MSc Econometrics, Autumn term notes

Ron Smith: r.smith@bbk.ac.uk

Autumn 2020

Department of Economics, Mathematics and Statistics,
Birkbeck, University of London

MSc, MRes & PGCert Econometrics

- These notes cover all the theoetical material potentially in the Autumn course. Because of timing the topics covered in each lecture may not match the notes exactly.
- Because of the unusual circumstances this year, it is also possible that we may not be able to cover all of the material.
- Doing the practical exercises will help you understand the econometric theory in these notes.
- If you see any typos or mistakes do tell me. Subsequent generations of students will thank you.

Contents

1. Introduction & bivariate LRM

1.1. Econometrics

Econometrics involves integrating

- data, which may be cross-section, Y_i, $i = 1, 2, ..., N$, or time-series, Y_t, $t = 1, 2, ..., T$. or panel, Y_{it},
- economic theory and background information: domain specific knowledge,
- statistical methods for estimation and inference (testing)
- to construct empirical models, usually using a computer package
- for some purpose. Purpose is central, you would use different models for different purposes such as forecasting, policy analysis, causal analysis, testing hypotheses etc.

We will spend considerable time on the statistical methods (econometric theory), particularly that for estimating linear regression models, LRM, from a sample of data, explaining some dependent variable Y by some independent variables, X. A cross-section bivariate regression model is of the form:

$$Y_i = \beta_1 + \beta_2 X_i + u_i, \ i = 1, 2, ..., N.$$

This is a set of N equations one for each observation, which explain the dependent variable, Y_i, by an explanatory variable X_i (which may be a non-linear function of some other variable, e.g. the logarithm) and we wish to obtain estimates of the intercept, $\widehat{\beta}_1$ and the slope $\widehat{\beta}_2$ and of the estimated errors $\widehat{u}_i = Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_i$. Notice we distinguish between the true, unknown values, like β_1, and our estimates from the sample $\widehat{\beta}_1$, which are random variables.

The first example in the practical uses data from Gapminder. Y_i is life expectancy in country i and X_i is log international dollar per-capita income.

A similar time-series equation is of the form:

$$Y_t = \beta_1 + \beta_2 X_t + u_t, \ t = 1, 2, ..., T. \tag{1.1}$$

for $t = 1, 2, ..., T$.

1.2. Bivariate regression

We will use three procedures to obtain estimators of β_j, denoted $\widehat{\beta}_j$, $j = 1, 2.$ in equations like (1.1). Estimators are formulae which tell you how to calculate an estimate from data for a particular sample.The procedures are (a) method of moments, (b) least squares and (c) maximum likelihood assuming normal errors. In this case the three procedures give the same answer. This is not generally the case. To obtain estimates we need to make some assumptions. First about the errors, u_t.

$$E(u_t) = 0, \tag{1.2}$$

the errors have mean zero. The intercept will pick up any non-zero mean.

$$E(u_t^2) = \sigma^2, \tag{1.3}$$

the errors have constant variance, are homoskedastic. If the assumption fails and the variance is not constant the errors are heteroskedastic.

$$E(u_t u_{t-i}) = 0, i \neq 0, \tag{1.4}$$

there is no serial correlation or autocorrelation in the errors.

We also require that the X_t vary and are not related to the u_t. The lack of relationship may arise because the explanatory variables in X are either (a) non stochastic (b) exogenous, distributed independently of the errors u_t or (c) predetermined, uncorrelated with the errors u_t. All of these imply that $E(u_t) = 0$ and $E(X_t u_t) = 0$. Note independence is a much stronger assumption than uncorrelated.

1.2.1. Method of moments

To get the method of moment estimators we find the estimators, $\widehat{\beta}_1$ and $\widehat{\beta}_2$, that make the assumptions we have made about the population: $E(u_t) = 0$ and $E(X_t u_t) = 0$ hold for their sample equivalents. The mean is the sample equivalent of expected value so these are: $T^{-1}\sum_t \widehat{u}_t = 0$ and $T^{-1}\sum_t X_t\widehat{u}_t = 0$, where $\widehat{u}_t = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_t$.

The first moment condition is

$$\begin{aligned} T^{-1}\sum_t \widehat{u}_t &= T^{-1}\sum_t \left(Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_t\right) = \overline{Y} - \widehat{\beta}_1 - \widehat{\beta}_2\overline{X} = 0 \\ \widehat{\beta}_1 &= \overline{Y} - \widehat{\beta}_2\overline{X} \end{aligned}$$

Use this to rewrite (1.1) in terms of the estimates

$$\begin{aligned} Y_t &= \widehat{\beta}_1 + \widehat{\beta}_2 X_t + \widehat{u} = (\overline{Y} - \widehat{\beta}_2 \overline{X}) + \widehat{\beta}_2 X_t + \widehat{u}_t \\ Y_t - \overline{Y} &= \widehat{\beta}_2 (X_t - \overline{X}) + \widehat{u}_t \\ y_t &= \widehat{\beta}_2 x_t + \widehat{u}_t \end{aligned}$$

where we define $y_t = Y_t - \overline{Y}$ and $x_t = (X_t - \overline{X})$. Working with deviations from the mean makes the algebra easier.

The second moment condition $T^{-1} \sum_t X_t \widehat{u}_t = 0$ is equivalent to $T^{-1} \sum_t x_t \widehat{u}_t = 0$ and

$$\begin{aligned} T^{-1} \sum_t x_t \widehat{u}_t &= T^{-1} \sum_t x_t (y_t - \widehat{\beta}_2 x_t) = (T^{-1} \sum_t x_t y_t) - \widehat{\beta}_2 (T^{-1} \sum_t x_t^2) = 0 \\ \widehat{\beta}_2 &= \frac{(T^{-1} \sum_t x_t y_t)}{(T^{-1} \sum_t x_t^2)} = \frac{\sum_t (X_t - \overline{X})(Y_t - \overline{Y})/T}{\sum_t (X_t - \overline{X})^2/T} = \frac{Cov(X_t Y_t)}{Var(X_t)}, \end{aligned}$$

as long as $Var(X_t) \neq 0$.

So we now have estimators for $\widehat{\beta}_2$ and $\widehat{\beta}_1$.

1.2.2. Least Squares

The least squares estimator minimises $S = \sum_t \widehat{u}_t^2$

$$\begin{aligned} S &= \sum_t (y_t - \widehat{\beta}_2 x_t)^2 = \sum_t y_t^2 + \widehat{\beta}_2^2 \sum_t x_t^2 - 2\widehat{\beta}_2 \sum_t x_t y_t \\ \frac{\partial S}{\partial \widehat{\beta}} &= 2\widehat{\beta}_2 \sum_t x_t^2 - 2 \sum_t x_t y_t = 0 \\ \widehat{\beta}_2 &= \frac{(\sum_t x_t y_t)}{(\sum_t x_t^2)}. \end{aligned}$$

The same as before. Note the second derivative is $2 \sum_t x_t^2 > 0$, so it is a minimum.

1.2.3. Properties of the estimator

For simplicity, we will write β for β_2 below. If the expected value of the random variable $\widehat{\beta}$ (it is different in every sample) equals its true value

$$E(\widehat{\beta}) = \beta$$

then $\widehat{\beta}$ is said to be unbiased.

Since

$$\widehat{\beta} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t(\beta x_t + u_t)}{\sum x_t^2} = \beta + \frac{\sum x_t u_t}{\sum x_t^2}$$

then $E(\widehat{\beta}) = \beta$, and it is unbiased if x_t and u_t are independent, allowing us to write $E(AB) = E(A)E(B)$:

$$E\left\{\frac{\sum x_t u_t}{\sum x_t^2}\right\} = E\left\{\frac{\sum x_t}{\sum x_t^2}\right\} E(u_t)$$

and $E(u_t) = 0$. To derive the variance of $\widehat{\beta}$, note since $\widehat{\beta}$ is unbiased and treating x_t as fixed

$$\begin{aligned} V(\widehat{\beta}) &= E(\widehat{\beta} - \beta)^2 = E(\frac{\sum x_t u_t}{\sum x_t^2})^2 \\ &= \frac{1}{(\sum x_t^2)^2} E\left(\sum x_t u_t\right)^2 \end{aligned} \tag{1.5}$$

We can write $E\left(\sum x_t u_t\right)^2$ as

$$\begin{aligned} &E(x_1u_1 + x_2u_2... + x_Tu_T)(x_1u_1 + x_2u_2... + x_Tu_T) \\ &E(x_1^2u_1^2 + x_2^2u_2^2 + ... + x_T^2u_T^2 + 2x_1u_1x_2u_2 + ...) \\ &x_1^2\sigma^2 + x_2^2\sigma^2 + ... + x_T^2\sigma^2 + 0 + ... \\ &\sigma^2 \sum x_t^2 \end{aligned}$$

then if for $t = 1, 2, ..., T$, $E(u_t^2) = \sigma^2$ and $E(u_t u_{t-i}) = 0.So$

$$\begin{aligned} V(\widehat{\beta}) &= \frac{1}{(\sum x_t^2)^2}\left(\sigma^2 \sum x_t^2\right) \\ &= \frac{\sigma^2}{\sum x_t^2} = \frac{\sigma^2}{T\sigma_X^2} \end{aligned} \tag{1.6}$$

Where $\sigma_X^2 = \sum_t (X_t - \overline{X})^2/T$ Note that $\sum x_t^2$ rises and $V(\widehat{\beta})$ and the standard error $\widehat{\beta}$ falls with T, the sample size, as in the case of the standard error of a mean.

Returning to the original notation, the residuals are

$$\widehat{u}_t = y_t - \widehat{\beta} x_t = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_t.$$

We prove later that the unbiased estimator of σ^2 is

$$s^2 = \sum \widehat{u}_t^2/(T-2)$$

because we estimate two parameters $\widehat{\beta}_1$ and $\widehat{\beta}_2$. Our estimator for the standard error of $\widehat{\beta}_2$ is the square root of $V(\widehat{\beta}_2)$ with σ replaced by s :

$$se(\widehat{\beta}_2) = s/\sqrt{\sum x_t^2}.$$

1.3. Revision material

This revises some background material not covered in lectures because it is assumed that you are familiar with it.

1.3.1. Differentiation with vectors and matrices

Consider the equation:

$$P = \underset{1\text{x}n}{x'} \; \underset{n\text{x}1}{a}$$

Then the derivatives of P with respect to x and x' are defined as :

$$\frac{dP}{dx} = a \text{ and } \frac{dP}{dx'} = a'$$

For $n = 2$:

$$\begin{aligned} P &= [x_1, x_2] \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \\ &= x_1 a_1 + x_2 a_2 \end{aligned}$$

Then

$$\frac{dP}{dx_1} = a_1 \text{ and } \frac{dP}{dx_2} = a_2$$

So

$$\frac{dP}{dx} = \begin{bmatrix} \frac{dP}{dx_1} \\ \frac{dP}{dx_2} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = a$$

and

$$\frac{dP}{dx'} = \left[\frac{dP}{dx_1}, \frac{dP}{dx_2}\right] = [a_1, a_2] = a'$$

Consider the quadratic form:

$$Q = \underset{1\text{x}n}{x'} \underset{n\text{x}n}{A} \underset{n\text{x}1}{x}$$

Then the derivative of Q with respect to x or x' is defined as :

$$\frac{dQ}{dx} = 2Ax \text{ and } \frac{dQ}{dx'} = 2x'A$$

For $n = 2$, assuming A is symmetric for simplicity:

$$Q = [x_1, x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\begin{aligned} Q &= [x_1, x_2] \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{12}x_1 + a_{22}x_2 \end{bmatrix} \\ &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \end{aligned}$$

So:

$$\frac{dQ}{dx_1} = 2a_{11}x_1 + 2a_{12}x_2 \text{ and } \frac{dQ}{dx_2} = 2a_{12}x_1 + 2a_{22}x_2$$

Then

$$\begin{aligned} \frac{dQ}{dx} &= \begin{bmatrix} \frac{dQ}{dx_1} \\ \frac{dQ}{dx_2} \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + 2a_{12}x_2 \\ 2a_{12}x_1 + 2a_{22}x_2 \end{bmatrix} = 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \underset{2\text{x}2}{2A} \underset{2\text{x}1}{x} \end{aligned}$$

and

$$\begin{aligned} \frac{dQ}{dx'} &= \left[\frac{dQ}{dx_1}, \frac{dQ}{dx_2}\right] = [2a_{11}x_1 + 2a_{12}x_2, 2a_{12}x_1 + 2a_{22}x_2] \\ &= 2[x_1, x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \\ &= \underset{1\text{x}2}{2x'} \underset{2\text{x}2}{A} \end{aligned}$$

1.3.2. Details of the bivariate case

Writing the sum of squared residuals $u'u = \sum_{t=1}^{T} u_t^2$ out explicitly we get the three terms as in (2.3)

$$\sum Y_t^2 - 2(\beta_1 \sum Y_t + \beta_2 \sum X_t Y_t) + [\beta_1^2 T + \beta_2^2 \sum X_t^2 + 2\beta_1\beta_2 \sum X_t] \quad (1.7)$$

you can see that the last term $\beta' X' X \beta$ in [..] is a quadratic.

In the bivariate model to minimise $u'u$ we have to differentiate the sum of squared residuals, (1.7) above, twice, with respect to β_1 and β_2, to get the 2×1 vector of derivatives and set them equal to zero. The two elements of the vector are

$$\frac{\partial u'u}{\partial \beta_1} = 2\widehat{\beta}_1 T + 2\widehat{\beta}_2 \sum X_t - 2\sum Y_t = 0 \quad (1.8)$$

$$\frac{\partial u'u}{\partial \beta_2} = 2\widehat{\beta}_2 \sum X_t^2 + 2\widehat{\beta}_1 \sum X_t - 2\sum X_t Y_t = 0 \quad (1.9)$$

Check that this corresponds to the matrix formula (2.4). We can also write these as

$$\begin{aligned} -2\sum (Y_t - [\widehat{\beta}_1 + \widehat{\beta}_2 X_t]) &= -2\sum \widehat{u}_t = 0 \\ -2\sum X_t (Y_t - [\widehat{\beta}_1 + \widehat{\beta}_2 X_t]) &= -2\sum x_t \widehat{u}_t = 0 \end{aligned}$$

So $\widehat{\beta} = (X'X)^{-1}X'y$ is a 2×1 vector.

$$\begin{aligned} (X'X) &= \begin{bmatrix} T & \sum X_t \\ \sum X_t & \sum X_t^2 \end{bmatrix} \\ (X'X)^{-1} &= \frac{1}{T\sum X_t^2 - (\sum X_t)^2} \begin{bmatrix} \sum X_t^2 & -\sum X_t \\ -\sum X_t & T \end{bmatrix} \\ \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} &= \frac{1}{T\sum X_t^2 - (\sum X_t)^2} \begin{bmatrix} \sum X_t^2 & -\sum x_t \\ -\sum X_t & T \end{bmatrix} \begin{bmatrix} \sum Y_t \\ \sum X_t Y_t \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum X_t^2 \sum Y_t - \sum X_t \sum X_t Y_t}{T\sum X_t^2 - (\sum X_t)^2} \\ \widehat{\beta}_2 &= \frac{-\sum X_t \sum Y_t + T\sum X_t Y_t}{T\sum X_t^2 - (\sum X_t)^2} \end{aligned}$$

These can be expressed in more intuitive form. From the first equation (1.8)

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum Y_t}{T} - \widehat{\beta}_2 \frac{\sum X_t}{T} \\ &= \overline{Y} - \widehat{\beta}_2 \overline{X}\end{aligned}$$

substituting for $\widehat{\beta}_1$ in the second equation (1.9) can be written

$$\begin{aligned}\widehat{\beta}_2 \sum X_t^2 + (\overline{Y} - \widehat{\beta}_2 \overline{X}) \sum X_t - \sum X_t Y_t &= 0 \\ \widehat{\beta}_2 \sum X_t (X_t - \overline{X}) - \sum X_t (Y_t - \overline{Y}) &= 0\end{aligned}$$

$$\widehat{\beta}_2 = \frac{\sum X_t (Y_t - \overline{Y})}{\sum X_t (X_t - \overline{X})} = \frac{\sum (X_t - \overline{X})(Y_t - \overline{Y})}{\sum (X_t - \overline{X})^2}$$

Dividing top and bottom by T, this is the ratio of the estimated covariance of X_t and Y_t to the estimated variance of X_t.

Note that

$$\begin{aligned}\sum (X_t - \overline{X})(Y_t - \overline{Y}) &= \sum X_t Y_t + T\overline{Y}\overline{X} - \sum X_t \overline{Y} - \sum Y_t \overline{X} \\ &= \sum X_t Y_t + T \frac{\sum X_t}{T} \frac{\sum Y_t}{T} - \sum X_t \frac{\sum Y_t}{T} - \sum Y_t \frac{\sum X_t}{T} \\ &= \sum X_t (Y_t - \overline{Y})\end{aligned}$$

2. LRM in matrix notation

2.1. The model

Multiple regression, with k explanatory variables takes the form

$$Y_t = \beta_1 + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t$$

where $X_{1t} = 1$ all t. This can be written in vector form as:

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

where β and X_t are $k \times 1$ vectors. Or in matrix form as

$$\underset{T\text{x}1}{y} = \underset{T\text{x}k}{X} \underset{k\text{x}1}{\beta} + \underset{T\text{x}1}{u}$$

where y and u are $T \times 1$ vectors and X is a $T \times k$ matrix. For the bivariate regression, this is

$$\begin{bmatrix} Y_1 \\ Y_2 \\ .. \\ Y_T \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ .. & .. \\ 1 & X_T \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ .. \\ u_T \end{bmatrix}.$$

2.2. The assumptions

The assumption about the errors made above, (1.2),

$$E(u_t) = 0$$

can be written in matrix notation $E(u) = 0$, while (1.3), (1.4),

$$\begin{aligned} E(u_t^2) &= \sigma^2 \\ E(u_t u_{t-i}) &= 0;\ i \neq 0 \end{aligned}$$

can be written as

$$E(u\ u') = E \begin{bmatrix} u_1^2 & u_1 u_2 & .. & u_1 u_T \\ u_2 u_1 & u_2^2 & .. & u_2 u_T \\ .. & .. & .. & .. \\ u_T u_1 & u_T u_2 & .. & u_{T^2} \end{bmatrix} = \sigma^2 I_T.$$

This is a $T \times T$ matrix with σ^2 on the diagonal and zeros on the off-diagonals. Distinguish uu' a $T \times T$ matrix and $u'u = \sum u_t^2$ the scalar sum of squared errors. In addition, we assume:

- no exact multicollinearity: the matrix of explanatory variables, X is of full rank k. In the bivariate case this implies that the variance of X_t is non-zero)., and that
- the explanatory variables in X are either (a) non stochastic (b) exogenous, distributed independently of the errors u_t or (c) pre-determined, uncorrelated with the errors u_t. This implies that $E(X'u) = 0$; which in the bivariate case is $E(u_t) = 0$ and $E(X_t u_t) = 0$. Exogeneity is discussed in more detail in section 15.

2.3. Estimators

We use 3 procedures. Method of moments chooses $\widehat{\beta}$ to make a property assumed to hold in the population hold in the sample. least squares chooses $\widehat{\beta}$ to minimise $\sum \widehat{u}_t^2$ and Maximum Likelihood, ML, chooses the $\widehat{\beta}$ most likely to have generated the observed sample. ML also requires an additional assumption about the conditional distribution of y (distribution of u). In the case of the linear regression model with normally distributed errors, the three procedures lead to the same estimator. This is not generally the case.

The method of moments estimator finds the estimator $\widehat{\beta}$ that makes the sample equivalent of $E(X'u) = 0$ which is $X'\widehat{u} = 0$, hold.

$$\begin{aligned} X'\widehat{u} &= X'(y - X\widehat{\beta}) = X'y - X'X\widehat{\beta} = 0 \\ \widehat{\beta} &= (X'X)^{-1}X'y \end{aligned}$$

as long as $(X'X)$ is non-singular, which is ensured by the assumption that the rank of $X = k$.

Least squares takes the estimated linear regression model

$$\underset{T\times 1}{y} = \underset{T\times k}{X}\underset{k\times 1}{\widehat{\beta}} + \underset{T\times 1}{\widehat{u}}$$

and finds the $\widehat{\beta}$ that minimizes the sum of squared residuals $\sum \widehat{u}_t^2 = \widehat{u}'\widehat{u}$:

$$\begin{aligned} \widehat{u}'\widehat{u} &= \left(y - X\widehat{\beta}\right)'\left(y - X\widehat{\beta}\right) && (2.1) \\ &= y'y - \widehat{\beta}'X'y - y'X\widehat{\beta} + \widehat{\beta}'X'X\widehat{\beta} && (2.2) \\ &= y'y - 2\widehat{\beta}'X'y + \widehat{\beta}'X'X\widehat{\beta}. && (2.3) \end{aligned}$$

If A is a $n \times m$ matrix, and B is an $m \times k$ matrix the transpose of the product $(AB)'$ is $B'A'$ a $k \times n$ matrix the product of a $k \times m$ matrix with a $m \times n$ matrix. $A'B'$ is not conformable. $y'X\beta = \beta'X'y$ because both are scalars (1×1 matrices). Scalars are always equal to their transpose. The term $\beta'X'X\beta$ is a quadratic form, i.e. of the form $x'Ax$ above.

The $k \times 1$ vector of derivativea is

$$\begin{aligned} \frac{\partial \widehat{u}'\widehat{u}}{\partial \widehat{\beta}} &= -2X'y + 2X'X\widehat{\beta} = 0 && (2.4) \\ \widehat{\beta} &= (X'X)^{-1}X'y \end{aligned}$$

as in the method of moments case. The second derivative is $2X'X$ which is a positive definite matrix, so this is a minimum. Matrix, A, is positive definite if for any a, $a'Aa > 0$. Matrices with the structure $X'X$ are always positive definite, since they can be written as a sum of squares. Define $z = Xa$, then $z'z = a'X'Xa = \sum z_t^2 > 0$.

2.4. Properties of the estimator

2.4.1. Expected Value of $\widehat{\beta}$

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) \\ \widehat{\beta} &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\ &= \beta + (X'X)^{-1}X'u \end{aligned} \tag{2.5}$$

$$E(\widehat{\beta}) = \beta + E((X'X)^{-1}X'u)$$

since β is not a random variable, and if X and u are independent

$$E((X'X)^{-1}X'u) = E((X'X)^{-1}X')E(u) = 0$$

since $E(u) = 0$. Thus $E(\widehat{\beta}) = \beta$ and $\widehat{\beta}$ is an ubiased estimator of β.

2.4.2. Variance Covariance matrix of $\widehat{\beta}$

From (2.5) we have

$$\widehat{\beta} - \beta = (X'X)^{-1}X'u$$

The variance-covariance matrix of $\widehat{\beta}$ is a $k \times k$ matrix

$$V(\widehat{\beta}) = E(\widehat{\beta} - E(\widehat{\beta}))(\widehat{\beta} - E(\widehat{\beta}))' = E(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'$$

since $\widehat{\beta}$ is unbiased. But from (2.5) we have

$$\widehat{\beta} - \beta = (X'X)^{-1}X'u$$

so

$$\begin{aligned} E(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)' &= E((X'X)^{-1}X'u)((X'X)^{-1}X'u)' \\ &= E((X'X)^{-1}X'u\ u'X(X'X)^{-1} \\ &= (X'X)^{-1}X'E(u\ u')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

since $E(uu') = \sigma^2 I$, σ^2 is a scalar, and $(X'X)^{-1}X'X = I$. Compare this to (1.6) above. We derive the variance covariance matrix conditional on the observed sample, which is why we can take the expected value inside in line 3.

We estimate $V(\widehat{\beta})$ by

$$\widehat{V(\beta)} = s^2(X'X)^{-1}$$

where $s^2 = \widehat{u}'\widehat{u}/(T-k)$.

The square roots of the ith diagonal element of $s^2(X'X)^{-1}$ gives the standard errors of $\widehat{\beta}_i$ the ith elements of $\widehat{\beta}$, which is reported by computer programs.

