08 Endogenous Right-Hand-Side Variables

Andrius Buteikis, andrius.buteikis@mif.vu.lt
http://web.vu.lt/mif/a.buteikis/

Introduction

Consider a simple regression model:

$$Y_t = \alpha + \beta X_t + u_t$$

- Under the classical Gauss-Markov conditions, the OLS estimators of its coefficients are BLUE (Best Linear Unbiased Estimator).
- However, if it happens that X_t is correlated with u_t, then the OLS estimator become biased, inconsistent and inefficient.

This situation often arises when one or more of the explanatory variables *is jointly determined* with the dependent variable, typically through an equilibrium mechanism (this is called a simultaneous equations model).

The leading method for estimating simultaneous equations models is the method of **instrumental variables (IV)** and we start its exposition in a *one equation case*.

One Equation

Consider a simple model written as:

$$Y_t = \alpha + \beta X_t + u_t,$$

where we think that X_t and u_t are correlated: $cov(X_t, u_t) \neq 0$ (thus, X_t is an *endogenous variable*). In order to obtain consistent estimators of α and β, suppose that we have an observable variable Z_t that satisfies two assumptions:

1. Z_t is uncorrelated with u_t, that is, $cov(Z_t, u_t) = 0$ (we say that Z is *exogenous variable*).
2. Z_t is correlated with X_t, that is, $cov(Z_t, X_t) \neq 0$ (we call Z_t an *instrumental variable* for X_t, or sometimes simply an *instrument* for X_t).

Recall that under the classical assumptions the usual

$$\widehat{\beta}_{LS} = \frac{\widehat{cov}(X_t, Y_t)}{\widehat{var}(X_t)}$$

is the solution of the following two moments equations:

$$\begin{cases} \sum(Y_t - (\alpha + \beta X_t)) = 0 & \sim \quad \mathbb{E}\epsilon = 0 \\ \sum(Y_t - (\alpha + \beta X_t))X_t = 0 & \sim \mathrm{Cov}(\epsilon, X) = 0 \end{cases}$$

Now, as the second equation fails, we replace X in it by the instrument Z and obtain the consistent instrumental variable estimator:

$$\widehat{\beta}_{IV} = \frac{\widehat{cov}(Z_t, Y_t)}{\widehat{cov}(Z_t, X_t)}$$

Note on Instrument selection

- Quite often the lag X_{t-1} serves as a good instrument to X_t
- X_{t-1} will be a "good" instrument if the correlation between X_{t-1} and X_t is sufficiently high or, what is almost the same, the coefficient δ_1 in the regression:

$$X_t = \delta_0 + \delta_1 X_{t-1}$$

is significant.

In what follows, we shall use Y to denote endogenous variables and Z exogenous. Thus our previous model can be rewritten as:

$$Y_{1,t} = \alpha + \beta Y_{2,t} + u_t$$

we also assume that we know Z_t.

The model can be generalized to (1 endo. & 1 exo var.):

$$Y_{1,t} = \alpha + \beta_1 Y_{2,t} + \beta_2 Z_{1,t} + u_t$$

In order to use the IV method, we need another exogenous variable, call it $Z_{2,t}$, that does not appear in our equation.

The last model can be further generalized to (1 endo. & k-1 exo. var.):

$$Y_{1,t} = \alpha + \beta_1 Y_{2,t} + \beta_2 Z_{1,t} + ... + \beta_k Z_{k-1,t} + u_t,$$

where we again assume that we have an instrument to $Y_{2,t}$, say $Z_{k,t}$.

Now, to get the IV estimators of β's, one has to solve the system:

$$\begin{cases} \sum(Y_t - (\alpha + \beta_1 Y_{2,t} + \beta_2 Z_{1,t} + ... + \beta_k Z_{k-1,t})) = 0 \\ \sum(Y_t - (\alpha + \beta_1 Y_{2,t} + \beta_2 Z_{1,t} + ... + \beta_k Z_{k-1,t}))Z_{1,t} = 0 \\ ... \\ \sum(Y_t - (\alpha + \beta_1 Y_{2,t} + \beta_2 Z_{1,t} + ... + \beta_k Z_{k-1,t}))Z_{k,t} = 0 \end{cases}$$

What if we have **two** instruments for $Y_{2,t}$: $Z_{k,t}$ and $Z_{k+1,t}$? We can get two IV estimators of β's, and neither of these would, in general, be efficient.

