Essex Summer School in Social Science Data Analysis and Collection

University of Essex

Analysis of Time Series Data: Non-Stationarity, Volatility, VARs, and Cointegration

Konstantinos Drakos

Professor
Department of Accounting and Finance
Athens University of Economics and Business
Greece

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

0. Basic Time Series Concepts

0.1 Introduction

A time series is a collection of observations made **sequentially** in time. Usually, these observations are taken at equally spaced intervals over time. An intrinsic feature of a time series is that, typically, adjacent observations are **dependent**. The nature of this dependence among observations of a time series is of considerable practical interest. Time Series Analysis is concerned with techniques for the analysis of this dependence.

0.2 Terminology

A time series is said to be **continuous** when observations are made continuously in time. Examples include the temperature at a given location, the price of a commodity, or the position of a projectile. A time series is said to be **discrete** when observations are taken only in specific times. Examples include the annual crop yields of harvest, monthly salaries or the majority of a political party at a general election. In this course we are exclusively concerned with discrete time series where the observations are taken at equal intervals. (i.e. monthly, quarterly, annually etc)

Much statistical theory is concerned with random samples of independent observations. The special feature of time series analysis is the fact that successive observations are usually not independent and that the analysis must take into account the time order of observations. When successive observations are dependent, future values may be predicted from past observations. If a time series can be predicted exactly, it is said to be **deterministic**. But most time series are **stochastic** in that future is only partly determined by past values, so that exact predictions are impossible and must be replaced by the idea that future values have a probability distribution, which is conditioned by knowledge of past values.

0.3 Stationary stochastic processes and their properties in the time domain

A statistical phenomenon that evolves in time according to probabilistic laws is called a **stochastic process**. The time series to be analysed may then be thought as a particular **realisation**, produced by the underlying probability mechanism of the system under study. In other words, in analysing a time series we regard it as a realisation of a stochastic process. To put more formally, a stochastic process is a family of random variables, defined in a probability space: $\{\boldsymbol{X_t}\}$, $t = ...-1, 0, +1, ...$ Now, time series is a sample path or realisation of a stochastic process, whose parameter (index) denotes time. In other words, it is an observation taken from a multivariate probability distribution.

A very special class of stochastic processes, called **stationary** processes, is based on the assumption that the process is in a particular state of **statistical equilibrium**. A stochastic process is said to be **strictly stationary** if its properties are unaffected by a change of time origin, that is, if the joint probability distribution associated with m observations $z_t, z_{t+1}, z_{t+2}, ...$made at any set of times t, $t+1, t+2,...$, is the same as that associated with m observations $z_{t+k}, z_{t+1+k}, z_{t+2+k}, …$made at times $t+k, t+1+k, t+2+k,...$

Thus, for a discrete process to be strictly stationary, the joint probability distribution of any set of observations must be unaffected by shifting all the times of observation forward or backward by any integer amount k. However, the notion of strict stationarity is very strong and is rarely satisfied by times series encountered in social sciences. A rather weaker notion related is the so-called **weak** or **second order** stationarity or wide sense stationarity. A time series is called weakly stationary if:

- $E(X_t) = \mu$ for every t (meaning independent of time)

- $Cov(X_t, X_{t+k}) = E[(X_t - \mu)(X_{t+k} - \mu)] = \gamma_k$ for every t (meaning independent of time and only a function of the time lag. (depends only on k, the length of time separating the observations, and not *t*, the date of observation)

No assumptions are made about higher order moments than those of second order. By letting $k = 0$, we note that the above assumption about the covariance function implies that the variance, as well as the mean, is constant.

A few remarks:

- A strictly stationary process is weakly stationary
- If the process is Gaussian, then weak stationarity implies strict stationarity
- By symmetry $\gamma_k = \gamma_{-k}$
- So, the graph of a stationary series will vary randomly around a constant (stable) mean value and also its variance will be constant through time.

0.4 Autocovariance

For a time series $\{X_t\}$ we define as **autocovariance** (ACV) of k-order the quantity: $Cov(X_t, X_{t+k}) = E[(X_t - \mu)(X_{t+k} - \mu)] = \gamma_k$, $k = \pm 1, \pm 2$

The term "auto" is prefixed because the members of the series are generated from the same stochastic process. The ACV function is an even function of *k*. that is because: $Cov(X_t, X_{t+k}) = Cov(X_{t+k}, X_t) = Cov(X_t, X_{t-k}) = \gamma_{-k}$

Obviously, $\gamma_0 = Cov(X_t, X_t) = Var(X_t)$

Remark:

Positive first-order autocovariance means that there is a tendency for the next observation to be towards the same side (sign) as the previous one with respect to the mean. For the above time series, the matrix:

$$\Gamma = V(X) = \begin{bmatrix} \gamma_o & \gamma_1 & \gamma_2 & \cdots & \gamma_{k-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{k-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{k-3} \\ . & . & . & \cdots & . \\ \gamma_{k-1} & \gamma_{k-2} & \gamma_{k-3} & \cdots & \gamma_0 \end{bmatrix},$$ is called the **autocovariance matrix**.

Obviously, this matrix is symmetric and positive-definite inheriting these properties form the autocovariance function as discussed above.