2.4.3. Predicted Values and residuals

The predicted values are $\widehat{y} = X\widehat{\beta}$.; The estimated residuals are uncorrelated with the explanatory variables by construction:

$X'\widehat{u} = X'(y - X\widehat{\beta}) = X'(y - X(X'X)^{-1}X'y) = X'y - X'y = 0.$

$X'\widehat{u}$ is a set of k equations of the form:

$$\sum_{t=1}^{T} \widehat{u}_t = 0; \sum_{t=1}^{T} x_{2t}\widehat{u}_t = 0; ...; \sum_{t=1}^{T} x_{kt}\widehat{u}_t = 0.$$

the residuals

$$\widehat{u} = y - \widehat{y} = y - X\widehat{\beta} = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = My;$$

where $M = I - P_x$, and $P_x = X(X'X)^{-1}X'$ where P_x is called a projection matrix. Both M and P_x are idempotent, equal to their product $P_xP_x = P_x$ and $MP_x = 0$. So

$$\widehat{u} = My = M(X\beta + u) = MX\beta + Mu = Mu,$$

since $MX\beta = (I - P_x)X\beta = X\beta - X(X'X)^{-1}X'X\beta = X\beta - X\beta = 0$.

So

$$y = P_x y + My$$

it is split into two orthogonal components, the projection of y on X and the orthogonal remainder.

Notice that while the estimated residuals are a transformation of the true disturbances, $\widehat{u} = Mu$,we cannot recover the true disturbances from this equation since M is singular, rank T-k.

2.4.4. Estimating the variance

The sum of squared residuals is:

$$\sum_{t=1}^{T} \widehat{u}_t^2 = \widehat{u}'\widehat{u} = u'M'Mu = u'Mu.$$

To calculate the expected value of the sum of squared residuals, note that $\widehat{u}'\widehat{u}$ is a scalar, thus equal to its trace, the sum of its diagonal elements. Thus using the properties of traces we can write

$$\begin{aligned} E(\widehat{u}'\widehat{u}) &= E(u'Mu) = E(tr(u'Mu)) = E(tr(Muu')) \\ &= tr(M\sigma^2 I) = \sigma^2 tr(M) = \sigma^2(T-k). \end{aligned}$$

Thus the unbiased estimate of σ^2 is $s^2 = \widehat{u}'\widehat{u}/(T-k)$. The last step uses the fact that the Trace of M is

$$\begin{aligned} tr\left[I_T - X(X'X)^{-1}X'\right] &= tr(I_T) - tr\left(X(X'X)^{-1}X'\right) \\ &= tr(I_T) - tr\left((X'X)^{-1}X'X\right) \\ &= tr(I_T) - tr(I_K) = T-k \end{aligned}$$

The standard error of regression, SER, $s = \sqrt{\widehat{u}'\widehat{u}/(T-k)}$, is a measure of the average size of the errors. Other measures that are used are

$$R^2 = 1 - \frac{\sum \widehat{u}_t^2}{\sum (Y_t - \overline{Y})^2}$$

which increases if you add another variable and the adjusted R^2

$$\overline{R}^2 = 1 - \frac{\sum \widehat{u}_t^2/(T-k)}{\sum (Y_t - \overline{Y})^2/(T-1)}$$

which increases if the SER reduces, which it will if you add a variable with a t ratio greater than one.

2.4.5. Gauss-Markov Theorem

This shows that among the class of linear, unbiased estimator $\widehat{\beta}$ has the smallest variance, if

$$\begin{aligned} E(u) &= 0, \\ E(u\ u') &= \sigma^2 I_T, \end{aligned}$$

X is of rank k and exogenous, independent of u. Notice we do not assume normality. This is sometimes expressed as $\widehat{\beta}$ is the best (minimum variance) linear unbiased estimator, BLUE. There may be non-linear or biased estimators with smaller variance. There are a number of different ways to prove this.

Consider any other linear estimator $\widetilde{\beta} = Cy$ where we assume that X and C are fixed (non-stochastic) matrices

$$\begin{aligned} \widetilde{\beta} &= Cy = C(X\beta + u) = CX\beta + Cu \\ E(\widetilde{\beta}) &= CX\beta + CE(u) \end{aligned}$$

so $\widetilde{\beta}$ will be unbiased as long as $CX = I$. Write $\widetilde{\beta} = Cy = ((X'X)^{-1}X' + W)y$, that is $W = C - (X'X)^{-1}X$. Then $CX = I$ implies $((X'X)^{-1}X' + W)X = I$ or $(X'X)^{-1}X'X + WX = I$ or $I + WX = I$. This can only be true if $WX = 0$. This also implies that $X'W' = 0$. Assume this is the case to ensure that $\widetilde{\beta}$ is unbiased.The variance covariance matrix of $\widetilde{\beta}$ is:

$$E(\widetilde{\beta} - E(\widetilde{\beta}))(\widetilde{\beta} - E(\widetilde{\beta}))' = E(\widetilde{\beta} - \beta)(\widetilde{\beta} - \beta)'$$

since $\widetilde{\beta}$ is unbiased by assumption. From above

$$\begin{aligned} \widetilde{\beta} &= \beta + Cu = \beta + ((X'X)^{-1}X' + W)u \\ \widetilde{\beta} - \beta &= (X'X)^{-1}X'u + Wu \end{aligned}$$

$$E(\widetilde{\beta} - \beta)(\widetilde{\beta} - \beta)' = E((X'X)^{-1}X'u + Wu)((X'X)^{-1}X'u + Wu)'$$

When we multiply out the brackets we have four terms:

$$\begin{aligned} E((X'X)^{-1}X'uu'X(X'X)^{-1}) &= \sigma^2(X'X)^{-1} \\ E(Wu\ u'W') &= \sigma^2 WW' \\ E((X'X)^{-1}X'uu'W') &= \sigma^2(X'X)^{-1}X'W' = 0 \\ E(Wuu'X(X'X)^{-1}) &= \sigma^2 WX(X'X)^{-1} = 0 \end{aligned}$$

The last two terms are zero since $WX = X'W' = 0$. So the Variance of any other linear unbiased estimator is

$$\begin{aligned} V(\widetilde{\beta}) &= E(\widetilde{\beta} - (\widetilde{\beta}))(\widetilde{\beta} - (\widetilde{\beta}))' = \sigma^2[(X'X)^{-1} + WW'] \\ &= V(\widehat{\beta}) + \sigma^2 WW' \end{aligned}$$

since WW' is a positive definite matrix for $W \neq 0$, we have shown that in the class of linear unbiased estimators the OLS estimator has the smallest variance.

As noted before Matrix, A, is positive definite if for any a, $a'Aa > 0$. Matrices with the structure $X'X$ are always positive definite, since they can be written as a sum of squares. Define $z = Xa$, then $z'z = a'X'Xa = \sum z_t^2 > 0$.

3. Distributions and maximum likelihood ML

Suppose we have a sample of data of observations on random variables Y_t a scalar and $\mathbf{X}_t$ a $k \times 1$ vector. The joint distribution of the random variables, $Y_t, \mathbf{X}_t$, (e.g. (3.2) below can always be written as the product of the distribution of Y_t conditional on $\mathbf{X}_t$, e.g (3.5) below and the marginal distribution of $\mathbf{X}_t$:

$$D_j(Y_t, \mathbf{X}_t; \theta_j) = D_c(Y_t \mid \mathbf{X}_t; \theta_c) D_m(\mathbf{X}_t; \theta_m) \tag{3.1}$$

θ_j is a vector of parameters of the joint distribution, θ_c of the conditional distribution, θ_m of the marginal. Suppose that we regard the causality as going from $\mathbf{X}_t$ to Y_t, then the parameters of interest are those of the conditional distribution of Y_t, θ_c, which we will usually denote by θ. We can say that $\mathbf{X}_t$ is weakly exogenous if there is no information in the marginal distribution of $\mathbf{X}$ about the parameters of the conditional distribution that we are interested in. Usually we are only interested in the first two moments of the distribution, the conditional expectation (the regression function) and the conditional variance, so in the LRM the parameters of the conditional distribution which we will want to estimate are $\theta_c = \theta = (\beta, \sigma^2)$. Note this definition of exogeneity, which we return to in section 15, is in terms of the distributions of the observables $Y_t, \mathbf{X}_t$ not in terms of the distribution of the unobservable u as in the definition of strict exogeneity that $\mathbf{X}$ is independent of u.

Consider the case where Y_t and $\mathbf{X}_t$ have a joint normal (Gaussian) distribution, so:

$$\begin{bmatrix} Y_t \\ \mathbf{X}_t \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right) \tag{3.2}$$

$\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}'_{yx}$ are $k \times 1$, $\boldsymbol{\Sigma}_{xx}$ is $k \times k$.

If the variables have a joint normal distribution, then the conditional expectation of Y_t is a linear function of $\mathbf{X}_t$:

$$E(Y_t \mid \mathbf{X}_t) = \mu_y + \left[\boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1}\right] (\mathbf{X}_t - \boldsymbol{\mu}_x)$$

Note $[\mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}] = \boldsymbol{\beta}'$ corresponds to $(X'X)^{-1}X'y$. We can decompose y_t into two components, the systematic part given by the conditional expectation and the unsystematic part, the error, which is uncorrelated with $\mathbf{X}_t$ from the properties of conditional expectations. The error is:

$$u_t = Y_t - E(Y_t \mid \mathbf{X}_t) = Y_t - \boldsymbol{\beta}'\mathbf{X}_t$$

which by the properties of conditional expectation is uncorrelated with $\mathbf{X}_t$. So:

$$Y_t = \boldsymbol{\beta}'\mathbf{X}_t + u_t;\ t = 1, 2, ..., T. \tag{3.3}$$

If the random variables are jointly normally distributed and the observations are independent, the conditional variance is a constant:

$$E(Y_t - E(Y_t \mid \mathbf{X}_t))^2 = E(u_t^2) = \sigma^2 = \sigma_y^2 - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}. \tag{3.4}$$

If the joint distribution of Y_t and $\mathbf{X}_t$ is normal, the conditional distribution is also normal, and if the sample is independent we can write the distribution for an observation:

$$\begin{aligned} D_c(Y_t \quad &| \quad \mathbf{X}_t; \theta) \sim IN(\beta'\mathbf{X}_t, \sigma^2) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}(\frac{Y_t - \beta'\mathbf{X}_t}{\sigma})^2\right\} \end{aligned} \tag{3.5}$$

or in matrix form for the whole sample:

$$\begin{aligned} D_c(y \quad &| \quad X; \theta) \sim N(X\beta, \sigma^2 I) \\ &= (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}. \end{aligned} \tag{3.6}$$

We do not need to specify conditional independence in the matrix form, the fact that the variance covariance matrix is $\sigma^2 I$ implies that the conditional covariances between Y_t and Y_{t-i} are zero. In the case of normally distributed variables zero covariance implies independence, this is not generally the case, independence is a stronger assumption than uncorrelated.

In the bivariate case:

$$E(Y_t \mid X_t) = \beta_1 + \beta_2 X_t$$

where $\beta_1 = \mu_y - \beta_2 \mu_X$, and $\beta_2 = \sigma_{xy}/\sigma_{xx}$, where σ_{xy} is the covariance of Y_t and X_t and σ_{xx} the variance of X_t. But we also have

$$E(X_t \mid Y_t) = \gamma_1 + \gamma_2 Y_t$$

where $\gamma_1 = \mu_X - \gamma_2 \mu_y$, $\gamma_2 = \sigma_{xy}/\sigma_{yy}$. Also $\beta_2 \gamma_2 = r^2$, the squared correlation coefficient. Notice these two conditional expectations need not have a causal interpretation. If Y_t was height and X_t weight; we could sensibly ask both what is the expected weight of someone of a particular height and what is the expected height of someone of a particular weight, without implying one caused the other. For prediction we do not need any exogeneity assumptions but to make causal statements we do.

3.1. ML estimation

3.1.1. Introduction

Suppose we have a random variable y with a known probability density function $f(y, \theta)$, where θ is a parameter or vector of parameters. We can use this function to tell us the probability of particular values of y, given known parameters. For instance, given that a coin has a probability of getting a head of $\theta = 0.5$, what is the probability of observing 10 heads in a row? Answer $(0.5)^{10}$. Alternatively, we can use the same formula to tell us the likelihood of particular values of the parameters, given that we have observed a sample of realisations of y, say $y_1, y_2, .., y_T$. If we observe y is ten heads in a row, how likely is it that this sample would be generated by an unbiased coin (i.e $\theta = 0.5$)? Again the answer is $(0.5)^{10}$. In the first case we interpret $f(y, \theta)$ as a function of y given θ. In the second case we interpret $f(y, \theta)$ as a function of θ given y. The maximum likelihood (ML) procedure estimates $\widehat{\theta}$ as the value most likely to have generated the observed sample. In the coin example, $\theta = 0.5$ is very unlikely to have generated the observed sample of 10 heads. $\theta = 1$ is more likely.

If the sample is random, the observations are independent and we can just multiply the probabilities for each observation together as we did in the coin example and write the Likelihood as:

$$L(\theta) = f(y_1, \theta) f(y_2, \theta) ... f(y_T, \theta)$$

We then choose θ that maximises this value for our observed sample $y_1, y_2, ..., y_T$. It is more convenient to work with the logarithm of the likelihood function. Since

logs are a monotonic function the value of θ that maximises the log-likelihood will also maximise the likelihood. Thus the log-likelihood is:

$$LL(\theta) = \sum_{t=1}^{T} \log f(y_t, \theta).$$

$LL(\theta)$ is a scalar, suppose θ is a $k \times 1$ vector. To find the maximum we take the k derivatives of $LL(\theta)$, this is called the score vector and set them to zero:

$$S(\widehat{\theta}) = \frac{\partial LL(\widehat{\theta})}{\partial \theta} = \frac{\partial \sum \log f(y_t, \widehat{\theta})}{\partial \theta} = 0$$

then solve for the value of $\theta, \widehat{\theta}$ that makes the derivatives equal to zero. For simple examples, like the LRM below we can solve these equations analytically, for more complicated examples we solve them numerically. To check that we have found a maximum, we need to check the second order conditions and calculate the kxk matrix of second derivatives:

$$\frac{\partial^2 LL(\theta)}{\partial \theta \ \partial \theta'},$$

evaluated at the true θ. For a maximum this matrix should be negative definite. The information in observation t is the negative of the expected value of the matrix of second derivatives:

$$I_t(\theta) = -E(\frac{\partial^2 LL_t(\theta)}{\partial \theta \ \partial \theta'})$$

which is a symmetric $k \times k$ matrix. The average information matrix in the sample of size T is:

$$I_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} I_t(\theta) = -E(\frac{1}{T} \frac{\partial^2 LL(\theta)}{\partial \theta \ \partial \theta'}).$$

3.1.2. Properties of the ML estimator

A useful result is that for any unbiased estimator (in small samples) or consistent estimator (asymptotically when $T \to \infty$) the inverse of the information matrix provides a lower bound (the Cramer-Rao lower bound) on the variance covariance matrix of the estimator

$$V(\widehat{\theta}) \geq I(\widehat{\theta})^{-1}.$$

where $\geq$ indicates that $V(\widehat{\theta}) - I(\widehat{\theta})^{-1}$ is a non-negative definite matrix.

Under certain regularity conditions the ML estimator $\widehat{\theta}$ is consistent, that is for some small number $\epsilon > 0$

$$\lim_{T \to \infty} \Pr(\mid \widehat{\theta}_T - \theta \mid > \epsilon) = 0.$$

When we evaluate asymptotic distributions we look at $\sqrt{T}(\widehat{\theta} - \theta)$ as $T \to \infty$, because since it is consistent the distribution of $\widehat{\theta}$ collapses to a point and scale the information matrix by T. The ML estimator is asymptotically normally distributed and asymptotically attains the Cramer-Rao lower bound (i.e. it is efficient), it is asymptotically

$$\begin{aligned} \sqrt{T}(\widehat{\theta}_T - \theta) &\to N(0, I(\theta)^{-1}), \\ I(\theta) &= \lim_{T \to \infty} -E(\frac{1}{T}\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'}). \end{aligned}$$

The scaled score $(\sqrt{T})^{-1}S(\theta)$ is also asymptotically normal $N(0, I(\theta))$. We will use these two asymptotic normality properties in testing. In addition, $E(S(\theta)S(\theta)') = T \times I(\theta)$.

ML estimators are also invariant in that for any function of θ, say $g(\theta)$, the ML estimator of $g(\theta)$ is $g(\widehat{\theta})$. Partly because of this ML estimators are not necessarily unbiased. Some are, many are not.

3.2. Non-linear estimation

We distinguish (1) equations which are non-linear in variables because of transformations, like logarithms or powers, but which can be estimated by a linear regression on the transformed data and (2) equations which are non-linear in parameters, where we need a non-linear estimation routine of the type discussed in 3.2. We first consider models that are linear in parameters then ones that are non-linear in parameters.

3.2.1. Logarithms

The most common transformation is logarithms. We often use logarithms of economic variables since

1. prices and quantities are non-negative so the logs are defined

2. the coefficients can be interpreted as elasticities, % change in the dependent variable in response to a 1% change in the independent variable, so the units of measurement of the variables do not matter

3. in many cases errors are proportional to the variable, so the variance is more likely to be constant in logs,

4. the logarithms of economic variables are often closer to being normally distributed

5. the change in the logarithm is approximately equal to the growth rate and

6. lots of interesting hypotheses can be tested in logarithmic models.

7. often effects are proportional, which is captured by logarithmic models.

Normally we use natural logarithms to the base e.

3.2.2. Interpreting regression coefficients with transformed data

Typically we interpret regression coefficients as derivatives or elasticities. Often the derivative or elasticity is not constant, but a function of the variables.

Linear In the standard linear regression between continuous untransformed variables

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

β measures the change in Y_t that result from a one unit change in $X_t : \Delta X_t = 1$. It corresponds to the derivative

$$\frac{\partial Y_t}{\partial X_t} = \beta.$$

β depends on the units that the variables are measured in. Suppose, X_t is per-capita GDP measured in dollars and Y_t is life expectancy in years, then β is the number of extra years of life bought by an extra dollar. The standard error of regression measures the size of a typical error and is in the same units as the dependent variable, here years.

The elasticity is the percentage (proportionate) change in Y_t that results from a one percent change in X_t.

$$\eta = \frac{\partial Y_t / Y_t}{\partial X_t / X_t} = \frac{\partial \log Y_t}{\partial \log X_t}$$

The elasticity is invariant to units, but in the linear case does depend on where we measure it.

$$\eta = \frac{\beta X_t}{Y_t}$$

For a linear relationship, the elasticity is different at every point on the line. A convenient place to measure it is at the typical values, the means of X_t and Y_t.

Log-Log. In a logarithmic regression

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

then β is the elasticity. The standard error of regression measures a typical proportional error (multiply by 100 to get percentage error). To provide a rough comparison with the fit of a linear model, divide the standard error of the linear model by the mean of the dependent variable (assuming the mean is positive and non-zero) which will also give a proportionate error.

Suppose we have a dummy variable in the equation

$$\log Y_t = \alpha + \beta \log X_t + \gamma D_t + u_t$$

where $D_t = 0$ or $D_t = 1$. The effect on Y_t of the dummy variable going from zero to one is $\exp(\gamma) - 1$.

Percentages. Suppose Y_t and X_t are both percentages. For instance, Y_t is the inflation rate, measured in percent, and X is the unemployment rate, also measured in percent. Then β measures the percentage **point** change in Y_t in response to a one percentage **point** change in X_t. If unemployment rises from 1% to 2%, it increases by one percentage point and 100%. The Phillips Curve relationship between inflation and unemployment may be non-linear: the effect on inflation of a one percentage point change in unemployment is much greater when unemployment is 1% than when it is 9%. We can represent this by using the reciprocal of unemployment

$$\begin{aligned} Y_t &= \alpha + \beta X_t^{-1} + u_t, \\ \frac{\partial Y_t}{\partial X_t} &= -\beta X_t^{-2}. \end{aligned}$$

So at 1% unemployment the effect is just $-\beta$, but at 9% unemployment it is $-\beta/9^2$, very small.

Semi-log. One may not log both dependent and independent variables one might regress life expectancy in years on log per-capita income (which is what

Gapminder does), i.e. an equation, sometimes called linear-log, of the form

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

If X_t changes by 1%, then $\Delta X_t / X_t = 0.01$, so Y_t changes by 0.01β.

One can have it the other way round a log-linear model

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

Here a unit change in X_t, $\Delta X_t = 1$, causes a $100\beta\%$ change in Y_t.

Standardised. Statisticians often remove the effect of units of measurement by standardising the data, subtracting the mean and dividing by the standard deviation so the new variables have mean zero and standard deviation one. The coefficients measure how many standard deviation changes in Y result from a one standard deviation change in X. One standard deviation measures a typical change, so these are natural units.

When you do a regression on standardised data it is just a reparameterisation. Observation subscripts have been dropped for clarity.

$$\begin{aligned}
Y &= a + bX_1 + cX_2 + u, \\
\frac{Y-\mu}{\sigma} &= \frac{a-\mu}{\sigma} + \frac{b}{\sigma}X_1 + \frac{c}{\sigma}X_2 + \frac{u}{\sigma}, \\
\frac{Y-\mu}{\sigma} &= \frac{a-\mu+b\mu_1+c\mu_2}{\sigma} + \frac{b}{\sigma}(X_1-\mu_1) + \frac{c}{\sigma}(X_2-\mu_2) + \frac{u}{\sigma}, \\
\frac{Y-\mu}{\sigma} &= \frac{a-\mu+b\mu_1+c\mu_2}{\sigma} + \frac{b\sigma_1}{\sigma}\left(\frac{X_1-\mu_1}{\sigma_1}\right) + \frac{c\sigma_2}{\sigma}(\frac{X_2-\mu_2}{\sigma_2}) + \frac{u}{\sigma}, \\
y &= \alpha + \beta x_1 + \gamma x_2 + \varepsilon.
\end{aligned}$$

One can put restrictions on the regression with the standardised variables, for instance $\beta = \gamma$, which gives

$$y = \alpha + \beta(x_1 + x_2) + \varepsilon$$

is equivalent to

$$\frac{b\sigma_1}{\sigma} = \frac{c\sigma_2}{\sigma}.$$

Whether this makes sense depends on context.

3.2.3. An S shaped function

If our dependent variable is a proportion, p_t taking values between zero and one, the logistic transformation is often used $\ln(p_t/(1-p_t)$. If this is made a function of time,

$$\ln\left(\frac{p_t}{1-p_t}\right) = a + bt + u_t$$

this gives an S shaped curve for p_t over time, which often gives a good description of the spread of a new good (e.g. the proportion of the population that have a mobile phone) and can be estimated by least squares, since it is linear in the parameters. The form of the relationship is

$$p_t = \frac{1}{1 + \exp -(a + bt + u_t)}$$

But we could also set it up as inherently non-linear. If we wanted to estimate a logistic with a saturation level so that $p_t = N_t/K$, where N_t is the number of mobile phone owners and K is the saturation level we could estimate

$$N_t = \frac{K}{1 + \exp -(a + bt)} + \varepsilon_t$$

directly by non-linear least squares. Notice the assumption about the errors is different. In the previous case the error was additive in the logit, here it is additive in the number. In practice, unless the market is very close to saturation it is difficult to estimate K precisely.

Most programs will estimate such non-linear models using an iterative method to find the minimum of the sum of squared residuals or the maximum of the likelihood function.

3.3. Iterative estimation procedures

For the LRM with normal (Gaussian) errors, the ML estimator has a closed form solution. For many of the models we will consider (MA errors, GARCH, Johansen) this is not the case. In such cases one typically requires some sort of iterative procedure to obtain estimates. There are examples in the practicals.

