Additional Cheat Sheet

By Marcelo Moreno - King Juan Carlos University
As part of the Econometrics Cheat Sheet Project

THIS IS A WORK IN PROGRESS
NOT INTENDEND FOR GENEAL PURPOSE

More about OLS in matrix notation

The general econometric model:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + u_i$$

Can be written in matrix notation:

$$y = X\beta + u$$

Let's call e the vector of estimated residuals ($e \neq u$):

$$e = y - X\hat{\beta}$$

The objective of OLS is to minimize the SSR:

$$Min \quad e^T e$$

Getting $e^T e$:

$$e^T e = (y - X\hat{\beta})^T (y - X\hat{\beta}) =$$
$$= y^T y - \hat{\beta}^T X^T y - y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta} =$$
$$= y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}$$

where there is the fact that: $y^T X\hat{\beta} = (y^T X\hat{\beta})^T = \hat{\beta}^T X^T y$

Minimizing $e^T e$:

$$\frac{\partial e^T e}{\partial \hat{\beta}} = -2X^T y + 2X^T X\hat{\beta} = 0$$
$$\hat{\beta} = (X^T X)^{-1}(X^T y)$$

The variance-covariance matrix:

$$Var(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$$

where:

$$\hat{\sigma} = \frac{e^T e}{n-k}$$

The standard errors are in the diagonal of:

$$se(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$$

Error measures:

- $SSR = e^T e = y^T y - \hat{\beta}^T X^T y = \sum (y_i - \hat{y}_i)^2$
- $SSE = \hat{\beta}^T X^T y - n\overline{y}^2 = \sum (\hat{y}_i - \overline{y})^2$
- $SST = SSR + SSE = y^T y - n\overline{y}^2 = \sum (y_i - \overline{y})^2$

The R-Squared:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Omission of variables

Most of the time, is hard to get all relevant variables for an analysis. For example, a true model with all variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

The estimated model (with the available variables):

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$$

Omitted variables provoke OLS bias and inconsistency. Depending of the correlation between x_1 and x_2 and the sign of β_2, the bias on $\tilde{\beta}_1$ could be:

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	$(+)$ bias	$(-)$ bias
$\beta_2 < 0$	$(-)$ bias	$(+)$ bias

$(+)$ bias $\rightarrow \tilde{\beta}_1$ will be higher than it should be (it includes the effect of x_2). $\tilde{\beta}_1 > \beta_1$

$(-)$ bias $\rightarrow \tilde{\beta}_1$ will be lower than it should be (it includes the effect of x_2). $\tilde{\beta}_1 < \beta_1$

If $Corr(x_1, x_2) = 0$, there is no bias on β_1, because the effect of x_2 will be picked up by the error term, u.

There are two approaches to solve the problem:

- Make use of proxy variables.
- Make use of instrumental variables (IV)

Proxy variables

Is the approach when a relevant variable is not available for the model because is non-observable, and there is no data. A proxy variable is something related with te non-observable variable that has data available. For example, the intellectual coefficient (IC) is a proxy variable for a subject's capacity (non-observable).

Instrumental variables (IV)

When proxy variables are not available, the alternative approach is to look for a variable, let's call it z, that has a relation with x. The z variable must meet the following requirements to be called an Instrumental Variable (IV):

$$Cov(z, u) = 0$$
$$Cov(z, x) \neq 0$$

TSLS

Can have multiple instrumental variables (is the IV, but with various instrumental variables at the same time). And $Cov(z, u) = 0$ can be relaxed, but there has to be a minimum of variables that satisfies it.

Can have multicollinearity problems.

Some tests:

- Endogeneity tests: is TSLS better than OLS when there are no endogenous variables? Do we really need TSLS? $\rightarrow$ Hausman test $\rightarrow H_0$: OLS is consistent (it is better to use OLS).
- Over-identification. An IV should meet:
  - $Corr(z, u) = 0$ (exogeneity)
  - $Corr(z, x) \neq 0$ (relevance)
- Is there too many IV? $\rightarrow$ Sagan test $\rightarrow H_0$: all IV seem ok

Incorrect functional forms

Ramsey RESET test: it test the specification errors of a regression. H_0: the model is correctly specified.

