Econometrics I - Problem Set 1

to be handed on February 29th 2012 at the beginning of the class

Instructors: Michele Pellizzari

Teaching Assistant: Marianna Battaglia

You should hand in a single problem set for each group. Please, write clearly all the names of the group members on the front page. For the questions that require computer work, you should hand in a complete do-file that shows all the commands implemented and a log-file with the complete output.

1. Assume the following IV simple setting:

$$y_i = \beta_0 + \beta x_i + u_i \tag{1}$$

$$x_i = \theta_0 + \theta z_i + v_i \tag{2}$$

The IV-TSLS estimator is consistent both under exogeneity (H_0: $E(u|x) = 0$) and under endogeneity (H_1: $E(u|x) \neq 0$) of the x variable, OLS only under H_0. Under H_0, as n goes to infinity, any difference in the value of the estimator would vanish but we should use OLS because it is more efficient. To see that, derive the formula of the IV asymptotic variance and check that it is larger than that of OLS. Based on that, Hausman has developed a test of endogeneity based on the significance of the difference between the two estimators. The relevant statistics is the following: $(\beta^{IV} - \beta^{OLS})/stdv(\beta^{IV} - \beta^{OLS})$. To implement the test, you need the variance of the beta's difference. Show that it is equal to the difference of the variances of the two estimators (i.e. their covariance is zero).

2. Consider the following general heterogeneous effects model, where you have a binary treatment D_i, a binary instrument Z_i and where the effect of the treatment on the outcome and of the instrument on the treatment are potentially heterogeneous. The effect of the treatment on the outcome is:

$$Y_{1i} - Y_{0i} = \rho_i \tag{3}$$

so that (written in random coefficient notation),

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = \alpha_0 + \rho_i D_i + \eta_i \tag{4}$$

The effect of the instrument on the treatment (first stage) is:

$$D_i = D_{0i} + (D_{1i} - D_{0i})Z_i = \pi_0 + \pi_{1i} + \epsilon_i \tag{5}$$

(a) Show that if the first stage comes from heterogeneous potential treatment assignment but treatment effects are constant so that $Y_{1i} - Y_{0i} = \rho$, then LATE=ρ with or without the monotonicity assumption.

(b) Show what happens in the heterogeneous effects model if the monotonocity assumption fails: derive a formula for the "bias" induced in the IV estimand by failure of this condition. Based on that, discuss in what cases failures of monotonicity need not to be fatal.

(c) Using the definitions of ATE (average treatment effect) and ATT (average treatment effect on the treated), show which is their relationship with the LATE for a given instrument Z_i.

(d) (The Bloom result) Suppose that the assumptions of the LATE theorem hold, and that $E(D_i|Z_i = 0) = Pr(D_i = 1|Z_i = 0) = 0$. Prove that, then:

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{P(D_i = 1|Z_i = 1)} = E(Y_{1i} - Y_{0i}|D_i = 1) \tag{6}$$

So, what is going on? Which is the meaning of the above assumption? Compare with the standard LATE framework and discuss what are you going to estimate by IV in this modified setting.

(e) In the general setting, a problem is that, unlike ATT, the LATE is defined only over the population of compliers, which is a function of the instrument Z_i. Because of that, it is important to be able to characterise who are the compliers. Using the results of the LATE theorem and some simple probability theory and algebra, discuss how to measure (in probabilty terms) the size of the group of compliers and of the compliers among the treated, i.e. give the formulas for $Pr(D_{1i} > D_{0i})$ and $Pr(D_{1i} > D_{0i}|D_i = 1)$.

3. The code *late.do* generates a simulated sample of 20,000 individual observations with the following characteristics: a gender indicator, age, unobserved ability, education, an instrument (z, which can be interpreted as college proximity) and log wages. Use the part of the code that generates data under the assumption of heterogenous returns to education. No modifications of the code are required at this stage, just use the original one to generate the simulated data.

(a) Use the simulated data to estimate the returns to education using a standard log wage equation with gender and age as exogenous control variables. Produce both OLS and IV estimates, for the entire sample and for compliers and non-compliers separately. Compare and comment your results.

(b) Modify the code that generates the simulated data by changing one (only one) parameter to aggravate the endogeneity problem. Reproduce the OLS and IV estimates for the entire sample and for compliers and non-compliers separately. Compare and comment, also in relation to your results to part a and to the size of the endogeneity problem.

(c) Modify the original code (i.e. the one you used in part a) by changing one (only one) parameter to make your estimates more precise. Reproduce the OLS and IV estimates for the entire sample and for compliers and non-compliers separately. Compare and comment the estimates, also in relation to your results to part a and to the precision of your estimates.

(d) Modify the original code (i.e. the one you used in part a) in such a way that the instrument z becomes non-monotonic. Provide an intuition for what the instrument might be and why it might be non-monotonic. Produce OLS and IV estimates of the returns to education using the entire sample and separately for the subgroups defined by the direction of the instrument (i.e those for whom the instrument affects education positively and those for whom it affects education negatively). Comment and compare the estimates, also in relation to your results to part a.

4. Read the following paper and write a critical summary of no more than 1000 words:

- Avraham Ebenstein. "When is the Local Average Treatment Close to the Average?: Evidence from Fertility and Labor Supply", *Journal of Human Resources*, Fall 2009, Vol. 44 Issue 4, pp. 955-975.

Your summary should specify clearly what is the aim of the paper (e.g. estimating the causal effect of education on earnings), what are the main empirical challenges (e.g. ability is unobservable and may directly influence both the amount of education that one acquires and one's earnings) and how the author addresses them (e.g. using college proximity as an instrument for education). Be as formal as possible in your discussion, i.e. write down the main equations that are estimated in the paper and the identification restrictions. Additionally, also provide your comments on the interest of the topic and the validity of empirical strategy (are the identification restrictions convincing? Provide examples of cases in which they would not be valid. Are such cases likely to be frequent?).