Consider maximising a quadratic function of a vector of k parameters $\boldsymbol{\theta}$, where $\mathbf{C}$ is a positive definite matrix

$$F(\boldsymbol{\theta}) = a + \mathbf{b}'\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}'\mathbf{C}\boldsymbol{\theta}$$

the first order conditions for a maximum and the closed form solution are

$$\begin{aligned} \frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \mathbf{b} - \mathbf{C}\boldsymbol{\theta} = \mathbf{0} \\ \boldsymbol{\theta} &= \mathbf{C}^{-1}\mathbf{b}. \end{aligned}$$

If $F(\boldsymbol{\theta})$ is the likelihood function, or GMM minimand, for a non-linear model, estimation is usually done using an iterative algorithm, where starting from some initial guesses, $\boldsymbol{\theta}_0$ the estimates are updated as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t \tag{3.7}$$

where λ_t is the step size and $\boldsymbol{\Delta}_t$ the direction and this continues until it converges to a maximum.

The most commonly used algorithms are gradient methods..Define the gradient and Hessian

$$\mathbf{g} = g(\boldsymbol{\theta}) = \frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}};\ \mathbf{H} = \frac{\partial^2 F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}.$$

The simplest gradient method is Newton's method based on a linear Taylor series expansion around $\boldsymbol{\theta}_0$

$$\begin{aligned} \frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &\simeq \mathbf{g}_0 + \mathbf{H}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = 0 \\ \boldsymbol{\theta} &\simeq \boldsymbol{\theta}_0 - \mathbf{H}_0^{-1}\mathbf{g}_0. \end{aligned}$$

In (3.7) this sets $\lambda_t = 1$ and $\boldsymbol{\Delta}_t = \mathbf{H}_t^{-1}\mathbf{g}_t$. This often works well, but may be improved by adjusting λ_t. It may be difficult to calculate $\mathbf{H}_t^{-1}$ and it may not be positive definite. In ML examples the outer product gradient, OPG, method uses $\left[\sum_{t=1}^{T} \mathbf{g}_t \mathbf{g}_t'\right]^{-1}$ instead of $(-H)^{-1}$. This is always positive definite and only requires calculating first derivatives. It is the basis of BHHH, Berndt, Hall, Hall & Hausman.

3.3.1. Issues

Thus the issues are: where you start, how you climb up hill and when you stop.

- Start: Try to choose sensible initial values, $\boldsymbol{\theta}_0$, e.g. based on linear approximations and try different values to check for local maxima.

- Climb depends on choice of λ_t and $\mathbf{\Delta}_t$, programs will often switch between procedures. You can choose between 4 in Stata for GARCH.

- Stop: determining whether it has converged to a maximum. $\mathbf{g}_t < \varepsilon$, .and $F_t - F_{t-1} < \varepsilon$ are sensitive to scaling, the units the variables are measured in, $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$ is less sensitive. In Stata GARCH you can set tolerances, ε, for the coefficients, log likelihood and Hessian scaled gradient.

- Whereas in the linear case if the parameter is not identified because $X'X$ is singular, it will be obvious, you get no estimates. This may not be so obvious in the non-linear case and the program may provide estimates even if the likelihood is very flat. This may occur if one has not identified the right sort of non-linearity.

For less well behaved functions there are algorithms like simulated annealing and genetic algorithms.

4. ML estimation of the LRM

For the LRM, the likelihood of the sample is given by (3.6) above, but now interpreted as a function of $\theta = (\beta, \sigma^2)$, the unknown parameters:

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\}.$$

The Log-likelihood function is :

$$LL(\beta, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

and to find the estimates that maximise this we differentiate it with respect to β and σ^2 and set the derivatives equal zero. Notice that

$$u'u = (y - X\beta)'(y - X\beta) = y'y + \beta'X'X\beta - 2\beta'X'y.$$

When we transpose we reverse the order to maintain the correct dimensions and $\beta'X'y = y'X\beta$ because both are scalars. Thus:

$$\frac{\partial LL(\beta, \sigma^2)}{\partial \beta} = -\frac{1}{2\sigma^2}(2X'X\beta - 2X'y) \tag{4.1}$$

and

$$\frac{\partial LL(\beta,\sigma^2)}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}u'u. \tag{4.2}$$

The derivative with respect to σ^2 of $\log(\sigma^2)$ is $1/\sigma^2$ and of $-1/2\sigma^2 = -(2\sigma^2)^{-1}$ is $(-1)(-(2\sigma^2)^{-2})$.

Setting (4.1) equal to zero gives one First Order Conditions, FOC

$$\begin{aligned} -\frac{1}{2\widehat{\sigma}^2}(2X'X\widehat{\beta} - 2X'y) &= 0 \\ \frac{1}{\widehat{\sigma}^2}(X'y - X'X\widehat{\beta}) &= 0 \end{aligned}$$

where the hats denote that these are the values of β and σ^2 that make the FOCs equal to zero. Notice that this can be written

$$\frac{1}{\widehat{\sigma}^2}X'(y - X\widehat{\beta}) = \frac{1}{\widehat{\sigma}^2}X'\widehat{u} = 0 \tag{4.3}$$

the first order conditions choose β that makes the estimated residuals, $\widehat{u} = y - X\widehat{\beta}$, uncorrelated with (orthogonal to) the explanatory variables. This estimate is

$$\widehat{\beta} = (X'X)^{-1}X'y.$$

Notice that as before we need X to be of full rank for the inverse of $(X'X)$ to exist.

Setting (4.2) equal to zero gives

$$-\frac{T}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\widehat{u}'\widehat{u} = 0$$

multiply through by $2\widehat{\sigma}^4$

$$-T\widehat{\sigma}^2 + \widehat{u}'\widehat{u} = 0$$

so our maximum likelihood estimator of the variance is:

$$\widehat{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{T}.$$

The ML estimator is biased and we usually use the unbiased estimator $s^2 = \widehat{u}'\widehat{u}/(T-k)$.

To check second order conditions and construct the information matrix we take derivatives of (4.1) and (4.2)

$$\frac{\partial^2 LL(\beta,\sigma^2)}{\partial\beta\ \partial\beta'} = -\frac{1}{\sigma^2}X'X \tag{4.4}$$

$$\frac{\partial^2 LL(\beta,\sigma^2)}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4}X'u. \tag{4.5}$$

Notice the derivative of $(\sigma^2)^{-1}X'u$ is $-(\sigma^2)^{-2}X'u$. Finally

$$\frac{\partial^2 LL(\beta,\sigma^2)}{\partial(\sigma^2)^2} = \frac{T}{2\sigma^4} - \frac{u'u}{\sigma^6}. \tag{4.6}$$

To get the information matrix we take the negative of the expected value of the second derivative matrix. Notice that $E(X'u) = 0,\ E(u'u) = T\sigma^2$ so the expected value of the final second derivative can be written:

$$\frac{T}{2\sigma^4} - \frac{T\sigma^2}{\sigma^6} = \frac{T}{2\sigma^4} - \frac{T}{\sigma^4} = -\frac{T}{2\sigma^4}$$

$$I(\theta) = -E(\frac{\partial^2 LL(\theta)}{\partial\theta\ \partial\theta'}) = \begin{bmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix}$$

$$I(\beta,\sigma^2)^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{T} \end{bmatrix}.$$

This gives the lower bound for the Variance-covariance matrix for estimators of β, σ^2. Notice that the estimators of β and σ^2 are independent, their covariances are zero. But there will be non-zero covariances between the elements of $\widehat{\beta}$.

4.1. Maximised log likelihood and model selection

We can put the ML estimates into the Log-likelihood function, to get the Maximised Log-Likelihood, MLL, reported by most programs

$$\begin{aligned} MLL &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2}\widehat{u}'\widehat{u} \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\widehat{\sigma}^2) - \frac{T\widehat{\sigma}^2}{2\widehat{\sigma}^2} \\ &= -\frac{T}{2}(\log(2\pi)+1) - \frac{T}{2}\log(\widehat{\sigma}^2) \end{aligned}$$

apart from the constant this is just the negative of half the sample size times the log of the ML estimate of the variance. This can be negative or positive.

The MLL has no interpretation in itself but it can be used in testing between models, see the Likelihood Ratio test below, and in choosing the 'best' model on some 'model selection' criterion. The most popular model selection are R^2 and $\overline{R}^2$, but they are not reliable and have many disadvantages.

Better criteria for choosing between various models are the Akaike Information Criterion ($AIC_i = MLL_i - k_i$); and the Schwarz Bayesian Information Criterion or Posterior Odds Criterion ($SBC = MLL_i - 0.5k_i \log T$); where MLL_i is the maximised log likelihood of model i, k_i is the number of parameters estimated in model i, and T is the sample size. On this definition you choose the model with the largest value.

The SBC tends to choose a more parsimonious model (fewer parameters) than the AIC.

About half of statistics programs (including Microfit) define the AIC in this way, in which case you choose the model with the largest value. The other half (including EViews and Stata) use-2 times these values in which case you choose the model with the smallest value. Be careful, which way they are defined.

5. Testing and exact procedures

Up to now we have focused on estimation, now we move to inference (testing). For that we are going to need distributions related to the normal.

5.1. Distributions; a reminder

By the central limit theorem, under quite general assumptions, many estimators are normally distributed in large samples, whatever the distribution of the original variables.

Linear functions of normally distributed variables are normally distributed. If $y \tilde{} IN(\mu, \sigma^2)$, where $\tilde{}$ means is distributed as then $a + by \tilde{} N(a + b\mu, b^2\sigma^2)$. From this $z = (y - \mu)/\sigma$ is standard normal $z \tilde{} IN(0, 1)$.

The multivariate version is that if the $T \times 1$ vector $Y \sim N(M, \Sigma)$, where M is $T \times 1$, Σ is a $T \times T$ variance covariance matrix, which specifies the dependence between observations. Then for given A and B of order $K \times 1$ and $K \times T$:

$$A + BY \sim N(A + BM, B\Sigma B'). \tag{5.1}$$

Secondly, quadratic forms (sums of squares) of T independent standard normal variables are distributed Chi-squared with T degrees of freedom:

$$\sum_{t=1}^{T}\left(\frac{y_t-\mu}{\sigma}\right)^2 \sim \chi^2(T).$$

The multivariate version for the $T \times 1$ vector $Y \sim N(M, \Sigma)$ is:

$$(Y-M)'\Sigma^{-1}(Y-M) \sim \chi^2(T). \tag{5.2}$$

If the errors are normally distributed, the sum of squared standardised errors $u'u/\sigma^2$ are distributed as $\chi^2(T)$, but the sum of squared standardised residuals $\widehat{u}'\widehat{u}/\sigma^2 = u'Mu/\sigma^2$ are $\chi^2(rankM) = \chi^2(T-k)$. Alternatively

$$(T-k)(\frac{s^2}{\sigma^2}) \sim \chi^2(T-k).$$

A (Student's) t distribution with n degrees of freedom is given by

$$t(n) = z/\sqrt{\frac{\chi^2(n)}{n}}.$$

Fisher's F distribution is the ratio of two independent Chi-squared divided by their degrees of freedom.

$$F(n_1, n_2) = \frac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2}.$$

5.2. Test procedures

Suppose that we have prior information on θ, which suggests that elements of the parameter vector take specified values, such as zero or one or are linked by other restrictions. We wish to decide whether to accept or reject this hypothesis, called H_0, not knowing whether it is true or false. The possible outcomes are:

	H_o True	H_o False
Accept H_0	$\surd$	Type II error
Reject H_0	Type I error	$\surd$

The Neyman-Pearson approach to testing involves:

(a) a null hypothesis usually called H_0; e.g. for a scalar parameter: $H_0 : \beta = 1$;

(b) an alternative hypothesis, e.g. $H_1 : \beta \neq 1$, this is a two sided alternative, a one sided alternative would be $\beta < 1$;

(c) a test statistic, which does not depend on the true value of the parameters (is pivotal), (e.g. $(\widehat{\beta} - 1)/SE(\widehat{\beta})$, where $SE(\widehat{\beta})$ is the estimated standard error of $\widehat{\beta}$) with a known distribution when the null hypothesis is true (e.g. a central t distribution);

(d) a specified size α, the chosen probability of Type I error (rejecting H_0 when it is true) usually 0.05;

(e) critical values so that if the null hypothesis is true the probability of lying outside the critical values is α;

(f) a power function which gives the probability of rejecting the null as a function of the true (unknown) value of β. The power of a test is the probability of rejecting H_0 when it is false (one minus the probability of type two error).

The procedure is: to accept (not reject) H_0 if the test statistic lies within the critical values and to reject H_0 if the test statistic lies outside the critical values.

The results can also be presented as p values, which can be (loosely) thought of as giving the probability of getting a test statistic of that value or greater if the model were correct and the hypothesis were true.[1] If the p value is small, less than the chosen size (probability of rejecting null when true), e.g. 0.05, then the null hypothesis is rejected: a true hypothesis is unlikely to have generated a test statistic of that value..

5.3. Joint and individual tests

Suppose that we have a model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t.$$

We could do two individual tests on the significance of the coefficients: $H_0^2 : \beta_2 = 0$ and $H_0^3 : \beta_3 = 0$, using their t statistics. We could also do a joint test that they are both zero; $H_0^J : \beta_2 = \beta_3 = 0$. This F test, for the hypothesis that all the slopes are zero, is done automatically by most prgograms.

[1] The American Statistical Association has a statement on p values and significance testing, since they are very often misunderstood. They say "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

Individual t tests and joint F test tests may give conflicting answers. The two t tests could reject each of $\beta_2 = 0$ and $\beta_3 = 0$, i.e. both are significant, but the F test not reject that they are both equal to zero: $\beta_2 = \beta_3 = 0$. The variables are cancelling each other out and if we have one we have to have the other. Conversely, they could be individually insignificant but jointly significant. This is more common and happens when they are highly correlated, we can drop one of them, but not both.

5.3.1. Distinguish significance and importance

The test asks whether the difference of the estimate from the null hypothesis could have arisen by chance, it does not tell you whether the difference is important, therefore you should distinguish substantive (economic) importance from statistical significance. A coefficient may be statistically significant because it is very precisely estimated but so small as to be of no economic importance. Conversely the coefficient may be large in economic terms but have large standard errors so not be statistically significant. You need to understand the units and the context to judge whether an estimate is large in economic terms.

The test statistic and p value are conditional on the model and data used. It is also useful to think of a test as informing a decision, accepting or rejecting the null and considering the costs of the two sorts of mistakes. The costs can be embodied in some form of loss function or utility function.

5.4. Exact Tests

In the LRM with linear restrictions we can derive small sample tests, where we know the exact distribution, rather than having to use asymptotic approximations. Suppose, our null hypothesis is a set of m linear restrictions of the form $R\beta = q$ or $R\beta - q = 0$, where R and q are known and of order $m \times k$ and $m \times 1$ respectively. The unrestricted model has k parameters, the restricted model k-m, each restriction reduces the number of parameters we estimate. In the case where m=k, all the parameters are specified, R is an identity matrix and the restrictions are $\beta = q$.

Since

$$\widehat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

and the restrictions are linear

$$(R\widehat{\beta} - q) \sim N(R\beta - q, \sigma^2 R(X'X)^{-1}R')$$

Under $H_0 : R\beta - q = 0$

$$(R\widehat{\beta} - q) \sim \ N(0, \sigma^2 R(X'X)^{-1}R')$$

and

$$(R\widehat{\beta} - q)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\widehat{\beta} - q) \ \sim \ \chi^2(m).$$

Notice that this is a special case of the Wald test statistic below and is of the same form. This is not yet a test statistic because it depends on the unknown σ^2, but we know $(T-k)s^2/\sigma^2 \sim \chi^2(T-k)$ and that for independent Chi-squares:

$$\frac{\chi^2(m)/m}{\chi^2(T-k)/(T-k)} \ \sim F(m, T-k)$$

so

$$\frac{(R\widehat{\beta} - q)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\widehat{\beta} - q)/m}{[(T-k)s^2/\sigma^2]/(T-k)} \ \sim \ F(m, T-k)$$

or since the two unknown σ^2 cancel:

$$\frac{(R\widehat{\beta} - q)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta} - q)/m}{s^2} \ \sim F(m, T-k).$$

This provides us with a test statistic. In practice it is easier to calculate it from another way of writing this formula. Define the unrestricted and restricted estimated equations as

$$y = X\widehat{\beta} + \widehat{u}; \quad \text{and} \quad y = X\beta^* + u^*$$

then

$$\frac{(u^{*\prime}u^* - \widehat{u}'\widehat{u})/m}{\widehat{u}'\widehat{u}/(T-k)} \sim F(m, T-k),$$

the ratio of (a) the difference between the restricted and unrestricted sum of squared residuals divided by the number of restrictions to (b) the unbiased estimate of the unrestricted variance. Computer programs automatically print out a test for the hypothesis that all the slope coefficients in a linear regression are zero, this is $F(k-1, T-k)$.

5.5. Non nested tests

Hypothesis tests require the two models being compared to be 'nested': one model (the restricted model) must be a special case of the other (the unrestricted or maintained model). In many cases we want to compare 'non-nested' models, e.g.

$$\begin{aligned} M_1 &: y_t = a_1 + b_1 x_t + u_{1t} \\ M_2 &: y_t = a_2 + c_2 z_t + u_{2t} \end{aligned}$$

where x_t and z_t are different scalar variables. There are no restrictions on M_1 that will give M_2 and vice-versa. We could nest them both in a general model:

$$M_3 : y_t = a_3 + b_3 x_t + c_3 z_t + u_{3t}.$$

The restriction $c_3 = 0$ gives M_1; so rejecting the restriction $c_3 = 0$ rejects M_1
The restriction $b_3 = 0$ gives M_2; so rejecting the restriction $b_3 = 0$ rejects M_2.
This gives four possible outcomes:

1. Reject M_1, do not reject $M_2 : c_3 \neq 0; b_3 = 0$;
2. Reject M_2, do not reject $M_1 : b_3 \neq 0; c_3 = 0$;
3. Reject both; $b_3 \neq 0; c_3 \neq 0$;
4. Do not reject either: $b_3 = 0; c_3 = 0$.

There are a range of other non-nested tests available (Microfit has a large selection) but they all give rise to the same four possibilities. If x_t and z_t are highly correlated case 4 is quite likely as noted above.

Model selection criteria like the AIC and BIC can be used both for nested or non-nested models. When comparing nested models the BIC can be interpreted as adjusting the size of the test (probability of type I error) with the number of observations. Suppose we have two models, $M1$ and $M2$, such that $M1$ has k parameters and is nested in $M2$ which has an extra variable and $k+1$ parameters. An LR test at the 5% level chooses $M2$ if $2(MLL_2 - MLL_1) > 3.84$. The BIC chooses $M2$ if $2(MLL_2 - MLL_1) > \ln T$.

This is sensible. As the sample size grows the standard error of the parameter falls and with a large enough sample any hypothesis will be rejected at a constant size even if the deviation from the hypothesis is tiny.

6. Asymptotic Test procedures: W, LR, LM

6.1. Principles

When we cannot derive exact, small sample standard errors or critical values for our tests, we often use asymptotic approximations using the central limit theorem to get asymptotic distributions. For instance, for t ratios the asymptotic distribution is normal with 5% critical value of 1.96 whereas for 10 degrees of freedom the exact critical value is 2.23. So we are more likely to reject using the asymptotic approximation.

We saw above that the ML estimates are those which maximise $LL(\theta)$, i.e. the $\widehat{\theta}$, which make

$$\frac{\partial LL(\theta)}{\partial \theta} = S(\widehat{\theta}) = 0$$

where $S(\widehat{\theta})$ is the score vector, the derivatives of the LL with respect to each of the k elements of the vector θ evaluated at the values, $\widehat{\theta}$, which make $S(\theta) = 0$. We will call these the unrestricted estimates and the value of the Log-likelihood at $\widehat{\theta}$, $LL(\widehat{\theta})$.

Suppose theory suggests $m \leq k$ prior restrictions (possibly non-linear) of the form $R(\theta) = 0$, where $R(\theta)$ is an $m \times 1$ vector. If $m = k$, theory specifies all the parameters and there are none to estimate. The restricted estimates maximises

$$\pounds = LL(\theta) - \lambda' R(\theta)$$

where λ is a $m \times 1$ vector of Lagrange Multipliers. The first order condition, FOC, is

$$\frac{\partial \pounds}{\partial \theta} = \frac{\partial LL(\theta)}{\partial \theta} - \frac{\partial R(\theta)}{\partial \theta}\lambda = 0$$

Write the $kx1$ vector $\partial LL(\theta)/\partial\theta$ as $S(\theta)$ and the $k \times m$ matrix $\partial R(\theta)/\partial\theta$ as $F(\theta)$ then at the restricted estimate θ^*, which makes the FOC hold

$$S(\theta^*) - F(\theta^*)\lambda^* = 0$$

Notice that at θ^* the derivative of the Log-likelihood function with respect to the parameters is not equal to zero but to $F(\theta^*)\lambda^*$. The value of the Log-likelihood at θ^* is $LL(\theta^*)$ which is less than or equal to $LL(\widehat{\theta})$.

If the hypotheses (restrictions) are true:

(a) the two log-likelihoods should be similar, i.e. $LL(\widehat{\theta}) - LL(\theta^*)$ should be close to zero;

(b) the unrestricted estimates should satisfy the restrictions $R(\widehat{\theta})$ should be close to zero (note $R(\theta^*)$ is exactly zero by construction);

(c) the restricted score, $S(\theta^*)$, should be close to zero (note $S(\widehat{\theta})$ is exactly zero by construction) or equivalently the Lagrange Multipliers λ^* should be close to zero, the restrictions should not be binding.

6.2. Test procedures

These implications are used as the basis for three types of test procedures. The issue is how to judge 'close to zero'? To judge this we use the asymptotic equivalents of the linear distributional results used above in the discussion of the properties of the LRM. Asymptotically, by the central limit theorem, the ML estimator is multivariate normal

$$\widehat{\theta} \ \widetilde{a} \ N(\theta, I(\theta)^{-1})$$

asymptotically the scalar quadratic form is chi-squared

$$(\widehat{\theta} - \theta)' I(\theta)(\widehat{\theta} - \theta) \ \widetilde{a} \ \chi^2(k).$$

and asymptotically $R(\widehat{\theta})$ is also normal

$$R(\widehat{\theta}) \ \widetilde{a} \ \ N(R(\theta), F(\theta)' I(\theta)^{-1} F(\theta))$$

where $F(\theta) = \partial R(\theta)/\partial\theta$. This gives us three procedures for generating asymptotic test statistics for the m restictions $H_0 : R(\theta) = 0$; each of which are asymptotically distributed $\chi^2(m)$, when the null hypothesis is true:

(a) Likelihood Ratio Tests

$$LR = 2(LL(\widehat{\theta}) - LL(\theta^*)) \sim \chi^2(m)$$

(b) Wald Tests

$$W = R(\widehat{\theta})'[F(\theta)' I(\theta)^{-1} F(\theta)]^{-1} R(\widehat{\theta}) \sim \chi^2(m)$$

where the term in [...] is an estimate of the variance of $R(\widehat{\theta})$ and $F(\theta) = \partial R(\theta)/\partial\theta$.

(c) Lagrange Multiplier (or Efficient Score) Tests where $\partial LL(\theta)/\partial\theta = S(\theta)$

$$LM = S(\theta^*)' I(\theta^*)^{-1} S(\theta^*) \sim \chi^2(m).$$

The Likelihood ratio test is straightforward to calculate when both the restricted and unrestricted models have been estimated. The Wald test only requires the unrestricted estimates. The Lagrange Multiplier test only requires the restricted estimates. For the LRM, the inequality W>LR>LM holds, so you are more likely to reject using W.

In the LRM, the LM test is usually calculated using regression residuals as is discussed below under diagnostic tests. The Wald test is not invariant to how you write non-linear restrictions. Suppose $m = 1$, and $R(\theta)$ is $\theta_1\theta_2 - \theta_3 = 0$. This could also be written $\theta_1 - \theta_3/\theta_2 = 0$ and these would give different values of the test statistic. The former form, using multiplication rather than division, is usually better.

