ECON 407: Problem Set 2 2/14/2013

This problem set is due by 5pm Friday February 21. Please slide stapled hardcopies of all materials under my door. Question (1) requires no work in stata, whereas (2) requires stata code with a strong dose of interpreting your results bridging the theory and the empirics. Recall that I *will not* grade responses that haven't been formatted properly.

1. (35 points) The moment generating condition for deriving the instrumental variables estimator is $\mathbf{z}'\mathbf{e} = 0$ where $\mathbf{z}$ is defined (as in class), as

$$\mathbf{z} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_2 & \dots & \mathbf{x}_{K-1} & \mathbf{z}_K \end{bmatrix} = \begin{bmatrix} 1 & x_{1,2} & \dots & x_{1,K-1} & z_{1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i,2} & \dots & x_{i,K-1} & z_{i,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N,2} & \dots & x_{N,K-1} & z_{N,1} \end{bmatrix}_{N\times(K+1)} \tag{1}$$

   Note, as I have written it, column K is the endogenous variable, and we replace it with the instrumental variable $\mathbf{z}$ (an $N \times 1$ column vector). This is exactly the setup we discussed in class for the case of 1 endogenous variable and exactly 1 instrumental variable.

   (a) Derive the instrumental variable estimator for β. In doing this, check for conformability of $\mathbf{z}'\mathbf{e} = 0$ and investigate the implied relationship between each column of $\mathbf{z}$ and 0.

   (b) Write the relevancy equation for this IV regression providing interpretation for variables, parameters, and errors. What hypothesis test do we perform to see if $\mathbf{z}$ is relevant (and if relevant, strong)?

   (c) Derive the variance/covariance matrix for the model coefficients $\mathbf{b}^{IV}$. [Hint: Proceed in an analogous way as the derivation for the variance/covariance of the parameters in the OLS case.]

2. (65 points) Consider the following data set on fertility which can be browsed here. The data dictionary describing the fields can be found here. In this example, we'll look at the relationship between fertility and education to consider the problem of endogeneity. While this data set is not accompanied by a paper, there probably are numerous discussions and even stata code out there that does what we'll consider here. Beyond the issue of copying work, it is in your best interest to struggle through and really think about what you are doing without relying on resources found on the web. The data can be accessed in stata by `webuse set http://rlhick.people.wm.edu/econ407/data` and then `webuse fertility`.

   (a) Estimate an equation of the effect of education on the age at which women have their first child. Include appropriate control variables as needed. Comment on the exogeneity assumptions inherent in this model and interpret your findings in this regression.

   (b) In your view, hypothesize which variable- particularly given your results from part (a)- might be endogenous? For convenience, I'll refer to this potentially endogenous variable as $\mathbf{x}_K$ in what follows. Provide an intuitive explanation of the nature of the hypothesized endogeneity problem.

(c) In light of your potential endogeneity problem, pick an Instrumental Variable (IV) and provide an argument for how your IV meets the requirements we set out in class.

(d) Using the IV identified, in the previous part consider an IV regression approach to the endogeneity problem where you use one instrumental variable for $\mathbf{x}_K$. Test for the relevancy of your instrumental variable. How does your parameter estimates in the instrumental variables regression compare to your OLS regression? Is your IV strong?

(e) Using `mata` replicate the parameters and standard errors of this model based on your findings in Question 1 parts (a) and (c).

(f) Conduct a test for the endogeneity of $\mathbf{x}_K$ and provide an intuitive explanation of how this test works. You might find the stata package `ivreg2` and its associated commands very useful in conducting this test. You will need to install this off of the web using the command `ssc install ivreg2`. Version 9 users might find that this command does not work for them, so I recommend using a lab computer having the latest version of stata for this part of the assignment. You may also use the stata command `ivreg` for this work. Can we test for the exogeneity of your IV?

(g) Now instrument for $\mathbf{x}_K$ using more than one instrumental variable. Bearing in mind that the requirements you laid out in part (c) of this question also apply here, justify the choice of your second IV. Test for the relevancy of your instruments. How do your results change from the earlier regression with only 1 instrument?

(h) Test the endogeneity of $\mathbf{x}_K$ in this new model.

(i) Conduct a test for overidentification. What is the intuition of this test?

(j) What is the rationale for using more than 1 instrument for $\mathbf{x}_k$ and risking overidentification?