

**ARE213****Econometrics****Spring 2006 UC Berkeley Department of Agricultural and Resource Economics**

## ENDOGENEITY II:

TWO-STAGE-LEAST-SQUARES, CONTROL FUNCTION,  
AND LIMITED-INFORMATION-MAXIMUM-LIKELIHOOD ESTIMATION

## 1. TWO-STAGE-LEAST-SQUARES

A more systematic way to combine the multiple instruments is through two-stage-least-squares estimation. Let us do this in more generality. The equation of interest is

$$Y_i = X_i' \beta + \varepsilon_i = X_{i1}' \beta_1 + X_{i2}' \beta_2 + \varepsilon_i.$$

Let  $\sigma^2$  be the variance of  $\varepsilon_i$ . The vector of covariates  $X_i$  can be split into two parts, a possibly endogenous part  $X_{i1}$  and an exogenous part  $X_{i2}$ . The vector of instruments is  $Z_i$ . It can be split into the excluded instruments  $Z_{i1}$  and the exogenous covariates  $X_{i2}$ , or  $Z_i = (Z_{i1}, X_{i2})$ . Typically the common part  $X_{i2}$  of the vectors  $Z_i$  and  $X_i$  will at least contain the intercept.

The TSLS estimation method consists of two stages. In the first stage all the endogenous regressors are regressed on all the instruments and exogenous variables. That is, we estimate

$$X_{i1} = Z_i' \Pi + \eta_i = Z_{i1}' \Pi_1 + X_{i2}' \Pi_2 + \eta_i.$$

Note that  $X_{i1}$  is a  $K$ -vector, so that with  $Z_i$  an  $L$ -vector,  $\Pi$  is a  $L \times K$  matrix of parameters.

Estimating this by least squares leads to

$$\hat{\Pi} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{X}_1.$$

We then calculate the predicted values for  $X$  based on this regression:

$$\hat{\mathbf{X}}_1 = \mathbf{Z}\hat{\Pi}.$$

Note that if we have a similar equation for  $X_{i2}$ ,

$$X_{i2} = Z'_i \Pi + \eta_i = Z'_{i1} \Pi_1 + X'_{i2} \Pi_2 + \eta_i,$$

the result would be  $\Pi_2 = I$  and  $\Pi_1 = 0$ , so that the predicted value is  $\hat{\mathbf{X}}_2 = \mathbf{X}_2$ . Hence in the end we could treat all regressors symmetrically and just regress  $\mathbf{X}$  on  $\mathbf{Z}$  to get

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.$$

In the second stage the outcome is regressed on the predicted regressors:

$$Y_i = \hat{X}'_i \beta + \nu = (Z'_i \hat{\Pi})' \beta + \nu_i.$$

We can write the estimator for  $\beta$  as:

$$\hat{\beta} = ((\mathbf{X}'\mathbf{Z}) \cdot (\mathbf{Z}'\mathbf{Z})^{-1} \cdot (\mathbf{Z}'\mathbf{X}))^{-1} \cdot (\mathbf{X}'\mathbf{Z}) \cdot (\mathbf{Z}'\mathbf{Z})^{-1} \cdot (\mathbf{Z}'\mathbf{Y}).$$

In large samples

$$\sqrt{N} \cdot (\hat{\beta} - \beta) \sim \mathcal{N}\left(0, \sigma^2 \cdot ((\mathbf{X}'\mathbf{Z}) \cdot (\mathbf{Z}'\mathbf{Z})^{-1} \cdot (\mathbf{Z}'\mathbf{X}))^{-1}\right).$$

The error variance is  $\mathbb{E}[(Y - X'\beta)^2]$ , estimated as  $\sum_i (Y_i - X'_i \hat{\beta})^2 / N$ . Note that this variance is not the variance you would get as the standard ols variance from regressing  $Y_i$  on  $\hat{X}_i$ .