0.5 Autocorrelation

For a time series $\{X_t\}$ we define as **autocorrelation** (ACR) of k-order the quantity: $\rho_k = \frac{Cov(X_t, X_{t+k})}{\sqrt{V(X_t)V(X_{t+k})}} = \frac{\gamma_k}{\sqrt{\gamma_0\gamma_0}} = \frac{\gamma_k}{\gamma_0}$. The graph of ρ_k is called correlogram, and will be shown later, provides vital information for the time series.

A few remarks:

- Because $\gamma_k = \gamma_{-k}$ it can be shown that $\rho_k = \rho_{-k}$
- Obviously the ACR function has the same properties with the ACV function, and furthermore satisfies the condition $\rho_0 = 1$
- Also, $-1 < \rho_k < 1$
- Note that the autocorrelation function is dimensionless, that is, independent of the scale of measurement of the time series.
- The correlogram of a stationary series will die out very fast, after the first k-lags, whereas in the case of a non-stationary series it will be very persistent, and die out very slowly.

For the above time series, the matrix:

$$R=\begin{bmatrix}\rho_o & \rho_1 & \rho_2 & \cdots & \rho_{k-1}\\ \rho_1 & \rho_0 & \rho_1 & \cdots & \rho_{k-2}\\ \rho_2 & \rho_1 & \rho_0 & \cdots & \rho_{k-3}\\ . & . & . & \cdots & .\\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_0\end{bmatrix}=\frac{1}{\gamma_0}\Gamma$$, is called the autocorrelation matrix.

0.6 Partial Autocorrelation

If X_1, X_2, X_3 are random variables, then we define as **partial correlation** the correlation between X_1 and X_2 when the linear effect of X_3 is subtracted. If $\rho_{12}, \rho_{13}, \rho_{23}$ are the correlation coefficients between the variables, taken pairwise, then it can be shown that the partial correlation coefficient between X_1 and X_2, when X_3 is kept fixed, is: $\rho_{12.3}=\dfrac{\rho_{12}-\rho_{13}-\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}$.

Similarly, for the case of a time series we define the **partial autocorrelation** of order k the partial autocorrelation between X_t, and X_{t+k}, when X_{t+1}, X_{t+2}, ..., X_{t+k-1} are kept fixed. Loosely, it is: $\rho_{k.}$ = Corr(X_t, $X_{t+k}/$ X_{t+1}, X_{t+2}, ..., X_{t+k-1} fixed). We denote partial autocorrelations as: $\rho_{k.}=\varphi_{kk}$.

0.7 White noise as a time series generator

A discrete time process is called a **purely random process** if it consists of a sequence of random variables $\{\varepsilon_t\}$, which are mutually independent and identically distributed (iid). From the definition it follows that the process has constant mean and variance and that: $\gamma(k)=Cov(\varepsilon_t\varepsilon_{t+k})=0,\ \text{for } k \text{ integer}$. Obviously, this process is weakly stationary and is better known as **white noise**. Typically, it is assumed that: $E(\varepsilon_t)=0$, $\begin{array}{l}\gamma_k=\sigma^2, k=0\\ \quad\;=0, k\neq 0\end{array}$, and $\begin{array}{l}\rho_k=\varphi_{kk}=1, k=0\\ \qquad\quad\;=0, k\neq 0\end{array}$.

0.8 The Wold decomposition Theorem

H. Wold (1938) proved a very important theorem illuminating the decomposition of a time series into a deterministic and an indeterministic part, which if put simply states that: *Every discrete stationary time series can be expressed as the sum of two uncorrelated series, a purely deterministic and a purely indeterministic*.

All the processes that will be considered in this course will fall to the category of purely indeterministic.

1. Review of Ordinary Least Squares

1.1 The Least Squares Principle and Assumptions

Model equation: $Y_t = \beta_1 + \beta_2 X_{2,t} + ... + \beta_j X_{j,t} + u_t$

The model expresses the value of a predictand variable as a linear function of one or more predictor variables and an error term, where: Y_t is the predictand in year t, $X_{j,t}$ is the predictor j in year t, β is the vector of unknown but estimable coefficients, u_t is the error term in year t.

Prediction/Fitted equation

$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2,t} + ... + \hat{\beta}_j X_{j,t}$

the model as it merges after applying Ordinary Least Squares, which yields parameter estimates such that the sum of squared errors is minimised.

Residuals

The error term is unobserved because the true (population) model is unknown. Once the model has been estimated, the regression residuals are defined as $\hat{u}_t = Y_t - \hat{Y}_t$. The residuals measure the closeness of fit of the predicted values and actual predictand values in the estimation period.

Note the differences between the two sets (u, β) and $(\hat{u}, \hat{\beta})$.

- Error/ disturbance versus residual
- Unknown parameters versus estimated coefficient

Define the residual sum of squares as: $S = S(\hat{\beta}) = \sum \hat{u}_t^2$, which is really the following:

$$\begin{bmatrix}\hat{u}_1 & & \hat{u}_T\end{bmatrix}\begin{bmatrix}\hat{u}_1 \\ \\ \hat{u}_T\end{bmatrix}=\hat{u}_1^2+\hat{u}_2^2+\ldots+\hat{u}_T^2$$

the OLS method consists of finding the values of the vector b that minimize the sum of squared residuals: $\hat{\beta}_{OLS}=\arg\min\left[\sum\hat{u}_t^2\right]$

Assumptions of OLS

The OLS model is based on several assumptions and when these are satisfied then the regression estimators are optimal in the sense that they are *unbiased* (the expected value of the estimator is equal to the true value of the parameter), *efficient* (the estimator has the smallest variance compared to any other linear estimator) and *consistent* (the bias and variance of the estimator approach zero as the sample size approaches infinity). The basic assumptions are the following:

A1) $E\left(u_t\right)=0$, A2) $Var\left(u_t\right)=\sigma^2$, A3) $E\left(u_t,u_s\right)\equiv Cov\left(u_t,u_s\right)=0,\ \forall s\neq t$, A4) $\left(u_t\right)$ □ $N\left(\bullet\right)$, A5) X "fixed".

