Practice 2

1. Consider a simple model to estimate the effect of personal computer (PC) ownership on college grade point average (GPA) for graduating seniors at a large public university

$$GPA = \beta_0 + \beta_1 PC + u,$$

where PC is a binary variable indicating PC ownership.

   (a) Is it possible that PC is correlated with u? Explain your answer

   (b) Explain why PC is probably related to parent's income. Does this mean the parent's income is an instrumental variable for PC? Explain your answer.

   (c) Suppose that, four years ago, the university gave grants to buy computers to roughly one-half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for PC.

2. The following is a simple model to measure the effect of a school choice program on standardized test performance,

$$score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1,$$

where *score* is the results of the standardized test at national level, *choice* is a binary variable indicating whether the student has attended a school of his/her choice in the last year, and $faminc$ is family's income. The IV for *choice* is *grant*,the dollar amount received by the student to pay for tuition. The grant received varies with family's income. That is why we control for $faminc$ in the equation

   (a) Even if we include $faminc$ in the equation, why would *choice* possibly be correlated with u_1?

   (b) If within each income class, the grant amounts were assigned randomly, would *grant* be correlated with u_1?

   (c) Write the reduced form equation for *choice*. What is needed for *grant* to be partially correlated with *choice*?

   (d) Write the reduced form equation for *score*. Explain why this is useful. [Hint: How do you interpret the coefficient on *grant*?]

3. Suppose you want to test whether girls who attend an all girls' high school do better in maths than girls who attend mixed-sex schools. Given a random sample of different high school girls in the USA, *score* is the score on a standardized math test, and *girlhs* is a dummy variable indicating whether a student attends an all girls' high school

   (a) What other factors would you control for in the equation?

   (b) Write an equation relating *score* with *girlhs* and the other factors you listed in part *(a)*.

(c) Suppose that parental support and motivation are unmeasured factors in the error term in part *(b)*. Are these likely to be correlated with *girlhs*? Explain your answer.

(d) Discuss the assumptions needed for the number of all girls' high schools within a twenty-mile radius of a girl's home to be a valid IV for *girlhs*.

4. The data FERTIL2 includes, for women in Botswana during 1988,information on the number of children (*children*), years of education (*educ*), age (*age*) and variables on economic status.

(a) Estimate the following model by OLS

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

and interpret the results. In particular, for a given age (*age*), what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?

(b) $frsthalf$ is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that $frsthalf$ is uncorrelated with the error term from part *(a)*,show that it is a valid instrument for *educ*.

(c) Estimate the model from part *(a)* using $frsthalf$ as an instrument for *educ*. Compare the estimated effect of *educ* with the OLS estimate from part *(a)* and test the possible enodegeneity of *educ*.

(d) Add the binary variables *electric*, *tv* and *bicycle* to the model and assume that they are exogenous. Estimate the equation by OLS and 2SLS and compare the estimated coefficients on *educ*. Interpret the coefficient of *tv* and explain why television ownership has a negative effect on fertility.

5. Use the data in `CARD.gdt` for this exercise.

(a) Consider the equation

$$\begin{aligned} \log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 \\ &\quad + \beta_4 black + \beta_5 smsa + \beta_6 south + u, \end{aligned}$$

For the IV estimator to be consistent, the instrument for *educ*, *nearc*4 (it is binary variable that indicates whether the individual grew up close to a four-year college) cannot be correlated with u. ¿Could *nearc*4 be correlated with any component or the error term such as unobserved ability? Explain your answer.

(b) For a subsample of the men in the data set, an IQ score is available. Regress IQ on *nearc*4 to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?

(c) Now regress IQ on *nearc*4, *smsa*66 and the regional variables $reg662, \ldots, reg669$. Does the relation between IQ and *nearc*4 still exist once the effect of the regional variables has been taken into account? How does your answer affect the conclusions drawn in part *(b)*?

(d) From parts *(a)* and *(b)*, what do you conclude about the importance of controlling for *smsa*66 and the 1996 regional variables in the equation for $\log(wage)$?

6. Suppose that annual earnings and alcohol consumption are determined by the following system of simultaneous equations:

$$\begin{aligned} \ln(earnings) &= \beta_0 + \beta_1 alcohol + \beta_2 educ + u_1, \\ alcohol &= \gamma_0 + \gamma_1 \ln(earnings) + \gamma_2 educ + \gamma_3 \log(price) + u_2, \end{aligned}$$

where *price* is a local price index for alcohol, which includes state and local taxes. Assume that *educ* (years of education of the individual) and *price* (price of alcohol) are exogenous, and that $\beta_1, \beta_2, \gamma_1, \gamma_2$ and γ_3 are all different from 0, which equation is identified? How would you estimate it?

7. Use the data in `SMOKE.gdt` for this exercise.

   (a) Consider the following model to estimate the effects of smoking on annual *income* (possibly through lost work days due to illness, or productivity effects):

$$\ln(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1$$

   where *cigs* is number of cigarettes smoked per day, on average, *educ* and *age* are the years of education and the age of the individual respectively. What is the interpretation of β_1?

   (b) To reflect the fact that cigarette consumption might be jointly determined with income, consider that the demand for cigarettes equation is

$$\begin{aligned} cigs &= \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 \\ &\quad + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2 \end{aligned}$$

   where *cigpric* is the price of a pack of cigarettes (in cents of a dollar), and *restaurn* is a binary variable equal to unity if the person lives in a state with restaurant smoking restrictions. Assuming these two last variables are exogenous to the individual, what signs do you expect for γ_5 and γ_6?

   (c) Under what assumptions is the income equation from part *(a)* identified?

   (d) Estimate the income equation by OLS and discuss the estimate of β_1.

   (e) Estimate the reduced form for *cigs*, regressing *cigs* on all the exogenous variables. Are $\log(cigpric)$ and *restaurn* significant in the reduced form?

   (f) Now estimate the income equation by 2SLS. Discuss how the estimate of β_1 compares with the OLS estimate.

   (g) Do you think that cigarette prices and restaurant smoking restrictions are exogenous in the income equation?