6.3. Monte Carlo and bootstrap methods

In some cases the asymptotic distribution provides a very poor approximation in samples of the size that are typically available or one cannot derive the asymptotic distribution analytically.[2] In this case one uses numerical procedures to provide better estimates for standard errors or critical values.

Monte Carlo methods simulate the distributions by drawing random numbers. For instance, the t statistic for testing $H_0 : \rho = 1$, against $H_1 : \rho < 1$ in the model $y_t = \rho y_{t-1} + \varepsilon_t$ has a non standard distribution. To get the critical values, you draw a sequence of $T + N$ random variables, ε_t^1, usually from a normal distribution, but it could be from some other distribution. Then starting from $y_0 = 0$, generate $y_1 = y_0 + \varepsilon_1$, $y_2 = y_1 + \varepsilon_2$, ...,$y_{T+N} = y_{T+N-1} + \varepsilon_t$. Discard the first N observations to remove the effect of the initial condition. This gives you a sample of T observations, y_t^1 from which you estimate $\widetilde{\rho}^1$.its standard error and t statistic for the hypothesis $\rho = 1$, $t\left(\widetilde{\rho}^1 = 1\right)$. You do this R times, getting estimates $t\left(\widetilde{\rho}^r = 1\right)$. From this empirical distribution you determine the critical value below which lie a proportion α, (e.g. 5%) of the t statistics fall. This is a one sided test.

Bootstrap methods use the estimates available. Suppose that we estimate a regression $Y_t = \alpha + \beta X_t + u_t$, from data $t = 1, 2, ..., T$. If $u_t \sim IN(0, \sigma^2)$ then $\widehat{\beta}$ is distributed as $t(T - 2)$ in small samples and normal in large samples with, $\widehat{\beta}\tilde{}N(\beta, \sigma^2/\sum x_t^2)$, where $x_t = X_t - \overline{X}$. If u_t is not normally distributed but is independent with expected value zero and constant variance σ^2, we can still use

[2] For instance the distributions of unit root and cointegration tests involve integrals of Brownian motions, for which no closed form solutions are available.

$\widehat{\beta}\tilde{}N(\beta, \sigma^2/\sum x_t^2)$ as an asymptotic approximation relying on the central limit theorem. If the constant variance assumption, $E(u_t^2) = \sigma^2$ fails, then the variance of $\widehat{\beta}$ is not $\sigma^2/\sum x_t^2$, but we can use a robust estimator, see 7.2.3. Again we are relying on asymptotic results and the robust estimators of the variance of $\widehat{\beta}$ may have very poor small sample properties.

An alternative procedure is to generate the sampling distribution numerically using the estimated residuals. From the estimated residuals $\widehat{u}_t$, you randomly choose, with replacement, a new sample,$\widetilde{u}_t^1$ of T observations. Because sampling is with replacement, some residuals may be duplicated and some residuals may not appear in this sample. You then construct a new sample of $\widetilde{Y}_t^1 = \widehat{\alpha} + \widehat{\beta} X_t + \widetilde{u}_t^1$. Using this sample, you get a new estimate of β, $\widetilde{\beta}^1$. You repeat this R times, getting $\widetilde{\beta}^r$, $r = 1, 2, ..., R$ where R is a large number. You then have a sampling distribution for your estimate from which you can calculate standard errors, confidence intervals, etc.

This procedure will reflect the effect of any non-normality or heteroskedasticity in the estimated residuals on the distribution of the estimates. However, because of the way the residuals are sampled it will not reflect any serial correlation in the residuals, because the order has been lost. There are other forms of bootstrap which can allow for that.

Typically you would need many more replications, a larger value of R, to determine the critical values for a test statistic than to estimate bootstrap standard errors.

7. Problems: consequences and cures

7.1. Regression Coefficients

In the Gaussian LRM, where y is a $T \times 1$ vector and X is a $T \times k$ full rank matrix of exogenous variables, then conditional on X,

$$y \sim N(X\beta, \sigma^2 I)$$

and the ML estimator is

$$\widehat{\beta} = (X'X)^{-1}X'y$$

a linear function of y, $\widehat{\beta}$ is normally distributed using (5.1):

$$\begin{aligned} \widehat{\beta} &\sim N\{(X'X)^{-1}X'X\beta,\ (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1}\} \\ &\sim N\{\beta, \sigma^2(X'X)^{-1}\} \end{aligned}$$

This indicates (1) $\widehat{\beta}$ is unbiased, $E(\widehat{\beta}) = \beta$, (2) it is fully efficient, its variance covariance matrix attains the lower bound obtained above $\sigma^2(X'X)^{-1}$. We generally estimate the variance covariance matrix by $s^2(X'X)^{-1}$, where $s^2 = \widehat{u}'\widehat{u}/(T-k)$, the unbiased estimator. The square roots of the diagonal elements of this matrix give the standard errors of the individual regression coefficients, e.g. β_i and the off diagonal elements give the covariances between regression coefficients, e.g. $Cov(\beta_i, \beta_j)$.

The estimates only have good properties if a number of assumptions hold and it is important to test those assumptions. The tests for whether the assumptions hold are called diagnostic or misspecification tests. Failure on a particular diagnostic test (rejection of the null that the model is well specified) only indicates that the model is sick, it does not tell you what the illness is. For instance, if you have chosen the wrong functional form you may fail tests for serial correlation. Apart from the structural stability tests most of these tests are Lagrange Multiplier tests which involve auxiliary regressions using the residuals from the first stage regressions. These tests ask whether the residuals have the properties we would expect if the assumptions were true. The null hypothesis is always that the assumptions are true, the model is well specified. Thus if the p value for the test is greater than 0.05, you can accept the hypothesis that the model is well specified at the 5% level.

There are a very large numbers of these tests. Those for serial correlation and non-linearity use the residuals as the dependent variable in an auxilliary regression. Those for heteroskedasticity use the squared residuals as the dependent variable. Those for normality which check that the third and fourth moments of the residuals have the values they should have under normality.

For each null, e.g. constant variance (homoskedasticity) there are a large number of different alternatives (ways that the variance changes) thus lots of different tests for heteroskedasticity of different forms. Although the justification of these tests is asymptotic, versions which use F tests and degrees of freedom adjustment seem to work well in practice.

Always inspect graphs of actual and predicted values and residuals.

Before considering the diagnostic tests, we first consider the consequences of the failures of the assumptions.

7.2. What happens when assumptions fail

7.2.1. Rank(X)$\neq k$

If X is not of full rank k, because there is an exact linear dependency between some of the variables, the OLS/ML estimates of β are not defined and there is said to be exact multicollinearity.[3] The model should be respecified to remove the exact dependency. When there is high, though not perfect, correlation between some of the variables there is said to be multicollinearity. This does not involve a failure of any assumption.

7.2.2. X not exogenous

If the X are not strictly exogenous, independent of u, the estimates of β are biased, though if the X are predetermined, uncorrelated with u, (e.g. lagged dependent variables where the disturbance term is not serially correlated), they will remain consistent. Otherwise, the estimator will be biased and inconsistent. Even if the exogeneity assumption fails least squares gives the best (minimum variance) linear predictor of y.

In certain circumstances endogeneity, the failure of the exogeneity assumptions can be dealt with by the method of Instrumental Variables discussed below in section 7. There are three man causes of endogeneity, two of them simultaneity and measurement errors in the independent variable are discussed later, the third omitted variables. is considered now.

Omitted variables Suppose the data are generated by

$$y_t = \alpha + \beta x_t + \gamma z_t + u_t \tag{7.1}$$

and you omit z_t, a $h \times 1$ vector and estimate

$$y_t = a + b x_t + v_t. \tag{7.2}$$

What is the relationship between the estimates? Suppose we describe the relation between the omitted and included right hand side variables by the regression:

$$z_t = c + d x_t + w_t \tag{7.3}$$

[3]Stata does this automatically by dropping variables, though it may not drop the most appropriate variable.

We can always do this, if they are unrelated $d = 0$. If you replace z_t in (7.1) by the right hand side of (7.3) you get:

$$y_t = \alpha + \beta x_t + \gamma(c + dx_t + w_t) + u_t$$
$$y_t = (\alpha + \gamma c) + (\beta + \gamma' d)x_t + (\gamma w_t + u_t).$$

Thus $b = (\beta + \gamma d)$ and $v_t = (\gamma w_t + u_t)$. The coefficient of x_t in (7.2) will only be an unbiased estimator of β, the coefficient of x_t in (7.1) if either $\gamma = 0$ (z_t really has no effect on y_t) or $d = 0$, (there is no correlation between the included and omitted variables). Notice that v_t also contains the part of z_t that is not correlated with x_t, w_t, and there is no reason to expect w_t to be serially uncorrelated or homoskedastic. Thus misspecification, omission of z_t, may cause the estimated residuals to show these problems. This generalises easily to x_t and z_t being vector

7.2.3. $E(uu') \neq \sigma^2 I$

If $Var(y \mid X) = E(uu') = \sigma^2\Omega$, that is its variance covariance matrix is not $\sigma^2 I$, there are two possible problems: the variances (diagonal terms of the matrix) are not constant and equal to σ^2 (heteroskedasticity) and/or the off diagonal terms, the covariances, are not equal to zero (failure of independence, serial correlation, autocorrelation). Under these circumstances, $\widehat{\beta}$ remains unbiased but is not minimum variance (efficient). Its variance-covariance matrix is not $\sigma^2(X'X)^{-1}$, but $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$. Corrected variance-covariance matrices are available in most packages (White Heteroskedasticity consistent covariance matrices or Newey-West heteroskedasticity and autocorrelation consistent, HAC, ones). These use estimates of $X'\Omega X$ in the formula. Use Options on the equation menu in EViews to get HAC robust standard errors. Notice that residual serial correlation or heteroskedasticity may indicate not that there is some covariances between the true disturbances but that the model is wrongly specified, e.g. variables are omitted, see below. When it is appropriate to model the disturbance structure in terms of Ω, Generalised Least Squares, discussed below, can be used. Often residual serial correlation or heteroskedasticity should lead you to respecify the model rather than to use Generalised Least Squares.

7.2.4. Generalised Least Squares

If $y \sim N(X\beta, \sigma^2\Omega)$ its distribution is given by:

$$2\pi^{-T/2} \mid \sigma^2\Omega \mid^{-1/2} \exp\left\{-\frac{1}{2}(y - X\beta)'(\sigma^2\Omega)^{-1}(y - X\beta)\right\}.$$

Notice that when $\Omega = I$, then the term in the determinant, $\mid \sigma^2\Omega \mid^{-1/2}$ is just $(\sigma^2)^{-T/2}$.

If Ω is a known matrix the Maximum Likelihood Estimator is the Generalised Least Squares estimator $\beta^{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, with variance-covariance matrix $V(\beta^{GLS}) = \sigma^2(X'\Omega^{-1}X)^{-1}$. Whereas the OLS estimator chooses β to make $(\sigma^2)^{-1}X'\widehat{u} = 0$, the GLS estimator chooses β to make $(\sigma^2)^{-1}X'\Omega^{-1}\widetilde{u} = 0$, where $\widetilde{u} = y - X\beta^{GLS}$. In practice GLS is implemented by finding a 'transformation matrix' P such that $P'P = \Omega^{-1}$ and $P\Omega P' = I$. This can always be done since Ω must be a positive-definite symmetric matrix. You then transform the data by premultiplying the equation by P

$$Py = PX\beta + Pu$$

$$y^* = X^*\beta + u^*$$

where $y^* = Py$, etc. OLS is then applied to the transformed data, which is fully efficient since

$$E(u^*u^{*\prime}) = E(Puu'P') = PE(uu')P' = P(\sigma^2\Omega)P' = \sigma^2 P\Omega P' = \sigma^2 I.$$

In practice, Ω is rarely known completely, but it may be known up to a few unknown parameters. These can be estimated and used to form an estimate of Ω, and P. This is known as the Feasible or Estimated GLS estimator. It generally differs from the exact ML estimator. The text books give large number of examples of FGLS estimators, differing in the assumed structure of Ω. But in many cases it is better to respecify the model or correct the standard errors than to apply FGLS to try and fix problems with the residuals.

7.2.5. u not Gaussian

If normality does not hold the Least Squares estimator, $\widehat{\beta} = (X'X)^{-1}X'y$, is no longer the Maximum Likelihood estimator and is not fully efficient, but it is the minimum variance estimator in the class of linear unbiased estimators (biased or non-linear estimators may have smaller variances). In small samples, the tests below will not have the stated distributions, though asymptotically they will be normal, because of the central limit theorem. If the form of the distribution is known (e.g. a t distribution) maximum likelihood estimators can be derived for that particular distribution and they will be different from the OLS estimators. EViews and Microfit will estimate model with errors distributed as t, under the GARCH options, see the applied exercise. For small degrees of freedom, the t has fatter tails, when the degrees of freedom are around 30 it is close to normal.

8. Diagnostic testing

8.1. Structural stability

The assumption that the parameters are constant over the sample is crucial and there are a variety of tests for constancy. Two are special cases of the F test for linear restrictions above.

Suppose that we have a sample of data for $t = 1, 2, ..., T$ and we believe that the relationship may have shifted at period T_1 within the sample, and both sub-samples have more than k observations. The unrestricted model estimates separate regressions for each sub period $t = 1, 2, ..., T_1$ and for $t = T_1 + 1, T_1 + 2, ..., T$; define $T_2 = T - T_1$: X_1 a $T_1 \times k$ matrix, X_2 a $T_2 \times k$ matrix, etc. Then the models for the two subperiods are:

$$\begin{aligned} y_1 &= X_1\beta_1 + u_1 \\ y_2 &= X_2\beta_2 + u_2 \end{aligned}$$

where we assume $u_i \sim IN(0, \sigma^2)$, $i = 1, 2$; the variances are the same in both periods. The unrestricted residual sum of squares is $(\widehat{u}_1'\widehat{u}_1 + \widehat{u}_2'\widehat{u}_2)$ with degrees of freedom $T - 2k$. The restricted model is

$$y = X\beta + u$$

where X is a $T \times k$ matrix. The restricted residual sum of squares is $\widehat{u}'\widehat{u}$ with degrees of freedom $T - k$. The null hypothesis is that $\beta_1 = \beta_2$, k restrictions and the test statistic is

$$\frac{[\widehat{u}'\widehat{u} - (\widehat{u}_1'\widehat{u}_1 + \widehat{u}_2'\widehat{u}_2)]/k}{(\widehat{u}_1'\widehat{u}_1 + \widehat{u}_2'\widehat{u}_2)/(T - 2k)} \sim F(k, T - 2k).$$

This is known as Chow's first or breakpoint test. He also suggested a second 'predictive failure' or forecast test for the case where there $T_2 < k$ though it can be used whether or not there are enough observations to estimate the second period model. The test statistic is:

$$\frac{[\widehat{u}'\widehat{u} - \widehat{u}_1'\widehat{u}_1)]/T_2}{\widehat{u}_1'\widehat{u}_1/(T_1 - k)} \sim F(T_2, T_1 - k).$$

This tests the hypothesis that in

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ X_2 & I \end{bmatrix} \begin{bmatrix} \beta_1 \\ \delta \end{bmatrix} + \begin{bmatrix} u_1 \\ 0 \end{bmatrix}$$

δ the $T_1 \times 1$ vector of forecast errors are not significantly different from zero. This has a dummy variable for each observation in the second period.

Chow's first test assumes that the variances in the two periods are the same. This can be tested using the Variance Ratio or 'Goldfeld-Quandt' test:

$$\frac{s_1^2}{s_2^2} = \frac{\widehat{u}_1'\widehat{u}_1/(T_1-k)}{\widehat{u}_2'\widehat{u}_2/(T_2-k)} \sim F(T_1-k, T_2-k).$$

You should put the larger variance on top so the F statistic is greater than unity. Notice that although this is an F test, it is not a test of linear restrictions on the regression parameters like the other F tests we have used. This is a test for a specific form of heteroskedasticity, tests for other types of heteroskedasticity are given below. If the variances are equal, the two equations can be estimated together using dummy variables

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

this will give the same estimates of the coefficients as running two separate regressions, but different estimators of the standard errors: this form imposes equality of variables, the separate regressions do not. For testing differences of individual coefficients, this can be rewritten

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ X_2 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 - \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{8.1}$$

Then some coefficients can be allowed to differ and others kept the same between periods. For instance, suppose that $k = 3$, and we define a dummy variable $D2 = 1$ for period 2, zero for period one; then (8.1) is equivalent to estimating on the whole period

$$y_t = \beta_{11} + \beta_{12}x_{2t} + \beta_{13}x_{3t} + \gamma_1 D2_t + \gamma_2 D2_t x_{2t} + \gamma_3 D2_t x_{3t} + u_t,$$

where $\gamma_1 = \beta_{21} - \beta_{11}$; $\gamma_2 = \beta_{22} - \beta_{12}$; $\gamma_3 = \beta_{23} - \beta_{13}$.; Then if $H_0 : \gamma_2 = \gamma_3 = 0$ was not rejected, you would just need to include the dummy variable for and intercept shift and not the two interaction terms.

Statistical packages also include a range of other ways to investigate structural stability of the parameters using recursive residuals such as the CUSUM and CUSUMSQ diagrams, which are particularly useful when one is uncertain about the breakpoint. These are presented as graphs of the statistics within two

lines. If the graphs cross the lines it indicates structural instability. They also present recursive estimates, where the parameters are estimated on the first $k+1$ observations, the first $k+2$ and so on up to T. Breaks may show up in the estimates. Thre are also Andrews-Quandt unknown breakpoint tests which identify the most likely place for a break. Tests with an unknown break-point will have much less power than tests with a known break-point.

8.2. Serial Correlation

Suppose the data were generated by:

$$\begin{aligned} y_t &= \beta' x_t + v_t;\ v_t = \rho v_{t-1} + u_t \\ y_t &= \beta' x_t + \rho v_{t-1} + u_t \end{aligned}$$

where u_t is a 'well-behaved' disturbance distributed $IN(0, \sigma^2)$; but we estimate

$$y_t = b' x_t + v_t$$

the estimated residuals

$$\widehat{v}_t = y_t - \widehat{b}' x_t = \beta' x_t + \rho v_{t-1} + u_t - \widehat{b}' x_t$$

$$\widehat{v}_t = (\beta - \widehat{b})' x_t + \rho v_{t-1} + u_t$$

we could test the hypothesis that $\rho = 0$, there is no serial correlation by running a regression of the estimated residuals on the regressors and the lagged residuals:

$$\widehat{v}_t = c' x_t + \rho \widehat{v}_{t-1} + u_t$$

and testing $\rho = 0$ with a t test. We replace the missing residuals (for period zero here) by their expected value zero. If we think there may be higher order correlations, we can add more lagged residuals and test the joint hypothesis that all the coefficients of the lagged residuals are zero, with an F test. For instance, if we have quarterly data, we would be interested in testing for up to fourth order serial correlation, i.e. all $\rho_i = 0, i = 1, 2, ..., 4$ in:

$$\widehat{v}_t = c' x_t + \rho_1 \widehat{v}_{t-1} + \rho_2 \widehat{v}_{t-2} + \rho_3 \widehat{v}_{t-3} + \rho_4 \widehat{v}_{t-4} + u_t$$

This is a different alternative hypothesis to that of no first order serial correlation, but the null hypothesis is the same.

8.3. Heteroskedasticity.

Suppose we estimate the first stage linear regression (8.3), then in heteroskedasticity tests we run second stage regressions using the squared residuals:

$$\widehat{u}_t^2 = \alpha + b'z_t + v_t$$

the null hypothesis is that the expected value of the squared residuals is a constant α, so $b = 0$, and this can be tested with an F test. On the alternative hypothesis, the variance, squared residuals, change with z_t. There are lots of ways that the variance could change, thus lots of possible candidates for z_t. It could be x_t, the regressors; it could be the squares and cross-products of the regressors, often called the White test; it could be the squared fitted values, the RESET version; it could be lagged squared residuals, testing for ARCH (autoregressive conditional heteroskedasticity); etc.

8.4. Normality

If the residuals are normal then their coefficient of skewness (third moment) should be zero and coefficient of kurtosis (fourth moment) three. This is tested by the Jarque-Bera test

$$T\left\{\frac{\mu_3^2}{6\mu_2^3} + \frac{1}{24}\left(\frac{\mu_4}{\mu_2^2} - 3\right)^2\right\} \sim \chi^2(2)$$

where $\mu_j = \sum_{t=1}^{T} \widehat{u}_t^j/T$:

8.5. Non-linearity

We distinguished (1) equations which are non-linear in variables because of transformations, like logarithms or powers, but which can be estimated by a linear regression on the transformed data and (2) equations which are non-linear in parameters, where we need a non-linear estimation routine of the type discussed in 3.2. iables

Suppose we are explaining the logarithm of wages, w_i, of a sample of men $i = 1, 2, ..., N$ by age, A, and years of education, E. This is certainly not linear, at some age wages peak and then fall with age thereafter, similarly with education: getting a PhD in the UK reduces expected lifetime earnings by about 20%. In addition, the variables interact, wages peak later in life for more educated people.

This suggests a model of the form:

$$w_i = a + bA_i + cA_i^2 + dE_i + eE_i^2 + fE_iA_i + u_i$$

This model is linear in parameters though it is non-linear in the variables and can be estimated by OLS on the transformed data. We expect, $b, d, f > 0$ and $c, e < 0$. The age at which earnings is maximised is given by the solution to:

$$\frac{\partial w}{\partial A} = b + 2cA + fE = 0 \tag{8.2}$$

$$A^* = -\frac{b + fE}{2c}.$$

which if the estimated coefficients have the expected signs is positive (since $c < 0$) and peak earning age increases with education.$\beta\%$ change in Y_t.

8.5.1. Testing for non-linearity

In the wage example above we had strong prior reasons to include squares and cross products. In other cases we do not, but just want to check whether there is a problem. Adding squares and cross-products can also use up degrees of freedom very fast. If there are k regressors, there are k(k+1)/2 squares and cross products, for k=5, 15 additional regressors. This is fine in large cross sections with thousands of observations, but in small samples it is a problem. Instead, we estimate a first stage linear regression:

$$y_t = \widehat{\beta}' x_t + \widehat{u}_t \tag{8.3}$$

with fitted values $\widehat{y}_t = \widehat{\beta}' x_t$; and run a second stage regression:

$$\widehat{u}_t = b' x_t + c\widehat{y}_t^2 + v_t$$

and test whether c is signficantly different from zero.

Eviews does this Ramsey RESET test slightly differently. It runs

$$y_t = d' x_t + e\widehat{y}_t^2 + v_t$$

and tests whether e is significantly different from zero. Noting that

$$y_t = \widehat{y}_t + \widehat{u}_t = \widehat{\beta}' x_t + \widehat{u}_t,$$

$d' = (b + \widehat{\beta})'$ so they give identifical test statistics.

Higher powers of $\widehat{y}_t$ can also be added. Notice that $\widehat{y}_t^2$ is being used as a measure of squares and cross-products of the x_t. For the simple model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$$

$$\begin{aligned} \widehat{y}_t^2 &= (\widehat{\beta}_1 + \widehat{\beta}_2 x_{2t} + \widehat{\beta}_3 x_{3t})^2 \\ &= \widehat{\beta}_1^2 + \widehat{\beta}_2^2 x_{2t}^2 + \widehat{\beta}_3^2 x_{3t}^2 + 2\widehat{\beta}_1 \widehat{\beta}_2 x_{2t} + ... \end{aligned}$$

Tests which use powers of the fitted values in this way are often known as RESET tests.

9. Univariate Stochastic Processes

Suppose we have a series of observations on some economic variable, $y_t, t = 1, 2, ..., T$, which may already have been transformed, e.g. the logarithm of GDP. It is useful to regard each y_t as a random variable with a density function, $f_t(y_t)$ and we observe one realisation from the distribution for that period. A family of random variables indexed by time is called a stochastic process, an observed sample is called a realisation of the stochastic process. A stochastic process is said to be 'strongly stationary' if its distribution is constant through time, i.e. $f_t(y_t) = f(y_t)$. It is first order stationary if it has a constant mean. It is second order, or weakly or covariance stationary if also has constant variances and constant covariances between y_t and y_{t-i}, i.e. the autocovariances (covariances with itself in previous periods) are only a function of i (the distance apart of the observations) not t, the time they are observed. These autocovariances summarise the dependence between the observations and they are often represented by the autocorrelation function or correlogram, the vector (graph against i) of the autocorrelations $r_i = Cov(y_t, y_{t-i})/Var(y_t)$. If the series is stationary, the correlogram converges to zero quickly.