Coefficient of determination

The explanatory power of the regression is summarized by its "R-squared" value, also called the *coefficient of determination,* and is often described as the proportion of variance "accounted for", "explained", or "described" by the regression. It is important to keep in mind that a high R^2 does not imply causation. The relative sizes of the sums-of-squares terms indicate how "good" the regression is in terms of fitting the calibration data. If the regression is "perfect", all residuals are zero, and R^2 is 1. If the regression is a total failure, the sum-of-squares of residuals equals the total sum-of-squares; **no** variance is accounted for by regression, and R^2 is zero.

A primer on residual analysis

Analysis of residuals consists of examining graphs and statistics of the regression residuals to check that model assumptions are satisfied. Some frequently used residuals tests are listed below.

- **Time series plot of residuals.** The time series plot of residuals can indicate such problems as non-constant variance of residuals, and trend or autocorrelation in residuals. A time-dependent variance might show, say, as an increasing scatter of the residuals about the zero line with time.
- The slope of the scatter plot of residuals on time can be tested for significance to identify trend in residuals.
- **Scatterplot of residuals against predicted values.** The residuals are assumed to be uncorrelated with the predicted values of the predictand. Violation is indicated by some noticeable pattern of dependence in the scatterplots. For example, the residual might flare out (increased scatter) with increasing value of the predictand; the remedy might be a transformation (e.g., log transform) of the predictand.
- **Scatterplots of residuals against individual predictors.** The residuals are assumed to be uncorrelated with the individual predictors. Violation of these assumptions would be indicated by some noticeable pattern of dependence in the scatterplots, and might suggest transformation of the predictors.
- **Histogram of residuals.** The residuals are assumed to be normally distributed. Accordingly, the histogram of residuals should resemble a normal pdf. But keep in mind that a random **sample** from a normal distribution will be only approximately normal, and so some departures from normality in the appearance of the histogram are expected – especially for small sample size.

- **Acf of residuals.** The residuals are assumed to be non-autocorrelated. If the assumption is satisfied, the acf of residuals should not be large at any non-zero lag. Special interest should be attached to the lowest lags, since physical systems are characterized by persistence from year to year.
- **Lag-1 scatterplot of residuals.** This plot also deals with the assumption of independence of residuals. The residuals at time *t* should be independent of the residuals at time *t-1*. The scatterplot should therefore resemble a formless cluster of points. Alignment in some direction might be evidence of autocorrelation of residuals at lag 1.

1.2 Hypothesis Testing and Types of Tests: Wald, Likelihood Ratio, Lagrange Multiplier

Once we have estimated a regression model we are usually interested in the significance and the values of the obtained coefficients. We use our theory as a guide as to whether the coefficient of a particular explanatory variable should be significant and/or what value it should attain. Furthermore, our theory may have something to say regarding a set of coefficients (how they should jointly behave). In order to test the validity of our theory, we first have to derive a set of testable implications which then will be confronted with the data. Based on our sample, rejecting these hypotheses would cast doubt on the validity of the underlying theory. In contrast, if these hypotheses are not rejected (i.e. the theory is consistent with the data) the theory has ‘survived’ (passed) the tests (Popperian principle).

Say we have a theoretical model that attempts to capture the production function of an economy: $Q = f\left(K, L\right)$, typically economic theory assumes the following: $Q\left(0\right) = 0, Q_i > 0, Q_{ii} < 0$. Let’s use a log-linear model of the form:

$q = \delta + \alpha k + \beta l + u$

What testable implications can be derived from the theory?

- α , β >0
- Additionally, the theory highlights certain combinations of these parameters that lead to major conclusions regarding the properties of the production function, known as *returns-to-scale*. For instance if $\alpha + \beta = 1$ then the production function exhibits *Constant-Returns-to-Scale (CRS).*
- If $\alpha + \beta < 1$ then the production function exhibits *Decreasing-Returns-to-Scale*
- If $\alpha + \beta > 1$ then the production function exhibits *Increasing-Returns-to-Scale*

Note that the log-linearity is handy since the estimated coefficients are equivalent to the respective elasticities: $\beta = \frac{d \ln Q}{d \ln L} = \varepsilon_{QL}$. Suppose we want to test whether CRS is a valid assumption. Then we test the null hypothesis: $H_0 : \alpha + \beta = 1$.

<u>Likelihood Ratio Test</u>

- Estimate the unrestricted model and obtain the maximised log-likelihood $L\left(\hat{\theta}\right)$.
- Estimate the restricted model and obtain the maximised log-likelihood $L\left(\tilde{\theta}\right)$.
- So, LR tests require the estimation both of the unrestricted and restricted models.

Note that it always be the case that $L\left(\hat{\theta}\right) \geq L\left(\tilde{\theta}\right)$. The question is whether their difference is statistically 'large enough' to justify rejection of the null hypothesis.

