PC project 2: Selection-on-observables

You are expected to solve this PC-Project until and including Exercise 2b until 7.3.2022. You may complete the remaining exercises during the session. Submit the PDF with your answers as well as your Python files with a reproducible code in pc2.py format. All functions should be submitted as pc2_functions.py. Upload your solution in a zip file named pc2_yournames.zip to the designated module in Canvas.

General information

Low birth weight is associated with negative labour market and educational outcomes during adult life (see for example Almond, Chay, and Lee, 2005). Therefore, the average effect of cigarette smoking during pregnancy on the child's birth weight is explored in the economics and epidemiology literature for example by Abrevaya (2006), da Veiga and Wilder (2008) and Walker, Tekin, and Wallace (2009) who find significantly negative effects. You get access to a random sample of the dataset. A detailed description of the variables can be found in the following table.

Variable name	**Description**
bweight	Infant birth weight (grams)
mhisp	Dummy: 1 if mother Hispanic
alcohol	Dummy: 1 if alcohol consumed during pregnancy
deadkids	Dummy: 1 if newborn died in previous birth
mage	Mother's age in years
medu	Mother's education attainment in years
nprenatal	# of prenatal care visits
monthslb	Months since last birth
order	Birth order of infant
msmoke	Intensity of cigarettes smoked during pregnancy (0-3 scale)
mbsmoke	Dummy: 1 if mother smoked during pregnancy
mrace	Dummy: 1 if mother is white
prenatal	Trimester of first prenatal care visit (possible values 1, 2, 3)

Part 1: Data preparation

Download the dataset *data_pc2.csv* and save it in a folder on your computer. Download also *pc2.py* and *pc2_functions.py* and save them in the same folder on your computer. You may re-use functions from *pc1_functions.py* and include them in *pc2_functions.py*.

a. Open *pc2.py* in Spyder. Specify the path variable accordingly and make sure that *pc2_functions* is imported. Load the *data_pc2.csv* into your environment as Pandas DataFrame object.

b. Code new, adjust or re-use the summary statistics and histogram functions developed in the PC project 1 to inspect the dataset. Can you detect any anomalies or missing values? Comment on the distribution of the outcome variable *bweight* and on the treatment shares based on the *mbsmoke* variable.

c. Drop the variables with multiple treatments, i.e., the variable *msmoke* from the dataset. Also, drop the variables that have too many missing values, i.e., the *monthslb* variable, from the dataset. Furthermore, delete the observations that have missing values in the variables *order* and *prenatal* (use the Pandas module for this).

d. Recode the variables *order* and *prenatal* into dummies indicating if the birth order of the infant was the first one and if the first prenatal care visit was in the first trimester, respectively, as such variable coding has been previously used by the studies mentioned above.

e. Print again the descriptive statistics of the cleaned dataset. Are there any missings left? Are the variable values suitable for a further econometric analysis? Save your final dataset as *data_pc2_clean.csv* (use the Pandas module for this).

f. Code new, adjust or re-use the *balance_check()* function from the PC project 1 and interpret the results. Examine the balance between treated and controls both for the covariates as well as the outcome. What do you think are the major channels for selection in this sample?

Part 2: Effect estimation

Continue further with your script *pc2.py* and the function file *pc2_functions.py*.

a. Code new, adjust or re-use your own OLS procedure from the PC project 1 to estimate ATE of smoking on the birthweight without covariates. Discuss the obtained results and the underlying identifying assumptions. Can you trust the estimated effect? Discuss how could you design an experiment that would remove confounding and thus allow you to identify a causal effect without conditioning on covariates.

b. Estimate ATE of smoking on birthweight by OLS with covariates and interpret the estimated effect. Which specification do you use and why? Should you include all covariates? Discuss the assumptions required to identify the ATE as well as their validity. Describe at least two unobserved covariates that could threaten your identification strategy.

c. Now you want to estimate the ATE using Inverse Probability Weighting (IPW). Predict the propensity score using a logit model. You can use the *.Logit()* function from the *statsmodels.api* module for this purpose. Write your own function that estimates the ATE via IPW and the standard error using the bootstrap. Code your own bootstrap function which draws subsamples with replacement of the size of the original sample. The number of replications should be specified as an option for the function input. Finally, estimate the ATE via IPW using 1000 bootstrap replications for the standard error and compare the results to the ones from 2b. Discuss the differences between OLS and IPW. Are there any notable differences (theoretically and empirically)?

d. Code your own function that draws the propensity score histograms for *D=0* and *D=1* in one plot. Include an optional argument for enforcing common support by dropping observations with no overlap. Graphically inspect the common support. Are there regions with no overlap?

e. Estimate the ATE on the sample with common support and compare the resulting IPW point estimates with your previous result. Is common support an issue here?

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. Journal of Applied Econometrics 21 (4), 489-519.

Almond, D., Chay, K., & Lee, D. (2005). The costs of low birth weight. The Quarterly Journal of Economics 120 (3), 1031-1083.

da Veiga, P. V. & Wilder, R. P. (2008). Maternal smoking during pregnancy and birthweight: a propensity score matching approach. Maternal and Child Health Journal 12 (2), 194-203.

Walker, M., Tekin, E., & Wallace, S. (2009). Teen smoking and birth outcomes. Southern Economic Journal 75 (3), 892-907.