum_{i \in K} x_i \end{bmatrix}_{K \times 1}$$

$$= \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \\ \vdots \\ \bar{x}_K \end{bmatrix}_{K \times 1}$$

And the first stage fitted values are

$$\hat{x}_i = z_i' \hat{\pi}$$

$$= [0, \ldots, 1, \ldots, 0] \cdot \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \\ \vdots \\ \bar{x}_K \end{bmatrix}$$

$$= \bar{x}_k$$

So the first-stage fitted values are group means! In our data, the fitted values look like this:

$$\hat{x} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_1 \\ \vdots \\ \bar{x}_K \\ \vdots \\ \bar{x}_K \end{bmatrix}$$

We have n_1 copies of $\bar{x}_1$, followed by n_2 copies of $\bar{x}_2$, etc.

Next, we want to run the second stage by regressing y_i on $\hat{x}_i$. Before doing this, we should note the following useful property of fitted values:

$$Cov\left(y_i, \hat{x}_i\right) = Cov\left(\hat{y}_i + \tilde{y}_i, \hat{x}_i\right) = Cov\left(\hat{y}_i, \hat{x}_i\right)$$

where $\hat{y}$ and $\tilde{y}$ are fitted values and residuals from a regression of y on the z's. Therefore, in the second stage, we can put the $\hat{y}$'s on the left-hand side instead of the y's and it doesn't make a difference. The second stage OLS regression is therefore

$$\bar{y}_k = \alpha + \beta \bar{x}_k + \epsilon_i,$$

where there are n_k copies of each $(\bar{y}_k, \bar{x}_k)$ pair in the data. This is a regression of group means on group means, with n_k copies of each pair. Therefore,

$$\begin{pmatrix} \hat{\alpha}_{2SLS} \\ \hat{\beta}_{2SLS} \end{pmatrix} = \arg\min_{\alpha,\beta} \sum_i (\bar{y}_k - \alpha - \beta \bar{x}_k)^2$$

$$\Longrightarrow \begin{pmatrix} \hat{\alpha}_{2SLS} \\ \hat{\beta}_{2SLS} \end{pmatrix} = \arg\min_{\alpha,\beta} \sum_k n_k (\bar{y}_k - \alpha - \beta \bar{x}_k)^2$$

This is a weighted least squares (WLS) regression using the group-means, with weights equal to group sizes!

Note that WLS using group-sizes as weights is generalized least squares (GLS) on the equation

$$\bar{y}_k = \alpha + \beta \bar{x}_k + \epsilon_k$$

where we now have 1 observation per group. If our original microdata model was

$$y_i = \alpha + \beta x_i + \epsilon_i, \; Var(\epsilon_i) = \sigma^2,$$

then in the grouped equation

$$\epsilon_k = \bar{\epsilon} \text{ for } i \text{ in group } k,$$

so

$$Var(\epsilon_k) = \frac{\sigma^2}{n_k}$$

We therefore have heteroskedasticity in the grouped equation, and the efficient thing to do is to weight each observation by the inverse of its variance, i.e., group size.

So, to interpret what we've shown: IV with group dummies as instruments is equivalent to GLS estimation of the group-mean equation. Note the contrast with fixed effects. When we do FE, we assume that permanent variation across groups is "contaminated" and therefore use only the within-group variation. When we do IV, we assume that the within-group variation is contaminated and use only the variation in means across groups.

Finally, one more fact: The minimized GLS minimand is

$$\hat{T} = \sum_k \frac{n_k}{\hat{\sigma}^2} \cdot \left(\bar{y}_k - \hat{\alpha} - \hat{\beta} \bar{x}_k\right)^2$$

This statistic doubles as an overidentification test of the validity of the instruments. Under the null hypothesis that all of the instruments are valid, it will have a χ^2_{K-1} distribution. (We can do an overid test whenever we have more instruments than endogenous variables). If it is larger than the relevant critical value from the χ^2 distribution, we will reject the hypothesis that all of our instruments are valid, though we will not know which ones are invalid.

2 Overidentification tests

Let's return to our simple (possibly endogenous) regression

$$y_i = \alpha + \beta x_i + \epsilon_i$$

As I noted in a previous recitation, if we have a single instrument z_i, the "exclusion restriction" that $Cov(z_i, \epsilon_i) = 0$ is not testable. However, if we have many instruments, we can jointly test all of the exclusion restrictions together. Suppose we've run 2SLS, gotten $\hat{\beta}$, and obtained the residuals $\hat{\epsilon}_i$. If we've gotten something close to the true β and the exclusion restrictions are valid, then intuitively the $\hat{\epsilon}$'s should not exhibit much correlation with the z's. To check this, let's think about running the regression

$$\hat{\epsilon}_i = z_i'\delta + \eta_i$$

We want to test $H_0: \ \delta = 0$. The OLS estimator of δ will be

$$\hat{\delta} = \left(\sum_i z_i z_i'\right)^{-1} \left(\sum_i z_i \hat{\epsilon}_i\right)$$

To perform tests on this object, we need to get its distribution. Let's multiply by $\sqrt{N}$:

$$\sqrt{N}\hat{\delta} = \left(\frac{1}{N}\sum_i z_i z_i'\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_i z_i \hat{\epsilon}_i\right)$$