The LR test is performed by calculating the following statistic: $LR(k) = 2\left[L\left(\hat{\theta}\right) - L\left(\tilde{\theta}\right)\right] \sim \chi_k^2$.

Wald test

- Calculate set of restrictions, say $r\left(\hat{\theta}\right)=0$, i.e. $\alpha+\beta-1=0$
- and test whether $r\left(\hat{\theta}\right)$ is 'close' to zero.
- Thus Wald-type tests require estimation only of the unrestricted model.

LM test

- The aim is to maximize $L\left(\theta\right)$ subject to the restrictions $r\left(\theta\right)=0$.
- Thus we seek to maximize $L\left(\theta\right)-\lambda r\left(\theta\right)=0$
- If restrictions are valid then λ, the Lagrange Multiplier, should be zero
- LM tests often based on the unrestricted equation, followed by an auxiliary equation relaxing the restriction.

1.3 Diagnostic Tests

- 'Diagnostic' as in medicine, where clues to illness would in our case indicate violation of assumptions. Failure of tests would suggest:
- Misspecification
- Estimates misleading
- Un-modelled information

1.3.1 Types of Tests

- F, t-tests of restrictions that will lead us to exclude sets of variables and ultimately deal with a less complex model that will enhance estimation accuracy (increase of degrees of freedom).
- Tests for the sphericity of residuals: autocorrelation tests (DW, LM); heteroscedasticty tests (LM).
- Normality of residuals: Jarque-Berra test
- Parameter Stability tests: Chow test, Predictive Failure test, CUSUM.

2. Deviations from OLS assumptions I: Autocorrelation

Multiple Regression was originally developed for cross-sectional data but Statisticians/Economists have been applying it (mostly incorrectly) to chronological or longitudinal data with little regard for the Gaussian assumptions. Recall that Time series = a sequence of observations taken on a variable or multiple variables at successive points in time. The objectives of time series analysis are to (i) understand the structure of the time series (how it depends on time, itself, and other time series variables) and (ii) forecast/predict future values of the time series. What is wrong then with using regression for modeling time series? Perhaps nothing, if the residuals satisfy the regression assumptions (linearity, Homoscedasticity, independence, and (if necessary) normality). So it is important to apply a battery of tests for Pulses or one-time unusual values and to either adjust the data or to incorporate a Pulse Intervention variable to account for the identified anomaly. Unusual values can often arise due to Seasonality, thus one has to identify and incorporate Seasonal Intervention variables. Unusual values can also often arise at successive points in time earmarking the need for either, a Level Shift Intervention to deal with the proven mean shift in the residuals. Additionally, time series analyzed by regression suffer from autocorrelated residuals. In practice, positive autocorrelation seems to occur much more frequently than negative. Positively autocorrelated residuals make regression tests more significant than they should be and confidence intervals too narrow; negatively autocorrelated residuals do the reverse. In some time series regression models, autocorrelation makes biased estimates, where the bias cannot be fixed no matter how many data points or observations that you have. As a rule before you use regression methods on time series data, first plot the data over time and study the plot for evidence of trend and seasonality. Use numerical tests for autocorrelation, if not apparent from the plot. Deterministic trend can be dealt with by using functions of time as predictors. Seasonality can be dealt with by using seasonal indicators (Seasonal Pulses) as predictors or by allowing specific auto-dependence or auto-projection such that the historical values (Y(t-s)) are used to predict Y(t). Autocorrelation can be dealt with by using lags of the response variable Y as predictors. In general, run the regression and diagnose how well the regression assumptions are met. In particular, the residuals should have approximately the same variance (homoscedasticity) otherwise some form of "weighted" analysis might be needed. Furthermore, the model form/parameters should be invariant i.e. unchanging over time. If not, then we perhaps have too much data and need to determine at what points in time the model form or parameters changed. In whatr follows we will discuss most of these issues in some more detail.

When the data are of the particular type of time series then autocorrelation is most likely to occur and the error from one period can affect the error in other time periods. In other words, the third assumption of the typical OLS is violated. This violation requires the modification, not the abandonment, of the framework of ordinary least squares estimation.

The violation of the non-autocorrelation assumption thus alters the effect of errors. An error that occurs in one time period does not exert its entire impact in that period; instead its influence carries forward to other time periods. As a result, the errors associated with the regression will be correlated. In other words, false predictions in one point in time will result in false predictions for the next point(s) in time. If autocorrelation is present, then it is misleading to think of the consecutive time points as independent observations. Autocorrelation implies that the number of independent observations is smaller than the number of time points.

The absence of autocorrelation requires that $Cov(\varepsilon_t, \varepsilon_{t-j}) = 0$, so that errors *j* periods apart are totally uncorrelated. Autocorrelation therefore, is defined as a nonzero error covariance (and consequently autocorrelation). Autocorrelation can either be positive or negative although positive autocorrelation occurs more frequently with economic data. When positive autocorrelation exists, so that the covariance between errors is positive, an above the average error in time *t* will tend to be associated with an above the average error in time period *t-j*. Since the average error is zero, this implies that the positive errors will tend to follow positive errors, while negative errors will tend to follow negative errors. For positive autocorrelation, the non-random pattern of errors thus manifests itself through strings of positive and negative errors. In general, the number of sign changes will be smaller than the number that would occur if autocorrelation were absent. Negative autocorrelation implies that above the average errors will tend to follow below average errors, so that negative errors will often follow positive errors, and the number of sign changes in the equation error will exceed the number that would exist without autocorrelation.

