Group Assignment 2:

DiD, IV and panel data analysis

The second group assignment is about applying your new skills on difference-in-difference, instrumental variable and panel data analysis. You will put these skills to use to examine several highly relevant questions: can financial incentives foster mothers to return to the labor market (DiD analyses); does compulsory schooling length influence your wages (IV analyses); and how to explain variation in time to export between countries and over time. The assignment consists of three main exercises with several sub-exercises. All parts of the assignment can be made with the material taught in the lectures and the tutorials of this and the previous week. Much of the tutorial code can be re-used. This group assignment is due on **Monday October 5, 2020, 8PM**.

Before starting the assignment it is suggested to make a designated folder with subfolders: *Data* for all data sets, *Programs* for all computer code, and *Results* for all results, e.g., figures and tables. This will enhance transparency of the work process and help collaboration within the group.

Submit your individual answers as a pdf-file on Canvas. The manuscript has the R-script in the appendix (in a not too large font). Alternatively, you may opt to write your assignment using R Markdown (or R Sweave), which integrates the code and your answers. The use of Latex for this submission, as stipulated in the course manual, is strongly encouraged (see *this* or *this* link for helpful resources)

1 Difference-in-Difference Analysis: Female Labor Force Participation

Introduction

Increasing female participation in the labor force is a goal of high importance to modern societies in order to promote equality and to stimulate the economy. In 1993, the US Government has installed an ‘Earned Income Tax Credit’ (EITC) in an attempt to stimulate female labor force participation. The EITC is a refundable tax credit for low-income workers, which applied to women with children. It basically serves as a financial incentive to foster female labor supply.

Table 1: Description of variables in *DiD_dataset.csv*

Variable	Variable Description
state	US State of Residence;
year	Year;
urate	State Unemployment Rate;
children	Number of children;
nonwhite	Indicator ethnicity (1 = Hispanic/Black, 0 = otherwise);
finc	Indicator annual family income;
earn	Annual Earnings;
age	Age of Woman;
ed	Years of Education;
work	Indicator wok status (1 = Employed, 0 = Otherwise);
unearn	Unearned Income.

Task

Download the *DiD_dataset.csv* from Canvas. You are tasked with estimating the effects of the 1993 expansion of this policy on labor supply for single women by whether or not they had children. You will be analyzing this question using a data set on women aged 20-54 with less than high-school education covering the years 1991 to 1996 in the US. Table 1 gives a description of the variables in *DID_dataset.csv*. Conduct the following analysis and report the results.

1. The effects of the 'Earned Income Tax Credit' (EITC) policy introduced in 1993 can be evaluated using multiple dependent variables. These are annual earnings (*earn*), annual family income (*finc*) and working/non-working (*work*). For each of these three variables, present a suitable plot to present visual evidence of the DiD effect of the EITC introduction. [About 0.5-1 page]

2. Use `stargazer` to make a table with summary statistics of the data, present a concise description of the data. [About 0.5 page]

3. What is is the difference-in-difference effect of EITC introduction if you evaluate the summary results in a matrix as discussed in the lecture (Session 3/slide 29) and the tutorial? Provide such a matrix for the three mentioned dependent variables. [About 0.5 page]

4. Conduct regression analyses of the DiD effect for the three dependent variables. Present these results in a proper table; and explain and interpret your findings. What is the effect of the policy introduction on the dependent variables? How does the effect change when adding control variables? Also, elaborate on whether robust standard errors seem necessary. [About 0.5-1 page]

Table 2: Description of selected variables in *IV_dataset.csv*

Variable	Variable Description
age	Age (in Years);
educ	Education (in Years);
lnwage	Log Weekly Earnings;
married	Marital status (1 = Married, 0 = Otherwise);
qob	Quarter of Birth;
SMSA	Indicator living situation (1 = Residents live in an urban Area, 0 = Otherwise);
yob	Year of Birth.