The law of large numbers and the continuous mapping theorem tell us that

$$\left(\frac{1}{N}\sum_i z_i z_i'\right)^{-1} \longrightarrow E\left[z_i z_i'\right]^{-1}$$

As long as our null hypothesis is true, ϵ and z are uncorrelated, and since we are working with plims we can replace $\hat{\epsilon}$ with ϵ. The CLT then gives us that

$$\frac{1}{\sqrt{N}}\sum_i z_i \epsilon_i \longrightarrow N\left(0, Var\left(z_i \epsilon_i\right)\right) = N\left(0, \sigma^2 E\left[z_i z_i'\right]\right)$$

Combining this with the first term yields

$$\sqrt{N}\hat{\delta} \longrightarrow N\left(0, \sigma^2 E\left[z_i z_i'\right]^{-1}\right)$$

Since we now know the distribution of $\hat{\delta}$ under H_0, we can form a test statistic. The usual way to do this with a vector is to construct the quadratic form

$$\hat{T} = \left(\sqrt{N}\hat{\delta}'\right) \widehat{Var}\left(\sqrt{N}\hat{\delta}\right)^{-1} \left(\sqrt{N}\hat{\delta}\right)$$

Under H_0, this is a quadratic form in a multivarite normal random vector, so it will have a χ^2 distribution. Filling in some terms, we have

$$\hat{T} = N \cdot \left(\sum_i z_i'\hat{\epsilon}_i\right)\left(\sum_i z_i z_i'\right)^{-1}\left[\hat{\sigma}^2 \cdot \left(\frac{1}{N}\sum_i z_i z_i'\right)^{-1}\right]^{-1} \cdot \left(\sum_i z_i z_i'\right)^{-1}\left(\sum_i z_i\hat{\epsilon}_i\right)$$

$$\implies \hat{T} = \frac{\left(\sum_i z_i'\hat{\epsilon}_i\right)\left(\sum_i z_i z_i'\right)^{-1}\left(\sum_i z_i\hat{\epsilon}_i\right)}{\hat{\sigma}^2}$$

This is an "omnibus" overidentification test for the validity of our instruments. It will have a χ^2 distribution with degrees of freedom equal to the degree of overidentification $(K-1)$. In the dummy instrument case above, the same manipulations we did before can be used to show that $\hat{T}$ just turns out to be the minimized minmand for grouped data GLS. If this statistic is very large, we will reject H_0, and we will therefore conclude that we have at least one invalid instrument; we can't know which one it is without further information. There are many other versions of overidentification testing, but usually they boil down to the same thing: Checking to see whether the residuals and instruments are "close enough" to orthogonal in a statistical sense, often by regressing the residuals on the instruments. You can also see from this formula why we can't do an overid test with only one instrument (hint: what first-order condition does univariate IV solve?)

3 Division Bias

It is worth going through the math of division bias a little more carefully. Suppose we are not worried about OVB for the moment, and we want to estimate the equation

$$\log h_i^* = \alpha + \delta \log w_i^* + \eta_i$$

If we had data on true hours and wages h_i^* and w_i^* we would be able to consistently estimate δ; in other words, $Cov(\log w_i^*, \eta_i) = 0$.

However, many people do not report an explicit hourly wage, so we have to compute w. If we had perfect measures of hours and earnings, we could successfully compute the true w_i^* :

$$w_i^* = \frac{y_i^*}{h_i^*}$$

Furthermore, hours are measured with multiplicative error v_i, which we can assume is independent of everything else in the model. We observe:

$$h_i = h_i^* \cdot v_i,\ y_i^*$$

So observed wages are

$$w_i = \frac{y_i^*}{h_i} = \frac{y_i^*}{v_i h_i^*}$$

$$\implies \log w_i = \log w_i^* - \log v_i$$

Suppose we regress observed hours on observed wages using OLS. Then we obtain

$$plim\hat{\delta} = \frac{Cov\left(\log h_i, \log w_i\right)}{Var\left(\log w_i\right)}$$

$$= \frac{Cov\left(\log h_i^* + v_i, \log w_i^* - \log v_i\right)}{Var\left(\log w_i^* - \log v_i\right)}$$

$$= \frac{Cov\left(\alpha + \delta \log w_i^* + \eta_i + \log v_i, \log w_i^* - \log v_i\right)}{Var\left(\log w_i^* - \log v_i\right)}$$

$$= \delta \frac{Var\left(\log w_i^*\right)}{Var\left(\log w_i^*\right) + Var\left(\log v_i\right)} - \frac{Var\left(\log v_i\right)}{Var\left(\log w_i^*\right) + Var\left(\log v_i\right)}$$

The first term shows the usual attenuation bias result – δ is multiplied by a positive number less than one, so the measurement error in w will pull our estimate towards zero. However, we also have a second term that is unambiguously negative; this term comes from the correlation between the measurement error on the left hand side and the measurement error in the denominator of our right hand side variable of interest. With a positive δ, this makes attenuation bias worse and can even reverse the sign of the coefficient.