

## Problem Set 7. Measurement Error, Endogeneity and Simultaneous Equation Estimation

Try the problems marked with a \* before the class and bring your answers with you.

Bring a calculator to answer the in-class problems.

\*1. Some economists believe that earnings are related positively to the number of years of work experience. Often years of experience is measured with error in household surveys. (Eg. Many women who have time out of work looking after children, so calculating years of experience as age – age left education is not an accurate measure of actual work experience). Read in the data set *school.dta* from the course web site. This file contains observations on the earnings and a measure of years of work experience, (calculated as age - year left full-time education –5).

Regress the log of hourly pay (*loghpay*) on experience (*exper*) for

- a) women
- b) men

(use the commands `regress logpay exper if female==1`  
`regress logpay exper if female==0`  
)

Is there are difference in the male and female coefficients on experience?  
Why might this be?

You decide to instrument years of experience with number of children, (*numkids*). Using the stata commands on the lecture handout, (*instruments.doc*), do an instrumental variable regression. What do you find? Why?

Do the Wu-Hausman test of endogeneity. What do you find?

2. Read in the data set *farms.dta* from the web site. This data set contains observations on the output and inputs, (capital, labour, years of managerial experience) from 75 factories. The managerial experience variable is a proxy for the unobserved variable managerial efficiency.

Estimate a production function relating output to the 3 inputs.

Interpret the estimated coefficient on the managerial experience variable.

You decide to instrument the proxy experience variable using the manager's age as the instrument for the badly measured experience variable. What do you find? Why?

Do the Wu-Hausman test of endogeneity. What do you find?

3. Given the following model, you suspect the presence of measurement error in the left hand side (dependent) variable on the number of cigarettes smoked, (Numcigs).

$$\text{Numcigs} = b_0 + b_1 \text{Age} + u$$

$$\text{ie Numcigs}^{\text{observed}} = \text{Numcigs}^{\text{true}} + e$$

where e is a (random) error term

Given the following information work out the consequences of this type of measurement error

$$\begin{aligned} N=40 \quad \text{Cov}(\text{Numcigs}, \text{Age}) &= 100 \quad \text{Var}(\text{Age}) = 50 \\ \text{Var}(\text{Numcigs}) &= 200 \quad \text{Var}(u) = 10 \quad \text{Var}(e) = 10 \\ \text{Cov}(e, u) &= 0 \quad E(u) = 0 \quad E(e) = 0 \end{aligned}$$

4. Given the following model, you suspect the presence of measurement error in the right hand side (explanatory) variable, income

$$\text{Food\_consumption} = b_0 + b_1 \text{Income} + u$$

(people typically lie more about income than expenditure)

$$\text{ie income}^{\text{observed}} = \text{income}^{\text{true}} + w \quad \text{where } w \text{ is a random error}$$

Given the following information work out

- the true (unobserved) OLS estimate of the effect of income on food expenditure and income in the absence of measurement error
- the OLS estimate in the presence of this type of measurement error
- How do the results compare?

$$\begin{aligned} \text{Cov}(\text{Food}, \text{Income}^{\text{true}}) &= 100 \quad \text{Cov}(\text{Food}, \text{Income}^{\text{observed}}) = 50 \\ \text{Var}(\text{Income}^{\text{true}}) &= 200 \quad \text{Var}(\text{Income}^{\text{observed}}) = 500 \\ \text{Var}(\text{Food}^{\text{true}}) &= 200 \quad \text{Var}(u) = 20 \quad \text{Var}(w) = 20 \quad \text{Cov}(w, u) = 0 \\ E(u) &= 0 \quad E(w) = 0 \end{aligned}$$

Turn over

5. Given the following model, you suspect the presence of measurement error in the right hand side (explanatory) variable

$$\text{Height} = b_0 + b_1 \text{Log}(\text{GDP}) + u$$

(Height is measured in metres )

Rich countries have, on average, taller people and GDP is often measured with error.

Because of this problem, you decide to instrument the variable GDP with its ranking in terms of size of GDP per capita (1 = smallest .... N = largest)

Given the following information:

$$\begin{aligned} N = 100 \quad \text{Cov}(\text{Height}, \text{Log}(\text{GDP})) = 50 \quad \text{Var}(\text{Height}) &= 30 \\ \text{Var}(\text{Log}(\text{GDP})) = 100 \\ \text{Cov}(\text{Height}, \text{RankGDP}) = 100 \quad \text{Cov}(\text{Log}(\text{GDP}), \text{Rank}(\text{GDP})) = 300 \quad \text{Var}(u) &= 100 \end{aligned}$$

- justify the choice of instrument
- work out the OLS estimate of the effect of GDP on height
- work out the IV estimate of the same effect
- find the variances of the OLS and IV estimates
- what would happen to IV estimates of the slope and variance if you found that the correlation coefficient between Rank(GDP) and Log(GDP) was instead 0.3 ?

6. A simple macroeconomic model consists of an aggregate income equation and a consumption function.

$$Y = a_1 C + a_2 I + a_3 G + e \quad (1)$$

$$C = b_0 + b_1 Y + u \quad (2)$$

where Y is income, C is consumption, I investment and G government expenditure

Assuming I and G are exogenous, determine whether each equation is identified, where e and u are disturbance terms. What would happen if you tried to estimate (1) by OLS? Find the reduced form for C. On this basis what instruments could you use for which variable and in which equation?

7. Given data on Profits (P), sales, (S), advertising, (A) and Firm size, (N) a researcher suggests the following model

$$S = a_0 + a_1 A + e \quad (1)$$

$$A = b_0 + b_1 S + b_2 N + b_3 P + u \quad (2)$$

Is either equation identified? Can you use OLS to estimate either equation?

Do your conclusions change if you allow this years sales to depend as well on last year's level of advertising?

$$S = a_0 + a_1A_t + a_2A_{t-1} + e$$

8.

Given the following simultaneous equation system

$$\text{Demand: } P_t = \gamma_1 Q_t + \gamma_2 I_t + u_t \quad (1)$$

$$\text{Supply: } Q_t = a + bP_t + v_t \quad (2)$$

Where P is price, Q is quantity, I is income and  $\gamma_1, \gamma_2, a$  and  $b$  are coefficients to be estimated.

- i) Determine the order condition for identification of each equation
- ii) Say what the order conditions imply about estimating the structural form parameters in either equation (ie can you use IV in either equation to get an unbiased estimates of the coefficients). If so, write down the equation for the IV/2SLS estimator

(Final Exam 2002)

9. Read in the data set *demsupp.dta* from the course web site. This file contains 30 annual observations of the market price, (*price*) and quantity (*output*) of a good along with the average income per head of the population, (*income*), the average wages of employees used to produce the good, (*wages*) and the price of a close substitute for the good, (*pricesub*).

Consider the following demand and supply equations.

$$\text{Demand: } \text{Price}_t = a_0 + a_1 \text{Output}_t + a_2 \text{Pricesub}_t + a_3 \text{Income}_t + e_t \quad (1)$$

$$\text{Supply: } \text{Output}_t = b_0 + b_1 \text{Price}_t + b_2 \text{Wages}_t + u_t \quad (2)$$

Estimate both these equations by OLS.

Why might you suspect the estimates were wrong?

Find the set of instruments for a) Price b) Output. Do OLS regressions of these variables on their instruments. Now use this to do a 2SLS estimates of (1) and (2).

Are your estimates different?

Do the test of over-identifying restrictions (instrumental validity) in (2)

Now do a Wu-Hausman test for each equation to check whether there was significant endogeneity bias in your original OLS estimates of the coefficients in (1) and (2).