Besides these differences in the sign of autocorrelation, a distinction also exists in the potential types of autocorrelation. **Quasi-autocorrelation** is the error correlation that occurs in a misspecified equation. The specification error that causes quasi-autocorrelation can result from either omitting an influential variable or the use of an incorrect functional form.

In a correctly specified regression, such temporal dependence of the errors is called **Pure autocorrelation**. When the correct specification is utilised, all the variables whose influence on the dependent variable is of secondary importance are omitted, and their joint influence is felt in the equation error. Pure positive autocorrelation can therefore arise as the result of positive temporal correlation among the set of omitted but non-influential variables. Measurement error in both the included and excluded variables can also cause autocorrelation. The frequency of the data is often reduced so that quarterly instead of monthly equations can be estimated. The method of converting monthly into quarterly data typically involves simple averaging, which results in estimating quarterly values with substantially less fluctuation than the monthly values upon which these are based. The use of this 'dampened' set of observations can itself cause pure autocorrelation in the equation errors.

2.1 Effects of Autocorrelation on OLS

The presence of autocorrelation has several consequences for least squares estimation. The OLS remain unbiased and consistent, but these are no longer best linear unbiased estimators, since they are not efficient. Furthermore, estimates of the residual sum of squares and coefficient variances obtained based on the assumption of white noise errors are biased. As a result statistical inference based on these values are invalid. In other words, carrying out hypotheses tests and/or constructing confidence intervals are likely to lead to incorrect inferences. According to Ostrom (1990, p. 26)

"*It should be noted that in most political and economic data the serial correlation is likely to be positive because the same random factors tend to operate on at least two successive periods' errors (and likely on more). Hence, we should be wary of the possibility of positive serial correlation in both the error terms and the independent variables.*"

The actual error term in a regression equation is unobserved. Residuals from the estimated equation must therefore serve as the basis for both the detection and correction of autocorrelation. The simplest method, and least foolproof, method for detecting autocorrelation among the residuals of an equation consists of visually inspecting them. Autocorrelation manifests itself through a non-random residual pattern, so any obvious pattern in a time plot of residuals is a signal that a potential problem exists. For positive autocorrelation, this plot will tend to show fairly infrequent sign changes, where a series of residuals with one sign will follow a series with the opposite sign. In contrast, negatively autocorrelated residuals will tend to display an inordinately large number of sign changes.

2.1.1 The Durbin-Watson test

The most frequently utilised statistical test for the presence of autocorrelation is the **Durbin-Watson test (DW** hereafter). This test is valid when the following conditions are met:

- The equation includes an intercept term
- The error process is first-order autoregressive
- The equation excludes a lagged dependent variable, and
- None of the explanatory variables are stochastic

Assuming the AR(1) process for the error term in mathematical form is: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$. The **DW** test utilises the residuals from an estimated model to test the null hypothesis: $H_0 : \rho = 0$. Against either a one-tailed or two-tailed alternative hypothesis. The test statistic for this test is given by: $DW = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}(\hat{\varepsilon}_t)^2}$.

Note that the summation in the numerator starts with the second observation, since the use of the lagged residual results in the loss of one observation. The validity of the **DW** statistic for testing first-order autocorrelation can be seen from an approximation derived from the above equation. Starting with the numerator we have:

$\sum\hat{\varepsilon}_t^2+\sum\hat{\varepsilon}_{t-1}^2-2\sum\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}$. Since the sums with squared residuals differ by only one term, then we use the approximation: $\sum\hat{\varepsilon}_t^2\approx\sum\hat{\varepsilon}_{t-1}^2$. The numerator can therefore be written: $2\sum\hat{\varepsilon}_t^2-2\sum\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}=2\left[\sum\hat{\varepsilon}_t^2-\sum\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}\right]$. Adding the denominator to this becomes: $DW\approx\frac{2\left[\sum\hat{\varepsilon}_t^2-\sum\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}\right]}{\sum\hat{\varepsilon}_t^2}$. Since the summations of the squared residuals differ only by one term, these terms are approximately equal. Dividing each term in the numerator by the denominator, the approximation for the **DW** becomes: $DW\approx2\left(1-\frac{\left[\sum\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}\right]}{\sum\hat{\varepsilon}_t^2}\right)$. The ratio of the summation terms is the estimated autocorrelation coefficient that is obtained by regressing the current period residuals on the lagged residuals. No wonder, this coefficient is equal to *ρ* appearing in the AR(1) error process. Utilising this information the operational form of the **DW** statistic is: $DW=2(1-\hat{\rho})$. Where $\hat{\rho}$ is the least squares estimate of *ρ.* Thus, the above equation links the **DW** statistic to the estimated coefficient of first-order autocorrelation. We can obtain values of the **DW** statistic for the different autocorrelation possibilities, since *ρ* takes values between -1 and +1. So a mapping of these possibilities would be:

- When *ρ* = 0 (absence of autocorrelation), then **DW** = 2
- When *ρ* = 1 (perfect positive autocorrelation), then **DW** = 0
- When *ρ* = -1 (perfect negative autocorrelation), then **DW** = 4

Therefore, because of this one-to-one mapping between *ρ* and **DW**, since *ρ* has an upper and a lower bound, the **DW** is bounded as well. The upper bound is given by 4 and the lower bound is given by 0. So, values of **DW** close to 2 constitute evidence supporting the absence of autocorrelation, in fact this value is a 'rule of thumb' for many researchers.