The order of integration is the number of times a series must be differenced to make it stationary (after perhaps removing deterministic elements like a linear trend). So a series, y_t, is said to be Integrated of order zero, I(0), if y_t is stationary; integrated of order one, I(1), if $\Delta y_t = y_t - y_{t-1}$ is stationary; integrated of order two, I(2), if

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

is stationary. Notice that $\Delta^2 y_t \neq \Delta_2 y_t = y_t - y_{t-2}$.

In examining dynamics it will be useful to use the Lag Operator, L, sometimes known as the backward shift operator B.

$$\begin{aligned} Ly_t &= y_{t-1}; \quad L^2 y_t = y_{t-2}; \quad etc \\ \Delta y_t &= (1-L)y_t, \ \Delta^2 y_t = (1-L)^2 y_t. \end{aligned}$$

9.1. White noise processes

A stochastic process is said to be White Noise if

$$\begin{aligned} E(\varepsilon_t) &= 0; \\ E(\varepsilon_t^2) &= \sigma^2; \\ E(\varepsilon_t \varepsilon_{t-i}) &= 0, \ i \neq 0 \end{aligned}$$

We will use ε_t below to denote white noise processes.

9.2. Autoregressive processes

A first order (one lag) autoregressive process (AR1) takes the form:

$$\begin{aligned} y_t &= \rho y_{t-1} + \varepsilon_t, \\ y_t(1-\rho L) &= \varepsilon_t, \end{aligned}$$

with $E(y_t) = 0$, and is stationary if $\mid \rho \mid < 1$. If it is stationary, then by repeated substitution, we get the sum of a geometric progression:

$$\begin{aligned} y_t &= \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3}.... \\ y_t &= (1-\rho L)^{-1}\varepsilon_t, \end{aligned} \tag{9.1}$$

the variance of y_t is $E(y_t^2) = \sigma^2/(1-\rho^2)$ and the correlations between y_t and y_{t-i}, $r_i = \rho^i$, so decline exponentially. A constant can be included $y_t = \alpha + \rho y_{t-1} + \varepsilon_t$, then $E(y_t) = \alpha/(1-\rho)$. If the process is stationary, the parameters of the AR model can be estimated consistently by Least Squares, though the estimates will not be unbiased (y_{t-1} is uncorrelated with ε_t but not independent since it is correlated with ε_{t-1}); the estimate of ρ will be biased downwards.

A p th order autoregression (ARp) takes the form:

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + ... + \rho_p y_{t-p} + \varepsilon_t$$

$$y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} - ... - \rho_p y_{t-p} = \varepsilon_t$$
$$(1 - \rho_1 L - \rho_2 L^2 - ... - \rho_p L^p) y_t = \varepsilon_t.$$

The last expression is a p th order polynomial in the lag operator, which we can write $A^p(L)$. y_t is stationary if all the roots (solutions), z_i, of $1 - \rho_1 z - \rho_2 z^2 - ... - \rho_p z^p = 0$ lie outside the unit circle (are greater than one in absolute value). If a root lies on the unit circle, some $z_i = 1$, the process is said to exhibit a unit root. The condition is sometimes expressed in terms of the inverse roots, which must lie inside the unit circle. Usually we just check that $\sum \rho_i < 1$ for stationarity.

Consider the case of an AR1 process

$$y_t = \rho y_{t-1} + \varepsilon_t.$$

For stability, the solution to $(1 - \rho z) = 0$, must be greater than unity in absolute value, since this implies $z = 1/\rho$ this requires $-1 < \rho < 1$. For an $AR2$ the real parts of solution to the two solutions to the quadratic $(1 - \rho_1 z - \rho_2 z^2)$ must be greater than unity.

9.3. Random Walks

If $\rho = 1$, there is said to be a unit root and the AR1 becomes a random walk:

$$y_t = y_{t-1} + \varepsilon_t;$$

or $\Delta y_t = \varepsilon_t$. Substituting back

$$y_t = \varepsilon_t + \varepsilon_{t-1} + ...\varepsilon_1 + y_0;$$

so shocks have permanent effects. A random walk with drift is of the form: $\Delta y_t = \alpha + \varepsilon_t$.

In both these cases, Δy_t is stationary, I(0), but y_t is non-stationary, I(1). If there is no drift the expected value of y_t will be constant at zero, if $y_0 = 0$, but the variance will increase with t. If there is a drift term the expected value of y_t, as well as the variance, will increase with t. Random walks appear very often in economics, e.g. the efficient market hypothesis implies that, to a first approximation, asset prices should be random walks.

9.4. Moving Average processes

A first order moving average process (MA1) takes the form

$$y_t = \varepsilon_t + \mu\varepsilon_{t-1};$$

a q th order moving average:

$$y_t = \varepsilon_t + \mu_1\varepsilon_{t-1} + \mu_2\varepsilon_{t-2} + ... + \mu_q\varepsilon_{t-q};$$

$$y_t = (1 + \mu_1 L + \mu_2 L^2 + ... + \mu_q L^q)\varepsilon_t = B^q(L)\varepsilon_t.$$

$Cov(y_t, y_{t-i}) = 0$, for $i > q$. A finite order moving average process is always stationary. Any stationary process can be represented by a (possibly infinite) moving average process. Notice that the AR1 is written as an infinite MA process in (9.1). The parameters of the MA model cannot be estimated by OLS, but maximum likelihood estimators are available. If a MA process is invertible we can write it as an AR, i.e. $B^q(L)^{-1}y_t = \varepsilon_t$. Notice that if we take a white noise process $y_t = \varepsilon_t$ and difference it we get

$$\Delta y_t = \varepsilon_t - \varepsilon_{t-1}$$

a moving average process with a unit coefficient.

10. ARIMA and unit roots

Combining the AR and MA processes, gives the ARMA process. The first order ARMA(1,1) with intercept is

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t + \mu\varepsilon_{t-1}$$

In practice, the data are differenced enough times, say d, to make them stationary and then modelled as an ARMA process of order p and q. This gives the Autoregressive Integrated Moving Average, ARIMA(p,d,q) process, which can be written using the lag polynomials above as:

$$A^p(L)\Delta^d y_t = \alpha + B^q(L)\varepsilon_t.$$

For instance, the ARIMA(1,1,1) process is

$$\Delta y_t = \alpha + \rho\Delta y_{t-1} + \varepsilon_t + \mu\varepsilon_{t-1}$$

ARIMA models often describe the univariate dynamics of a single economic time-series quite well and are widely used for forecasting.

10.1. Trend and difference stationary processes

Most economic time-series, e.g. log GDP, are non-stationary, trended. The trend can be generated in two ways. First, the traditional assumption was that the series could be regarded as stationary once a deterministic trend was removed. For instance:

$$y_t = \alpha + \rho y_{t-1} + \gamma t + \varepsilon_t \tag{10.1}$$

with $\mid \rho \mid < 1$ is a trend stationary process. The effects of the shocks ε_t are transitory and die away through time, since ε_{t-i} is multiplied by ρ^i when you substitute back, see (9.1) above. If the variables are in logs, the long run growth rate is $g = \gamma/(1-\rho)$. Second the series could be regarded as a random walk with drift, difference stationary:

$$\begin{aligned} \Delta y_t &= \alpha + \varepsilon_t \\ y_t &= \alpha + y_{t-1} + \varepsilon_t \end{aligned}$$

The long run growth rate is α.

We want to test the null of a difference stationary process (one with a unit root) against the alternative of a trend stationary process. Substitute $\gamma = g(1-\rho)$ then subtract y_{t-1} from both sides, so we can write the trend stationary process as:

$$\Delta y_t = \alpha + (\rho - 1)(y_{t-1} - gt) + \varepsilon_t \tag{10.2}$$

$$\Delta y_t = \alpha + \beta(y_{t-1} - gt) + \varepsilon_t \tag{10.3}$$

where $\beta = \rho - 1$. If $\rho = 1$ or equivalently $\beta = 0$, we get the random walk with drift:with growth rate α. Substituting back we get

$$\begin{aligned} y_t &= \alpha + (\alpha + y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ y_t &= y_{t-2} + 2\alpha + \varepsilon_t + \varepsilon_{t-1} \end{aligned}$$

continuing the process to period zero, we get:

$$y_t = y_0 + \alpha t + \sum_{i=0}^{t-1} \varepsilon_{t-i}. \tag{10.4}$$

In this case, the difference stationary process, the efects of the shocks are permanent or persistent, they last for ever, and the series is determined by an initial

value, y_0, a deterministic trend αt, and a 'stochastic trend', $\sum_{i=0}^{t-1} \varepsilon_{t-i}$, the sum of past errors.

If we had not restricted (10.2) so that the trend term dropped out when $\beta = 0$, there would be a quadratic trend in y_t. Show this by substituting back in

$$\begin{aligned} y_t &= \alpha + y_{t-1} + \gamma t + \varepsilon_t \\ y_t &= \alpha + (\alpha + y_{t-2} + \gamma(t-1) + \varepsilon_{t-1}) + \gamma t + \varepsilon_t \end{aligned} \tag{10.5}$$

etc.

10.2. Testing for unit roots

Choosing between the trend stationary and difference stationary model is a matter of determining whether $\beta = 0$ or equivalently $\rho = 1$; whether there is a 'unit root' in y_t. To do this we can estimate (10.2) by running a regression of Δy_t on a constant, y_{t-1} and a linear trend; estimate $\widehat{\beta}$ the coefficient on the lagged level of y_{t-1}; construct the 't statistic' $\tau_\beta = \widehat{\beta}/SE(\widehat{\beta})$ to test $H_0 : \beta = 0$; against $H_1 : \beta < 0$. If we do not reject the null we conclude that there is a unit root in y_t, it is integrated of order one, I(1), stationary after being differenced once. If we reject the null we conclude that y_t is trend stationary I(0). This is a one-sided test and if $\widehat{\beta} > 0$, we do not reject the null of a unit root. The test statistic τ_β does not have a standard t distribution, but a Dickey Fuller distribution and the critical value is -3.5 at the 5% significance level, when a trend is included. This is because under H_0 the regressor, y_{t-1} is non-stationary. If there is no trend included in the regression the 5% critical value is -2.9. Most programs will give you these critical values or the relevant p values.

To get good estimates of (10.2) we require that ε_t is white noise. Often this will not be the case and the error will be serially correlated. To remove this serial correlation, lags of the dependent variable are added to give the 'Augmented Dickey Fuller' (ADF) regression:

$$\Delta y_t = \alpha + \beta y_{t-1} + \gamma t + \sum_{i=1}^{p} \delta_i \Delta y_{t_{-i}} + \varepsilon_t$$

where p is chosen to try to make the residual white noise. Show that this is a reparameterisation of a AR(p+1) with trend. Again the procedure is to use the t

ratio on β with the non standard critical values to test the null hypothesis $\beta = 0$ against the alternative $\beta < 0$.

To test for $I(2)$ versus $I(1)$ you just take a further difference:

$$\Delta^2 y_t = \alpha + \beta \Delta y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta^2 y_{t_{-i}} + \varepsilon_t$$

if it was thought that there might be a trend in the change (not common for economic series) it could be included also. Again $H_0 : \beta = 0$; against $H_1 : \beta < 0$.

There are a range of other procedures for determining whether there is a unit root. They differ, for instance, in how they correct for serial correlation (in a parametric way like the ADF where you allow for lags or in a non-parametric way like Phillips Peron where you allow for window size); whether they include other variables; whether they use the null of a unit root like the ADF or the null of stationarity, like KPSS; whether they use GLS detrending; and whether they use both forward and backward regressions. EViews gives you a lot of choices.

Most of these tests have low power, it is very difficult to distinguish $\rho = 1$ from a stationary alternative in which ρ is close to unity. The power of the tests depends on the span of the data not the number of observations. For instance UK unemployment rates 1945-1985 appear I(1), UK unemployment rates 1855-1985 appear I(0). The tests are also sensitive to step changes, an I(0) process with a single change in level will appear I(1), as it should since the shock (change in level) is permanent. The order of integration is a univariate statistical summary of how the time series moves over the sample, it is not an inherent structural property of the series. Whether you treat a variable as $I(0)$ or $I(1)$ depends on the purpose of the exercise, for estimation it is often safer to treat it as $I(1)$.

11. Dynamic Linear Regression, ARDL & ECM models

When we look at the relationship between variables, there are usually lags between the change in one variable and the effect on another. The distributed lag of order q, DL(q) regression model takes the form:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + ... + \beta_q x_{t-q} + u_t$$

notice that it is similar to a moving average, except that here the shocks are observed, x_t, rather than being unobserved. We can combine the distributed lag

with an autoregressive component to give the ARDL(p,q) process:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \beta_0 x_t + \beta_1 x_{t-1} + ... + \beta_q x_{t-q} + u_t \qquad (11.1)$$

where u_t is usually a white noise error, though it could be moving average.

If the error u_t is white noise, the parameters can be estimated consistently by OLS, though the estimates are not unbiased. y_{t-1} is not independent of u_t, though it is predetermined and uncorrelated with u_t as long as u_t is not serially correlated. Independence is a much stronger concept which implies it is uncorrelated with all leads and lags, but since y_{t-1} is determined by u_{t-1}, it is correlated with it, so they are not independent.

A simple version is ARDL(1,0)

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta x_t + u_t \qquad (11.2)$$

which can be derived from a partial adjustment model (PAM) where there is a long run relationship determining an equilibrium or target value:

$$y_t^* = \theta_0 + \theta_x x_t$$

and slow adjustment to equilibrium

$$\Delta y_t = \lambda(y_t^* - y_{t-1}) + u_t,$$

where λ measures the speed of adjustment, the proportion of the deviation from equilibrium made up in any period. Then

$$y_t = \lambda\theta_0 + \lambda\theta_x x_t + (1-\lambda) y_{t-1} + u_t. \qquad (11.3)$$

The parameters of (11.3), which can be given a theoretical interpretation, can be recovered from the estimates of (11.2) : $\lambda = 1 - \alpha_1$; $\theta_x = \beta/(1-\alpha_1)$; $\theta_0 = \alpha_0/(1-\alpha_1)$.

The process (11.1) is stationary, (conditional on x_t the process is stable), if all the roots (solutions), z_i, of the characteristic equation

$$1 - \alpha_1 z - \alpha_2 z^2 - ... - \alpha_p z^p = 0 \qquad (11.4)$$

lie outside the unit circle (are greater than one in absolute value). We usually check that $\sum \alpha_i < 1$. In this case, if x_t is constant, say at x, then y_t will converge to a constant, say y, and the long run relation between them will be:

$$y = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i} + \frac{\sum_{i=0}^{q} \beta_i}{1 - \sum_{i=1}^{p} \alpha_i} x = \theta_0 + \theta_x x.$$

This can be obtained by setting $y_{t-i} = y$ and $x_{t-i} = x$ for all i. This long-run solution is usually interpreted as the long-run equilibrium or target value for y_t and can be calculated from the estimated regression coefficients. Standard errors for the long-run coefficients can be calculated by the delta method, which is available in most programmes.

This procedure is appropriate in quite a wide variety of circumstances including if all the variables are I(0) and x_t is exogenous; or if all the variables are I(1), there is a single cointegrating relationship and x_t is exogenous; or if there are mixtures of I(0) and cointegrating I(1) variables such that u_t is white noise. See below for more details.

11.1. ARDL(1,1) and Error Correction Models

ARDL models or dynamic linear regressions are widely used to examine the relationship between economic variables. We will use the ARDL(1,1) for illustration, this is:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t. \tag{11.5}$$

It is stable if $-1 < \alpha_1 < 1$, and then has a long run solution:

$$y_t^* = \frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0 + \beta_1}{1-\alpha_1} x_t = \theta_0 + \theta_x x_t.$$

Where y_t^* is the target or long run equilibrium value for y_t to which it would tend in the absence of further shocks to x_t and u_t. There are a number of other useful ways of rewriting (reparameterizing) (11.5).

We can estimate the long run effect of x_t on y_t from the OLS estimates of (11.5) as

$$\widehat{\theta}_x = \frac{\widehat{\beta}_0 + \widehat{\beta}_1}{1-\widehat{\alpha}_1}.$$

Since the estimated $\widehat{\alpha}_1$ can be close to one, causing the estimate of $\widehat{\theta}_x$ to blow up, it can have poor properties.

Write (11.5) as

$$y_t - y_{t-1} = \alpha_0 + (\alpha_1 - 1) y_{t-1} + \beta_0 (x_t - x_{t-1}) + (\beta_0 + \beta_1) x_{t-1} + u_t$$

$$\Delta y_t = a_0 + b_0 \Delta x_t + a_1 y_{t-1} + b_1 x_{t-1} + u_t \tag{11.6}$$

where $a_0 = \alpha_0; b_0 = \beta_0; a_1 = (\alpha_1 - 1); b_1 = \beta_0 + \beta_1$; or in terms of adjustment to a long-run target:

$$\Delta y_t = \lambda_1 \Delta y_t^* + \lambda_2 (y_{t-1}^* - y_{t-1}) + u_t$$

where the long-run target or equilibrium (as calculated above) is

$$y_t^* = \theta_0 + \theta_x x_t,$$

and the λ_i are adjustment coefficients which measure how y adjusts to changes in the target and deviations from the target. Notice $a_0 = \lambda_2 \theta_0; a_1 = -\lambda_2; b_0 = \lambda_1 \theta_x; b_1 = \lambda_2 \theta_x$. This form is usually known as an 'Error (or equilibrium) Correction Model' ECM. The dependent variable changes in response to changes in the target and to the error, the deviation of the actual from the equilibrium in the previous period: $(y_{t-1}^* - y_{t-1})$.

An alternative parameterization, which unlike the ECM nests the partial adjustment model (11.2) is:

$$\Delta y_t = \alpha_0 + (\alpha_1 - 1) y_{t-1} + (\beta_0 + \beta_1) x_t - \beta_1 \Delta x_t + u_t.$$

11.2. Reparameterizations & Restrictions

When you **reparameterize** a model, as we did above going from (11.5) to (11.6), you estimate exactly the same number of parameters (4 in this case), just written in different ways.You will get identical estimates of say, the long-run coefficient, whether you estimate it as an ARDL, ECM or by a non-linear procedure. The statistical properties of the model do not change, the estimated residuals, standard error of the regression and the maximised log-likelihood are identical between the different versions. R^2 will change, because the proportion of variation explained is measured in terms of a different dependent variable, Δy_t in the ECM rather than y_t in the ARDL. Any RESET tests that use fitted values of the dependent variable will also change. Use the misspecification tests which use the fitted values of Δy_t.

When you **restrict** a model, you reduce the number of parameters estimated and such restrictions are testable. The ARDL(1,1) nests a number of interesting restricted special cases, including:

(a) Static: $\alpha_1 = 0; \beta_1 = 0$.
(b) First difference: $\alpha_1 = 1; \beta_1 = -\beta_0$
(c) Partial Adjustment Model:$\beta_1 = 0$
(d) First order disturbance serial correlation: $\beta_1 = -\beta_0 \alpha_1$

(e) Unit long-run coefficient: $\beta_1 + \beta_0 + \alpha_1 = 1$

(f) Random Walk with drift: $\alpha_1 = 1; \beta_1 = \beta_0 = 0$.

A useful procedure in many circumstances is to start with a general model, e.g. the ARDL(1,1) and test down to specific restricted cases. This general to specific procedure has the advantage that any tests on the general model are valid. Whereas if you start from the restricted model, the tests will not be valid if the model is misspecified.

Case (d) is got by assuming that the model is:

$$y_t = \alpha + \beta x_t + v_t; \; v_t = \rho v_{t-1} + \varepsilon_t$$

where ε_t is white noise, this can be written:

$$y_t = \alpha + \beta x_t + \rho v_{t-1} + \varepsilon_t$$

noting that

$$v_t = y_t - \alpha - \beta x_t; \text{ and } \; v_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}$$

$$y_t = \alpha + \beta x_t + \rho(y_{t-1} - \alpha - \beta x_{t-1}) + \varepsilon_t$$

$$y_t = \alpha(1-\rho) + \beta x_t + \rho y_{t-1} - \beta\rho x_{t-1} + \varepsilon_t$$

which is of the same form as (11.5) with the restriction that the coefficient of x_{t-1} equals the negative of the product of the coefficients of x_t and y_{t-1}, i.e. $\beta_1 = -\beta_0\alpha_1$ in terms of the parameters of the unrestricted model. This is sometimes called the common factor model, since it can be written $(1-\rho L)y_t = (1-\rho L)(\alpha + \beta x_t) + \varepsilon_t$, both sides of the static model are multiplied by the common factor $(1-\rho L)$. The restricted model (with AR1 errors) is not linear in the parameters and is estimated by Generalised Least Squares or Maximum Likelihood.

In case (e) the restricted model can be written:

$$\Delta y_t = a_0 + b_0 \Delta x_t + a_1(y_{t-1} - x_{t-1}) + e_t.$$

and the restriction is equivalent to assuming $\mathrm{b}_1 = -a_1$ in (11.6).

11.3. ARDL(1,1,1)

The structure generalises to more explanatory variables, e.g. the ARDL(1,1,1) model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \gamma_0 z_t + \gamma_1 z_{t-1} + u_t. \qquad (11.7)$$

has a long run solution:

$$y = \frac{\alpha_0}{1-\alpha_1} + \frac{\beta_0+\beta_1}{1-\alpha_1}x + \frac{\gamma_0+\gamma_1}{1-\alpha_1}z = \theta_0 + \theta_x x + \theta_z z.$$

Notice that our error correction adjustment process

$$\Delta y_t = \lambda_1 \Delta y_t^* + \lambda_2(y_{t-1}^* - y_{t-1}) + u_t$$

$$y_t^* = \theta_0 + \theta_x x_t + \theta_z z_t,$$

now imposes restrictions. In the case of one exogenous variable, there were four ARDL parameters $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ and four theoretical parameters $(\lambda_1, \lambda_2, \theta_0, \theta_x)$ so no restrictions. In the case of two exogenous variables there are six ARDL parameters, $(\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma_0, \gamma_1)$ but only five theoretical parameters $(\lambda_1, \lambda_2, \theta_0, \theta_x, \theta_z)$. What is the restriction?

11.4. Adaptive Expectations

Define the expected value of x_{t+1} conditional on information available at time t, as:

$$E(x_{t+1} \mid I_t) = x_t^e;$$

agents determine their actions according to:

$$y_t = \beta x_t^e + u_t \tag{11.8}$$

and determine their expectations according to:

$$x_t^e - x_{t-1}^e = \phi(x_t - x_{t-1}^e)$$

they adjust their forecast proportional to the forecast error they made in the previous period (note x_{t-1}^e is the forecast of x_t made in the previous period). This can be written:

$$\begin{aligned} x_t^e &= \phi x_t + (1-\phi)x_{t-1}^e \\ (1-(1-\phi)L)x_t^e &= \phi x_t \\ x_t^e &= \frac{\phi x_t}{(1-(1-\phi)L)} = \phi \sum_{i=0}^{\infty}(1-\phi)^i x_{t-i} \end{aligned}$$

substituting this exponentially weighted moving average of past x_t in (11.8) gives:

$$y_t = \beta \left(\frac{\phi x_t}{(1-(1-\phi)L)} \right) + u_t \tag{11.9}$$

premultiply by $(1-(1-\phi)L)$ to give

$$(1-(1-\phi)L)y_t = \beta\phi x_t + (1-(1-\phi)L)u_t$$

$$y_t = \beta\phi x_t + (1-\phi)y_{t-1} + u_t - (1-\phi)u_{t-1}$$

an ARDL(1,0) with a MA1 error, with a restriction that the AR and MA coefficients should be equal and of opposite sign.

