Econometrics: Methods and Applications

Erasmus
University
Rotterdam

Peer-graded Assignment: Test Exercise 4

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, . . ., 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.

Answer:
The OLS estimator for β is possibly not consistent as the variable d_i may be endogenous (where di = 1 if the diet was followed). However, OLS requires exogenous regressors. Presumably, there is a correlation between d_i and the vector xi, 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consistent and the conventional results (t-test, F-tests, diagnostic tests in previous sections of this chapter) are no longer valid (Textbook p.398).

The OLS is bias toward zero and the true effect of β on y_i is underestimated.

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 $\frac{1}{n}Z'\varepsilon \to 0$ and $\frac{1}{n}Z'X \to Q \neq 0$ 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:
In a situation where one or more of the regressors are endogenous random variation in X are correlated with the random variation ε in y.

$$\text{plim}\left(\frac{1}{n}X'\varepsilon\right) \neq 0.$$

The condition ($\frac{1}{n}Z'\varepsilon \to 0$) 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

$$E[z_i\varepsilon_i] = 0, \qquad i = 1, \cdots, n.$$

The condition ($\frac{1}{n}Z'X \to Q \neq 0$) means that the instruments should be sufficiently correlated with the regressors. This is called the rank condition. As Q_{zx} is an m × k matrix, this requires that m ≥ k —that is, the number of instruments should be at least as large as the number of regressors. (Textbook p.398)

(b) For both assumptions in (b), indicate whether it can be tested statistically given the available variables. If yes, indicate how. If no, why not?

Yes, it can be tested statistically with the available variables.

Test condition **conditions are** $\frac{1}{n}Z'\varepsilon \to 0$

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

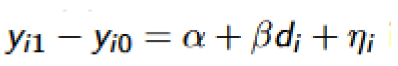
Test condition $\frac{1}{n}Z'X \to Q \neq 0$:

$$\operatorname{plim}\left(\frac{1}{n}Z'X\right) = Q_{zx}, \qquad \operatorname{rank}(Q_{zx}) = k,$$

$$\operatorname{plim}\left(\frac{1}{n}Z'Z\right) = Q_{zz}, \qquad \operatorname{rank}(Q_{zz}) = m.$$

Source: Testbook p.398

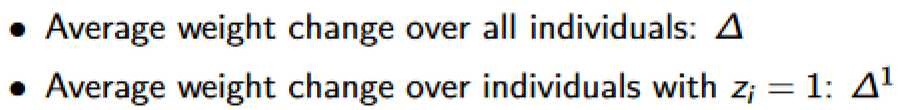
(c) Suppose that z_i satisfies the conditions in (b) and suppose that z_i is uncorrelated with y_{i0} and x_i .
In this case the 2SLS-estimator for β in the model



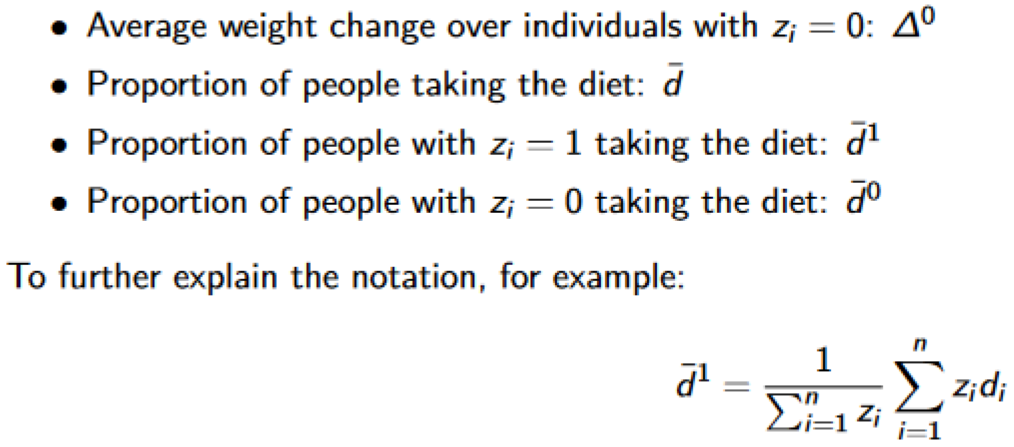
$$y_{i1} - y_{i0} = \alpha + \beta d_i + \eta_i$$

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:



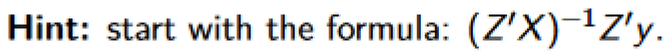
- Average weight change over all individuals: Δ
- Average weight change over individuals with $z_i = 1$: Δ^1



- 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: $(Z'X)^{-1}Z'y$.

Answer:

We assume

$$E[z_i\varepsilon_i] = 0, \qquad i = 1, \cdots, n.$$

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

$\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i' b_{IV}) = 0$—that is,

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

Source: Textbook p.399