Introduction to Econometrics - MGT581

Submission on December 18 by 23:59 PM

Please submit your R (.Rmd file), Python, or Stata code and PDF report file. Details on how to transform your R code to a pdf are in the announcements section on Moodle.

Exercise 1

In this exercise, you will work with the **Fertility** dataset which contains data for married women from the 1980 U.S. Census. The dataset contains information on married women aged 21–35 with two or more children. Both Excel and Stata versions of the dataset are available on Moodle (use the one you are more comfortable with).[1]

(a) Regress *weeksm1* only on the indicator variable *morekids*, using OLS. On average, do women with more than two children work less than women with two children? If so, how much?

(b) Explain potential endogeneity issue in the OLS regression estimated in (a).

(c) The data set contains the variable *samesex*, which is equal to 1 if the first two children are of the same sex (boy–boy or girl–girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child? Justify your answer based on appropriate regression model.

(d) Is *samesex* a weak instrument? If so, why? Justify your answer with appropriate statistics.

(e) How can the exclusion restriction condition be violated when using *samesex* as an instrument? Discuss it.

(f) Estimate the IV regression of *weeksm1* on *morekids*, using *samesex* as an instrument. How large is the fertility effect on labor supply?

(g) Do the magnitude and significancy of the IV estimation results (the coefficient for *morekids*) change when you include the variables *agem1*, *black*, *hispan*, and *othrace* in the regression in part f (treating these variables as exogenous)? Explain why or why not.

[1] A detailed description is given in **Fertility_Description**, which is available on Moodle.

Exercise 2

In this exercise, you will work with a simplified version of **SchoolingReturns** dataset from the U.S. National Longitudinal Survey of Young Men (NLSYM) conducted in 1976, with some variables dating back to earlier years. The dataset captures information on individuals' wages, education levels, labor market experience, ethnicity, and other demographic factors (we limited dataset to only necessary variables for the exercise). This dataset has been utilized in studies examining the causal relationship between education and earnings, notably by David Card (1995)[2], who employed geographical proximity to colleges as an instrumental variable to estimate the returns to schooling. Both Excel and Stata versions of the dataset are available on Moodle (use the one you are more comfortable with).

The variable description is as follows:

- *educ*: Education level
- *momeduc*: Mothers' education level
- *dadeduc*: Dads' education level
- *lwage*: Log wage
- *smsa*: Indicator for whether an individual resided in an urban area (as defined by the U.S. Census Bureau) at the time of the survey. The value is one for individuals living in an urban/metropolitan area and zero for individuals living in a non-urban/rural area
- *black_dummy*: Indicator for black individuals
- *south_dummy*: Indicates whether an individual resided in the Southern United States at the time of the survey. We use it as a control to capture regional differences in labor markets, education access, and wage structures
- *nearc4_dummy*: Indicator for whether an individual lived near a four-year college during their youth
- *age*: individual age
- *age_squared*: individual age squared

In all models, specify individuals and their parent education as a continuous variable not a categorical variable.

(a) Run an OLS regression using *lwage* as the dependent variable and *educ*, *black_dummy*, *south_dummy*, *age*, *age_squared*, and *smsa* as independent variables. Interpret the education level significancy and the estimated coefficient.

[2]Card, David. "Estimating the return to schooling: Progress on some persistent econometric problems." Econometrica 69.5 (2001): 1127-1160.

We know that the estimated effect of schooling on future wage can be endogenous. One example is 'ability bias' (see Griliches, 1977). Suppose that some individuals have unobserved characteristics (ability) that enable them to get higher earnings. If these individuals also have above-average schooling levels, this implies a positive correlation between schooling and error term and an OLS estimator that is upward biased. Another reason for endogeneity is the existence of measurement error in the schooling measure which leads to downward bias in the OLS estimator. Therefore, we decide to implement IV estimation to alleviate the potential bias.

(b) Implement TSLS using *nearc4_dummy* as instrument for *educ* (Do not perform this manually. Use the packages!). Put the same exogenous variables as part a. Compare the estimated coefficient and standard error for *educ* of this model with part a. Specifically, is the OLS regression upwardly or downwardly biased? which model suggests higher standard error and why?

(c) Now implement TSLS again, using *momeduc* and *dadeduc* as instruments for *educ* and report the results.

Now we want to compare model in part b and c.

(d) Which model has weaker instrument(s), model b or c? Justify your answer using appropriate statistics?

(e) Compare standard error of *educ* in model b and c. Why are they very different?

(f) Test the overidentifying restriction using Hansen J-test for model c (use the steps discussed in the exercise sessions). Is it significant at the 1% level? What does significancy or insignificancy of the J-statistic mean?

(g) Which model do you prefer? b or c? Explain based on what you have found in previous steps.