This type of transformation (known as the Koyck transform) can be used to get rid of a variety of exponentially weighted infinite distributed lags.

12. Cointegration

12.1. Introduction

Suppose y_t and x_t are I(1) then in general any linear combination of them will also be I(1). If there is a linear combination that is I(0), they are said to cointegrate. If they cointegrate, they have a common stochastic trend, random walk type component, which is cancelled out by the linear combination; and this linear combination is called the cointegrating vector, which is often interpreted as an equilibrium relationship.

Suppose we have data on s_t, p_t, p_t^*, the logarithms of the spot exchange rate (domestic currency per unit foreign), domestic and foreign price indexes and that each of these are I(1). Purchasing Power Parity says that the real exchange rate $e_t = s_t - p_t + p_t^*$ should be stationary, i.e. $e_t = e + u_t$ where e is the equilibrium real exchange rate and u_t is a stationary (not necessarily white noise) error. The cointegrating vector is then $(1, -1, 1)$. It is quite common in economics to get ratios of non-stationary variables being approximately stationary. These 'great ratios' include the real exchange rate, the savings ratio, the velocity of circulation of money, the capital-output ratio, the share of wages in output, the profit rate, etc. In each case a linear combination of the logarithm of the variables with cointegrating vectors of plus and minus ones should be stationary and this can be tested using the unit root tests described above.

The coefficient does not need to be unity. If

$$y_t = \alpha + \beta x_t + u_t \tag{12.1}$$

and u_t is stationary, the cointegrating vector is $(1, -\beta)$ since $(y_t - \beta x_t = \alpha + u_t)$ is I(0).

If y_t and x_t are I(1) and do not cointegrate, say they are independent unrelated random walks, the error in (12.1) will be I(1) and this will be a 'spurious' regression. As $T \to \infty$, the R^2 of this regression will go to unity and the t ratio for $\widehat{\beta}$ will go to a non-zero random variable. Thus even if there is no relationship, the regression would indicate a close relationship. Therefore it is important to test for cointegration, but this can be better done in the context of an ARDL/ECM than a levels relationship like (12.1).

12.2. Consistency and super-consistency

An estimator $\widehat{\theta}$ is consistent, if for some small number $\epsilon > 0$

$$\lim_{T \to \infty} \Pr(\mid \widehat{\theta}_T - \theta \mid > \epsilon) = 0.$$

As the sample size gets large $\widehat{\theta}_T$ converges to θ, they become exactly the same, so the both the bias and the variance go to zero. This means that we cannot compare the variances of two consistent estimators, they both have variance zero.

Consider $X_t \sim iid(\mu, \sigma^2)$ that is independently and identically distributed with $E(X_t) = \mu$ and $Var(X_t) = E(X_t - \mu)^2 = \sigma_x^2$. The mean $\overline{X} = \sum_{t=1}^{T} X_t$ is unbiased and has variance σ_x^2/T which goes to zero as $T \to \infty$. But so does the variance of another estimator $\widetilde{X}$ with variance $2\sigma_x^2/T$.

To deal with the problem of the variance going to zero when we evaluate asymptotic distributions we scale the difference and we usually look at $\sqrt{T}(\widehat{\theta} - \theta)$ as $T \to \infty$, so the asymptotic variance of the mean is σ_x^2 not $0 = \sigma_x^2/T$. When the variance falls at rate T then $\mid \widehat{\theta}_T - \theta \mid \to 0$ at rate $\sqrt{T}$ and $\widehat{\theta}$ is said to be $\sqrt{T}$ consistent.

In the linear regression model

$$y_t = \beta x_t + u_t$$

where the variables are measured as deviations from their means so $E(y_t) = E(x_t) = E(u_t) = 0$ with $E(u_t^2) = \sigma_u^2$, $E(u_t u_{t-i}) = 0$; $Var(x_t) = E(x_t^2) = \sigma_x^2$. The

OLS estimator and its variance are given by

$$\begin{aligned}\widehat{\beta} &= \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t (\beta x_t + u_t)}{\sum x_t^2} = \beta + \frac{\sum x_t u_t}{\sum x_t^2}. \\ V\left(\widehat{\beta}\right) &= E(\widehat{\beta} - \beta)^2 = E\left(\frac{\sum x_t u_t}{\sum x_t^2}\right)^2 = \frac{\sigma_u^2}{\sum x_t^2} = \frac{\sigma_u^2}{T\sigma_x^2}\end{aligned}$$

So in this case $\widehat{\beta}$ is also $\sqrt{T}$ consistent. Notice that if each individual x_t has variance $E(x_t^2) = \sigma_x^2$, then the variance of the sum is

$$E(\sum_{t=1}^{T} x_t^2) = T\sigma_x^2.$$

Now suppose x_t follows a mean zero random walk

$$x_t = x_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$ then from (10.4) above

$$\begin{aligned}x_t &= x_0 + \sum_{i=0}^{T-1} \varepsilon_{t-i} \\ \sigma_x^2 &= E(x_t^2) = T\sigma_\varepsilon^2\end{aligned}$$

where $E(x_0) = 0$.

Then the variance of $\widehat{\beta}$ is

$$V\left(\widehat{\beta}\right) = \frac{\sigma_u^2}{\sum x_t^2} = \frac{\sigma_u^2}{T\sigma_x^2} = \frac{\sigma_u^2}{T(T\sigma_\varepsilon^2)} = \frac{\sigma_u^2}{T^2\sigma_\varepsilon^2}.$$

So when x_t is a random walk (or more generally $I(1)$) $\widehat{\beta}$ converges to its true value at rate T rather than $\sqrt{T}$, it is said to be T consistent or super-consistent. Of course, even if it converges rapidly to its true value with T, it may have a large bias in small samples.

If we have a regression on a time trend

$$y_t = a + bt + u_t$$

then $\widehat{b}$ is $T^{3/2}$ consistent, since we have $\sum t^2$ in the denominator of the formula for its variance.

12.3. Cointegration and the ARDL/ECM

In most cases it is better to estimate single equation relations not in levels form like (12.1) but in ARDL or ECM form, which is robust to whether the variables are I(0) or I(1) and whether or not they are cointegrated. Write the ARDL(1,1) in ECM form:

$$\Delta y_t = a_0 + b_0 \Delta x_t + a_1 y_{t-1} + b_1 x_{t-1} + u_t. \tag{12.2}$$

This equation does not seem to balance, the left hand side Δy_t is I(0) and there are two I(1) terms y_{t-1} and x_{t-1} on the right hand side. It only balances if a linear combination of the I(1) terms is I(0), that is if y_t and x_t cointegrate so that $y_t - \theta_x x_t$ is I(0) with cointegrating vector $(1, -\theta_x)$, in:

$$\Delta y_t = a_0 + b_0 \Delta x_t + \lambda(y_{t-1} - \theta_x x_{t-1}) + u_t \tag{12.3}$$

Notice that if they are I(1) and cointegrate λ must be non-zero and negative (this is the feedback that keeps y_t and x_t from diverging. We can test for this, though the critical values are non standard, see below. Notice we are free to normalise the cointegrating vector, since $a_1 y_{t-1} + b_1 x_{t-1}$ is I(0), we could also have called the cointegrating vector $(a_1, b_1) = (\lambda, -\lambda\theta_x)$.

Notice this equation works in that OLS provides consistent estimates whether the variables are

- $I(0)$, the standard case, where there may be a long-run relationship, though it is not a cointegrating relationship since the variables are not $I(1)$.
- $I(1)$ and cointegrated, like (12.3) $\lambda \neq 0$.
- $I(1)$ and not cointegrated (12.3) with $\lambda = 0$, when a first difference model is appropriate,
- $I(0)$ y_t and $I(1)$ x_t, with $b_1 = 0$
- $I(1)$ y_t and $I(0)$ x_t, $a_1 = 0$.

In all these cases the OLS estimates can make it balance. Although the estimation is standard, testing is not. The critical values for testing for a long run relationship are different depending on whether the variables are $I(0)$ or $I(1)$. The PSS Bounds Test[4], provides $I(0)$ and $I(1)$ critical values for the F statistic for

[4]Pesaran, Shin and R.J. Smith, Journal of Applied Econometrics, 2001, p289-326, Bounds Testing Approaches to the Analysis of Level Relationships.

the no levels relationship hypothesis: $a_1 = b_1 = 0$ in (12.2). If the test statistic is below the $I(0)$ critical value, there is definitely no long run relation. If it is above the $I(1)$ critical value, there definitely is a long run relation. If it is in between it depends on the order of integration of the variables.

With 2 lags the same arguments can be applied to include $I(2)$ cases. If the variables are $I(2)$ and cointegrate to $I(1)$ then both sides of the equation below are $I(0)$

$$\Delta^2 y_t = a_0 + b_0 \Delta^2 x_t + \lambda(\Delta y_{t-1} - \theta_x \Delta x_{t-1}) + u_t$$

and this equation is a special case of the ARDL(2,2) with 2 restrictions:

$$\begin{aligned} y_t &= a_0 + (2+\lambda) y_{t-1} - (1+\lambda) y_{t-2} \\ &\quad + b x_t - (2b + \lambda\theta_x) x_{t-1} + (b + \lambda\theta_x) x_{t-2} + u_t. \end{aligned}$$

Thus OLS on an ARDL(2,2) will be able to capture the relationship.

12.4. Ways to test for cointegration.

With only two I(1) variables there can only be a single cointegrating vector, but with more than two variables there can be more than one cointegrating vector and any linear combination of these cointegrating vectors will also be a cointegrating vector. Suppose that we have data on domestic and foreign interest rates and inflation $(r_t, r_t^*, \Delta p_t, \Delta p_t^*)$ and all are I(1) (this implies that p_t is I(2)). If real interest rates ($r_t - \Delta p_t$ and $r_t^* - \Delta p_t^*$) are I(0) with cointegrating vectors $(1, 0, -1, 0)$ and $(0, 1, 0, -1)$; then the real interest rate differential $(r_t - \Delta p_t) - (r_t^* - \Delta p_t^*)$ would also be I(0), with cointegrating vector $(1, -1, -1, 1)$. Thus we need to consider a number of cases.

12.4.1. Known cointegrating vector

If the cointegrating vector is known a priori (as with the real exchange rate or real interest rate examples above) we can form the hypothesised I(0) linear combination (the log of the real exchange rate or the real interest rates) and use an ADF test to determine whether it is in fact I(0).

12.4.2. Single unknown cointegrating vector

There are three procedures here.

(a) Those that can be used for multiple unknown cointegrating vectors discussed below.

(b) Estimating an ARDL model and testing for the existence of a long-run relationship, i.e. test the null hypothesis that the levels x_{t-1} and y_{t-1} should not appear in the equation or equivalently that $\lambda = 0$ in (12.3) above, using the appropriate (non-standard) critical values of the PSS Bounds test discussed above.

(c) Running the levels equation (12.1) above and testing whether the residuals are I(1), using an ADF test and the appropriate critical values, which are different from those for an ADF on an ordinary variable. This is the original Engle-Granger procedure. Although the estimates of (12.1) are 'super-consistent' (converge to their true values at rate T rather than $\sqrt{T}$), (12.1) is clearly misspecified because it omits the dynamics, thus the estimates can be badly biased in small samples. In addition doing a unit root test on the residuals, imposes very strong restrictions on the short-run dynamics, which may not be appropriate. Thus the original Engle-Granger procedure is not recommended in most cases. If you know that one variable is exogenous use (b), if you do not know which is the exogenous variable start with (a) and test for exogeneity.

12.4.3. Multiple unknown cointegrating vectors

Again there are a variety of procedures, but the most commonly used is the Johansen procedure discussed below. This procedure operates in the context of a VAR, which we consider first.

13. Vector Autoregressions and cointegration

13.1. VARs

The generalisation of an AR2 to a vector is the VAR2:

$$y_t = a + A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t$$

where y_t is now a $m \times 1$ vector, a a $m \times 1$ vector, A_1 and A_2 are $m \times m$ matrices and $\varepsilon_t \sim N(0, \Sigma)$, where Σ is a $m \times m$ matrix with elements σ_{ij}.

For $m = 2, ... y_t = (y_{1t}, y_{2t})'$ the VAR is:

$$\begin{aligned} y_{1t} &= a_1^0 + a_{11}^1 y_{1t-1} + a_{12}^1 y_{2t-1} + a_{11}^2 y_{1t-2} + a_{12}^2 y_{2t-2} + \varepsilon_{1t}, \\ y_{2t} &= a_2^0 + a_{21}^1 y_{1t-1} + a_{22}^1 y_{2t-1} + a_{21}^2 y_{1t-2} + a_{22}^2 y_{2t-2} + \varepsilon_{2t}. \end{aligned}$$

Each equation of the VAR can be estimated consistently by OLS and the covariance matrix Σ can be estimated from the OLS residuals,

$$\widehat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\varepsilon}_{it}\widehat{\varepsilon}_{jt}$$

where $\widehat{\sigma}_{11}$ is the estimated variance of ε_{1t}, $\widehat{\sigma}_{12}$ the estimated covariance of ε_{1t} and ε_{2t}.

A variable y_{2t} is said to Granger cause y_{1t} if knowing current values of y_{2t} helps you to predict future values of y_{1t} equivalently, current y_{1t} is explained by past y_{2t}. In this case, y_{2t} is Granger causal with respect to y_{1t} if either a_{12}^1 or a_{12}^2 are non zero. You can test that they are both zero with a standard F test of linear restrictions. The restricted model just excludes $y_{2,t-1}$ and $y_{2,t-2}$ from the equation for y_{1t}. Granger causality is rarely the same as economic causality, particularly because expectations cause consequences to precede their cause: weather forecasts Granger Cause the weather.

More lags can be included and you can decide the appropriate lag length by Likelihood Ratio tests or model selection criteria like the AIC or SBC. Make sure that you use the same sample for the restricted and unrestricted model; i.e. do not use the extra observation that becomes available when you shorten the lag length. If the variables are I(1), the usual tests for Granger Causality are no longer valid, but Toda and Yamamoto (1995) suggest that the problem can be dealt with by adding extra lags, beyond the optimal number, which you do not use in the tests.[5]

A pth order VAR

$$y_t = a + \sum_{i=1}^{p} A_i y_{t-i} + \varepsilon_t$$

is stationary if all the roots of the determinantal equation $\mid I - A_1 z - A_2 z^2 - ... - A_p z^p \mid = 0$ lie outside the unit circle. When you estimate a VAR, EViews will give you a graph of the inverse roots, which should lie inside the unit circle for the variables to all be stationary.

If the lag length is p, each equation of the VAR with intercept has $1 + mp$ parameters. This can get large, 4 lags in a 4 variable VAR gives 17 parameters in each equation. Be careful, you can easily run out of degrees of freedom with

[5] Dave Giles blog has a good discussion of this and many other topics: http://davegiles.blogspot.ca/2011/04/testing-for-granger-causality.html. Unfortunately the blog has now stopped.

over-parameterised VARs. Bayesian VARs discussed below are one way of dealing with the problem.

13.2. Cointegration in VARs

We can reparameterise the VAR2:

$$y_t = A_0 + A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t$$

as:

$$\begin{aligned} y_t - y_{t-1} &= A_0 - (I - A_1 - A_2) y_{t-1} - A_2 (y_{t-1} - y_{t-2}) + \varepsilon_t \\ \Delta y_t &= A_0 - \Pi y_{t-1} + \Gamma \Delta y_{t-1} + \varepsilon_t \end{aligned}$$

and the VARp as:

$$\Delta y_t = A_0 - \Pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t.$$

Notice that this is the vector equivalent of the Augmented Dickey Fuller regression that we used above for testing for unit roots. Express the Γ_i in terms of the A_i.

If all the variables, the m elements of y_t, are I(0), Π is a full rank matrix. If all the variables are I(1) and not cointegrated, $\Pi = 0$, and a VAR in first differences is appropriate. If the variables are I(1) and cointegrated, with r cointegrating vectors, then there are r cointegrating relations, combinations of y_t that are I(0),

$$z_t = \beta' y_t$$

where z_t is a $r \times 1$ vector and β' is a $r \times m$ matrix. Then we can write the model as:

$$\Delta y_t = A_0 - \alpha z_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t,$$

in which the I(0) dependent variable is only explained by I(0) variables and α is a $m \times r$ matrix of 'adjustment coefficients' which measure how the deviations from equilibrium (the r I(0) variables z_{t-1}) feed back on the changes. This can also be written:

$$\Delta y_t = A_0 - \alpha \beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t,$$

so $\Pi = \alpha\beta'$ has rank $r < m$ if there are r cointegrating vectors. If there are $r < m$ cointegrating vectors, then y_t will also be determined by $m - r$ stochastic trends, and will have $m - r$ roots on the unit circle and m roots outside the unit circle. If there is cointegration, some of the α must be non-zero, there must be some feedback on the y_t to keep them from diverging, i.e. there must be some Granger causality in the system.

If there are r cointegrating vectors and Π has rank r, it will have r non-zero eigenvalues and Johansen provided a way of estimating the eigenvalues and two tests for determining how many of the eigenvalues are different from zero. These allow us to determine r, though the two tests may give different answers. The Johansen estimates of the cointegrating vectors β are the associated eigenvectors.

There is an 'identification' problem, since the α and β are not uniquely determined. We can always choose a non-singular $r \times r$ matrix P such that $(\alpha P)(P^{-1}\beta) = \Pi$ and the new estimates $\alpha^* = (\alpha P)$ and $\beta^* = (P^{-1}\beta)$ would be equivalent, though they might have very different economic interpretations. Put differently, if $z_{t-1} = \beta' y_{t-1}$ are I(0) so are $z^*_{t-1} = P^{-1}\beta' y_{t-1}$, since any linear combination of I(0) variables is I(0). We need to choose the appropriate P matrix to allow us to interpret the estimates. This requires r^2 restrictions, r on each cointegrating vector. One of these is provided by normalisation, we set the coefficient of the 'dependent variable' to unity, so if $r = 1$ this is straightforward (though it requires the coefficient set to unity to be non-zero). If there is more than one cointegrating vector it requires prior economic assumptions. The Johansen identification assumption, that the β are eigenvectors with unit length and orthogonal, do not allow an economic interpretation. Programs like EViews or Microfit allow you to specify the r^2 just identifying restrictions and test any extra 'over-identifying' restrictions.

As we saw above with the Dickey Fuller regression, there is also a problem with the treatment of the deterministic elements. If we have a linear trend in the VAR, and do not restrict the trends, the variables will be determined by $m - r$ quadratic trends. To avoid this (economic variables tend to show linear not quadratic trends), we enter the trends in the cointegrating vectors,

$$\Delta y_t = A_0 - \alpha(\beta' y_{t-1} + ct) + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \varepsilon_t,$$

so if an element of α is zero the trend drops out. Most programs give you a choice of how you enter trends and intercepts; unrestricted intercepts and restricted trends, option 4 in Eviews, is a good choice for trended economic data.

13.3. Example: money demand

Consider a VAR1 in the logarithms of real money, m_t, and income, y_t, which are both I(1) with a linear trend:

$$\begin{aligned} m_t &= a_{10} + a_{11}m_{t-1} + a_{12}y_{t-1} + \gamma_1 t + \varepsilon_{1t} \\ y_t &= a_{20} + a_{21}m_{t-1} + a_{22}y_{t-1} + \gamma_2 t + \varepsilon_{2t} \end{aligned}$$

The VECM is:

$$\begin{aligned} \Delta m_t &= a_{10} + (a_{11} - 1)m_{t-1} + a_{12}y_{t-1} + \gamma_1 t + \varepsilon_{1t} \\ \Delta y_t &= a_{20} + a_{21}m_{t-1} + (a_{22} - 1)y_{t-1} + \gamma_2 t + \varepsilon_{2t}, \end{aligned}$$

or

$$\begin{aligned} \Delta m_t &= a_{10} + \pi_{11}m_{t-1} + \pi_{12}y_{t-1} + \gamma_1 t + \varepsilon_{1t} \\ \Delta y_t &= a_{20} + \pi_{21}m_{t-1} + \pi_{22}y_{t-1} + \gamma_2 t + \varepsilon_{2t}. \end{aligned}$$

Cointegration Suppose that they cointegrate so that $z_t = m_t - \beta y_t$ is I(0). The cointegrating vector is $(1, -\beta)$ and we have normalised the equation by setting the coefficient of m_t to unity, which is natural if we treat it as a demand for money function. This single restriction just identifies the cointegrating vector for r=1.We can write this

$$\begin{aligned} \Delta m_t &= a_{10} + \pi_{11}(m_{t-1} + \frac{\pi_{12}}{\pi_{11}}y_{t-1}) + \gamma_1 t + \varepsilon_{1t} \\ \Delta y_t &= a_{20} + \pi_{21}(m_{t-1} + \frac{\pi_{22}}{\pi_{21}}y_{t-1}) + \gamma_2 t + \varepsilon_{2t}. \end{aligned} \tag{13.1}$$

The cointegration restriction is that the long-run coefficient is the same in both equations,

$$\frac{\pi_{12}}{\pi_{11}} = \frac{\pi_{22}}{\pi_{21}}.$$

This says ($\pi_{11}\pi_{22} - \pi_{12}\pi_{21} = 0$) the determinant of Π is zero, so Π is singular, not of full rank. With this restriction (13.1) becomes:

$$\begin{aligned} \Delta m_t &= a_{10} - \alpha_1(m_{t-1} - \beta y_{t-1}) + \gamma_1 t + u_{1t} \\ \Delta y_t &= a_{20} - \alpha_2(m_{t-1} - \beta y_{t-1}) + \gamma_2 t + u_{2t} \end{aligned} \tag{13.2}$$

where $-\alpha_1 = \pi_{11}$ etc. Thus

$$\Pi = \begin{bmatrix} -\alpha_1 & +\alpha_1\beta \\ -\alpha_2 & +\alpha_2\beta \end{bmatrix}$$

which is clearly of rank 1, since a multiple of the first column equals the second column. A natural over-identifying restriction to test in this context would be that $\beta = 1$.

Restricted trend The equation has unrestricted trend and intercept, to restrict the trend we put it in the cointegrating vector, saving one further parameter:

$$\begin{aligned} \Delta m_t &= a_{10} - \alpha_1(m_{t-1} - \beta y_{t-1} + \gamma t) + u_{1t} \\ \Delta y_t &= a_{20} - \alpha_2(m_{t-1} - \beta y_{t-1} + \gamma t) + u_{2t} \end{aligned}$$

Exogenous income If y_t is weakly exogenous then $\alpha_2 = 0$, which can be tested and if accepted means income is a random walk with drift and (13.2) becomes

$$\begin{aligned} \Delta m_t &= a_{10} - \alpha_2(m_{t-1} - \beta y_{t-1}) + \gamma_1 t + u_{1t} \\ \Delta y_t &= a_{20} + u_{2t} \end{aligned}$$

Define $E(u_{it}u_{jt}) = \sigma_{ij}$, $i, j = 1, 2$, and noting that

$$E(u_{1t} \mid u_{2t}) = \frac{\sigma_{12}}{\sigma_{22}} u_{2t} = \frac{\sigma_{12}}{\sigma_{22}} (\Delta y_t - a_{20})$$

and defining $v_t = u_{1t} - E(u_{1t} \mid u_{2t})$, we can get the ECM treating y_t as exogenous:

$$\Delta m_t = (a_{10} - a_{20}) + \frac{\sigma_{12}}{\sigma_{22}} \Delta y_t - \alpha_2(m_{t-1} - \beta y_{t-1}) + \gamma_1 t + v_t.$$

An alternative just identifying restriction Rather than normalising (13.1) on m_t we could normalise it on y_t

$$\begin{aligned} \Delta m_t &= a_{10} + \pi_{12}\left(\frac{\pi_{11}}{\pi_{12}} m_{t-1} + y_{t-1}\right) + \gamma_1 t + \varepsilon_{1t} \\ \Delta y_t &= a_{20} + \pi_{22}\left(\frac{\pi_{21}}{\pi_{22}} m_{t-1} + y_{t-1}\right) + \gamma_2 t + \varepsilon_{2t}. \end{aligned}$$

which is observationally equivalent to (13.1), involves the same cross equation restriction ($\pi_{11}\pi_{22} - \pi_{12}\pi_{21} = 0$), but gives us a new cointegrating relation $z_t^* = \beta^* m_t - y_t$ and cointegrating vector $(\beta^*, -1)$.