However, there is a more formal way of testing the significance of the **DW** statistic, which is slightly different than the standard t and F tests because in the case of the **DW** one has to use a critical range (given by an upper critical value and a lower one) instead that of a critical value. Furthermore, apart from the number of observations used in the estimation stage one has also to take into account the number of regressors used. The way to use the statistic in order to test for the presence of serial correlation is the following:

- Estimate the original equation with least squares
- Obtain the value of the **DW** (all econometric packages report it in their standard estimation output)
- Consult the table for the critical values, based on the number of observations, the number of explanatory variables and the desired level of significance

Then follow these decision rules:

In the case of positive autocorrelation;

- If **DW** <d_L, *reject* the null of no autocorrelation
- If **DW** >d_U, *do not reject* the null of no autocorrelation
- If **DW** belongs to [d_L , d_U], then the test is *inconclusive*

In the case of negative autocorrelation;

- If **DW** > 4-d_L, *reject* the null of no autocorrelation
- If **DW** < 4-d_U, *do not reject* the null of no autocorrelation
- If **DW** belongs to [4-d_U , 4-d_L], then the test is *inconclusive*

In conclusion, the **DW** test is a very useful tool but it has a number of limitations: it cannot be used when the model does not include an intercept, it is not valid if the endogenous variable is lagged (later we will show that one can overcome this limitation), it is not robust to alternative processes for the error term (apart from the AR(1) process) and finally in some cases it fails to produce inference (inconclusiveness).

2.1.2 The Breusch-Godfrey test

As discussed above, the **DW** test can be applied only when one tests for an AR(1) error term process. This is obviously somewhat restrictive since the autocorrelation present could still be an AR process but of higher order. **Breusch** and **Godfrey** (**BG** hereafter) proposed a more general test for autocorrelation that can be employed when the order of the error's autoregressive dependence extends beyond the first order. The **BG** test is essentially a statistical test for the joint significance of a set of autocorrelation coefficients.

If the error process is assumed to be of the k-th order autoregressive: $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + ... + \rho_k \varepsilon_{t-k} + u_t$. The **BG** test ascertains whether the set of coefficients ρ_1 to ρ_k are significantly different form zero. The null hypothesis is then: $H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0$. The basis for conducting this test is the likely correlation of the current residual with its own lags when the error process is serially correlated. The way to use the statistic in order to test for the presence of serial correlation is the following:

- Estimate the original equation with least squares
- Obtain the residuals and calculate the set of lagged residuals that correspond to the order of autoregressive process postulated in the null hypothesis.
- Regress the current residual on the set of explanatory variables included in the original equation plus the lagged residuals.

The test statistic is given by *(n-k)* R^2 and follows a chi-square distribution with k degrees of freedom and n-k is the number observations used to estimate this equation. If the value of the statistic exceeds the critical chi-square value for the selected level of significance, then reject the null of no serial correlation.

There are two potential difficulties with the use of this test. First, the order of the AR process is unknown and therefore the researcher must decide on an appropriate value for k. Second, when the original model contains a large number of explanatory variables, the degrees of freedom used to estimate the second (auxiliary) equation will be small and even negative. As the frequency of the data increases the **BG** (say from monthly to annual), this becomes less important. In spite of these difficulties, a major strength of the test is that it does not have an inconclusive region, and therefore can be employed when the **DW** is inconclusive.

2.1.3 The Ljung-Box test

A method originating from the **Box-Jenkins** *methodology* for testing the presence of autocorrelation also exists and constitutes of a visual inspection of information obtainable from the residuals and a statistical test of the null hypothesis of no autocorrelation. Essentially, what one does is to inspect the sample autocorrelation function and the sample partial autocorrelation function of the residuals in the search for clues about the specific type of process that might have generated the error term. The null hypothesis for the **Ljung-Box** (**LB** hereafter) is: $H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0$. The suggested lag length is n/4 and the statistic is: $LB = n(n+2)\sum \frac{\rho_i^2}{n-k}$. Where the summation runs from 1 to k. The statistic follows a chi-square distribution with k degrees of freedom under the null hypothesis.

Therefore, the null of no autocorrelation is rejected when the statistic exceeds the chi-square critical value for the chosen significance level.

2.2 Modelling with autocorrelation

In the previous topic we discussed the consequences of autocorrelated errors and also a number of methods that can be used in order to test whether errors are indeed autocorrelated. However, after detecting the problem the challenge is to cope with it; that is how one can proceed with the estimation of the underlying model by surpassing the 'problem' of autocorrelation.

