PC project 5: Instrumental Variables

You are expected to solve this PC-Project until and including Exercise 2b prior to 28.03.2022. Submit the PDF with your answers as well as your Python files with a reproducible code in pc5.py format. All functions should be submitted as pc5_functions.py. Upload your solution in a zip file named pc5_yournames.zip to the designated module in Canvas. Please, provide all 3 files, namely the pc5.py, pc5_functions.py and pc5.pdf. Both the codes and the written answers are graded.

General information

Angrist and Evans (1998) investigate the causal effect of fertility on labor supply of women. They exploit parental preferences for a heterogeneous sibling-sex composition to construct instrumental variables estimates of the effect of childbearing on labor-market outcomes. The file *data_pc5.csv* is a reduced and modified sample of US women in 1980 and 1990 based on data of their study. The dataset includes the following variables:

Variable name	**Description**
weeks_work	Number of weeks worked per year
employed	Dummy: 1 if employed at least 1 week per year
kidcount	Number of kids ever born (2-12)
morekids	Dummy: 1 if woman has more than two kids
multi2nd	Dummy: 1 if second birth was a multiple birth (twins, …)
samesex	Dummy: 1 if first two children are of the same sex
black	Dummy: 1 if mother is African-American
hisp	Dummy: 1 if mother is Hispanic
age_mother	Age of mother in years (21-35)
hsgrad	Dummy: 1 if mother graduated from a high-school
colgrad	Dummy: 1 if mother graduated from a college

Exercise 1: Descriptive statistics

Angrist and Evans (1998) investigate the labour market effects of having more than two kids instead of having exactly two. Therefore, observations with women having one or no child have been excluded from the dataset. Download the dataset *data_pc5.csv* and save it in a folder on your computer. Download also the *pc5.py* and *pc5_functions.py* and save them in the same folder on your computer. You may re-use the functions from the previous PC projects and include them in *pc5_functions.py*.

a. Open the *pc5.py* in Spyder. Specify the path variable accordingly and make sure that *pc5_functions* is imported. Load the *data_pc5.csv* into your environment as Pandas DataFrame object.

b. Code new, adjust or re-use the summary statistics function and report the descriptives. Additionally, plot the distribution of *kidcount* and *weeks_work*. Comment on the distribution of these variables. Are there any specific patterns? Does the data appear to be plausible?

c. Code a new function to print the mean of *employed* for each number of kids (*kidcount*) together with the corresponding sample size and comment on this table. Are these numbers indicative of potential effects? Why could they be misleading?

d. Code a function to generate a cross table to investigate the relation of *morekids* and *multi2nd*. Comment if the obtained results are plausible.

Exercise 2: Homogeneous effects

a. Consider the following model:

$$y_i = \beta_0 + \beta_1 d_i + \gamma x_i + \zeta_i, \quad E[\zeta_i | x_i] = 0$$

Estimate $\hat{\beta}_1$ by OLS, where y_i denotes *weeks_work*, d_i stands for *morekids*, and the regressors x_i include *age, black, hisp, hsgrad,* and *colgrad*. What is the estimated average effect of having more than two kids on weeks worked per year? Is it plausible that the estimated coefficient $\hat{\beta}_1$ describes a causal relationship? Discuss the underlying assumptions and their validity.

b. Now use the model equations:

$$y_i = \beta_0 + \beta_1 d_i + \gamma x_i + \zeta_i, E[\zeta_i | x_i, z_i] = 0$$
$$d_i = \delta_0 + z_i'\theta + \delta x_i + v_i, E[v_i | x_i, z_i] = 0$$

The instrumental variable z_i refers to *multi2nd*, which is an instrument for the endogenous variable d_i. Aside from z_i, all variables remain as defined in Exercise 2a. Compute the 2SLS estimator by plugging the fitted OLS values of the first stage $\hat{d}_i$into the second stage. Code your own function for the 2SLS estimation. Report the estimation results and interpret the causal effect. Discuss the two major assumptions of an instrument and their validity in this example.

c. Apply again the model in Exercise 2b and use *samesex* as the instrument for *morekids* to estimate the first stage once without and once with covariates x_i. Interpret the coefficients. Is *samesex* a strong instrument? What could be a potential explanation for your conclusion?

Exercise 3: Allowing for heterogeneous effects[1]

a. Code your own function to calculate the Wald-estimator without regressors. Use the function to estimate the local average treatment effect (LATE) for both instruments, namely *multi2nd* and *samesex* and estimate also the size of the complier population for each instrument. Compare the estimated LATEs and discuss what could cause potential differences.

b. Comment on the policy relevance of the estimated effects.

References

Angrist, J., & Evans, W. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *The American Economic Review*, 88(3), 450-477.

Angrist, J. D., & Pischke, J. S. (2009*). Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

[1] For more details on this issue compare Angrist & Pischke (2009), chapter 4.4.