Advanced time-series analysis (University of Lund, Economic History Department)

30 an-3 February and 26-30 March 2012

Lecture 8 Optional (depending on previously acquired knowledge): Fundamentals of system estimation. Problems of identification. ILS, 2SLS, GMM.

8.a. When is the exogeneity assumption violated?

From the Classical Linear Model, we know that when the orthogonality or exogeneity condition is violated, that is, $E(xu) \neq 0$, the parameters will be estimated with bias.

Such violation may occur in several cases:

1. Measurement error in a right-hand side variable, even if it has a zero mean and is IID, can lead to biased estimates.
2. Omitting an important variable from the model.
3. Having simultaneity, that is, a two-way causal relationship between the explanatory and the dependent variables.

We are going to look at option 3 now.

Let us take a classical case of the Keynesian cross:

$C_t = \alpha_0 + \alpha_1 Y_t + u_t$ which is the standard Keynesian absolute income hypothesis

$Y_t = C_t + I_t$, which is an equality, true for a closed economy.

In this system of equations we have two endogenous variables: Y_t and C_t and a single exogenous variable, I_t. Endogenous variables are determined within the system, while the value of exogenous variables is given by some un-modeled process.

If you were to estimate the first equation with an OLS, that would be tantamount with assuming the lack of the second relationship. Such a case leads to simultaneity bias.

Why?

Let us express the two endogenous variables using the two equations:

$$C_t = \frac{1}{1-\alpha_1}\left(\alpha_0 + \alpha_1 I_t + u_t\right)$$

$$Y_t = \frac{1}{1-\alpha_1}\left(\alpha_0 + u_t + I_t\right)$$

These are the so-called reduced form equations. Please observe that the equation for Y_t also includes u_t at the right hand side. That is, Y_t and u_t are correlated and this is a violation of the orthogonality assumption.

But let us think further about the reduced form equations:

What here happened is simply that we expressed each endogenous variable as a function of the exogenous variables. This can be done for all systems.

Since endogenous variables are determined within the system, they cannot be responsible for long-run changes in the system. Simply, if you had no exogenous factors (no exogenous variables or shocks) the system should sooner or later settle at its equilibrium (or steady state). The endogenous variables would have a constant value and that is it. It is the exogenous factors (variables and shocks) that finally determine the long-run movements of the endogenous variables. The reduced form equations express this feature of the endogenous variables. For this reason the coefficients from the reduced-form equations are the long-run effects of a change in an exogenous variable on the endogenous variable.

Any additional unit of investment will, for example, finally lead to $\frac{1}{1-\alpha_1}$ unit increase in output,

where α_1 is the marginal propensity to consumption (this is the Keynesian investment multiplier).

What if we had no exogenous effect in our system?

$C_t = \alpha_0 + \alpha_1 Y_t$ and $Y_t = \beta_0 + \beta_1 C_t$

Obviously our reduced form equations would have constant value, showing that the variables should settle to a constant equilibrium value.

$C_t = \frac{\alpha_0 + \alpha_1\beta_0}{1-\alpha_1\beta_1}$ and $Y_t = \frac{\beta_0 + \beta_1\alpha_0}{1-\beta_1\alpha_1}$

8.b Identification

We, of course, wish to estimate the parameters in the equations of the system. This cannot be done with OLS because of simultaneity bias, so you need to find an alternative way.

One possible way could be to use the reduced-form equations.

For the above problem, we could estimate one of the reduced form equations and so have an estimate of its coefficients:

$C_t = \pi_0 + \pi_1 I_t + v_t$ where $\pi_0 = \frac{\alpha_0}{1-\alpha_1}, \pi_1 = \frac{\alpha_1}{1-\alpha_1}$

As such, the coefficients could be estimated form these coefficients as follows:

$\frac{\pi_1}{1+\pi_1} = \alpha_1$ and $\pi_0\left(1-\frac{\pi_1}{1+\pi_1}\right) = \alpha_0$.

This is a clear case: using the parameters of the reduced form equations we could arrive at unique solution for the parameters of the system.

When you can arrive at unique estimates for the parameters of an equation based on the reduced form equation, the equation is **exactly identified**.

The estimation method that uses the parameter estimates from the reduced form equations to express the coefficients of an equation of a system is called the **Indirect Least Squares (ILS)**. ILS can only be used for exactly identified equations.