To find the best IV, we choose the *linear combination* of *all* exogenous variables that are best correlated with $Y_{2,t}$. This turns our to be given by:

$$\widehat{Y}_{2,t} = \widehat{\pi}_0 + \widehat{\pi}_1 Z_{1,t} + ... + \widehat{\pi}_{k-1} Z_{k-1,t} + \widehat{\pi}_k Z_{k,t} + \widehat{\pi}_{k+1} Z_{k+1,t},$$

where $\widehat{\pi}$'s are the OLS estimates in respective model.

Two-Stage Least Squares Estimation

We can use $\widehat{Y}_{2,t}$ as an instrument to $Y_{2,t}$ or, **alternatively**, apply the following **two stage least squares** (2SLS) procedure:

1. Obtain the above mentioned estimator $\widehat{Y}_2$.
2. Replace Y_2 with $\widehat{Y}_2$ in:

$$Y_{1,t} = \alpha + \beta_1 Y_{2,t} + \beta_2 Z_{1,t} + ... + \beta_k Z_{k-1,t} + u_t$$

   and once again apply OLS.

The 2SLS estimator is less efficient than OLS when the explanatory variables are exogenous. Therefore, it is useful to have a test for endogeneity of an explanatory variable that shows whether 2SLS is necessary.

It is common to use **the Hausman test** to test for exogeneity.

Note that 2SLS can also be used in models with more than one endogenous explanatory variable. For example, consider the model:

$$Y_{1,t} = \alpha + \beta_1 Y_{2,t} + \beta_2 Y_{3,t} + \beta_3 Z_{1,t} + \beta_4 Z_{2,t} + \beta_5 Z_{3,t} + u_t$$

To estimate β's we need at least two more exogenous variables $Z_{4,t}$ and $Z_{5,t}$ that do not appear in this equation but that are correlated with $Y_{2,t}$ and $Y_{3,t}$.

1. On the first stage, we apply OLS and estimate:

$$\widehat{Y}_{2,t} = \widehat{\pi}_0^{(2)} + \widehat{\pi}_1^{(2)} Z_{1,t} + ... + \widehat{\pi}_5^{(2)} Z_{5,t}$$

   and

$$\widehat{Y}_{3,t} = \widehat{\pi}_0^{(3)} + \widehat{\pi}_1^{(3)} Z_{1,t} + ... + \widehat{\pi}_5^{(3)} Z_{5,t}$$

2. On the second stage, replace $Y_{2,t}$ and $Y_{3,t}$ with, respectively, $\widehat{Y}_{2,t}$ and $\widehat{Y}_{3,t}$ and estimate α and β's with OLS.

Endogeneity test

Lets say, we want to estimate the following model:

$$Y_{1,t} = \alpha + \beta_1 Y_{2,t} + \beta_2 Z_t + \epsilon_t$$

where: $Y_{1,t}$ and $Y_{2,t}$ are endogenous variables; Z_t - exogenous variable; I_t - instrumental variables, which are not included in the regression.

To perform the **Housman test**:

1. Regress $Y_{2,t}$ on Z_t and the instruments I_t and save the residuals:

$$Y_{2,t} = \pi_0 + \pi_1 I_t + \pi_2 Z_t + \nu_t$$

   Because Z_t and I_t are exogenous (i.e. uncorrelated with ϵ_t), $Y_{2,t}$ is exogenous if, and only if, ϵ_t and ν_t are uncorrelated (this is what we need to test). This leads to the next step:

2. Run the Hausman test by regressing $Y_{1,t}$ on $Y_{2,t}$, Z_t and $\widehat{\nu_t}$:

$$Y_{1,t} = \alpha + \beta_1 Y_{2,t} + \beta_2 Z_t + \delta \widehat{\nu_t} + \omega_t$$

For this test $H_0 : \delta = 0$. If δ is significant, we can conclude, that $Y_{2,t}$ was in fact endogenous, because the error terms **were** correlated.

The first case where endogenous variable on the right-hand-side emerges is a *measurement error in explanatory variable*.

Assume that the right model is

$$Y_t = \alpha + \beta X_t + u_t$$

but instead of X_t we observe $X_t^* = X_t + \nu_t$ where $\mathbb{E}\nu_t = 0$ and ν_t does not depend on u_t. Thus, our regression model is of the form:

$$Y_t = \alpha + \beta X_t^* + (u_t - \beta\nu_t) = \alpha + \beta X_t^* + \epsilon_t^*$$

where X_t^* is endogenous because

$$cor(X_t^*, \epsilon_t^*) = \mathbb{E}(X_t + \nu_t)(u_t - \beta\nu_t) = -\beta\mathbb{E}\nu_t^2 \neq 0$$

Recall that we want to estimate β in $Y_t = \alpha + \beta X_t + u_t$ but since we do not have X, we replace it by X^*. Consequently, we have to look for an instrument for X^*.

