Exercise-Week5

Part A. Omitted Variable and Proxy Variable

1. Load the data set, called WAGE2.DTA. The data contains 935 men in 1980 from the Young Men's Cohort of the National Longitudinal Survey (NLSY), USA.

2. Type **browse** and have a look at the data.

3. Summarize the data.

4. Perform the following regression:

$$\log(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + u_i.$$

What do you find about the estimated return to education?

5. Now you consider that the individual ability related to intelligence may have an effect on the wage determination. Thus, you want to modify the previous regression equation into:

$$\log(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \beta_3 Abil_i + u_i.$$

Unfortunately, you realize that there is no ability variable. Do you expect the OLS estimate from the previous equation is unbiased?

6. However, there is a variable, called IQ, which can serve as a proxy variable for the ability. Perform the following regression:

$$\log(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \beta_3 IQ_i + u_i.$$

How does the estimated return to education change?

Part B. Measurement Error

1. Load the data set, called wtp.dta. This is the survey data on the willingness to pay for a better waste service in Kuala Lumpur.

2. Browse the data and summarize them.

3. Construct the log of the following variables:

- income (called y in the data)
- willingness to pay (called wtp in the data)

Construct dummies for education by grouping in three broad categories:

- **gen educ = g_6**

- **recode educ 1/2 = 1 3/4 = 2 5/7 = 3**

The data set contains variables which code for the ownership of a TV (coded g_121), a washing machine (g_122), an air conditioning (g_124), and a car (g_11). As the names are not straightforward, rename them in an appropriate way by typing: e.g., **rename g_121 TV**

4. Run the regression of the following equation:

$$\ln wtp_i = \beta_0 + \beta_1 \ln income_i + u_i$$

What is the elasticity of WTP with respect to the income level? Is it significanly different from zero (5% significance level)? Correct for heteroskedasticity.

5. Income might be measured with error. In this case, the estimated relationship in the previous equation is biase (why?). To correct for this, we want to find variables correlated with income but not with the measurement error. The data set reports several variables on ownership (TV, washing machine, air conditioning and car) as well as education. We want to instrument income with these variables. Thus, run the following regression:

$$\ln income_i = \beta_0 + \beta_1 TV_i + \beta_2 WM_i + \beta_3 Car_i + \beta_4 AC_i + \beta_5 Educ1_i + \beta_6 Educ2_i + v_i$$

How strong is the correlation between income and these variables?

6. Estimate the regression model in question 4 using instrument variables. To this end, you need to type **ivreg lnwtp (lny = TV WM Car AC Educ1 Educ2)**.

Part 3. Simultaneous Equations Models

1. Load the data set, called MROZ.DTA, which is the data on working, married women already in the workforce. Among them are

- Age: woman's age
- Kidslt6: the number of children less than six years old
- Nwifeinc: the woman's nonwage income which includes husband's earnings

2. Browse and summarize the data.

3. The equilibrium conditions in the labor market are expressed by the following system of equations:

$$\begin{cases} \log(Hours_i) = \beta_0 + \beta_1 \log(wage_i) + \beta_2 Educ_i \\ \quad +\beta_3 Age_i + \beta_4 Kidslt6_i + \beta_5 Nwifeinc_i + u_i \\ \log(wage_i) = \alpha_0 + \alpha_1 \log(Hours_i) + \alpha_2 Educ_i \\ \quad +\alpha_3 Exper_i + \alpha_4 Exper_i^2 + v_i \end{cases}$$

4. First, run the OLS regression for each equation with ignoring the issue of endogeneity. The first equation is the labor supply function. What is the effect of log(wage) on the labor supply hours? Is it statistically significant?

5. Now, in order to control endogeneity in the estimation, we want to use two-stage least squares with instrument variables. Are both equations identified? If so, what are the instrument variables for each equation?

6. Run the two-stage least squares estimation. To this end, you need to type as follows:

- ivreg lhours (lwage = exper expersq educ age kidslt6 nwifeinc) educ age kidslt6 nwifeinc
- ivreg lwage (lhours = age kidslt6 nwifeinc educ exper expersq) educ exper expersq

Does the effect of log(wage) on the labor supply hours change?

7. Given the estimation results from OLS and 2SLS, we want to do Hausman's exogeneity test.

- In the first equation (labor supply), we want to test whether log(wage) is endogenous.
- First, we regress the following equation to get the residual $\widehat{v}_i$:

$$\begin{aligned}\log(wage_i) &= \alpha_0 + \alpha_1 Exper_i + \alpha_1 Exper_i^2 + \alpha_2 Educ_i \\ &\quad + \alpha_3 Age_i + \alpha_4 Kidslt6_i + \alpha_5 Nwifeinc_i + v_i\end{aligned}$$

- Then add $\widehat{v}_i$ in the first equation and do OLS:

$$\begin{aligned}\log(Hours_i) = \beta_0 + \beta_1 \log(wage_i) + \beta_2 Educ_i + \beta_3 Age_i \\ + \beta_4 Kidslt6_i + \beta_5 Nwifeinc_i + \gamma \widehat{v}_i + u_i\end{aligned}$$

What are the OLS estimate and t-statistic on $\widehat{v}_i$? What's your conclusition?