Is it possible to have an equation for which you cannot have a solution? Yes it is, then the equation is non-identified or unidentified.
Let us see a case for this:
We have a classic equilibrium model for a good:
$Q_t^S = \alpha_0 + \alpha_1 P_t + u_t$ for the supply and

$Q_t^D = \beta_0 + \beta_1 P_t + v_t$ for the demand.
This becomes a system if we believe that there is an equilibrium:
$Q_t^S = Q_t^D = Q_t$.

$$Q_t = \alpha_0 + \alpha_1 P_t + u_t$$
$$Q_t = \beta_0 + \beta_1 P_t + v_t$$
The endogenous variables are Q and P and there are no exogenous variables. The reduced-form equations yield:
$P_t = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{u_t - v_t}{\beta_1 - \alpha_1}$ and $Q_t = \beta_0 + \frac{\beta_1(\alpha_0 - \beta_0)}{\beta_1 - \alpha_1} + \frac{\beta_1(u_t - v_t)}{\beta_1 - \alpha_1} + v_t$
The problem is that now you could not arrive at a solution to the parameters. Neither of the equations is identified.

What would happen if we were to introduce an exogenous variable, say income (Y), to the demand equation?
$$Q_t = \alpha_0 + \alpha_1 P_t + u_t$$
$$Q_t = \beta_0 + \beta_1 P_t + \beta_2 Y_t + v_t$$
Then you would obtain the following reduced form equations:
$P_t = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} - \frac{\beta_2}{\beta_1 - \alpha_1} Y_t + \frac{u_t - v_t}{\beta_1 - \alpha_1}$ and $Q_t = \alpha_0 + \frac{\alpha_1(\alpha_0 - \beta_0)}{\beta_1 - \alpha_1} - \frac{\alpha_1 \beta_2}{\beta_1 - \alpha_1} Y_t + \frac{\alpha_1(u_t - v_t)}{\beta_1 - \alpha_1} + u_t$
Now, by dividing the Y coefficient of the reduced form for Q by the coefficient of Y from the reduced form equation for P, you could get an estimate for α_1. Using this you could also calculate α_0. Still you could still not estimate the coefficients of the demand equation. So by introducing an exogenous variable to one equation we could exactly identify the **other** equation.

It is also possible to have equations where you could have more than one possible solutions for the coefficients: these are **over-identified** equations. You could add another exogenous variable to the demand equation, like wealth (W):
$$Q_t = \alpha_0 + \alpha_1 P_t + u_t$$
$$Q_t = \beta_0 + \beta_1 P_t + \beta_2 Y_t + \beta_3 W_t + v_t$$
Now the first equation is over-identified, since you could get two solutions for both parameters, but the second equation is still underidentified.

How can we make both equations exactly identified? Simply by adding a single exogenous variable to each equation.

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + u_t$$

and

$$Q_t = \beta_0 + \beta_1 P_t + \beta_2 Y_t + v_t$$

where P_{t-1} is a predetermined variable (past value of an endogenous variable).

Let us see the general rule:
An equation with m number of endogenous variables is exactly identified if it does not contain at least m-1 of the exogenous and predetermined variables of the system.
If the number of omitted exogenous or predetermined variable is less than m-1, the equation is underidentified. If it does not contain more than m-1 exogenous or predetermined variables then it is over-identified.

Let us take an example:

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + u_t$$

$$Q_t = \beta_0 + \beta_1 P_t + \beta_2 Y_t + \beta_3 W_t + v_t$$

The number of endogenous variables in the system is 2 (Q_t and P_t).
The number of exogenous or predetermined variable sin the system is 3 (Y_t, W_t, P_{t-1}).
The first equation has 2 endogenous variables (the dependent variable should also be added!), but it does not have 2 of the exogenous or predetermined variables (Y_t and W_t). Since m-1=1 this equation is overidentified.
The second equation also has two endogenous variables, and it does not contain P_{t-1} of the exogenous or predetermined variables. As a result it is exactly identified.

