

Test Exercise 4

Notes:

- See website for how to submit your answers and how feedback is organized

Goals and skills being used:

- Practice with identifying causes of endogeneity
- Practice with identifying valid instruments
- Obtain insight in the logic behind the 2SLS estimator

Questions

To run a study on the effect of a new diet a researcher runs a survey. The three most important questions in this survey are: (1) What was your weight one year ago? (2) What is your current weight? (3) Did you follow this diet in the past year?

We denote the answers of individual $i = 1, \dots, n$ to the first two questions with y_{i0} and y_{i1} , the answer to the third question is denoted by d_i (where $d_i = 1$ if the diet was followed). Furthermore, some background characteristics of the respondents are collected. These characteristics are combined in the vector x_i . Assume that all respondents are perfectly able to correctly answer the questions in the survey and do so truthfully.

(a) First of all the researcher uses OLS to estimate the parameters of the model

$$y_{i1} - y_{i0} = \alpha + \beta d_i + \gamma y_{i0} + x_i' \delta + \varepsilon_i.$$

The OLS estimator for β is possibly not consistent as the variable d_i may be endogenous. Clearly explain why this may be the case. Indicate whether your reason would lead OLS to overestimate or underestimate the true effect of the diet.

The researcher finds out that in some regions of the country the diet was promoted via door-to-door advertising. The researcher manages to construct a variable z_i that indicates whether individual i does ($z_i = 1$) or does not ($z_i = 0$) live in a region in which the diet was advertised.

- (b) In general there are two important conditions for variables Z to be useful as instruments. In formal terms these conditions are $\frac{1}{n} Z' \varepsilon \rightarrow 0$ and $\frac{1}{n} Z' X \rightarrow Q \neq 0$ as the sample size n grows large. Rephrase these two conditions in words in the context of this application for the above mentioned advertising variable (no formulas!).
- (c) For both assumptions in (b), indicate whether it can be tested statistically given the available variables. If yes, indicate how. If no, why not?
- (d) Suppose that z_i satisfies the conditions in (b) and suppose that z_i is uncorrelated with y_{i0} and x_i . In this case the 2SLS-estimator for β in the model $y_{i1} - y_{i0} = \alpha + \beta d_i + \eta_i$ is consistent when a constant and z_i are used as instruments.

Show that we can write this 2SLS estimator for β in terms of simple sample averages. You can use the following averages:

- Average weight change over all individuals: Δ
- Average weight change over individuals with $z_i = 1$: Δ^1

- Average weight change over individuals with $z_i = 0$: Δ^0
- Proportion of people taking the diet: \bar{d}
- Proportion of people with $z_i = 1$ taking the diet: \bar{d}^1
- Proportion of people with $z_i = 0$ taking the diet: \bar{d}^0

To further explain the notation, for example:

$$\bar{d}^1 = \frac{1}{\sum_{i=1}^n z_i} \sum_{i=1}^n z_i d_i$$

Hint: start with the formula: $(Z'X)^{-1}Z'y$.