BM01BAM - Advanced Statistics & Programming

Individual Assignment 2:

Female Labor Force Participation and the Effects of Schooling on Wages

The second individual assignment is about applying your newly acquired knowledge and skills to difference-in-difference and instrumental variable analysis. You will put these skills to use to examine highly relevant questions, such as: can financial incentives foster mothers to return to the labor market (DiD analyses), and does compulsory schooling length influence wages (IV analyses)? The assignment consists of two 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. The assignment is due on **Friday September 30, 2022, 11:59AM**. [Expected length of your submission is around 6 to 9 pages, all inclusive.]

Before starting the assignment, make a designated folder with sub-folders: *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, adds to developing hygienic coding skill, and helps us to help you in the case of questions.

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 explained in the course manual, is strongly encouraged (see *this* or *this* link for helpful resources). A decent layout is part of the assessment, as well as a proper motivation of your claims and findings.

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.

1 Difference-in-Difference analysis: female labor force participation

In 1993, the US Government 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, and basically served as a financial incentive to foster female labor supply.

Task

Download the *DiD_dataset.csv* from Canvas. You are tasked with estimating the effects of the 1993 policy intervention 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*. Use these data to elaborate the following questions.

1. Consider the canonical difference-in-difference equation as presented in the lecture:

$$y_{it} = \beta_0 + \beta_1 D_i + \beta_2 T_t + \beta_3 D_i T_t + \epsilon_{it} \tag{1}$$

   Indicate which of the coefficients(s) from equation (1) yield the following outcomes:

   - $E = (y_{T=1}|D = 1)$
   - $E = (y_{T=0}|D = 1)$
   - $E = (y_{T=1}|D = 0)$
   - $E = (y_{T=0}|D = 0)$

Insert the coefficients into the following equation and simplify the term as much as possible in order to obtain a condensed (formal) expression of the difference-in-difference effect:

$$[E(y_{T=1}|D=1) - E(y_{T=0}|D=1)] - [E(y_{T=1}|D=0) - E(y_{T=0}|D=0)]$$

[About 0.5 page]

2. The effects of the 'Earned Income Tax Credit' (EITC) policy intervention 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 provide visual evidence of the DiD effect of the EITC introduction. Use function `ggplot` to make the plots. [About 0.5-1 page]

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

4. 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 55) and the tutorial? Provide such matrices for the three mentioned dependent variables. [About 0.5 page]

5. Analyze the DiD effect with appropriate regression models 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]

6. Further examine the presence of effect heterogeneity. Specifically, analyze if high-education mothers react differently to the EITC policy measure than low education mothers. Define low education as less than nine years of education. Explore the following effects of the EITC measure:

   - Single mothers with high education and with children. Compare them with single mothers with low education and with children.
   - Single women with low education, without children. Compare them with single women with low education, with children.

   [About 0.5-1 pages]

2 Instrumental Variable analysis: effect of compulsory schooling on wages

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

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.

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 characteristics: (*i*) the minimum legal school dropout age and (*ii*) the annual quarter of birth of a person. All 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. 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 education. Give two concrete examples of conditions that could bias the estimated education effect. For 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]

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

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

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 decent 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 102 and following. [About 0.5-1 page]

6. Elaborate on possible causes of concern that could render your instrumental variable identification strategy invalid. Base your argumentation on the required assumptions of instrumental variables as taught in the lecture. 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]