8.c. Estimating the parameters of a simultaneous system

One solution could be to estimate **all equations simultaneously**. The most popular estimator of this kind is the **Full-Information Maximum Likelihood** (FIML) estimator. This assumes that the error terms are jointly normally distributed. (It assumes a multivariate normal distribution for the vector of errors.) This allows for the errors to be correlated between equations.
The FIML method is efficient if the assumption regarding the distribution of the residuals is correct.

Let us see an example:
We have a simple system:

$$\ln Y_t = \alpha_{10} + \alpha_{11} \ln I_t + \alpha_{12} \ln Y_{t-1} + u_t$$

$$\ln I_t = \alpha_{20} + \alpha_{21} \ln I_{t-1} + \alpha_{22} \ln Y_t + v_t$$

where Y is the real GDP and I is the real investments.
Logically, even though investments cause real GDP to grow, more income leads to more investment.
The predetermined variables are the first lags of the endogenous variables.

Now with a simple OLS by equation we would obtain:

System: SYS
Estimation Method: Least Squares
Date: 02/02/12 Time: 21:20
Sample: 1950Q2 2000Q4
Included observations: 203
Total system (balanced) observations 406

	Coefficient	Std. Error	t-Statistic	Prob.
C(10)	0.209051	0.027363	7.639996	0.0000
C(11)	0.047853	0.006588	7.263999	0.0000
C(12)	0.939537	0.008136	115.4727	0.0000
C(20)	-0.602152	0.156723	-3.842154	0.0001
C(21)	0.189234	0.046633	4.057944	0.0001
C(22)	0.847448	0.037779	22.43177	0.0000

Determinant residual covariance 1.18E-07

Equation: LOG(REALGDP)=C(10)+C(11)*LOG(REALINVS)+C(12)
*LOG(REALGDP(-1))
Observations: 203

R-squared	0.999668	Mean dependent var	8.316882
Adjusted R-squared	0.999665	S.D. dependent var	0.484569
S.E. of regression	0.008867	Sum squared resid	0.015724
Durbin-Watson stat	1.287174		

Equation: LOG(REALINVS)=C(20)+C(21)*LOG(REALGDP)+C(22)
*LOG(REALINVS(-1))
Observations: 203

R-squared	0.993619	Mean dependent var	6.309467
Adjusted R-squared	0.993556	S.D. dependent var	0.599625
S.E. of regression	0.048136	Sum squared resid	0.463415
Durbin-Watson stat	1.537421		

With an FIML the results are different:

System: SYS
Estimation Method: Full Information Maximum Likelihood (Marquardt)
Date: 02/02/12 Time: 21:25
Sample: 1950Q2 2000Q4
Included observations: 203
Total system (balanced) observations 406
Convergence achieved after 55 iterations

	Coefficient	Std. Error	z-Statistic	Prob.
C(10)	0.047541	0.040089	1.185909	0.2357
C(11)	0.005605	0.010171	0.551098	0.5816
C(12)	0.991061	0.012401	79.91503	0.0000
C(20)	-0.365898	0.154338	-2.370765	0.0178
C(21)	0.117040	0.048627	2.406877	0.0161
C(22)	0.905266	0.041030	22.06365	0.0000

Log likelihood	-1889.131	Schwarz criterion	18.76917
Avg. log likelihood	-4.653031	Hannan-Quinn criter.	18.71086
Akaike info criterion	18.67124		
Determinant residual covariance		8.19E-08	

Equation: LOG(REALGDP)=C(10)+C(11)*LOG(REALINVS)+C(12)
*LOG(REALGDP(-1))
Observations: 203

R-squared	0.999600	Mean dependent var	8.316882
Adjusted R-squared	0.999596	S.D. dependent var	0.484569
S.E. of regression	0.009736	Sum squared resid	0.018958
Durbin-Watson stat	1.312035		

Equation: LOG(REALINVS)=C(20)+C(21)*LOG(REALGDP)+C(22)
*LOG(REALINVS(-1))
Observations: 203

R-squared	0.993543	Mean dependent var	6.309467
Adjusted R-squared	0.993478	S.D. dependent var	0.599625
S.E. of regression	0.048424	Sum squared resid	0.468969
Durbin-Watson stat	1.626699		

Now we find that the immediate impact of a growth in investment is statistically insignificant on real GDP, and the magnitude of the effect of real GDP on investments also reduced.

Another solution can be to estimate the system by equation.

