Assignment 3

ECN 620 - Applied Economic Analysis – Fall 2024

Instructions:

1. The assignment is worth 10% of the final grade.
2. Assignment is due on **December 2nd at 5 PM**
3. Late submissions will not be accepted.
4. Students may work in groups of up to 4 individuals.
5. **But each student must write and submit their own do_file**.
6. Group members' names must be listed in the do_file; failure to do so will result in a score of zero.
7. There are two exercises in this assignment. The first exercise worth 60 marks, and exercise 2 is worth 40 marks
8. Within each exercise, all sub-questions are equally weighted.
9. You are required to enter your code and responses in the do-file provided on D2L Brightspace.
10. Remember to save your do-file using the following convention: **A3_YourStudentNumber_YourName**
11. For questions where applicable, you will receive full marks if your code runs successfully and produces the correct answer. If the answer is correct but the code does not work, you will receive half the points.

Exercise 1: Panel data models (each question is worth 10 points)

There is a large literature on the rentier states theory and the resource curse. The theory states that rentier capitalism can be a curse on the systemic level. States that extract rents from easily lootable resources instead of taxing their people develop institutions that become unresponsive to their citizens and provide less public goods. In this exercise, you will use panel data models to test this theory. This exercise makes use of the dataset **Panel_data.csv** which contains a balanced panel of data from 59 countries worldwide for the years 1996 through 2010. A detailed description of the main variables is provided in the table below.

Name	Description
country	Country name
countrycode	3 letter country abbreviation
year	
aid	Net aid flow (in per cent of GDP)
oil	Oil rents (in per cent of GDP)
gdpcapita	GDP per capita in constant 2000 US dollars
institutions	World governance indicator index for quality of institutions
polity2	Polity IV project index
population	Population size
mortality	Rate (per 1000 live births)

Load the dataset on Stata and answer the following questions.

1. Data cleaning: The variables *oil*, *gdpcapita*, *aid*, *polity2*, and *mortality* are coded as factor variables, but they should be numeric.

    a. Convert these variables to numeric variables. (**Hint**: Note that for some variables, missing values are coded as ".." or "NA", first change the value of the missing into "." and use the Stata command destring to convert in numeric)

    b. Log-transform variables gdp per capita and population size

2. This question has three parts:

    a. Estimate (i) a regression of quality of institutions against oil rents and (ii) a regression of quality of institutions against oil rents, aid flow, log of gdp per capita, polity2, log of population and mortality

    b. Interpret the coefficient on oil rents in regression (ii). Does adding control variables in regression (ii) change the estimated effect of oil rents as measured by statistical significance?

    c. Suggest a variable that varies across countries but plausibly varies little—or not at all—over time and that could cause omitted variable bias in regression (ii).

3. Repeat the analysis in regression (ii) but add country fixed effects, using the **dummy variable approach**. Do the results change when you add country fixed effects (compare to regression (ii))? If so, which set of regression results is more credible, and why?

4. Add year fixed effects to the regression in question 3, again using the dummy variables approach [**Note:** Do not use robust or cluster standard errors.] Do the results change when you add time fixed effects? If so, which set of regression results is more credible, and why?

5. Replicate the analysis in Question 4 using Stata's **xtreg** command with standard errors clustered at the country level. Do the results change when using clustered standard error (compare to question 4)? Why?

6. Run a **first difference** regression of the change in the quality of institutions between 1996 and 2010 on the change in oil rents between 1996 and 2010 (**Hint**: You should first convert your data from wide to long format as we did in the empirical exercise in module 8).

Exercise 2: Instrumental Variables (each question is worth 10 points)

In this exercise you will use IV to explore how the number of children affects the labor supply of parents. This exercise is based on Josh Angrist and Bill Evans' paper called "*Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size*.". This exercise uses the **instrumental_variable.dta** which is a cross-section data from the 1980 US Census on married women aged 21–35 with two or more children. The main variables are described in the table below. In this exercise, you should always use robust standard errors in your regression analysis unless otherwise specified

Name	Description
morekids	=1 if mother had more than 2 children
gender1	=1 if first child was a boy
gender2	=1 if second child was a boy
age	Age of mom at the time of census
black	=1 if mother is black
hispanic	=1 if mother is hispanic
other	=1 if mother is not black, Hispanic or white
work	Number of weeks in which the mother worked in 1979
instrument1	=1 if the first two children are of the same sex

1. Load the instrumental_variable.dta dataset

    a. Run an OLS regression to estimate the effect of having more than two children on weeks worked by mothers, controlling for if the first and second children were boys (gender 1 and gender 2), indicators for race/ethnicity of the mother.

b. Interpret the coefficient of having more than two children.

c. Do you think that this multivariate analysis provides an unbiased estimate of the causal effect of having more than two children on mother's labor supply? Explain. Provide an example of omitted variables that is likely to cause bias in the results

To avoid the endogeneity issue of total number of children, Angrist and Evans instrument for whether families have three or more children using whether or not the first two children were of the same sex (instrument1). The hypothesis is that both children being of the same sex is a relevant predictor of having one or more additional children because parents are often thought to have preferences for child gender or gender mix. If parents want a girl and they have only had boys, they may be more likely to have an additional child. To the extent that these preferences do exist, then having the first two children from the same sex should be a relevant predictor of having 3 or more children. The argument for exogeneity of the instrument is that birth gender itself is already as good as random, hence having two children of the same sex should also be as good as random. Since birth gender is random, only a couple of possibilities would seem to pose a potential problem to instrument exogeneity. You will now use this instrument to estimate the 2SLS effect of having more children on weeks worked.

2. First, generate the 2SLS estimates manually using the multivariate specification including controls from question 2. (**Hint**: That is, estimate the first stage regression and generate predicted values of having more than two children. Then use the predicted values in the second stage regression to estimate the effect of having more than two children on weeks worked by mothers.)

3. Generate the 2SLS estimate again, but this time using the **ivreg2** command. How do the estimates change between OLS and IV? And the standard errors?

4. Evaluate the instrument relevance and weakness using the ivreg2 output. Can the test of overidentifying restrictions tell us anything about the instrument exogeneity in this case?