2 Instrumental Variable Analysis: Effect of Compulsory Schooling on Wages

Introduction

Education is the backbone of society. The quality and quantity of education in modern societies is on a steady rise. Yet, it is difficult to measure if and how much education contributes to future earnings on the labor market, because many unobservable factors exist, which bias OLS regression of wages on years of education. To circumvent these biases, scholars have come up with nifty instrumental variables techniques to tease out the causal effect of education on wages. One such technique makes use of a combination of two facts: The minimum legal school dropout age and the annual quarter of birth of a person. All the students born in the same year are admitted to school in the same cohort (i.e., the same class). However, a student born in January reaches the legal school dropout age earlier than a student born in September, for instance.[1] In essence, the idea is to randomize school exposure to students, assuming that in each year, a constant fraction of students drops out of school and this dropout pattern is unrelated to when a student is born.

Task

Download the *IV_dataset.csv* from Canvas. You are tasked with estimating the effects of the years of education on the (log) wages. You will be analyzing this question using a data set on people in the United States born between 1930 and 1939. Focus on the variables from *IV_dataset.csv*, described in table 2. Please conduct the following analysis and report the results.

1. Suppose that OLS regression of wages on education is performed in order to determine the gains of extra years of eduction. Give two concrete examples of conditions that could bias the estimated education effect. For

[1] The minimum legal school dropout age differences by state and year. In some it is 16 years, in some 17, and in others even 18. You do not need to consider this in your analyses.

example, students preference for education may influence how long they stay in school and how much they earn on the labor market (because people who are open to learning arguably make good employees), which thus biases the OLS estimator of the education effect.[2] [About 0.5 page]

2. Use stargazer to present summary statistics of the data and give a concise description of the data. [About 0.5 page]

3. An important question to address is if the quarter of birth (*qob*) would be a good instrument for years of education. Explore if variable *qob* meets the relevance criterion by means of evidence in the form of regressions and plots. [About 0.5-1 page]

4. Conduct IV regression analysis of the effect of education on log wages, using quarter of birth as the instrument. Present these results in proper tables and explain and interpret your findings. What is the effect of an additional year of education on wages? How does the effect change when adding control variables? Examine if the use of robust standard errors critically affects your statistical inferences. [About 0.5-1 page]

5. Following the previous analysis, conduct the same regressions with OLS as with IV. Choose and apply a formal test to decide between the two OLS- and IV-estimated models. Examine if over-identification can be an issue in the IV analysis. Hint: you may want to check Session 3/slides 67 and following. [About 0.5-1 page]

6. Elaborate on possible causes of concern that could render your instrumental variable identification strategy invalid. An arbitrarily made-up example would be: “Birthdays in the US have been recorded with a varying delay between 1 and 6 months. Hence, the measurement of school years for dropout students could be inaccurate.” [About 0.5 page]

3 Panel data modeling: time to export

Introduction

The time to export, a measure of border compliance, is an important indicator of the quality of trade conditions of countries. The longer the time to export the larger the burden and costs on behalf of exporting firms, the larger the risks of obsolescence and deteriorated goods. This purpose of the concluding task is to estimate the effects of potential determinants of variation in time to export between countries and over time. All analyses are based on data provided by the World Bank.

[2] Just to be clear: this illustrative example does not count towards the two desired concrete examples.

Task

The following specific tasks need to be addressed.

- Data from the World Bank's open data portal can be dowloaded with the functionality of package *wbstats*. Install and load the package. Use function `wb_indicators()` to obtain a table with available variables. The column name of the dependent variable 'time to export' is *IC.EXP.TMBC*. Decide about *at most four* other variables that hypothetically influence variation in the time to export. Download the data. Assign meaningful column names to the dependent and independent variables. Present the population model that reflects the assumption. [About 0.5 page]

- Construct a balanced data set of the downloaded data by implementing the following steps: (*i*) remove from the downloaded data all records with missing values of the dependent variable 'time to export'; (*ii*) select all records that have been completely observed during the period 2014-2019. Use function `stargazer` to make a table with summary statistics of the dependent and independent variables of the model, and give a concise description of the insights obtained from this table. [About 0.5-1 page]

- Use function `ggplot` to make a scatter plot of the dependent variable against an explanatory variable of choice. Present the plot, and describe the main insights obtained from the plot. [About 0.5-1 page]

- Estimate the previously formulated model based on pooled regression, between regression, fixed effect regression, and random effect regression. Use function `stargazer` to make a table of the combined results. Concisely discuss the main insights obtained from the table. [About 0.75-1.25 page]

- Use appropriate statistical tests to decide which of the estimated models is the preferred specification for the analysis of 'time to export'. [About 0.5 page]