INTERMEDIATE ECONOMETRICS M1 - TSE - FALL 2023
HOMEWORK 3: INSTRUMENTAL VARIABLES

Evaluation of R code [1 point]
You get 1 point for your code runs without errors (outputs must align with your answers) and for code readability.

Evaluation of writing [1 point]
You earn 1 point for presenting your answers concisely and clearly. Make sure all answers to the questions are written in a PDF file, **NOT** in your code script.

Note:

- All statistical tests should be conducted at 5% level of significance.
- We assume that **all** regression errors are homoscedastic.
- The dataset is free of measurement errors.

Exercise 1

The aim of this exercise is to explore the economic returns associated with schooling. We will utilize the `CollegeDistance` dataset available in the `AER` package. This dataset originates from a survey of high school graduates and contains variables that capture wages, education level, average tuition, and various socio-economic factors. Additionally, it provides information on the distance from a college at the time these survey participants were in high school. The dataset comprises 4,739 observations. The variables we will focus on in this exercise are:

- *gender* - factor indicating gender (Female, Male)
- *ethnicity* - factor indicating ethnicity (African-American, Hispanic or other)
- *mcollege* - factor. Is the mother a college graduate? (Yes, No)
- *urban* - factor. Is the school in an urban area? (Yes, No)
- *unemp* - county unemployment rate (in %)
- *wage* - state hourly wage (in US dollars)
- *distance* - distance from 4-year college (in 10 miles)
- *education* - number of years of education.

Q0 Attach the `AER` package and load the `CollegeDistance` data.

Before starting your analysis, make sure to execute the following line:

```
CollegeDistance$mcollege <- ifelse(CollegeDistance$mcollege=="yes",1,0)
CollegeDistance$ethnicity <- ifelse(CollegeDistance$ethnicity =="afam", 1, 0)
```

ethnicity is now a binary variable (1 = African-American, 0 = otherwise).

Ordinary least squares approach

Q1 [4 points] Consider the regression model given by

$$\log(wage_i) = \beta_0 + \beta_1 education_i + \beta_2\, unemp_i + \beta_3\, ethnicity_i + \beta_4\, gender_i + \beta_5\, urban_i + \varepsilon_i, \tag{1}$$

where we treat $(unemp, ethnicity, gender, urban)$ as **exogenous** regressors.

1. Estimate the model (1) using the OLS method and report the estimate and standard error for β_1.
2. Conduct a test of significance for β_1. Your answer should include the following: (a) null and alternative hypotheses, (b) the test statistic and its asymptotic distribution under the null, and (c) decision rule to reject/do not reject the null hypothesis. Based on the test result, interpret the impact of *education* on *wage*.
3. Give a reason why *education* in the model (1) may be endogenous. (Max. 3 sentences)

Instrumental Variable approach I

Q2 [9 points] Considering the endogeneity issue, we propose using *distance* as an instrumental variable.

1. Argue why *distance* is a valid instrumental variable. (Max. 3 sentences)
2. State the first stage equation for *education*.
3. Estimate the first stage using the OLS method. Report the OLS estimates and their corresponding standard errors.
4. Using the OLS results from Q2.3, test the significance of the coefficient for *distance* and explain how this observation is connected to the instrument's validity discussed in Q2.1.
5. Estimate β_1 using the simple IV method, and report the estimated value and standard error for β_1. Is the effect of *education* on *wage* statistically different from zero? Based on the significance test result, interpret the effect of *education* on *wage*.
6. (Test for Endogeneity) Conduct a Hausman test to check the endogeneity of *education*. Your answer should include the following: (a) null and alternative hypotheses, (b) the test statistic and its asymptotic distribution under the null, (c) decision rule to reject/do not reject the null hypothesis, and (d) consequences of rejecting/not rejecting the null hypothesis.

Instrumental Variable approach II

Q3 [5 points] In this second approach, we propose using both *distance* **and** *mcollege* as instrumental variables.

1. Estimate β_1 using the 2SLS method, and report the estimated value and standard error for β_1.
2. (Test for Overidentification) Conduct a Sargan test to check the exogeneity of the two instruments. Your answer should include the following: (a) null and alternative hypotheses, (b) the test statistic and its asymptotic distribution under the null, (c) decision rule to reject/do not reject the null hypothesis, and (d) consequences of rejecting/not rejecting the null hypothesis.
3. Suppose *distance* is a valid instrument. Based on your test result on Q3.2, give a reason why *mcollege* may or may not satisfy the exogeneity condition. (Max. 4 sentences)

Exercise 2 (Optional)

This exercise discusses a subject that is not addressed in lectures, so answers should be provided in words only. Exercise 2 offers bonus points but your total score cannot exceed 20.

Q1 [1 point] Explain the weak instruments problem and its effect on the 2SLS estimator. (Max. 5 sentences)

Q2 [1 point] Suppose we have a just-identified model with a single endogenous variable and no exogenous regressors. How can we test if the instrument is weak? (Max. 5 sentences)