MOOC Econometrics: Test Exercise #4

Greg Faletto

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 anwers of individual $i = 1, \ldots, n$ 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.

Problem (a)

First of all the researcher uses OLS to estimate the parameters of the model

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

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.

Solution

Individuals decide for themselves whether they will go on the diet. Whether or not an individual followed this diet might be correlated with how health-conscious the individual is in general, which would also be correlated with how much weight they lost. I.e., if an individual is more health-conscious in general, they may have followed this diet but also engaged in other activities that would help their weight loss (like exercise).

My reason would lead OLS to *overestimate* the true effect of the diet. The causal benefits of the other activities these individuals would be engaging in (e.g. exercise) would be "absorbed" into the coefficient for whether they followed the diet. Including a variable like this in an OLS model would likely decrease the absolute value of the OLS coefficient for whether they followed the diet.

Problem (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.

In general there are two important conditions for variables Z to be useful as instruments. In formal terms these conditions are $\frac{1}{n}Z'\epsilon \to 0$ 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!).

Problem (b) continued on next page. . .

Solution

$\frac{1}{n}Z'\epsilon \to 0$: this means that Z and ϵ are not correlated. In the context of this application, it means that whether or not an individual lives in a region where the diet was promoted does not explain any of the variation in the weight loss data that is left unexplained by a 2SLS model using a constant, an instrument for d_i based on Z, y_i, and x_i (ϵ). In other words, advertisers did not select the regions where the ads were placed based on the average height or weight of the people who live in the region. We can verify this assumption using a Sargan test.

$\frac{1}{n}Z'X \to Q \neq 0$: This means that X and Z are sufficiently correlated. In the context of this experiment, it means that whether or not an individual lives in a region where the diet was promoted is correlated at least somewhat with whether they followed the diet (the stronger the correlation, the better). Individuals who live in regions where the diet was advertised should follow the diet at a higher rate than individuals elsewhere.

Problem (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?

Solution

We would test the first condition ($\frac{1}{n}Z'\epsilon \to 0$) by using a Sargan test. In this case that would mean the following procedure:

- Setup
  - Model: $y = X\beta + \epsilon$
  - Explanatory variables: $X = d_i$ (potentially endogenous), a constant, y_{i0}, x_i (exogenous)
  - Instruments: $Z = z_i$, a constant, y_{i0}, x_i
  - Null hypothesis (H_0): Correlation of Z and ϵ equals 0.
  - Test Procedure: Rewrite H_0: $\delta = 0$ in $\epsilon = Z\delta + \xi$. ϵ cannot be observed, so estimate ϵ using 2SLS.
- Use z_i (whether or no the individual lives in a region where the diet was advertised) to obtain the 2SLS estimator b_{2SLS} for β.
  - Regress d_i on Z (including z_i, a constant, y_{i0}, and x_i) to get what part of d_i is explained by Z. (i.e., in the model $d_i = Z\gamma + \eta$, so the part of d_i that is explained by Z is $Z\gamma = \hat{d}$.
  - Obtain b_{2SLS} by regressing y on $\hat{d}$.
- Calculate $e_{2SLS} = y - \hat{d}b_{2SLS}$.
- Regress e_{2SLS} on Z.
- Use a significance test: $nR^2 \approx \chi^2(m-k)$

Unfortunately, **this test cannot be completed for z_i**. One reason is because **we don't have enough instruments for a Sargan test**. We would need "too many" instruments in order to carry out the Sargan test (the number of instruments m would have to be greater than the number of explanatory variables k in

Problem (c) continued on next page. . .

order to calculate the value of a χ^2 statistic with $m-k$ degrees of freedom).

Another issue arises in the step of regressing d_i on Z to obtain b_{2SLS}. d_i **is a categorical (dummy) variable, so it doesn't make sense to use a linear regression to predict** d_i. Specifically, this would violate assumption A1; clearly, the data generating process for d_i is not $d_i = \alpha + \beta x_i + \epsilon_i$ since d_i is not continuous.

We would test the second condition ($\frac{1}{n}Z'X \to Q \neq 0$) by checking the significance of instruments in the first stage regression of 2SLS (Step 1 above). Again, this regression is not possible to complete because it violates assumption A1, since d_i is a categorical variable. Therefore, **the second condition can also not be tested statistically given the available variables.**

Problem (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:

1. Average weight change over all individuals: Δ
2. Average weight change over individuals with $z_i = 1$: Δ^1
3. Average weight change over individuals with $z_i = 0$: Δ^0
4. Proportion of people taking the diet: $\bar{d}$
5. Proportion of people with $z_i = 1$ taking the diet: $\bar{d}^1$
6. 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: $(Z'X)^{-1}Z'y$.