2.2.1 Incorporation of the data generation process characteristics

Suppose that a researcher is dealing with the simplest possible model of the form:

$Y_t = \alpha + \beta X_t + \varepsilon_t \qquad (6.1)$

$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \qquad (6.2)$

$-1 < \rho < 1, \quad u \sim \text{white noise}$

For the time being assume that ρ is known. Take the lagged form of the above model and multiply through by ρ, which produces: $\rho Y_{t-1} = \rho\alpha + \rho\beta X_{t-1} + \rho\varepsilon_{t-1} \quad (6.3)$. Now, subtract the lagged model from the initial model: $(Y_t - \rho Y_{t-1}) = (\alpha - \rho\alpha) + (\beta X_t - \rho\beta X_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1}) \quad (6.4)$. Given (6.2), (6.4) can be written as: $(Y_t - \rho Y_{t-1}) = (\alpha - \rho\alpha) + (\beta X_t - \rho\beta X_{t-1}) + u_t \quad (6.5)$ or $Y_t^* = \alpha^* + \beta X_t^* + u_t \quad (6.6)$. Where, the variables with asterisks correspond to the following: $Y_t^* \equiv Y_t - \rho Y_{t-1}$, $X_t^* = X_t - \rho X_{t-1}$, and $\alpha^* = \alpha(1-\rho)$.

Model (6.6) is also known as ***generalised difference equation***. Notice that the 'new' model, after the transformation has similar characteristics with the initial model but also has a significant difference. The transformed model is associated with a new error term (*u*), which is autocorrelation-free, since by its construction it is a white noise process. In fact, that was the goal of the whole transformation; we wanted a 'new' model whose error term would be 'well-behaved'. This type of transformation is known in the literature as the ***Cohrane-Orcutt*** transformation. However, the new model has a serious drawback. In the differencing procedure we lose one observation because the first sample observation has no antecedent. To avoid the loss of one observation, which might be costly in relatively small samples, the first observation of Y and X are transformed as follows: $Y_1^* = \sqrt{1-\rho^2}(Y_1)$ and $X_1^* = \sqrt{1-\rho^2}(X_1)$.

The above transformation is known as the ***Prais-Winsten*** transformation. In practice, if the sample size is large enough this transformation is not generally applied and one simply uses *n-1* observations. Now, if model (6.6) is treated as time series regression and its parameters estimated by OLS, then the procedure is called ***Generalised Least Squares*** (GLS). However, it should be noted that GLS assumes that both the process generating the error term and the parameters of this process are known. In our context, requires the knowledge of the particular generation mechanism [for instance AR(1)] and also the value of ρ (for example ρ = 0.7). Therefore, if this information is available then GLS can be applied in order to obtain consistent estimates of the parameters.

As you suspect it is usually the case that not only we do not know the parameters of the generating model but also, which is even worse, we are agnostic for the model itself. So before applying the transformation one has to obtain information for the model and its parameters that govern the time series behaviour of the error term. As far as the model is concerned one can employ the Box-Jenkins methodology in order to assess the time series properties of the residuals obtained from the original model (that is 6.1), and then somehow estimate the involved parameters.

2.2.2 How to estimate ρ

This section will discuss the case where the error term is generated from an AR(1) process. So suppose that we applied the BJ methodology on the residuals and decided that the appropriate model for the error term is that of an AR(1). How can we obtain a reliable estimate for the parameter ρ?

There are two ways to achieve that. The first is by using the **DW** statistic. Recall that the following relationship between the **DW** statistic and the autocorrelation coefficient is true: $dw \approx 2\left(1-\hat{\rho}\right) \Rightarrow \hat{\rho} \approx 1-\frac{dw}{2}$. The other way is simply to run a regression of the residual on its lag in order to obtain a point estimate of ρ. That is run the regression: $\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + v_t$.

However, although these are relatively simple and intuitive ways of estimating the autocorrelation coefficient, other methods are employed which appear in econometric software as standard routines. Here is a list of these methods[1]:

- Cohrane-Orcutt iterative procedure
- Cohrane-Orcutt two-step method
- Durbin two-step method
- Hildreth-Lu search procedure
- Maximum Likelihood method

[1] Detailed exposition of the methods is beyond the scope of the course. However, their use will be demonstrated in the lab session.

Appendix-2

A2.1 Autoregressive Models

Consider the time series $\{X_t\}$, which behaves according to the following model:

$X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + \varepsilon_t$ (3.1)

Where c is a constant (independent of time) and ε_t a white noise process.

This process is called an **Autoregressive** model of order p, or more succinctly $AR(p)$. Autoregressive models have always been popular, partly because they are very intuitive, and partly because they are easier to estimate (as will be shown later). So, the idea is that an $AR(p)$ model is generated by a weighted average of past observations going back p periods, together with a random disturbance in the current period. Equation (3.1) can be written as: $X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} - \ldots - \varphi_p X_{t-p} = c + \varepsilon_t$ or $X_t - \varphi_1 B X_t - \varphi_2 B^2 X_t - \ldots - \varphi_p B^p X_t = c + \varepsilon_t$ or $(1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p) X_t = c + \varepsilon_t$ or $\varphi_{\text{p}}(B) X_t = c + \varepsilon_t$, where $\varphi_{\text{p}}(B) = (1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p)$ is the **linear filter** and B^j is the backward operator. In this case the linear filter is a polynomial of pth order. It can be shown that a necessary and sufficient condition for the stationarity of series X_t is that the roots of the above polynomial must lie outside the unit circle.

