MICROECONOMETRICS
CLASS 5

Marek Giergiczny
Wiktor Budziński

ENDOGENEITY

In OLS we estimate equation $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$

One of the crucial assumptions is that independent variables are not related to the error term: $E\left(\varepsilon \middle| \mathbf{X}\right) = 0$

- If this condition does not hold then we will say that our covariates are endogenous

It can be caused by different issues:

- Missing variables
- Simultaneity
- Measurement errors

ENDOGENEITY

The basic solution involves finding so called instrumental variables to 'filter out' the correlated error terms

- IVs have to be correlated with an endogenous variable, and have to be independent from the error term:

$$E\left(\varepsilon \middle| \mathbf{Z}\right) = 0$$

$$E\left(X^{e} \middle| \mathbf{Z}\right) \neq 0$$

Usual estimator is called Two stage least square method (2SLS)

2SLS

The name of the method stems from the fact that it can be estimated in two steps

- Although usually is not in modern statistical packages

In the first step independent variables are explained by instrumental variables, and then fitted values from these regressions are used: $\mathbf{X} = \mathbf{ZB}$

In the second step these fitted values are used instead of actual covariates:

- Because of that the correlated error terms are 'filtered out'

$$\boldsymbol{\beta}_{2SLS} = \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{Y}$$

There is needed at least one additional instrumental variable for each endogenous variable

EXERCISE 1: 2SLS

1. Read *me.twins.rds* data and built a regression model in which logarithm of income is explained by indicator of whether women has more than two kids
2. Fertility decision is likely to be endogenous in the model (e.g. because of simultaneity). Estimate 2SLS model with *twins* variable as an instrument
3. Compare the results with model in which *samesex* variable is used as an instrument

ENDOGENEITY DIAGNOSTICS

Hausman test can be used to test for endogeneity

- Intuitively, it tests whether OLS and 2SLS estimates differ

Formally test statistic is given by:

$$H = \left(\boldsymbol{\beta}_{2SLS} - \boldsymbol{\beta}_{OLS}\right)' \left(\mathbf{V}_{2SLS} - \mathbf{V}_{OLS}\right)^{-1} \left(\boldsymbol{\beta}_{2SLS} - \boldsymbol{\beta}_{OLS}\right)$$

H0 is that there is no endogeneity, then both 2SLS and OLS estimates will be consistent. If H0 does not hold, then OLS will no longer be consistent

ENDOGENEITY DIAGNOSTICS

Wu-Hausman is another popular test

It is based on the model: $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e + \rho v_i + \varepsilon_i$

- V is assumed to be a missing covariate which causes X_i^e to be endogenous

If there is no endogeneity (H0), then $\rho = 0$

To test it, we first estimate auxillary regression: $X_i^e = \mathbf{Z}_\mathbf{i}\boldsymbol{\theta} + \mathbf{X}_i\boldsymbol{\beta} + u_i$

We then use residuals from it as approximation of v: $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e + \rho \hat{u}_i + \varepsilon_i$

- We can then test significance of rho

ENDOGENEITY DIAGNOSTICS

Sargan test for overidentification can be used if there is more ivs than one for each endogenous covariates

- It checks whether residuals from the regression are correlated with instruments

If H0 is rejected then the instruments are not correct, and one should use different method than 2SLS

- Could be also caused by model misspecification: heteroscedasticity or wrong functional form

EXERCISE 2: 2SLS

1. Conduct Hausman and Wu-Hausman tests for endogeneity
2. Estimate 2SLS with two instrumental variables to use Sargan test

WEAK IVS

2SLS is consistent but biased in a finite sample

- More instrumental variables may actually increase the bias, especially in small samples

When instrumental variables are weakly correlated with endogenous variable then we call them '*weak instruments*'

- Weak instruments may cause larger bias than endogeneity (cure worse than disease)

WEAK IVS

Less formal ways of testing whether IV's are weak involve

- Checking their correlation with endogenous regressor
- Checking their joint significance in first-step regression

Conditional F-test can be used to judge the strength of instruments

- Sanderson, E., & Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of econometrics*, 190(2), 212-221.
- This F-statistics arise in the denominator of bias formula for IV estimates

WEAK IVS

There is really no easy solution if IVs are weak

- It is recommended to at least report corrected confidence intervals for endogenous regressors
- Anderson-Rubin confidence sets are often employed

Some researchers recommend using Limited Information Maximum Likelihood estimates

- There is also Fuller's modification of LIML
- This study:
  - Hahn, J., & Hausman, J. (2003). Weak instruments: Diagnosis and cures in empirical econometrics. American Economic Review, 93(2), 118-125.
- recommends using Fuller's modification - they find that LIML does not work that well with weak ivs