Let us see what we get from this for the Angrist-Krueger data. The first stage is the same regression we did before:

$$\widehat{\text{educ}}_i = 12.6881 + 0.0566 \cdot \text{qob}_{2i} + 0.1173 \cdot \text{qob}_{3i} + 0.1514 \cdot \text{qob}_{4i}$$

$$(0.0115) \quad (0.0163) \quad (0.0160) \quad (0.0163)$$

With the estimated coefficients from this regression the predicted value for the endogenous regressor for each observation is estimated as  $\widehat{\text{educ}}_i$ . In the second stage the regression of

interest is estimated using the predicted value for the endogenous regressor. For the AK data we get:

$$\begin{array}{rcc} \log(\widehat{\text{earnings}})_i & = & 4.5898 + 0.1026 \cdot \widehat{\text{educ}}_i \\ \text{tsls se} & & (0.2490) \quad (0.0195) \\ \text{ols se} & & (0.2616) \quad (0.0205) \end{array}$$

Note that the ols standard errors differ from the tsls ones. The tsls ones are correct.

## 2. CONTROL FUNCTIONS

We can get the tsls estimates in a different way that gives some additional insights into the method. Again do the first stage by regressing the endogenous regressors on all the exogenous variables and the instruments. Now calculate not the predicted values but the residuals from these regressions:

$$\hat{\eta} = \mathbf{X} - \mathbf{Z}\hat{\Pi}.$$

Note that some of these  $\eta$ , namely those corresponding to exogenous regressors that are included in the set of instruments, will be identically equal to zero. Then in the second stage regress the outcome on all the regressors and add the first stage residuals as additional control variables:

$$Y_i = X_i'\beta + \hat{\eta}_i'\gamma + \nu_i.$$

Mechanically this gives us the same results as tsls, but in a different way. We now deal with the endogeneity by including an additional regressor, the control function such that conditional on the additional regressor those regressors that were originally endogenous are now exogenous. Before, we purged the endogeneity by replacing the endogenous regressors by exogenous predictors.

When we do this for the AK data we get the following estimates.

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 4.5898 + 0.1026 \cdot \text{educ}_i - 0.0318 \cdot \hat{\eta}_{3i} \\ &\quad (0.2458) \quad (0.0192) \quad (0.0192) \end{aligned}$$

Again the standard errors we get from just doing ols here are not necessarily correct.

### 3. LIMITED-INFORMATION-MAXIMUM-LIKELIHOOD

A third way of estimating the parameters in linear regressions with endogenous regressors is limited-information-maximum-likelihood estimation. The likelihood is based on normality for the reduced form errors  $\nu$  and  $\eta$  with covariance matrix  $\Omega$ , although consistency and asymptotic normality of the estimator do not rely on this assumption. The log likelihood function is

$$L = \sum_{i=1}^N -\ln(2\pi) - \frac{1}{2} \ln |\Omega| - \frac{1}{2} \begin{pmatrix} Y_i - (Z_i' \Pi)' \beta \\ X_i - Z_i' \Pi \end{pmatrix}' \Omega^{-1} \begin{pmatrix} Y_i - (Z_i' \Pi)' \beta \\ X_i - Z_i' \Pi \end{pmatrix}.$$

This is a standard log likelihood function, with all the standard properties. In particular the asymptotic variance for the estimator can be found using the second derivatives of the log likelihood function. The main issue is that its maximization is a little tricky, involving eigen values.

Formally, define for any matrix  $X$  the residual generating matrix  $\mathbf{M}_X$  as

$$\mathbf{M}_X = I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Also define

$$\mathbf{W} = (\mathbf{Y}, \mathbf{X}_1)' \mathbf{M}_Z (\mathbf{Y}, \mathbf{X}_1),$$

and

$$\mathbf{W}_1 = (\mathbf{Y}, \mathbf{X}_1)' \mathbf{M}_{\mathbf{X}_2} (\mathbf{Y}, \mathbf{X}_1).$$

Then let  $\lambda$  be the characteristic root, or the smallest eigenvalue, of  $W_1 \cdot W^{-1}$ . This parameter can also be characterized as

$$\lambda = \min_{\delta} \frac{\delta' W_1 \delta}{\delta' W \delta}.$$

Finally

$$\hat{\beta}_{liml} = (\mathbf{X}(\mathbf{I} - \lambda \mathbf{M}_Z)\mathbf{X})^{-1} (\mathbf{X}(\mathbf{I} - \lambda \mathbf{M}_Z)\mathbf{Y}).$$

In large samples  $\sqrt{N}(\lambda - 1) \rightarrow 0$ .