14. Cointegration examples,

14.1. Old exam question

A second-order cointegrating vector error-correction model (VECM), with unrestricted intercepts and restricted trends, was estimated on quarterly US data from 1947Q3 to 1988Q4. The variables included were the logarithm of real consumption (c_t), the logarithm of real investment (i_t), and the logarithm of real income (q_t). The Johansen maximal eigenvalue tests for, r, the number of cointegrating vectors, were:

H_o	H_1	$Statistic$	$10\%CV$
$r = 0$	$r = 1$	34.6	23.1
$r \leq 1$	$r = 2$	15.8	17.2
$r \leq 2$	$r = 3$	3.3	10.5

The Johansen Trace Tests were:

H_o	H_1	$Statistic$	$10\%CV$
$r = 0$	$r \geq 1$	53.7	39.3
$r \leq 1$	$r \geq 2$	19.1	23.1
$r \leq 2$	$r = 3$	3.3	10.5

Assuming that $r = 2$, the following two just-identified cointegrating vectors $Z1_t$ and $Z2_t$ (standard errors in parentheses) were estimated:

c	i	q	t
1	0	-1.13	0.0003
		(0.16)	(0.0006)
0	1	-1.14	0.0007
		(0.26)	(0.001)

The system maximised log-likelihood (MLL) was 1552.9. The system was then estimated subject to the over-identifying restrictions that: (i) both coefficients of income were unity, giving a MLL of 1552.3; and (ii) not only were the income coefficients unity, but that the trend coefficients were also zero, giving a MLL of 1548.1.

The Vector Error Correction Estimates [t statistics] for the just identified system (constants included but not reported) were

	Δc_t	Δi_t	Δq_t
$Z1_{t-1}$	0.075068	0.262958	0.192686
	[2.74240]	[3.20914]	[4.63684]
$Z2_{t-1}$	−0.011232	−0.171416	0.009323
	[−0.67114]	[−3.42157]	[0.36694]
Δc_{t-1}	−0.209469	−0.171819	0.094535
	[−2.31259]	[−0.63368]	[0.68749]
Δi_{t-1}	0.022574	0.334330	0.156990
	[0.72374]	[3.58069]	[3.31537]
Δq_{t-1}	0.212411	0.697502	0.126186
	[3.17484]	[3.48267]	[1.24236]
R^2	0.146291	0.405637	0.320507
SER	0.007527	0.022533	0.011427

(a) How many cointegrating vectors do the tests indicate?

(b) If there are r cointegrating vectors, how many restrictions on each vector do you need to identify it.

(c) Interpret the just identifying restrictions used above.

(d) Test the two sets of overidentifying restrictions. 5% asymptotic (bootstrap) critical values are $\chi^2(2) = 5.99$ $(8.46), \chi^2(4) = 9.49$ (14.23). Comment on the difference between the results using the two sets of critical values.

(e) The VECM was estimated with unrestricted intercepts and restricted trends. What does this mean?

(f) Do you think investment is Granger Causal for Consumption.

Answer

(a) one (b) r

(c) Investment does not appear in the consumption function and consumption does not appear in the investment function.

(d) (i) 2(1552.9-1552.3)=1.2$<\chi^2(2)$, do not reject H_0 (ii) 2(1552.9-1548.1)=9.6$>\chi^2(4)$ reject H_0 . With bootstrap critical values you would not reject H_0 in case (ii). Small sample critical values given by the bootstrap are bigger than the asymptotic values and so one is less likely to reject using the bootstrap critical values.

(e) The VECM for a $m \times 1$ vector y_t with unrestricted intercepts and restricted trends is

$$\Delta y_t = \mu + \alpha(\beta y_{t-1} + \gamma t) + u_t$$

where the intercepts μ lie outside the error correction term and the trends γt are restricted to lie within it. Whereas one estimates m intercepts, one only estimates

r trend coefficients, giving $m - r$ restrictions.

(f) The fact that both $Z2_{t-1}$ (which is a function of lagged investment) and Δi_{t-1} are individually insignificant in the consumption equation suggests that investment may be Granger non-causal for consumption, though the two terms could be jointly significant.

15. Exogeneity, simultaneity and identification

15.1. Exogeneity

Originally a variable was said to be endogenous if it was simultaneously determined within a system. But the terminology was widened to include any case where the assumption of exogeneity failed including omitted variables, discussed above, and measurement error.

Exogeneity is a difficult concept, Hendry's text Dynamic Econometrics is a good treatment. There are a number of different definitions, which fall into two classes of approach in terms of (a) the relationship between the unobserved error term and the regressors (b) the joint distribution of the observed random variables. We have used the former above we now look at the latter.

The joint distribution of the random variables, y_t, x_t, can be written as the product of the distribution of y_t conditional on x_t and the marginal distribution of x_t :

$$D_j(y_t, x_t; \theta_j) = D_c(y_t \mid x_t; \theta_c) D_m(x_t; \theta_m) \tag{15.1}$$

θ_j is a vector of parameters of the joint distribution, θ_c of the conditional distribution, θ_m of the marginal. The distribution that we will be interested in is the distribution of y_t conditional on x_t and the parameters that we will be interested in are the parameters of the conditional distribution θ_c which we will usually denote by θ. **Weak exogeneity** requires that the parameters of interest should be functions only of the parameters of the conditional distribution, $\theta_c = \theta = (\beta, \sigma^2)$, and that the parameters of the conditional and marginal distributions should be 'variation free': there are no restrictions linking them. Essentially this says that we can ignore the information in the marginal distribution of x for the purpose of estimating particular parameters. Notice that exogeneity is not an inherent property of x, it is only defined relative to the parameters you want to estimate. x may be exogenous for some parameters and not for others. This is the assumption that we need for efficient inference about the parameters of interest.

Notice that in section (13.3) above income was weakly exogenous if $\alpha_2 = 0$, because the parameter of interest β did not appear in the the equation determining income, which gave the marginal distribution for income.

The main reasons for the exogeneity assumption failing in economics are: (a) omitted variables discussed above, where the parameters of interest are not of the distribution conditional on x_t, but on the distribution conditional on x_t and the omitted variable; (b) simultaneity, where the regressors are jointly determined with the dependent variable (prices and quantities are simultaneously determined by demand and supply), and (c) measurement errors in the regressors. In each case we need information about the processes generating the regressors to consistently estimate the parameters of interest. **Strong exogeneity** is weak exogeneity plus Granger Non-causality of y_t with respect to x_t, we need this assumption for forecasting. **Super exogeneity** requires that the parameters of the conditional distribution, θ_c, should be invariant to changes in the parameters of the marginal distribution of x_t. In this case even if the process generating x_t changes, the parameters of our regression do not change. We need this for policy analysis which usually involves changing right hand side policy variables and essentially this assumption precludes the Lucas Critique. Notice that these three definitions are presented in terms of the distributions of the observables, y_t and x_t, not the unobservable u_t.

The second approach, used above and very common in the text books, presents the assumptions in terms of the unobserved error or disturbance u_t. Notice that our assumption, in terms of the conditional distribution of y, $D_c(y \mid X; \theta) \sim N(X\beta, \sigma^2 I)$, is equivalent to an assumption in terms of the unconditional distribution of the disturbance $u \sim N(0, \sigma^2 I)$. In this framework, there are three types of exogeneity assumptions that are made about X. Firstly, it may be a set of fixed numbers, **non-stochastic** or deterministic. These phrases are all equivalent ways to describe the fact that X is not a random variable. Apart from trends and seasonals non-stochastic variables are rare in economics. Secondly, it may be **strictly exogenous,** a set of random variables which are distributed independently of the disturbance term. Thirdly, it may be **predetermined** a set of random variables which are uncorrelated with the disturbance term. If X is strictly exogenous, x_t is uncorrelated with the whole sequence of u_t, $t = 1, 2, ..., T$. If it is predetermined, it is only uncorrelated with the current value of u_t. Typically predetermined variables are lagged (past) values of y_t which are included in the x_t.

15.2. The Simultaneous Equations Model

Consider a vector of m endogenous variables, $\boldsymbol{y}_t$, jointly determined by k exogenous variables $\boldsymbol{x}_t$

$$\boldsymbol{B}\boldsymbol{y}_t = \boldsymbol{\Gamma}\boldsymbol{x}_t + \boldsymbol{u}_t; \quad \mathbf{E}(\mathbf{u}_t\mathbf{u}_t') = \boldsymbol{\Omega} \tag{15.2}$$

Where B is an $m \times m$ matrix, that describes how the endogenous variables influence each other and Γ is a $m \times k$ matrix. This is known as the structural form. The reduced form is

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{x}_t + \boldsymbol{B}^{-1}\mathbf{u}_t & (15.3)\\ \boldsymbol{y}_t &= \boldsymbol{\Pi}\boldsymbol{x}_t + \mathbf{v}_t & (15.4)\\ \mathbf{E}(\mathbf{v}_t\mathbf{v}_t') = \boldsymbol{\Sigma} &= \boldsymbol{B}^{-1}\boldsymbol{\Omega}\boldsymbol{B}^{-1\prime} \end{aligned}$$

We can estimate the $m \times k$ matrix $\boldsymbol{\Pi}$**,** since the $\mathbf{x}_t$ are exogenous, this is just m equations estimated by OLS. However we want to estimate both B an $m \times m$ matrix and Γ a $m \times k$ matrix, so we are m^2 elements short. This is the "identification problem", we need to specify m^2 prior restrictions to obtain all the parameters of the structural form. Another way to see this is that since, $\boldsymbol{\Pi} = \boldsymbol{B}^{-1}\boldsymbol{\Gamma} = (\boldsymbol{B}\boldsymbol{P})^{-1}\mathbf{P}\boldsymbol{\Gamma}$ for any $m \times m$ non-singular matrix $\boldsymbol{P}$, then $\boldsymbol{B}$ and $\boldsymbol{B}^* = (\boldsymbol{B}\boldsymbol{P})$ and $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}^* = \mathbf{P}\boldsymbol{\Gamma}$ are observationally equivalent. Note that this is the same problem we had with identifying cointegrating vectors above. In each case we need to specify $\mathbf{P}$ *a priori.* Here this again involves specifying m^2 a priori "just identifying" restrictions in order to obtain unique estimates.

The m^2 restrictions on the system mean that we need at least m restrictions on each of the m equations. If there are d restrictions available for an equation, when $d < m$, the equation is said to be underidentified or not identified and cannot be estimated; when $d = m$ it is said to be exactly identified or just identified; when $d \geqslant m$ it is said to be overidentified. You can have a system with some equations identified and others not identified. Over identifying restrictions can be tested.

One restriction on each equation will come from normalisation. Normalisation, specifies that the coefficient of a dependent variable equals unity. The requirement that $d \geqslant m$ for each equation is called the order condition, a necessary but not sufficient condition for identification. The sufficient condition is the rank condition. The order condition can be written in lots of different but equivalent ways. One way of expressing it for a particular equation is that the number of excluded exogenous variables (not appearing in that equation) must be greater or equal to the number of included right hand side endogenous variables.

Equation (15.3) is called the restricted reduced form, since it reflects any restrictions on $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, while (15.4) is called the unrestricted reduced form.

15.3. A Demand and supply example

Consider a simple demand and supply model for an agricultural product in structural form as

$$\begin{aligned} q_t^d &= \gamma_{10} + \beta_{12} p_t + \gamma_{11} y_t + u_{1t}, \\ q_t^s &= \gamma_{20} + \beta_{22} p_t + \gamma_{22} w_t + u_{2t}. \end{aligned}$$

Demand is determined by price and income, supply is determined by price and the weather and price adjusts so that demand equals supply $q_t^d = q_t^s = q_t$. This system simultaneously determines the two endogenous variables price and quantity in terms of the exogenous variables income and the weather and the errors, so the restricted reduced form is:

$$\begin{aligned} p_t &= [\beta_{12} - \beta_{22}]^{-1} \{(\gamma_{20} - \gamma_{10}) - \gamma_{11} y_t + \gamma_{22} w_t + (u_{2t} - u_{1t})\}' \\ q_t &= [\beta_{12} - \beta_{22}]^{-1} \{(\beta_{12}\gamma_{20} - \beta_{22}\gamma_{10}\} - \beta_{22}\gamma_{11} y_t + \beta_{12}\gamma_{22} w_t + (\beta_{12} u_{2t} - \beta_{22} u_{1t})\}, \end{aligned}$$

and the unrestricted reduced form is

$$\begin{aligned} p_t &= \pi_{10} + \pi_{11} y_t + \pi_{12} w_t + v_{1t} \\ q_t &= \pi_{20} + \pi_{21} y_t + \pi_{22} w_t + v_{2t} \end{aligned}$$

The standard demand-supply system in economics is normalised in an unusual form, making quantity the dependent variable in both equations. Usually systems are normalised so that each endogenous variable is the dependent variable in each equation.

We can write the system in matrix notation as in (15.2), with $m = 2$, $k = 3$ as

$$\begin{bmatrix} 1 & -\beta_{12} \\ 1 & -\beta_{22} \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \gamma_{10} & \gamma_{11} & 0 \\ \gamma_{20} & 0 & \gamma_{22} \end{bmatrix} \begin{bmatrix} 1 \\ y_t \\ w_t \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

In the demand-supply case

In general OLS estimates of the structural form will be inconsistent since in the demand equation u_{1t} will be correlated with p_t (which is a function of u_{1t} as the reduced form equations show), so the exogeneity assumption fails. The reduced form can be consistently estimated by OLS.

In the demand and supply example, both equations are exactly identified because we have two restrictions in each case, $d = 2, m = 2$. In the demand equation we have $\beta_{11} = 1; \gamma_{12} = 0$. In the supply equation $\beta_{21} = 1; \gamma_{21} = 0$. Here the rank condition is that $\gamma_{11} \neq 0$ and $\gamma_{22} \neq 0$, so that income and the weather do influence the system and are correlated with the endogenous variables.

Estimation can be done by Two stage Least Squares, which is the same as Instrumental Variables, IV, discussed below. First estimate the reduced form and obtain the predicted values for p_t and q_t :

$$\begin{aligned} \widehat{p}_t &= \widehat{\pi}_{10} + \widehat{\pi}_{11} y_t + \widehat{\pi}_{12} w_t \\ \widehat{q}_t &= \widehat{\pi}_{20} + \widehat{\pi}_{21} y_t + \widehat{\pi}_{22} w_t \end{aligned}$$

these are just functions of the exogenous variables and so are not correlated with u_{1t} and u_{2t} and can be used in two second stage regressions estimating the structural equations

$$\begin{aligned} q_t &= \gamma_{10} + \beta_{12} \widehat{p}_t + \gamma_{11} y_t + e_{1t} \\ q_t &= \gamma_{20} + \beta_{22} \widehat{p}_t + \gamma_{22} w_t + e_{2t} \end{aligned}$$

where $e_{1t} = u_{1t} + \beta_{i2} \widehat{v}_{1t}$ neither of which are correlated with $\widehat{p}_t$.

15.4. A Keynesian system example

Consider another example, the simple Keynesian model of an identity and a consumption function:

$$\begin{aligned} Y_t &= C_t + I_t, \\ C_t &= \alpha + \beta Y_t + u_t. \end{aligned}$$

Notice that identification is not an issue for the identity, since there are no coefficients to estimate. The restricted reduced form is

$$\begin{aligned} Y_t &= \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{u_t}{1-\beta} \\ C_t &= \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} I_t + \frac{u_t}{1-\beta}. \end{aligned}$$

Notice the coefficient of investment in the income equation is the standard Keynesian multiplier. The unrestricted reduced form which we can estimate is

$$\begin{aligned} Y_t &= \pi_{10} + \pi_{11} I_t + v_{1t} \\ C_t &= \pi_{20} + \pi_{21} I_t + v_{2t}, \end{aligned}$$

where $v_{1t} = v_{2t}$, etc. Clearly Y_t is correlated with u_t in the consumption function, since as the reduced form shows u_t determines Y_t through consumption. What is the covariance between Y_t and u_t? However we could estimate β by "indirect least squares", ILS, as the ratio of the two reduced form coefficients of investment, where the lower case letters indicate deviations from the mean:

$$\widehat{\beta}^{ILS} = \frac{\widehat{\pi}_{21}}{\widehat{\pi}_{11}} = \frac{\sum c_t i_t / \sum i_t^2}{\sum y_t i_t / \sum i_t^2} = \frac{\sum c_t i_t}{\sum y_t i_t}.$$

This can only be done in exactly identified cases, like this, where all the various estimators (2SLS, IV, ILS and others) give the same estimates. Note that the fourth term in the equation is exactly the same as the just identified IV estimator that appears below.

Suppose that we had two exogenous variables, investment and government expenditure

$$\begin{aligned} Y_t &= C_t + I_t + G_t, \\ C_t &= \alpha + \beta Y_t + u_t. \end{aligned}$$

the restricted and unrestricted reduced form equations for Y_t are

$$\begin{aligned} Y_t &= \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{1}{1-\beta} G_t + \frac{u_t}{1-\beta} \\ Y_t &= \pi_{10} + \pi_{11} I_t + \pi_{12} G_t + v_{1t}. \end{aligned}$$

The testable over-identifying restriction is that $\pi_{11} = \pi_{12}$.

15.5. Recursive systems and Impulse response functions

15.5.1. Identification through the covariance matrix

Above we considered identification by restrictions on the coefficient matrices, $\mathbf{B}$ and $\mathbf{\Gamma}$. But we can also get identification through restrictions on the covariance matrix $\mathbf{\Omega}$. If we assume that $\mathbf{\Omega}$ is diagonal, this gives us $m(m-1)/2$ restrictions, that all the off diagonal elements are zero. If we also assume that $\mathbf{B}$ is triangular, all the elements above the diagonal are zero, this gives us another $m(m-1)/2$ restrictions. Together with the m normalisation restrictions, this totals m^2 and the system is just identified. Such a system is called recursive and can be estimated

by OLS on each equation. An example is:

$$\begin{aligned} y_{1t} &= \gamma_1 x_t + u_{1t} \\ y_{2t} &= \beta_{21} y_{1t} + \gamma_2 x_t + u_{2t} \\ y_{3t} &= \beta_{31} y_{1t} + \beta_{32} y_{2t} + \gamma_3 x_t + u_{3t} \end{aligned}$$

with $E(u_{it}u_{jt}) = 0$. u_{2t} is not correlated with y_{1t} because there is no direct link, y_{2t} does not influence y_{1t}, and no indirect link, u_{2t} is not correlated with u_{1t}.Since none of the right hand side variables are correlated with the errors OLS is consistent. The system in the form (15.2) is

$$\begin{bmatrix} 1 & 0 & 0 \\ -\beta_{21} & 1 & 0 \\ -\beta_{31} & -\beta_{32} & 1 \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} x_t + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix}.$$

15.5.2. Impulse response functions

A VAR is the reduced form of a structural system in which instead of exogenous variables there appears the predetermined, lagged endogenous variables. In the case of a VAR(1) $\mathbf{x}_t = \mathbf{y}_{t-1}$ in (15.2):

$$\begin{aligned} \boldsymbol{B}\boldsymbol{y}_t &= \boldsymbol{\Gamma}\boldsymbol{y}_{t-1} + \boldsymbol{u}_t; \quad \mathbf{E}(\mathbf{u}_t\mathbf{u}_t') = \boldsymbol{\Omega} \\ \boldsymbol{y}_t &= \boldsymbol{B}^{-1}\boldsymbol{\Gamma}\boldsymbol{y}_{t-1} + \boldsymbol{B}^{-1}\,\mathbf{u}_t \\ \boldsymbol{y}_t &= \boldsymbol{\Pi}\boldsymbol{y}_{t-1} + \mathbf{v}_t \\ \mathbf{E}(\mathbf{v}_t\mathbf{v}_t') = \boldsymbol{\Sigma} &= \boldsymbol{B}^{-1}\boldsymbol{\Omega}\boldsymbol{B}^{-1\prime} \end{aligned}$$

To analyse the interaction between the variables in a VAR, it is common to use impulse response functions, IRFs, these measure the time profile of the effect of a shock (usually a one standard error shock) on the expected future values of the variables in the system. They are presented graphically plotting the expected effect on $y_{i,t+h}$ of a shock to an element of $\mathbf{v}_t$, say v_{jt}. See the applied exercise.

IRFs are derived from the moving average representation, which for a stationary VAR(1) is;

$$\boldsymbol{y}_t = \mathbf{v}_t + \boldsymbol{\Pi}\boldsymbol{v}_{t-1} + \boldsymbol{\Pi}^2\boldsymbol{v}_{t-2} + \boldsymbol{\Pi}^3\boldsymbol{v}_{t-3} + ...$$

This is similar to (9.1) above. In the general case $\mathbf{A}(L)\mathbf{y}_t = \mathbf{v}_t$, the MA representation is $\mathbf{y}_t = \mathbf{A}(L)^{-1}\mathbf{v}_t = \boldsymbol{\Psi}(L)\mathbf{v}_t$, as long as $\mathbf{A}(L)$ is invertible.

When we shock an element of $\mathbf{v}_t$, say v_{it}, the other reduced form errors will also change because $\boldsymbol{\Sigma}$ is not a diagonal matrix and $E(v_{it}v_{jt}) \neq 0$. Generalised IRFs,

which are an option in many programs, allow for this, using the reduced form covariance matrix to estimate how a shock to one error will be associated with changes in the other errors. These cannot be identified with a particular structural shock, since the reduced form errors are a transformation of the structural errors and even if the structural errors, $\mathbf{u}_t$ were uncorrelated, $\mathbf{\Omega}$ was diagonal, the reduced form errors would be correlated.

15.5.3. Orthogonalised IRFs and the Cholesky decomposition

A common identification assumption is that the structural system is recursive, with a causal ordering and orthogonal shocks, as in the previous sub-section. This is the basis of the Orthogonalised IRFs obtained by Cholesky decomposition. many programs will impose these restrictions using the order in which you list the variables. Write the MA representation

$$\mathbf{y}_t = \mathbf{v}_t + \mathbf{\Psi}_1\mathbf{v}_{t-1} + \mathbf{\Psi}_2\mathbf{v}_{t-2} + ... = \sum_{j=0}^{\infty} \mathbf{\Psi}_j\mathbf{v}_{t-j},$$

with $\mathbf{\Psi}_0 = I$. The Cholesky decomposition of $\mathbf{\Sigma} = \mathbf{P}\mathbf{P}'$ where $\mathbf{P}$ is lower triangular. This decomposition is not unique it depends on the ordering of the variables. Unlike Cholesky IRFs, generalised IRFs are invariant to ordering. The MA representation can then be written

$$\mathbf{y}_t = \sum_{j=0}^{\infty} \mathbf{\Psi}_j\mathbf{P}\mathbf{P}^{-1}\mathbf{v}_{t-j} = \sum_{j=0}^{\infty} \mathbf{B}_j\boldsymbol{\eta}_{t-j}$$

where $\mathbf{B}_j = \mathbf{\Psi}_j\mathbf{P}$ and $\boldsymbol{\eta}_{t-j} = \mathbf{P}^{-1}\mathbf{v}_{t-j}$, so $E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t') = \mathbf{P}^{-1}\mathbf{\Sigma}\mathbf{P}^{-1\prime} = I_m$. The new errors are orthogonal. Whether one would expect structural shocks, e.g. demand and supply shocks, to be orthogonal is a matter of debate. There are a variety of other ways to identify structural VARs.