The ILS will only work for equations that are exactly identified. For equations that are overidentified you need to use either a 2SLS or a GMM estimator.

2-stage least squares 2SLS:

Let us use the first example for the explanation.

$$C_t = \alpha_0 + \alpha_1 Y_t + u_t$$
$$Y_t = C_t + I_t$$

The OLS could not be used to estimate the equation for consumption, since Y_t was correlated with u_t. Now we need to turn to instrumentation.

An instrument is a variable that is correlated with the endogenous variable, but does not correlate with the residual. Such an instrument is the exogenous variable I_t. Basically, what we should do is to project Y_t on the column vector space defined by I_t, which is orthogonal to (uncorrelated with) u_t. We can use the reduced form equation for Y for this purpose:

$$Y_t = \lambda_0 + \lambda_1 I_t + e_t$$

Now here I_t is an instrument, and the fitted value $\hat{y}_t = \lambda_0 + \lambda_1 I_t$ is the instrumented Y_t. This cannot be correlated with u_t, since I_t is uncorrelated with u_t. The residual term e_t will hold that component of Y_t which correlates with the residual term u_t.

Running this regression is the **first-step**.

The second step is to estimate the original equation but this time with the instrumented variable instead of the endogenous variable:

$$C_t = \alpha_0 + \alpha_1 \hat{Y}_t + u_t$$

Now you should obtain the unbiased estimates of the alpha parameters.

A great advantage of the 2SLS is that it can estimate overidentified equations as well, which would not be possible to be done with ILS.

Let us try this procedure in Eviews withTable f5.1 with realinvst as instrument!

2SLS output from the Eviews

Dependent Variable: REALCONS
Method: Two-Stage Least Squares
Date: 02/02/12 Time: 22:04
Sample: 1950Q1 2000Q4
Included observations: 204
Instrument specification: REALINVS C

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-147.3792	6.631673	-22.22354	0.0000
REALGDP	0.689691	0.001325	520.4592	0.0000

R-squared	0.999293	Mean dependent var	2999.436
Adjusted R-squared	0.999289	S.D. dependent var	1459.707
S.E. of regression	38.91183	Sum squared resid	305854.3
F-statistic	270877.8	Durbin-Watson stat	0.323007
Prob(F-statistic)	0.000000	Second-Stage SSR	22396714
J-statistic	3.88E-41	Instrument rank	2

Could we overidentify the equation? Yes!

Let us rewrite our model:

$$C_t = \alpha_0 + \alpha_1 Y_t + u_t$$

$$Y_t = C_t + I_t + G_t$$

where G_t is the government expenditures.

The first-step (reduced form equation) becomes:

$$Y_t = \lambda_0 + \lambda_1 I_t + \lambda_2 G_t + e_t$$

Dependent Variable: REALCONS
Method: Two-Stage Least Squares
Date: 02/02/12 Time: 22:05
Sample: 1950Q1 2000Q4
Included observations: 204
Instrument specification: REALINVS C REALGOVT

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-146.4722	6.518678	-22.46961	0.0000
REALGDP	0.689492	0.001298	531.2619	0.0000

R-squared	0.999292	Mean dependent var	2999.436
Adjusted R-squared	0.999289	S.D. dependent var	1459.707
S.E. of regression	38.92724	Sum squared resid	306096.7
F-statistic	282239.2	Durbin-Watson stat	0.322612
Prob(F-statistic)	0.000000	Second-Stage SSR	4855381.
J-statistic	0.540997	Instrument rank	3
Prob(J-statistic)	0.462020		

Over-identification of the equation makes it possible that you test the validity of your instruments, namely if they are really uncorrelated with the residual. This is done with the Sargan and the Hansen tests.

The null-hypothesis of these tests is that the instruments are uncorrelated with the residual, so they are indeed exogenous and valid. In this case we cannot reject the null-hypothesis at 10% level of significance, so our estimates are acceptable.

How does the Sargan test work?

First you estimate the 2SLS. Now you save the residuals from it. If these residuals can be explained by your set of instruments in a statistical significant way, you have invalid instruments.