EXERCISE 3: WEAK IVS

1. Check whether employed instrumental variables are weak
2. Calculate Anderson-Rubin confidence intervals
3. Compare estimates from 2SLS with LIML and Fuller's version of LIML
4. Use simulation to see how these estimates differ when IVs are weak

WORKBOOK 5

Now try to conduct a similar analysis for the exercises in Workbook5b.R

- Exercise 1

PROXIES

If endogeneity is caused by missing variable then researcher could try to substitute for it with some proxy variable (q is unobserved in equation below)

$$y = X\beta + \alpha q + \varepsilon$$

Proxy variable *z* should fulfil two conditions:

- It should be "*redundant*": $E(y \mid X, q, z) = E(y \mid X, q)$
- It should be a predictor of missing variable $q = \beta z + \eta$, such that endogenous variable is not correlated with eta

Example: we can substitute for '*ability*' variable with results of IQ score

MULTIPLE INDICATORS SOLUTION

Often we have variables which only fulfill the first condition for being a proxy

- Examples include some attitudinal questions in the survey
- They are usually considered to be a functions of the "*original*" variable

$$z_1 = \alpha_1 q + \eta_1$$
$$z_2 = \alpha_2 q + \eta_2$$

Using one of such variables would led to biased estimates, but if we have at least two of them, then we could use one of them as an instrument for the other

- The crucial assumption is that etas (measurement errors) have to be uncorrelated

Basically we estimate 2SLS but treating indicator variable as endogenous

- Results may differ depending which indicator is an instrument and which is in regression

SEM

The same objective as with MIS could be obtained by using Structural Equation Models

SEM is a combination of multivariate regression analysis with factor analysis

It consists of so called measurement equations:

$$\begin{cases} y_1 = X\beta_1 + U\alpha_1 + \varepsilon_1 \\ \vdots \\ y_n = X\beta_n + U\alpha_n + \varepsilon_n \end{cases}$$

And structural equations:

- U are some latent (unobserved factors)

$$\begin{cases} U_1 = X\gamma_1 + \eta_1 \\ \vdots \\ U_m = X\gamma_m + \eta_m \end{cases}$$

SEM

In our case we have 3 measurement equations:

- Unobserved variable (*u*) measures the missing covariate

$$y = X\beta + u\lambda + \varepsilon$$

$$z_1 = \alpha_1 u + \eta_1$$

$$z_2 = \alpha_2 u + \eta_2$$

We could add structural equation to account for dependence between *u* and X

SEM is more flexible than MIS, although probably harder to estimate

- More assumptions about distributions of different variables
- Can use more indicator variables

EXERCISE 4: MIS

1. Read *nls80.xls* into R and estimate wage equation
2. Try to use IQ as proxy variable for *ability*
3. Use MIS treating IQ and Knowledge of the world of work score (*KWW*) as indicators of ability
4. Estimate SEM with one latent factor

LATENT IVS

There are few methods which do not require neither instrumental variables nor proxies/indicators

Such methods usually rely on some distributional assumptions to identify the model

- As these methods do not use any additional information they can lead to high standard errors

LATENT IVS

One of such method is Latent Instrumental Variable method

- Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, 3(4), 365-392.

It assumes a following model, with two error terms being correlated

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e + \varepsilon_i$$

$$X_i^e = \pi_i + \eta_i$$

- π_i is LIV which is assumed to be a categorical variable with *m > 1* categories
- One needs to estimate *m* values this variable takes in each category, and *m-1* probabilities of belonging to each category

$$P\left(\pi_i = v_j\right) = p_j$$

- It is basically a mixture model.
- Assumption about π_i being discrete allow us to estimate covariance of error terms

COPULA CORRECTION

Another method uses copulas to correct for endogeneity

- Park, S., and Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4), 567-586.

This method is based on the assumption that endogenous variable is not normally distributed

It is estimated in two steps

- First non-parametric estimation of X_i^e distribution is conducted: $\hat{H}\left(x\right)$
- Second the copula function is estimated such that $C\left(H\left(X_i^e\right), G\left(\varepsilon_i\right)\right)$
- Epsilon is assumed to be normally distributed and C() is assumed to be a Gaussian copula

COPULA CORRECTION

Basically $P^* = \Phi^{-1}\left(H\left(X_i^e\right)\right)$ and epsilon will both have normal distribution, and therefore a maximum likelihood function can be specified assuming that they come from a bivariate normal distribution

Alternatively, the model can be thought as in the form $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \alpha X_i^e + \gamma P^* + \omega_i$

- P* is used to filter out the correlation, it works as an instrument
- If X_i^e is close to normal distribution, then $P^* \approx X_i^e$, and the model will suffer from collinearity

EXERCISE 5: LIV AND COPULAS

1. Use simulated data to test LIV estimation in R
2. Use simulated data to test copulas correction estimation in R