System of Equations

Another important source of endogeneity is simultaneity. The reason that there are two equations in a supply and demand model is that there are two variables - Q for equilibrium quantity and P for equilibrium price - whose values the model explains:

$$\begin{cases} Q_t^D = \beta_0 + \beta_1 P_t + \beta_2 I_t + \epsilon_t^D & \text{(demand equation)} \\ Q_t^S = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \epsilon_t^S & \text{(supply equation)} \\ Q_t^D = Q_t^S = Q_t & \text{(equilibrium condition)} \end{cases}$$

the model can also contain some extra variables, it is I (the income of buyers) and W (the wage rate of seller's employees). Note that the supply equation can be rewritten as:

$$\begin{cases} Q_t = \beta_0 + \beta_1 P_t + \beta_2 I_t + \epsilon_t^D & \text{(demand equation)} \\ Q_t = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \epsilon_t^S & \text{(supply equation)} \end{cases}$$

- A variable (such as Q and P) is *endogenous* to an economic model if its value is defined within the model.
- A variable (such as I and W) is *exogenous* to the model if its value is taken as given (i.e., is treated as a fixed parameter) by the model (the market forces bring Q and P to equilibrium *together*, but market forces do not influence neither I or W).

These definitions are equivalent to the following ones: the right-hand-side variable of an equation is called endogenous if it is *correlated* (and exogenous if it is *uncorrelated*) with the error term ϵ.

Recall that the OLS estimates of the coefficients of an equation are BLUE only if certain (Gauss -Markov) conditions are met, in particular, if all the right-hand-side variables are exogenous. But what happens if one of the right-hand-side variables is endogenous?

- If all the Gauss-Markov assumptions are true except the one of exogeneity then the OLS estimators of the coefficients become biased, inconsistent and inefficient.

Thus, we cannot apply the OLS to neither demand nor supply equation. to cure the **structural** (or economic) system, solve the model for its endogenous variables - the new **reduced** (or econometric) system of the model will take the form of:

$$\begin{cases} Q_t = \delta_0 + \delta_1 W_t + \delta_2 I_t + \epsilon_{Q,t} \\ P_t = \pi_0 + \pi_1 W_t + \pi_2 I_t + \epsilon_{P,t} \end{cases}$$

Since W and I are exogenous, respective estimates are BLUE. They provide a simple description of the equilibrium of the model and of how it changes when the exogenous variables change.

However, δ's and π's are not the slopes of the supply and demand lines. One possibility is to work backward from these values to slopes but sometimes it is rather complicated or even impossible (in any case, it will not provide estimates of the standard errors of the β and γ parameters which are necessary to test hypothesis about them).

Therefore, to estimate the coefficients of the **original** equation, we apply a *two-stage least squares* procedure:

Let K be the number of all the exogenous variables in the model (including a constant) and H_j the number of (unknown) coefficients in the j th structural equation. The necessary condition for the equation to be identified (or estimable) is:

$$K \geq H_j$$

We shall explain the procedure by means of example.

Example

In the first system

$$\begin{cases} Q_t^D = \beta_0 + \beta_1 P_t + \beta_2 I_t + \epsilon_t^D & \text{(demand equation)} \\ Q_t^S = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \epsilon_t^S & \text{(supply equation)} \\ Q_t^D = Q_t^S = Q_t & \text{(equilibrium condition)} \end{cases}$$

the list of exogenous variables consists of a constant, I, and W, therefore, $K = 3$. In the demand equation, we have three β's, in the supply equation three γ's, thus according to the order condition, we can proceed with both equations.

Stage 1. Using OLS, regress the endogenous variables on all of the exogenous variables (you have to estimate both equations in the system):

$$\begin{cases} Q_t = \delta_0 + \delta_1 W_t + \delta_2 I_t + \epsilon_{Q,t} \\ P_t = \pi_0 + \pi_1 W_t + \pi_2 I_t + \epsilon_{P,t} \end{cases}$$

Stage 2. Now estimate the structural equations in the first system by OLS, replacing the endogenous variables with their predicted values, $\widehat{Q}_t$ and $\widehat{P}_t$, from Stage 1:

$$\begin{cases} \widehat{Q}_t = \beta_0 + \beta_1 \widehat{P}_t + \beta_2 I_t + \epsilon_t^{(1)} \\ \widehat{Q}_t = \gamma_0 + \gamma_1 \widehat{P}_t + \gamma_2 W_t + \epsilon_t^{(2)} \end{cases}$$

It can be proved that the estimated $\widehat{\beta}$'s and $\widehat{\gamma}$'s from the second-stage regression are consistent estimators of the true β and γ parameters. They are biased, but the bias diminishes as the sample grows larger.

