

# TEST4

2023-04-25

## 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)

- 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  $\beta$  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.

## Answer (a)

The OLS is bias toward zero and the true effect of  $\beta$  on  $y_i$  is underestimated.

The OLS estimator for  $\beta$  is possibly not consistent as the variable  $d_i$  may be endogenous (where  $d_i = 1$  if the diet was followed). However, OLS requires exogenous regressors. Presumably, there is a correlation between  $d_i$  and the vector  $x_i$ , which contains background characteristics of the respondents. It is likely that those characteristics have an impact on the probability to follow this diet in the past year. If one or more of the regressors is endogenous, then OLS is no longer consistent and the conventional results (t-test, F-tests, diagnostic tests in previous sections of this chapter) are no longer valid.

### (b)

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!).

## Answer (b)

First of all, in a situation where one or more of the regressors are endogenous random variation in  $X$  are correlated with the random variation  $\varepsilon$  in  $y$ .

$$plim(\frac{1}{n}X'\varepsilon) \neq 0$$

The condition ( $\frac{1}{n}Z'\varepsilon \rightarrow 0$ ) means that the instruments (i.e., variables  $Z$ ) should be exogenous. This is satisfied (under weak additional conditions) when the instruments are uncorrelated with the disturbances in the sense that

$$E[z_i \varepsilon_i] = 0$$

( $i = 1, \dots, n.$ )

The condition ( $\frac{1}{n}Z'X \rightarrow Q \neq 0$ ) means that the instruments should be sufficiently correlated with the regressors. This is called the rank condition. As  $Q_{zx}$  is an  $m \times k$  matrix, this requires that  $m \geq k$  —that is, the number of instruments should be at least as large as the number of regressors.

### (c)

- 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?

## Answer (c)

Of course, it can be tested statistically with the available variables. Test condition conditions are  $\frac{1}{n}Z'\varepsilon \rightarrow 0$

$$plim(\frac{1}{n}Z'\varepsilon) = 0$$

Test condition :  $\frac{1}{n}Z'X \rightarrow Q \neq 0$

$$plim(\frac{1}{n}Z'X) = Q_{zx}$$

$$rank(Q_{zx}) = k$$

$$plim(\frac{1}{n}Z'Z) = Q_{zz}$$

$$rank(Q_{zz}) = m$$

### (d)

- 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  $\beta$  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  $\beta$  in terms of simple sample averages. You can use the following averages:

- Average weight change over all individuals:  $\Delta$
- Average weight change over individuals with  $z_i = 1$  :  $\Delta^1$
- Average weight change over individuals with  $z_i = 0$  :  $\Delta^0$
- 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:  $(Z'X)^{-1}Z'y$ .

## Answer (d)

before,

$$E[z_i \varepsilon_i] = 0$$

( $i = 1, \dots, n.$ )

This corresponds to  $m$  moment conditions. The IV estimator is defined as the GMM estimator corresponding to these moment conditions. In the exactly identified case ( $m = k$ ), the IV estimator  $b_{IV}$  is given by the solution of the  $m$  equations

$$\frac{1}{n} \sum_{i=1}^n z_i (y_i - x_i' b_{IV}) = 0$$
$$b_{IV} = (\sum_{i=1}^n z_i x_i')^{-1} \sum_{i=1}^n z_i y_i = (Z'X)^{-1}Z'y$$