Instrumental Variables: Two-Stage Least Squares (2SLS) – The basics **[DATE]**

The impact evaluation model is:

$$Y_{ij} = \alpha_0 + \alpha_1 X1_{ij} + \alpha_2 P_{ij} + \varepsilon 1_{ij} \qquad (1)$$

The problem is that P_{ij} is correlated with the error term $\varepsilon 1_{ij}$, that is, $corr\left(P_{ij}, \varepsilon 1_{ij}\right) \neq 0$

If we use simple OLS in equation (1), we will get a biased and inconsistent program impact estimate, that is, $\hat{\alpha}_2$ will be biased and inconsistent.

This situation occurs when there are unobserved factors influencing both program participation (P_{ij}) and the outcome of interest (Y_{ij}). In this case, we say that P_{ij} is endogenous.

Solution: Instrumental variables - The 2SLS procedure

Stage 1:

- Specify a model for P_{ij}: $\qquad P_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2 Z_{ij} + \varepsilon 2_{ij} \qquad (2)$

Where Z_{ij} has the following characteristics:

- It affects P_{ij}
- It does not affect Y_{ij} directly, only through P_{ij} (Z_{ij} is not in equation (1))
- It is not affected by other factors, it is exogenous ($corr\left(Z_{ij}, \varepsilon 2_{ij}\right) = 0$).

Z_{ij} is called the instrumental variable.

- Then, run equation (2) using OLS, obtain the estimated coefficients, and generate predicted program participation:

$$\hat{P}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 X1_{ij} + \hat{\beta}_2 Z_{ij}$$

Notice that $\hat{P}_{ij}$ is not influenced by the error term, $\varepsilon 2_{ij}$, that is, it is not influenced by the unobservables that are the source of the endogeneity.

Stage 2:

- Remember that: $\quad P_{ij} = \hat{P}_{ij} + \varepsilon 2_{ij} \qquad (3)$
- Substitute P_{ij} , using (3) in the main equation (1):

$$\begin{aligned} Y_{ij} &= \alpha_0 + \alpha_1 X1_{ij} + \alpha_2 P_{ij} + \varepsilon 1_{ij} \qquad (1) \\ Y_{ij} &= \alpha_0 + \alpha_1 X1_{ij} + \alpha_2 (\hat{P}_{ij} + \varepsilon 2_{ij}) + \varepsilon 1_{ij} \\ Y_{ij} &= \alpha_0 + \alpha_1 X1_{ij} + \alpha_2 \hat{P}_{ij} + (\alpha_2 \varepsilon 2_{ij} + \varepsilon 1_{ij}) \end{aligned}$$

$$Y_{ij} = \alpha_0 + \alpha_1 X1_{ij} + \alpha_2 \hat{P}_{ij} + \varepsilon 1^{*}_{ij} \qquad (1')$$

Note: The steps above are equivalent to just substituting P_{ij} by its predicted value $\hat{P}_{ij}$ in the main equation (1).

- Estimate the modified equation (1’) using OLS
- The estimated parameter $\hat{\alpha}_2$ is a consistent estimate of program impact
- You need to correct the standard errors, because the substitution of P_{ij} by $\hat{P}_{ij}$. Stata has the command `ivregress` which implements 2SLS with corrected standard errors.

Cases when to use 2SLS

You can use 2SLS in the following cases:

1) Y_{ij} and P_{ij} are both continuous variables
2) Y_{ij} is discrete, and P_{ij} is continuous
3) Y_{ij} is continuous, and P_{ij} is discrete, but it is better to use maximum likelihood estimation (MLE) procedures. The Stata command `treatreg` can do that easily.

If Y_{ij} and $P_{ij}$are both discrete, you should not use 2SLS, you should use a MLE method that estimates both equations simultaneously. You can use `biprobit` or `mvprobit` commands in Stata.

Exogeneity test

The intuition of the test is the following: Since endogeneity is created by unobservables that influence both participation and the outcome, one solution would be to measure those unobservables and include them in the model. In most cases we cannot measure those unobservables, but what if we approximate them? Remember that those unobservables are also in $\varepsilon 2_{ij}$, the error term of equation (2). You can obtain the predicted residuals from equation (2), and then include them in equation (1), as an additional variable; if they are statistically significant, it means that there are unobservables influencing program participation and also the outcome, and you have endogeneity.

The test:

1) Run equation (2), and obtain the predicted residuals: $\widehat{\varepsilon 2}_{ij}$
   Remember that $\widehat{\varepsilon 2}_{ij} = P_{ij} - \hat{P}_{ij}$

2) Include the predicted residuals into equation (1) as an additional variable,

$$Y_{ij} = \alpha_0 + \alpha_1 X1_{ij} + \alpha_2 P_{ij} + \delta\,\widehat{\varepsilon 2}_{ij} + \varepsilon 1_{ij}$$

If the estimated parameter of the residuals, $\hat{\delta}$, is significant, there is endogeneity

If the estimated parameter of the residuals, $\hat{\delta}$, is not significant, there is no endogeneity.