For the AK data the liml estimator leads to  $\lambda = 1.000008632$ .

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 4.5782 + 0.1035 \cdot \text{educ}_i \\ &\quad (0.2528) \quad (0.0198) \end{aligned}$$

This is close to, but not identical to tsls. In large samples the asymptotic distribution is identical to that of tsls.

#### 4. FEW WEAK INSTRUMENTS

For the asymptotic properties of tsls and liml we need the model to be identified. This requires that the instruments have some explanatory power for the endogenous variables. Traditionally the requirement is stated in terms of full rank of  $\Pi$ . If this is not satisfied we cannot consistently estimate all elements of  $\beta$  (although sometimes we may be able to estimate some elements but not others.) If the rank of  $\Pi$  is close to deficient, we have weak instruments. In that case the estimator may not be close to normally distributed. To see this, think about what happens if there is a single endogenous variable and a single instrument. In that case liml and tsls are identical, and both equal to the ratio of reduced form coefficients. If the true reduced form coefficients are both zero, the estimator is the ratio of two normal random variables with zero mean, leading to a Cauchy ratio that has much thicker tails than the normal distribution, and no moments.

To see what happens in practice, let us look at the AK data. We make one modification. Instead of using the actual quarter of birth data, we use a random number drawn from a uniform distribution over the integers 1 through 4. That makes the instruments completely uninformative, and we should get a wide confidence interval for the parameter of interest, the returns to education. This idea of looking to see what happens if the instrument is completely uninformative was first used in Bound, Jaeger and Baker (1995).

First the tsls estimates:

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 5.1403 + 0.0595 \cdot \widehat{\text{educ}}_i \\ &\quad (3.1823) \quad (0.2492) \end{aligned}$$

Next, the liml estimates:

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 5.1656 + 0.0575 \cdot \text{educ}_i \\ &\quad (0.3.4510) \quad (0.2702) \end{aligned}$$

In both cases the estimator has fairly wide confidence intervals, including all reasonable values for the returns to education. (We are pretty sure a priori that the returns are positive and less than 20%, so a confidence interval that includes  $(-0.50, 0.50)$  cannot be misleading, even if it is based on an incorrect assumption of normality.)

In practice, one should look at the F-statistic on the instruments in the first stage. The F-statistic has the form

$$F = \frac{(SSR_r - SSR_{ur}) \cdot (N - K)}{SSR_{ur} \cdot K_2},$$

where  $N$  is the sample size,  $K$  the total number of regressors in the regression (dimension of  $Z$ ),  $K_2$  the number of excluded instruments (dimension of  $Z_1$ ),  $SSR_r$  the restricted sum of squares, and  $SSR_{ur}$  the unrestricted sum of squares. If the F-statistic is at least 5 or larger, weak instrument problems are unlikely to be severe. With the real data, 4 regressors

and the excluded 3 instruments the F-statistic is 34.0094, suggesting normality is a good approximation. With the random qob data the F-statistic is 0.2035, suggesting normality for the estimator is unlikely to be a good approximation.

### 5. MANY WEAK INSTRUMENTS

With many weak instruments things can be more misleading. Following some of the regressions in the original AK paper, we look again at estimates based on the real data and estimates based on data with the quarter of birth variable replaced by a random number. To illustrate how misleading tsls and liml can be, we look at a case with lots of instruments. The exogenous variables are all interactions of state of birth and year of birth, leading to about 500 exogenous variables, and all of those interacted with an indicator for quarter of birth equal to 4 serve as instruments, so that there are 500 instruments. We look at tsls and liml for the real data and the random quarter of birth data. The results are in the following table.

500 Instruments	TOLS	LIML
Real QOB	0.073 (.057,.089)	0.094 (.061,.129)
Random QOB	0.059 (.042,.076)	-0.330 (-.999,-.149)

Note that with the random qob data tsls is much more misleading than liml. In fact, one can show that with many weak instruments tsls get very close to ols estimates, whereas liml is much less biased in that case. Even for liml the 95% confidence intervals are misleading with the random instruments, excluding all positive values for the returns to education.