VAR

The VAR model general form:
$y_t = A_1 y_{t-1} + ... + A_p y_{t-p} + B_0 x_t + ... + B_q x_{t-q} + CD_t + u_t$
where:

- $y_t = (y_{1t}, ..., y_{Kt})'$ is a vector of K observable endogenous variables.
- $x_t = (x_{1t}, ..., x_{Mt})'$ is a vector of M observable exogenous or unmodelled variables.
- D_t contains all deterministic variables which may consist of a constant, a linear trend, seasonal dummy variables...
- u_t is a K-dimensional unobservable zero mean white noise process with positive definite covariance matrix $E(u_t u_t' = \Sigma_u)$.
- The A_i, B_j and C are parameter matrices of suitable dimension.

For example, a model with two endogenous variables (with two lags), an exogenous contemporaneous variable, a constant ($const$) and a trend ($Trend_t$):

$$y_{1t} = a_{11,1} y_{1,t-1} + a_{12,1} y_{2,t-1} + a_{11,2} y_{1,t-2} + a_{12,2} y_{2,t-2} + b_{11} x_t + c_{11} + c_{12} Trend_t + u_{1t}$$

$$y_{2t} = a_{21,1} y_{2,t-1} + a_{22,1} y_{1,t-1} + a_{21,2} y_{2,t-2} + a_{22,2} y_{1,t-2} + b_{21} x_t + c_{21} + c_{22} Trend_t + u_{2t}$$

For example, the equations:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = A_1 \cdot \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + A_2 \cdot \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + B_0 \cdot \begin{bmatrix} x_t \end{bmatrix} + C \cdot \begin{bmatrix} const \\ Trend_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} a_{11,1} & a_{12,1} \\ a_{21,1} & a_{22,1} \end{bmatrix} \cdot \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{11,2} & a_{12,2} \\ a_{21,2} & a_{22,2} \end{bmatrix} \cdot \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} \cdot \begin{bmatrix} x_t \end{bmatrix} + \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \cdot \begin{bmatrix} const \\ Trend_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

Information criterias

VECM

$$\Delta y_t = \Pi^* \begin{bmatrix} y_{t-1} \\ D_{t-1}^{co} \end{bmatrix} + \Gamma_1 \Delta y_{t-1} + ... + \Gamma_p \Delta y_{t-p} + B_0 x_t + ... + B_q x_{t-q} + CD_t + u_t$$

where:

- $\Pi^* = \alpha \beta^{T*}$.
- D_t^{co} contains all deterministic terms included in the cointegration relations.
- D_t contains all remaining deterministic variables.

For example. The matrix form:

$$\begin{bmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{bmatrix} = \Pi^* \begin{bmatrix} y_{t-1} \\ const \end{bmatrix} + \Gamma_1 \cdot \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \end{bmatrix} + B_0 \cdot \begin{bmatrix} x_t \end{bmatrix} + C \cdot \begin{bmatrix} Trend_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$\begin{bmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{bmatrix} = \alpha \begin{bmatrix} \beta^T \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + C^* \begin{bmatrix} const \end{bmatrix} \end{bmatrix} + \Gamma_1 \cdot \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \end{bmatrix} + B_0 \cdot \begin{bmatrix} x_t \end{bmatrix} + C \cdot \begin{bmatrix} Trend_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

$$\begin{bmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{bmatrix} = \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} c^* \end{bmatrix} \begin{bmatrix} const \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \cdot \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} \cdot \begin{bmatrix} x_t \end{bmatrix} + \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} \cdot \begin{bmatrix} Trend_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

The VECM model general form:

$$\Delta y_t = \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + ... + \Gamma_p \Delta y_{t-p} + B_0 x_t + ... + B_q x_{t-q} + CD_t + u_t$$

where:

- $y_t = (y_{1t}, ..., y_{Kt})'$ is a vector of K observable endogenous variables.
- $x_t = (x_{1t}, ..., x_{Mt})'$ is a vector of M observable exogenous or unmodelled variables.
- $\Pi = \alpha \beta'$. Suppose $rk(\Pi) = r = rk(\alpha) = rk(\beta)$
- D_t contains all deterministic variables which may consist of a constant, a linear trend, seasonal dummy variables...
- u_t is a K-dimensional unobservable zero mean white noise process with positive definite covariance matrix $E(u_t u_t' = \Sigma_u)$.
- The A_i, B_j and C are parameter matrices of suitable dimension.