Exercise 1: instrumental variables

Let y_i be a measure of good health, and D_i an indicator for smoking ($D_i = 1$ if person i smokes, $D_i = 0$ otherwise). We want to estimate the impact of somking on health.

For this means, you collect data on y_i and D_i from randomly selected individuals in Strangetown. To your astonishment you find that y_i and D_i are positively correlated in your dataset. However, you also notice that both cigarettes and health services are very expensive in Strangetown, because of high taxes on cigarettes and a shortage of medical personal. You therefore suspect that D_i is endogenous, because only people with high income can afford medical services and smoking (usually low income individuals have a higher propensity to smoke, but Strangetown is a little strange).

The true model reads:

$$y_i = \alpha + \beta_1^{true} D_i + \beta_2 w_i + \epsilon_i \tag{1}$$

with w_i denoting individual income. But since you do not observe income you can only estimate the following model:

$$y_i = \alpha + \beta_1 D_i + u_i \tag{2}$$

(a) Do you overestimate or underestimate β_1^{true} if you cannot observe income? Give a verbal explanation in a first step. In a second step, derive the inconsistency of $\hat{\beta}_1^{OLS}$ that you obtain if you estimate model (2) by OLS. Show the direction of the bias. How do researchers usually call this type of bias?

Since you want to obtain an unbiased estimate of β_1 you intensify your investigations in Strangetown. You figure out that a couple of years ago a big tobacco company randomly selected individuals as part of a publicity campain providing each selected individual with 100 packs of cigarettes for free. More precisely, you observe an indicator z_i of whether individual i was selected ($z_i = 1$) or not ($z_i = 0$).

(b) Serves z_i as an instrument? Discuss the assumptions under which $\hat{\beta}^{IV}$ is consistent.

Your data collection takes some time, but after a few weeks you have surveyed a total sample of 70 individuals. You observe $n_{00} = 30$ individuals with $D_i = z_i = 0$; the average health outcome y_i in that group is observed to be $\bar{y}_{00} = 1.0$. You also observe $n_{01} = 10$ individuals with $D_i = 0$ and $z_i = 1$; the average outcome in that group is $\bar{y}_{01} = 0.8$. We observe $n_{10} = 20$ individuals with $D_i = 1$ and $z_i = 0$; the average outcome in that group is $\bar{y}_{10} = 1.5$. Finally, we observe $n_{11} = 10$ individuals with $D_i = z_i = 1$; the average outcome in that group is $\bar{y}_{11} = 1.2$.

(c) Calculate β_1^{OLS} that you obtain by estimating equation (2) by OLS. Interpret the coefficient.

(d) Calculate β_1^{IV} and discuss your result wrt. to the previous findings.

(e) Assume that $Var(u_i|z_i) = 1/10$. Calculate the estimated standard error of the IV estimator.

(f) Test the hypothesis H_0: $\beta_1^{IV} = 0$. Do you reject the hypothesis when employing a large sample t-test with size 5%?

(g) Under heterogeneous treatment effects we need further assumptions to interpret β^{IV} as the local average treatment effect (LATE). State the LATE assumptions and discuss them. Explain how we can interpret the LATE. Refer the LATE assumptions and the LATE interpretation to the present setting of Strangetown.

(h) Given that the LATE assumptions hold, calculate the share of always taker, never taker, and compliers in the Strangetown sample.

(i) Given that the LATE assumptions hold, we identify an internally valid estimate for a specific subgroup. Considering the Strangetown example and treatment effect heterogeneity, discuss the external validity of the LATE.

(j) Can we estimate the average treatment effect on the treated (ATT) in the present setting? Out of which groups does the ATT consist?

(k) What effect does the LATE correspond to in case of homogenous treatment effects?

(l) Which of the LATE assumptions is/are sufficient for a causal interpretation of the reduced form? Which effect does the reduced form identify?

(m) Explain why the present setting can be interpreted as an „encouragement design“.

Exercise 2: applied exercise

This question is based on the paper from Bonjour, Cherkas, Haskel, Hawkes, Spector („Returns to education: Evidence from UK twins“, The American Economic Review, 2003), which in the following we will refer to as BCHHS. Start by reading the paper.

The dataset of BCHHS is available on the website of the american economic association (`http://www.aeaweb.org/articles.php?doi=10.1257/000282803322655554`), i.e. it is exactly the dataset that BCHHS have submitted to the journal, together with their paper. It contains the following variables: family (family number), twinno (twin number within family: 1 or 2), earning (hourly wage), highqua (estimated years of schooling, we will also refer to this as education), twihigh (twin's estimated years of schooling), age (age), and some other variables that are not relevant for us here. Open the dataset in stata (e.g. command: „use filename, clear“), look at the dataset (e.g. command: „br“) and familiarize yourself with the variables listed above. Generate the variable log-earnings („gen lnearn=ln(earning)“) and generate the variable age-squared („gen agesq=age*age“).

(a) Use the dataset to reproduce the results in columns (2) and (3) of table 2 in BCHHS. Do you find any discrepancy between your results and the results reported in BCHHS? Is the discrepancy in the regression coefficients serious?

(b) The main coefficient of interest is the coefficient on education (= years of schooling). Explain the interpretation of this coefficient.

(c) Which two concerns do the the authors mention with regard to education ("highqua")? Dou you have further concerns in mind?

(d) Discuss the proposed instrument critically?

(e) Which of the previously mentioned concerns is addressed by the instrument?

(f) How were the standard errors calculated in column (2) of table 2 in BCHHS? Estimate heteroscedasticity robust standard errors for this regression (using the "robust" option in the command "reg"). What do you find?

Next, reshape the dataset such that the unit of observation becomes the family (i.e. a twin pair) instead of a twin (command: „reshape wide earning-agesq, i(family) j(twinno)“). This reduces the number of observations by a factor of two. Furthermore, generate new variables that contain the difference of log earnings between the twins („gen dlnearn = lnearn1-lnearn2“), the difference of years of schooling between the twins („gen dhigh = highqua1-highqua2“), and the difference of twin's estimated years of schooling between the twins („gen dtwihi=twihigh1-twihigh2“).

(g) Now, reproduce the results in columns (4) and (5) of table 2 in BCHHS. Do you find any discrepancy between your results and the results reported in BCHHS?

(h) Comparing the estimations you performed previously, what is the advantage of taking differences between the observations of identical twins before estimating returns to schooling? Does this address any of the endogeneity problems?