In essence this is a test regression of the residuals form you 2SLS on your instruments:

Dependent Variable: RESID01
Method: Least Squares
Date: 02/02/12 Time: 22:15
Sample: 1950Q1 2000Q4
Included observations: 204

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7.872781	13.02563	0.604407	0.5463
REALINVS	0.013392	0.018531	0.722678	0.4707
REALGOVT	-0.016658	0.023160	-0.719255	0.4728

R-squared	0.002678	Mean dependent var	-6.20E-13
Adjusted R-squared	-0.007245	S.D. dependent var	38.83125
S.E. of regression	38.97167	Akaike info criterion	10.17814
Sum squared resid	305276.9	Schwarz criterion	10.22694
Log likelihood	-1035.171	Hannan-Quinn criter.	10.19788
F-statistic	0.269882	Durbin-Watson stat	0.330543
Prob(F-statistic)	0.763746		

The joint insignificance of the model is indicative that the instruments are indeed exogenous.

<u>Generalized Method of Moments (GMM):</u>

First you need to understand what the method of moments is.
The idea is simple. Instead of minimizing the residual sum of squares, the coefficients of the regression model is estimated based on assumption regarding its moments.
For example, if you have the model:
$y_t = \beta_0 + \beta_1 x_t + u_t$, then our assumptions regarding the moments are:

$E(u_t) = 0$ and $E(u_t \cdot x_t) = 0$ this are moment restrictions, and two restrictions are enough to estimate two coefficients. This is a case of exact identification.
All that we need to do is to create an objective function, like this:
$V(\beta_0, \beta_1, x_t, y_t) = \left(E(u_t)\right)^2 + \left(E(u_t \cdot x_t)\right)^2$ and choose those coefficient which bring this as close to zero as possible.

What is x and u are correlated? Then we need an instrument z, which is correlated with x but uncorrelated with u: $E(u_t \cdot z_t) = 0$, $Cov(x_t, z_t) \neq 0$.
The new moment restrictions are:
$E(u_t) = 0$
$E(u_t \cdot z_t) = 0$
Two restrictions can also lead to two parameter estimates. Now the objective function is something like this:
$V(\beta_0, \beta_1, x_t, y_t, z_t) = \left(E(u_t)\right)^2 + \left(E(u_t \cdot z_t)\right)^2$

But what happens if we have an over-identified equation? Then we have more than one instrument for x, which yields additional moment conditions. Say, we have another instrument denoted by w_t. Then our moment conditions are:
$E(u_t) = 0$
$E(u_t \cdot z_t) = 0$
$E(u_t \cdot w_t) = 0$
Three conditions for two parameters: a clear case of over-identification. You could, of course, simply minimize the following objective function:
$V(\beta_0, \beta_1, x_t, y_t, z_t) = \left(E(u_t)\right)^2 + \left(E(u_t \cdot z_t)\right)^2 + \left(E(u_t \cdot w_t)\right)^2$
but there is no guarantee that this would lead to the lowest possible value. Instead, you should attach weights to each condition. This makes the method a Generalized Method of Moments, that has been suggested by Hansen. The weighting gets important only when there is a case of over-identification.
The weights themselves are also parameters to be estimated. This is possible by successive steps. This is the reason why you can have one-step, two-step or n-step GMM estimation.
The value of the objective function is the J-statistics. If you have an exact identification, its value can only be zero. If you have overidentification, its value can actually be more than zero. If your instruments are good, however, it should be statistically indifferent from zero. This is the J-test, which is equivalent with Sargan-test.

Results from an iterative n-step GMM procedure

Dependent Variable: REALCONS
Method: Generalized Method of Moments
Date: 02/02/12 Time: 22:36
Sample: 1950Q1 2000Q4
Included observations: 204
Linear estimation with 1 weight update
Estimation weighting matrix: HAC (Bartlett kernel, Newey-West fixed bandwidth = 5.0000)
Standard errors & covariance computed using estimation weighting matrix
Instrument specification: REALINVS C REALGOVT

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-146.3493	11.95132	-12.24545	0.0000
REALGDP	0.689476	0.002259	305.1839	0.0000

R-squared	0.999292	Mean dependent var	2999.436
Adjusted R-squared	0.999289	S.D. dependent var	1459.707
S.E. of regression	38.92874	Sum squared resid	306120.3
Durbin-Watson stat	0.322575	J-statistic	0.132056
Instrument rank	3	Prob(J-statistic)	0.716310