Problem (d) continued on next page...

Solution

$$\Delta = \frac{1}{n}\sum_{i=1}^{n} y_{i1} - y_{i0}$$

$$\Delta^1 = \frac{1}{\sum_{i=1}^{n} z_i}\sum_{i=1}^{n} z_i(y_{i1} - y_{i0})$$

$$\Delta^0 = \frac{1}{\sum_{i=1}^{n} 1 - z_i}\sum_{i=1}^{n}(1 - z_i)(y_{i1} - y_{i0})$$

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

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

$$\bar{d^0} = \frac{1}{\sum_{i=1}^{n}(1 - z_i)}\sum_{i=1}^{n}(1 - z_i)d_i$$

$$Z = \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & d_1 \\ 1 & d_2 \\ \vdots & \vdots \\ 1 & d_n \end{bmatrix}$$

$$y = \begin{bmatrix} y_{11} - y_{10} \\ y_{21} - y_{20} \\ \vdots \\ y_{n1} - y_{n0} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = (Z'X)^{-1}Z'y = \begin{bmatrix} n & \sum_{i=1}^{n} d_i \\ \sum_{i=1}^{n} z_i & \sum_{i=1}^{n} z_i d_i \end{bmatrix}^{-1} Z'y$$

$$= \begin{bmatrix} n & \sum_{i=1}^{n} d_i \\ \sum_{i=1}^{n} z_i & \sum_{i=1}^{n} z_i d_i \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_{i1} - y_{i0} \\ \sum_{i=1}^{n} z_i(y_{i1} - y_{i0}) \end{bmatrix}$$

$$\begin{bmatrix} n & \sum_{i=1}^{n} d_i \\ \sum_{i=1}^{n} z_i & \sum_{i=1}^{n} z_i d_i \end{bmatrix}^{-1} = \frac{1}{n\sum_{i=1}^{n} z_i d_i - \sum_{i=1}^{n} d_i \sum_{i=1}^{n} z_i} \begin{bmatrix} \sum_{i=1}^{n} z_i d_i & -\sum_{i=1}^{n} d_i \\ -\sum_{i=1}^{n} z_i & n \end{bmatrix}$$

$$\implies \begin{bmatrix} n & \sum_{i=1}^n d_i \\ \sum_{i=1}^n z_i & \sum_{i=1}^n z_i d_i \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_{i1} - y_{i0} \\ \sum_{i=1}^n z_i(y_{i1} - y_{i0}) \end{bmatrix} =$$

$$\frac{1}{n\sum_{i=1}^n z_i d_i - \sum_{i=1}^n d_i \sum_{i=1}^n z_i} \begin{bmatrix} \sum_{i=1}^n z_i d_i & -\sum_{i=1}^n d_i \\ -\sum_{i=1}^n z_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_{i1} - y_{i0} \\ \sum_{i=1}^n z_i(y_{i1} - y_{i0}) \end{bmatrix}$$

The first row of this matrix is $\hat{\alpha}$ and the second row of this matrix is $\hat{\beta}$. Hence we are only interested in the second row.

$$\begin{aligned} \hat{\beta} &= \frac{1}{n\sum_{i=1}^n z_i d_i - \sum_{i=1}^n d_i \sum_{i=1}^n z_i} \sum_{i=1}^n -z_i \sum_{i=1}^n (y_{i1} - y_{i0}) + n \sum_{i=1}^n z_i(y_{i1} - y_{i0}) \\ &= \frac{n\sum_{i=1}^n z_i(y_{i1} - y_{i0}) - \sum_{i=1}^n z_i \sum_{i=1}^n (y_{i1} - y_{i0})}{n\sum_{i=1}^n z_i d_i - \sum_{i=1}^n d_i \sum_{i=1}^n z_i} \end{aligned}$$

Dividing the numerator and denominator by $n\sum_{i=1}^n z_i$ yields

$$\frac{\frac{1}{\sum z_i}\sum_{i=1}^n z_i(y_{i1} - y_{i0}) - \frac{1}{n}\sum_{i=1}^n (y_{i1} - y_{i0})}{\frac{1}{\sum z_i}\sum_{i=1}^n z_i d_i - \frac{1}{n}\sum_{i=1}^n d_i}$$

$$\boxed{= (\Delta^1 - \Delta)/(\bar{d}^1 - \bar{d})}$$