16. Instrumental Variable estimation

Let us return to the LRM

$$y = X\beta + u$$

where X is a $T \times k$ matrix, but the X are not exogenous, so $E(X'u) \neq 0$. This may happen because of simultaneity (some of the X are jointly determined with the y)

or because of omitted variables (u contains variables correlated with X) or because some of the X are measured with error. In consequence, the OLS estimates will be biased and inconsistent. Suppose that there exists a $T \times i$, matrix of 'Instruments', W, where $i \geq k$, which are correlated with X so that $E(W'X) \neq 0$ but are not correlated with the disturbances so that $E(W'u) = 0$. W will include the elements of X that are exogenous (including the column of ones for the constant), but we need at least one instrument for each endogenous X. If $i = k$, the model is said to be just-identified, if $i > k$ it is said to be over-identified. The condition $i \geq k$ is the same order condition, we encountered in simultaneous systems. There is also the rank condition from $E(W'X) \neq 0$ to ensure that $(W'X)$ is of full rank and $(W'X)^{-1}$ exists. Notice $(W'X)' = (X'W)$.

If the model is just or exactly identified, the consistent instrumental variable estimator is

$$\widetilde{\beta} = (W'X)^{-1}W'y$$

with variance-covariance matrix $\sigma^2(W'X)^{-1}W'W(X'W)^{-1}$. The efficiency of the estimator will increase (the size of the standard errors reduce) with the correlation between W and X. Notice this estimator chooses the β that imposes the orthogonality condition:

$$\begin{aligned} W'\widetilde{u} &= 0 \\ W'(y - X\widetilde{\beta}) &= 0 \\ W'y &= W'X\widetilde{\beta} \\ (W'X)^{-1}W'y &= \widetilde{\beta}. \end{aligned}$$

Notice that in the case of a single right hand side endogenous variable, like the Keynesian consumption function above (where y corresponds to C_t, X to Y_t and W to I_t) the IV estimator is the ratio of the coefficient of the regression of y_t on w_t to the coefficient of the regression of x_t on w_t.

If the model is over identified, the Generalised Instrumental Variable Estimator (GIVE), which is the same as the Two Stage Least Squares Estimator (2SLS) is obtained by first regressing each of the X on the W;

$$X = WB + V$$

to give the $i \times k$ matrix of coefficients $\widehat{B} = (W'W)^{-1}W'X$, then calculating the predicted values of X as: $\widehat{X} = W\widehat{B} = W(W'W)^{-1}W'X$. Substituting $X = \widehat{X} + \widehat{V}$ into the original regression we get:

$$y = (\widehat{X} + \widehat{V})\beta + u = \widehat{X}\beta + (\widehat{V}\beta + u).$$

Now $\widehat{X}$ is uncorrelated with u since it is only a function of the W which are uncorrelated with u, and is uncorrelated with $\widehat{V}$ by construction. Therefore it satisfies our exogeneity conditions. The GIVE estimator is

$$\begin{aligned}\widetilde{\beta} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y \\ &= (X'W(W'W)^{-1}W'X)^{-1}X'W(W'W)^{-1}W'y \\ &= (X'P_wX)^{-1}X'P_wy\end{aligned}$$

with $P_w = W(W'W)^{-1}W'$ being a projection matrix. Its variance covariance matrix is $\sigma^2(X'P_wX)^{-1}$ and we estimate the residuals using the actual X not their fitted values:

$$s^2_{IV} = (y - X\widetilde{\beta})'(y - X\widetilde{\beta})/(T-k)$$

This estimator chooses $\widetilde{\beta}$ to make $X'P_w\widetilde{u} = 0$. It minimises the estimate of $\widetilde{u}'P_w\widetilde{u}$, the IV minimand, rather than $\widehat{u}'\widehat{u}$ as OLS does. Many programs report the IV minimand, which will be zero when the model is just identified. Show this by multiplying out

$$(y - X\widetilde{\beta})'W(W'W)^{-1}W'(y - X\widetilde{\beta})$$

for the just identified case where $\widetilde{\beta} = (W'X)^{-1}W'y$.

16.1. Example: Testing

Below we use a simple example to illustrate estimation with potentially endogenous regressors and testing for endogeneity, over-identifying restrictions and weak instruments.

Suppose x_{it} denote potentially endogenous variables, w_{it} exogenous variables and the structural model is

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 w_{1t} + u_t$$

with w_{2t}, w_{3t}, w_{4t} as potential instruments. Note that $X = [1\ x_{1t}\ x_{2t}\ w_{1t}]$, and $W = [1\ w_{1t}\ w_{2t}\ w_{3t}\ w_{4t}]$ so $k = 4$, $i = 5$ and the degree of overidentification is one. To get the fitted values, you run the two 'reduced form' regressions of $X = WB + v$:

$$\begin{aligned}x_{1t} &= b_{10} + b_{11}w_{1t} + b_{12}w_{2t} + b_{13}w_{3t} + b_{14}w_{4t} + v_{1t} \\ x_{2t} &= b_{20} + b_{21}w_{1t} + b_{22}w_{2t} + b_{23}w_{3t} + b_{24}w_{4t} + v_{2t}\end{aligned} \tag{16.1}$$

to give you estimates of $\widehat{x}_{1t}, \widehat{x}_{2t}, \widehat{v}_{1t}, \widehat{v}_{2t}$; for GIVE/2SLS you use the fitted values in the regression:

$$y_t = \beta_0 + \beta_1 \widehat{x}_{1t} + \beta_2 \widehat{x}_{2t} + \beta_3 w_{1t} + e_t \tag{16.2}$$

where $e_t = u_t + \beta_1 \widehat{v}_{1t} + \beta_2 \widehat{v}_{2t}$. The OLS estimates from this regression give the GIVE estimates of β_i and the residuals are estimated as:

$$\widetilde{u}_t = y_t - (\widetilde{\beta}_0 + \widetilde{\beta}_1 x_{1t} + \widetilde{\beta}_2 x_{2t} + \widetilde{\beta}_3 w_{1t})$$

i.e. not using the fitted values.

You do not have to do GIVE/2SLS estimation in two stages in practice, since it is programmed into most packages. You just choose the option and list the instruments in addition to the model. Do not forget to include constant and right hand side exogenous variables among the instruments.

However, it is usually a good idea to look at the F statistic on the reduced form regressions in (16.1). A rule of thumb is that this should be greater than about 10. If the instruments are weak, do not explain x_{it} very well, then the GIVE estimates will be badly biased and have large variance even in large samples.

If the instruments (or more precisely the over-identifying restrictions which exclude w_{2t}, w_{3t}, w_{4t} from the structural model) are valid, these GIVE or 2SLS residuals should be uncorrelated with the instruments. This can be tested by a Sargan (Bassman) test. The GMM version is the Hansen J test. Eviews calls it a J test. It involves regressing the GIVE residuals on all the instruments:

$$\widetilde{u}_t = c_0 + c_1 w_{1t} + c_2 w_{2t} + c_3 w_{3t} + c_4 w_{4t} + \varepsilon_t$$

and testing the hypothesis $c_1 = c_2 = c_3 = c_4 = 0$, this will be distributed $\chi^2(i-k)$, i.e. with degrees of freedom equal to the number of overidentifying restrictions, one in this case. This can also be expressed as the ratio of the IV minimand (see above) to the GIVE variance. When the model is just identified, the IV minimand is zero, so the test is not defined. .

To test whether the x_{it} are in fact exogenous you can use the Wu-Hausman test. To do this you save the residuals from the reduced form regressions, $\widehat{v}_{1t}, \widehat{v}_{2t}$, and include them in the original regression, i.e. run by OLS:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 w_{1t} + \delta_1 \widehat{v}_{1t} + \delta_2 \widehat{v}_{2t} + u_t$$

then test the null that they are exogenous $H_0 : \delta_1 = \delta_2 = 0$. Rejection of the null (significant reduced form residuals) indicates that one or both of them are

endogenous and GIVE should be used. This tests whether there is a significant difference between the OLS and GIVE estimates.

This is a convenient form for the Hausman test in this case. In general, suppose we have a $k \times 1$ vector of estimates, $\widehat{\delta}$ from an estimator which is efficient under the null, but inconsistent under the alternative, and a set of estimates $\widetilde{\delta}$ from an estimator which is less efficient under the null but consistent under both the null and alternative. Then the Hausman test statistic is the quadratic form

$$\left(\widehat{\delta} - \widetilde{\delta}\right)' \left[V(\widetilde{\delta}) - V(\widehat{\delta})\right]^{-1} \left(\widehat{\delta} - \widetilde{\delta}\right) \tilde{} \chi^2(k)$$

In this case the null is that x_{it} are exogenous, $\widehat{\delta}$ is the OLS estimator and $\widetilde{\delta}$ the GIVE estimator.

17. Bayesian Estimation

17.1. Introduction

Frequentist or classical statistics

- Regards probabilities as the limits of relative frequencies as the sample size goes to infinity
- regards parameters as fixed numbers;
- imagines some sampling distribution over lots of hypothetical samples of which the data is just one;
- Uses the Neyman-Pearson hypothesis testing framework

Bayesian Statistics

- Regards probabilities as measuring degrees of belief
- Regards Parameters as random variables
- Uses prior distributions for the parameters based on past experience and uses Bayes rule to provides a systematic way to update beliefs
- Estimation and inference.is done conditional on the observed data not sets of hypothetical samples

- Has an explicit loss function based on decision criteria and do not test

Gary Koop, *Bayesian Econometrics*, Wiley 2003 is very good. Sharon Bertsch McGrayne, *The theory that would not die*, Yale University Press 2011. Non-technical account of the history of Bayes' rule

In the literature distinguish (a) committed Bayesians from (b) pragmatic Bayesians who only use it for particular problems where frequentist methods do not work, e.g. Model Selection, DSGEs and VARS and (c) ad hoc Bayesians who use priors and Bayesian interpretations when doing frequentist statistics.

17.2. Bayes Theorem/rule

Bayes Theorem for continuous variables A and B follows from the definitions of conditional probability in terms of the joint and marginal probabilities

$$f(A \mid B) = \frac{f(A,B)}{f(B)}$$
$$f(B \mid A) = \frac{f(A,B)}{f(A)}$$

so

$$f(A,B) = f(B \mid A)f(A)$$

giving Bayes Theorem

$$f(A \mid B) = \frac{f(B \mid A)f(A)}{f(B)}$$

Bayes Rule

Treat A as the parameter θ and B as the data Y.

$$f(\theta \mid Y) = \frac{f(Y \mid \theta)f(\theta)}{f(Y)}$$

For data Y and random parameter θ, Bayesian statistics derives the posterior distribution, $f(\theta \mid Y)$, as proportional to the product of the likelihood, $f(Y \mid \theta)$ and the prior distribution, $f(\theta)$, treating $f(Y)$ as a constant (it has no information about θ, so can be ignored).

$$f(\theta \mid Y) \propto f(Y \mid \theta)f(\theta).$$

We will usually write the $\propto$ as $=$.

Need priors, $f(\theta)$. Can use uninformative priors but they are likely to be improper: $f(\theta)$ does not integrate to one. Conjugate priors are widely used because when combined with the likelihood they give a posterior with the same form of distribution.

Bayes rule gives us a posterior distribution for the parameter conditional on the data. Choice of an estimator is a decision problem. To choose an estimator we need a loss function. Quadratic loss function gives mean, absolute loss function gives median. Suppose that we choose the mean, this is

$$E(\theta) = \int \theta f(\theta \mid Y) d(\theta)$$

which involves integration, over the support of θ. Similarly, estimating the posterior variance to get a standard error involves integration. These integrals can rarely be worked out analytically, instead done numerically through Markov Chain Monte Carlo, MCMC, methods. Increased computing power has made Bayesian methods easier to apply.

17.3. Regression

One case where we can get analytical results is the normal linear regression model

$$y \sim N(X\beta, \sigma^2 I) = f(y \mid X\beta, \sigma^2 I)$$

Bayes Rule is

$$f(\beta, \sigma^2 \mid y, X) = \frac{f(y, X \mid \beta, \sigma^2) f(\beta, \sigma^2)}{f(y, X)}$$

If X is distributed independently of β, σ^2 we can condition on it

$$\begin{aligned} f(\beta, \sigma^2 \quad \mid \quad y, X) &= \frac{f(y \mid X\beta, \sigma^2 I) f(X) f(\beta, \sigma^2)}{f(y \mid X) f(X)}, \\ &= \frac{f(y \mid X\beta, \sigma^2 I) f(\beta, \sigma^2)}{f(y \mid X)} \end{aligned}$$

In much Bayesian analysis it is more convenient to work with the precision, the inverse of the variance, $h = \sigma^{-2}$ or covariance matrix: $H = V^{-1}$. Normal gamma prior

$$f(\beta, h) = N(\beta \mid \underline{\beta}, \underline{H}) f_\gamma(h \mid \underline{\sigma^2}, \underline{\nu}_\sigma)$$

with for β prior mean $\underline{\beta}$, prior precision $\underline{H}$ and for σ^2 $\underline{\sigma^2}, \underline{\nu_\sigma}$. Together with the normal likelihood this gives a normal gamma posterior

For regression model $y = X\beta + u$ with least squares estimates $\hat{\beta} = (X'X)^{-1}X'y$, $H(\hat{\beta}) = \hat{\sigma}^{-2}(X'X)$ prior mean $\underline{\beta}$, prior precision $\underline{H}$

The posterior is normally distributed with a mean which is a matrix weighted average

$$\overline{\beta} = (H(\hat{\beta}) + \underline{H})^{-1}(H(\hat{\beta})\hat{\beta} + \underline{H}\underline{\beta})$$

and precision $\overline{H} = (H(\hat{\beta}) + \underline{H})$.

Elements of a matrix weighted average vector, do not have to lie between the prior and OLS, $\overline{\beta_i}$ need not be between $\underline{\beta}_i$ and $\hat{\beta}_i$.As $T \to \infty$ $H(\hat{\beta})$ gets larger, while $\underline{H}$ is constant so asymptotically $\overline{\beta}$ goes to the Maximum Likelihood estimator $\hat{\beta}$.

This form of the Bayesian regression estimator

$$\overline{\beta} = (H(\hat{\beta}) + \underline{H})^{-1}(H(\hat{\beta})\hat{\beta} + \underline{H}\underline{\beta})$$

can be interpreted as a shrinkage estimator (like Ridge Regression or Lasso), shrinking the least squares estimator to the prior $\underline{\beta}$, which could be zero or as combining estimates of β from two different samples.

You can implement Bayesian estimation by generating data from the prior (e.g. a DSGE model) and adding it to the real data.

17.4. Bayesian VARs

VARs are just linear regressions so the simple analytical results for Bayesian regression can be applied. The big problem with VARs (Very Awful Regressions according to Zellner) is that they have too many parameters, this means that they do not forecast well. The Bayesian regression estimator can be regarded as a shrinkage estimator, shrinking the least squares estimator to priors $\underline{\beta}$, which could be zero. The rery popular Minnesota (Litterman) Priors, treat all the variables as random walks.So the prior is to put the coefficient of the lagged dependent variable to one if I(1) and all other coefficients to zero.

In EViews Minnesota/Litterman prior, Mu1 is the prior for the AR1 coefficient, lambda1 the tightness of the prior for the AR1 coefficient, the smaller lambda1 the tighter the prior, zero imposes the prior,.inf gets you an uniformative prior, lambda2 the tightness of the other variables, lambda3 the tightness of the lags

Issues in forecasting using a VAR

- How many variables to include in the VAR?
- Transformations of variables? e.g. differences of logs,
- How long a sample? When to start?
- Observations to save for ex post forecasts for evaluation:
- Lag lengths?
- Allow for cointegration?
- Frequentist or Bayesian VAR.

18. Measurement error

One cause of correlation between errors and regressors is measurement error. Suppose the model is

$$y_t = \beta x_t^* + \varepsilon_t; \tag{18.1}$$

where the variables are measured as deviations from their means and the true value x_t^* is not observed, but we observe

$$x_t = x_t^* + v_t \tag{18.2}$$

where

$$\begin{aligned} E(\varepsilon_t) &= E(v_t) = 0 \\ E(\varepsilon_t) &= \sigma_\varepsilon^2;\ E(v_t) = \sigma_v^2 \end{aligned}$$

and ε_t and v_t are independent of each other and x_t^*. In some cases, e.g. where x_t^* was the expected value of x_t we may have suitable instruments and can apply instrumental variables, but suppose we do not.

Now

$$\begin{aligned} y_t &= \beta x_t^* + \varepsilon_t \\ &= \beta(x_t - v_t) + \varepsilon_t \\ &= \beta x_t + (\varepsilon_t - \beta v_t) \\ &= b x_t + u_t \end{aligned}$$

Clearly x_t and u_t are correlated $E(x_t, u_t) = E((x_t^* + v_t)(\varepsilon_t - \beta v_t) = -\beta\sigma_v^2$, hence b will be an inconsistent estimator for β. x_t is not weakly exogeous for β, because we need to know information about the marginal distribution of x_t, i.e. σ_v^2. We can observe the variances for y_t and x_t and their covariance:

$$S_{xx} = \frac{1}{T}\sum x_t^2;\ S_{yy} = \frac{1}{T}\sum y_t^2;\ S_{xy} = \frac{1}{T}\sum x_t y_t.$$

The variables are defined as deviations from their means. Assuming large samples, we can match these up with their theoretical values; defining the variance of x_t^* as σ_*^2

$$\begin{aligned} S_{xx} &= \sigma_*^2 + \sigma_v^2 \\ S_{yy} &= \beta^2\sigma_*^2 + \sigma_\varepsilon^2 \\ S_{xy} &= \beta\sigma_*^2 \end{aligned}$$

The first line is got by squaring (18.2), and using the fact that the covariance of v_t and x_t^* is zero; the second line is got by squaring (18.1) ; the third line is got by multiplying (18.1) by (18.2). The OLS estimator from a regression of y_t on x_t is

$$\widehat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\beta\sigma_*^2}{\sigma_*^2 + \sigma_v^2} < \beta.$$

So unless $\sigma_v^2 = 0$, the direct least squares estimator is biased downwards. Consider taking the inverse of the coefficient of the reverse regression of x_t on y_t which is

$$\widehat{d} = \frac{S_{yy}}{S_{xy}} = \frac{\beta^2\sigma_*^2 + \sigma_\varepsilon^2}{\beta\sigma_*^2} > \beta$$

so unless $\sigma_\varepsilon^2 = 0$, this reverse least squares estimator is biased upwards. This gives us a bound in large samples

$$\widehat{b} < \beta < \widehat{d}$$

This may be useful in seeing the size of the effect of the possible measurement error. Unfortunately this does not generalise to more than two variables in any simple way; but with more variables there may be other ways to deal with measurement error.

Up to now we have considered point identification, a parameter is either identified or not identified. In this case the parameter is identified as being within a bound, β is between $\widehat{b}$ and $\widehat{d}$. There are other cases of identification within bounds.

The model is not point identified because we have three pieces of information S_{ij} and four theoretical parameters. One extra piece of information would identify it. If we knew that the errors in measurement were the same size as the errors in equation $\sigma_\varepsilon^2 = \sigma_v^2$ (or any other known ratio) this would identify it. In the case where $\sigma_\varepsilon^2 = \sigma_v^2 = \sigma^2$ then

$$\begin{aligned} S_{xx} &= \sigma_*^2 + \sigma^2 \\ S_{yy} &= \beta^2\sigma_*^2 + \sigma^2 \\ S_{xy} &= \beta\sigma_*^2 \end{aligned}$$

From the third equation, $\sigma_*^2 = S_{xy}/\beta$; from the first equation

$$\sigma^2 = S_{xx} - \sigma_*^2 = S_{xx} - S_{xy}/\beta$$

substituting these in the second equation gives

$$S_{yy} = \beta^2 \left(S_{xy}/\beta\right) + S_{xx} - S_{xy}/\beta.$$

Rearranging this shows β is a solution to the quadratic equation

$$\beta^2 S_{xy} + \beta(S_{xx} - S_{yy}) - S_{xy} = 0.$$

Which by the usual formula

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

gives

$$\frac{-(S_{xx} - S_{yy}) \pm \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}.$$

18.1. Expectations

Suppose that we have a Phillips Curve where inflation depends on the expected output gap and past inflation:

$$\pi_t = \alpha + \beta E(y_{t+1} \mid I_t) + \gamma\pi_{t-1} + u_t$$

where I_t indicates information available in t and expected output is given by

$$E(y_{t+1} \mid I_t) = \rho_0 + \rho_1 y_t + \rho_2 y_{t-1}$$

then we could estimate by IV

$$\pi_t = \alpha + \beta y_{t+1} + \gamma \pi_{t-1} + u_t$$

instrumenting y_{t+1} by y_t and y_{t-1}. That is we replace y_{t+1} by its prediction $\widehat{y}_{t+1} = \widehat{\rho}_0 + \widehat{\rho}_1 y_t + \widehat{\rho}_2 y_{t-1}$. This model is over-identified and the over-identifying restriction can be tested.

18.2. Principal Components

If you have a number of indicators of an unobserved variable or factor, then Principal Components, PC, can provide an estimate. The PC of a $T \times N$ data matrix $\mathbf{X}$ (which is usually standardised by subtracting the mean of the variable and dividing by the standard deviation) are the linear combination which explains as much as possible of the variance of all the $\mathbf{X}$. The first principal component is $f_1 = \mathbf{X}a_1$ where the variance of f_1 i.e. $f_1'f_1 = \sum f_{1t}^2 = a_1'\mathbf{X}'\mathbf{X}a_1$ is maximised. Notice that if the data are standardised, $\mathbf{X}'\mathbf{X}$ is the correlation matrix of the data, otherwise it is the covariance matrix. This $z_1'z_1$ can be made as large as you like depending on the units of a_1 so we need to choose a normalisation that determines scale, it is usual to use $a_1'a_1 = 1$. Set this up as a Lagrangian,

$$\begin{aligned} \pounds &= a_1'\mathbf{X}'\mathbf{X}a_1 - \lambda_1(a_1'a_1 - 1) \\ \frac{\partial \pounds}{\partial a_1} &= \mathbf{X}'\mathbf{X}a_1 - \lambda_1 a_1 = 0. \end{aligned}$$

Thus λ_1 is the largest eigenvalue and a_1 the corresponding eigenvector. If the data are standardised λ_1 tells you the proportion of the variation in $\mathbf{X}$ explained by the first PC. One can get the other PCs in a similar way and they will be orthogonal (uncorrelated). This gives you N new variables which are linear combinations of the $\mathbf{X}$ that is $\mathbf{F} = \mathbf{XA}$. One uses a subset of these PCs corresponding to the r largest eigenvalues. There are various ways to choose r, one is to use any PCs where the eigenvalues from standardised data are greater than one. In EViews to get PCs define a group, the variables in $\mathbf{X}$, open the group; choose View and one of the options will be to calculate the PCs for the group. These are known as static PCs or factors, dynamic factors take the PCs of the long-run covariance matrix (spectral density matrix) of $\mathbf{X}$. It can be difficult to give an interpretation to the PCs, but often in time series the first PC is very similar to the mean.

PCs are used for various purposes including factor augmented VARs, FAVARs. Suppose we have a $N \times 1$ vector $\mathbf{X}_t$, where N is large, e.g. 400, and we can express

this as determined by a $r \times 1$ vector of factors $\mathbf{F}_t$, estimated by the first r PCs, where r is small, e.g. 3:

$$\mathbf{X}_t = \Lambda \mathbf{F}_t + v_t. \tag{18.3}$$

The factors are then included with the observed endogenous variables, $\mathbf{y}_t$, in a VAR, such as the VAR(1)

$$\begin{aligned} \mathbf{y}_t &= \mathbf{a}_{10} + \mathbf{A}_{11}\mathbf{y}_{t-1} + \mathbf{A}_{12}\mathbf{F}_{t-1} + \mathbf{u}_{1t}, \\ \mathbf{F}_t &= \mathbf{a}_{20} + \mathbf{A}_{21}\mathbf{y}_{t-1} + \mathbf{A}_{22}\mathbf{F}_{t-1} + \mathbf{u}_{2t}. \end{aligned} \tag{18.4}$$

We can estimate this as usual and calculate IRFs not only for $\mathbf{y}_t$ and $\mathbf{F}_t$ but using (18.3) all the $\mathbf{X}_t$.