

# Additional Cheat Sheet

By Marcelo Moreno - King Juan Carlos University

As part of the Econometrics Cheat Sheet Project

THIS IS A WORK IN PROGRESS

## OLS in matrix notation

The general econometric model:

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

Can be written in matrix notation as:

$$y = X\beta + u$$

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

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

The objective of OLS is to minimize the SSR:

$$\text{Min } \sum \hat{u}_i^2 = \text{Min } \hat{u}^T \hat{u}$$

- Defining  $\hat{u}^T \hat{u}$ :

$$\begin{aligned} \hat{u}^T \hat{u} &= (y - X\hat{\beta})^T (y - X\hat{\beta}) = \\ &= y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X \hat{\beta} \end{aligned}$$

- Minimizing  $\hat{u}^T \hat{u}$ :

$$\frac{\partial \hat{u}^T \hat{u}}{\partial \hat{\beta}} = -2X^T y + 2X^T X \hat{\beta} = 0$$

$$\hat{\beta} = (X^T X)^{-1} (X^T y)$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} n & \sum x_1 & \dots & \sum x_k \\ \sum x_1 & \sum x_1^2 & \dots & \sum x_1 x_k \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_k & \sum x_k x_1 & \dots & \sum x_k^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum y x_1 \\ \vdots \\ \sum y x_k \end{bmatrix}$$

## Variance-covariance matrix

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1} = \begin{bmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \dots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_k) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{Var}(\hat{\beta}_1) & \dots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_k, \hat{\beta}_0) & \text{Cov}(\hat{\beta}_k, \hat{\beta}_1) & \dots & \text{Var}(\hat{\beta}_k) \end{bmatrix}$$

$$\text{where: } \hat{\sigma} = \frac{\hat{u}^T \hat{u}}{n-k-1}$$

The standard errors are in the diagonal of:

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

## Error measures

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

## R-squared

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

The r-squared corrected by degrees of freedom:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

## Variables omission

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):

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

Omission of variables provoke OLS bias and inconsistency. Depending of the correlation between  $x_1$  and  $x_2$  and the sign of  $\beta_2$ , the bias on  $\tilde{\beta}_1$  could be:

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	(+) 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  $\text{Corr}(x_1, x_2) = 0$ , there is no bias on  $\beta_1$ , because the effect of  $x_2$  will be fully 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 the non-observable variable that has data available.

For example, the intellectual coefficient 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):

$$\text{Cov}(z, u) = 0$$

$$\text{Cov}(z, x) \neq 0$$

This method let the omitted variable in the error term, but instead of estimate the model by OLS, it utilizes a method that recognize the presence of an omitted variable. This method can also solve error measurements.

## TSLS

Can have multiple instrumental variables (is the IV, but with various instrumental variables at the same time). And  $\text{Cov}(z, u) = 0$  can be relaxed, but there has to be a mini-

mum 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:
  - $\text{Corr}(z, u) = 0$  (exogeneity)
  - $\text{Corr}(z, x) \neq 0$  (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:

$y_t = A_1 y_{t-1} + \dots + A_p y_{t-p} + B_0 x_t + \dots + B_q x_{t-q} + CD_t + u_t$   
where:

- $y_t = (y_{1t}, \dots, y_{Kt})'$  is a vector of  $K$  observable endogenous variables.
- $x_t = (x_{1t}, \dots, x_{Mt})'$  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:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = A_1 \cdot \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + A_2 \cdot \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + B_0 \cdot [x_t] + C \cdot \begin{bmatrix} const \\ Trend_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$
$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} a_{11,1} & a_{12,1} \\ a_{21,1} & a_{22,1} \end{bmatrix} \cdot \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{11,2} & a_{12,2} \\ a_{21,2} & a_{22,2} \end{bmatrix} \cdot \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} \cdot [x_t] + \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \cdot \begin{bmatrix} const \\ Trend_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

## Information criterias

### VECM

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

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:

$$\begin{bmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{bmatrix} = \Pi^* \begin{bmatrix} y_{t-1} \\ const \end{bmatrix} + \Gamma_1 \cdot \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \end{bmatrix} + B_0 \cdot [x_t] + C \cdot [Trend_t] + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$
$$\begin{bmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{bmatrix} = \alpha \begin{bmatrix} \beta^T & y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + C^* [const] + \Gamma_1 \cdot \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \end{bmatrix} + B_0 \cdot [x_t] + C \cdot [Trend_t] + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$
$$\begin{bmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{21} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + [c^*] [const] + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \cdot \begin{bmatrix} \Delta y_{1,t-1} \\ \Delta y_{2,t-1} \end{bmatrix} + \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} \cdot [x_t] + \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} \cdot [Trend_t] + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

The VECM model general form:

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

where:

- $y_t = (y_{1t}, \dots, y_{Kt})'$  is a vector of  $K$  observable endogenous variables.
- $x_t = (x_{1t}, \dots, x_{Mt})'$  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.