Note that in a similar system

$$\begin{cases} Q_t = \beta_0 + \beta_1 P_t + \epsilon_t^D \\ Q_t = \gamma_0 + \gamma_1 P_t + \gamma_2 W_t + \epsilon_t^D \end{cases}$$

$K = 2$, $H_1 = 2$, and $H_2 = 3$, thus the second equation is not identified (i.e., we **cannot** consistently estimate γ's from our data by **any** estimation method).

Finally, in the system

$$\begin{cases} Q_t = \beta_0 + \beta_1 P_t + \epsilon_t^D \\ Q_t = \gamma_0 + \gamma_1 P_t + \epsilon_t^D \end{cases}$$

both equations are unidentified (i.e., if our data consists of the equilibrium data (Q_1, P_1),...,(Q_T, P_T) only, there is no way to estimate β's and γ's). Indeed, we can solve the system as

$$\begin{cases} Q_t = \delta_Q + \epsilon_{Q,t} \\ P_t = \delta_P + \epsilon_{P,t} \end{cases}$$

but we cannot restore **four** parameters β_0, β_1, γ_0, and γ_1 from **two** parameters δ_Q and δ_P.

Example

Consider the following IS-LM model:

$$\begin{cases} R_t = \beta_{11} + \beta_{12} M_t + \beta_{13} Y_t + \beta_{14} M_{t-1} + u_{1t} \\ Y_t = \beta_{21} + \beta_{22} R_t + \beta_{23} I_t + u_{2t} \end{cases}$$

Where R_t denotes the interest rates; M_t denotes the money stock; Y_t is GDP; I_t is investment expenditure.

In this model, R_t and Y_t are endogenous variables and M_t, M_{t-1} and I_t are exogenous variables. The first (LM) equation is exactly identified and the second (IS) one is overidentified (i.e. more exogenous variables than coefficients to estimate).

The example data are annual time series from 1969 to 1977 for the UK economy:

```
suppressPackageStartupMessages({
  library(readxl)
  require(systemfit)
  require(AER)
})
txt1 <- "http://uosis.mif.vu.lt/~rlapinskas/(data%20R&GRETL/"
txt2 <- "simult.xls"
tmp = tempfile(fileext = ".xls")
download.file(url = paste0(txt1, txt2),
              destfile = tmp, mode = "wb")
sim.dt <- data.frame(read_excel(path = tmp))
data1  <- ts(sim.dt[, -1], frequency = 1, start = 1969)
data1 <- data.frame(data1,
                    M.L1 = c(NA, data1[-nrow(data1), 2]))
```

Because some of the equations have endogenous variables on the right-hand-side, we also define the instrumental variables (in this case these are the exogenous variables) after the model equation:

```
lm.R <- ivreg(R ~ M + Y + M.L1 | M + M.L1 + I, data = data1)
lm.Y <- ivreg(Y ~ R + I | M + M.L1 + I, data = data1)
```

```r
round(summary(lm.R)$coefficients, 4)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.5275    11.1348  2.4722   0.0209
## M             0.0019     0.0019  1.0091   0.3230
## Y            -0.2647     0.2241 -1.1809   0.2492
## M.L1         -0.0017     0.0018 -0.9855   0.3342
```

The coefficients of Y as well as M are insignificant, suggesting that the LM function is very flat.

```r
round(summary(lm.Y)$coefficients, 4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.7996    68.7067  1.4380   0.1628
## R            -4.0430     3.1306 -1.2914   0.2084
## I             0.0002     0.0004  0.5329   0.5988
```

Interpreting this model, we can say that income and the rate of interest are negatively related, according to the theoretical prediction, and income is quite sensitive to changes in the rate of interest. Also, a change in investment is would cause the function to shift to the right, again as theory suggests.

However, the p-values indicate that the coefficients are not statistically significant from zero.

Alternatively, `systemfit` can be used to estimate the equation system (more on the eq. system estimation in the next lecture slides).