A2.1.1 First-order Autoregressive (Markov) Process

For p = 1 model (3.1) becomes: $X_t = c + \varphi_1 X_{t-1} + \varepsilon_t$ (3.1.1) According to the above discussion, necessary and sufficient condition for the stationarity of the model is that the root of the polynomial $\varphi_1(B) = 1 - \varphi_1 B$, must lie outside the unit circle. So,

$\varphi_1(B) = 0 \Leftrightarrow 1 - \varphi_1 B = 0 \Leftrightarrow B = \frac{1}{\varphi_1}$. Thus $\frac{1}{|\varphi_1|} > 1 \Leftrightarrow |\varphi_1| < 1 \quad (3.2)$. If (3.2) is satisfied then the series is stationary and $E(X_t) = \mu$

From (3.1.1) it is: $E(X_t) = c + \varphi_1 \, E(X_{t-1}) + E(\varepsilon_t) \Leftrightarrow \mu = c + \varphi_1 \mu \Leftrightarrow \mu = \frac{c}{1 - \varphi_1} \quad (3)$. If we now define the deviation from the mean as $\tilde{X}_t = X_t - \mu$, then we have $\tilde{X}_t = \varphi_1 \tilde{X}_{t-1} + \varepsilon_t \quad (3.4)$. From (4) we can obtain $\tilde{X}_t = \varphi_1 \tilde{X}_{t-1} + \varepsilon_t = \varepsilon_t + \varphi_1 \left(\varphi_1 \tilde{X}_{t-2} + \varepsilon_{t-1} \right) = ... = \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_1^2 \varepsilon_{t-2} + ... = \sum_{j=0}^{\infty} \varphi_1^j \varepsilon_{t-j} \quad (3.5)$.

From (3.5) it is obvious that $E(\tilde{X}_t) = 0$, given that the power series in (3.5) converges. Furthermore,

$$\gamma_0 = Var(X_t) = E(\tilde{X}_t^2) = E\left[\sum_{j=0}^{\infty} \varphi_1^j \varepsilon_{t-j} \right]^2 = \sum_{j=0}^{\infty} \varphi_1^{2j} E(\varepsilon_{t-j}^2) + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \varphi_1^{i+j} E(\varepsilon_{t-i} \varepsilon_{t-j}) =$$

$\sum (\varphi_1^2)^j \sigma^2 + 0 = \frac{\sigma^2}{1 - \varphi_1^2} \quad (3.6)$. Now if we multiply (3.4) by $\tilde{X}_{t-k}$ we obtain $\tilde{X}_t \tilde{X}_{t-k} = \varphi_1 \tilde{X}_{t-1} \, \tilde{X}_{t-k} + \varepsilon_t \, \tilde{X}_{t-k}$. Thus $E(\tilde{X}_t \tilde{X}_{t-k}) = \varphi_1 E(\tilde{X}_{t-1} \, \tilde{X}_{t-k}) + E(\varepsilon_t \, \tilde{X}_{t-k}) \quad (3.7)$.

However, the expectation of the product of the error term and the lagged process is zero for values of k larger than zero because according to (3.5) the only disturbances affecting it are past ones and not future ones. Back to (3.7), we see that it can be written as $\gamma_k = \varphi_1 \gamma_{k-1}$. This recursive relationship combined with (3.6) produce: $\gamma_{\mathrm{k}} = \varphi_1^{\mathrm{k}} \frac{\sigma^2}{1 - \varphi_1^2} = \varphi_1^{\mathrm{k}} \gamma_0$. Dividing both sides by γ_0 gives: $\rho_k = \varphi_1^k \quad (3.8)$.

According to (3.8) if $0 < \varphi_1 < 1$ then the correlogram should decay exponentially, taking values on the positive axis, whereas if $-1 < \varphi_1 < 0$, the correlogram will decay and oscillate in sign. As far as the partial autocorrelations are concerned, we have: $\varphi_{11.} = \rho_{1.} = \rho_1 = \varphi_1$ and $\varphi_{22} = \rho_{2.} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = \frac{\varphi_1^2 - \varphi_1^2}{1 - \varphi_1^2} = 0$. So, it is obvious that for an AR(1) model the following holds: $\varphi_{kk} = 0$ for all *k > 1,* which means that the only non-zero partial autocorrelation is that of the first order.

Remark

Inspecting (3.8) implies that the process has an infinite memory. The current value of the process depends on all past values, although the magnitude of dependence declines with time.

A2.2 Moving Average Models

In the **Moving Average** process of order q each observation X_t is generated by a weighted average of random disturbances going back q periods. We denote this process as MA(q) and represent it as: $X_t = b + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}$ (3.16). Where the parameters θ may be positive or negative.

In the MA model (and also in the AR model) the random disturbances are assumed to be independently distributed across time. In particular, each disturbance term is assumed to be a normal random variable with mean equal to zero and a constant variance, and a zero covariance. Now, (3.16) can be more compactly written as: $X_t = b + (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q) \varepsilon_t$ or $X_t = b + \theta_q(B) \varepsilon_t$

Remark

All MA models are stationary (by their construction). However, another concept is relevant for MA, called **invertibility condition**. As you suspected, a necessary and sufficient condition for invertibility is that the roots of the polynomial $\theta_q(B)$ must lie outside the unit circle.

A2.2.1 First-order Moving Average process

For *q = 1* model (3.16) becomes:

$X_t = b + \varepsilon_t - \theta_1 \varepsilon_{t-1}$ (3.16.1) or

$X_t = b + (1 - \theta_1 B) \varepsilon_t$ (3.16.2) or

$X_t = b + \theta_1(B) \varepsilon_t$ (3.16.3)

According to the condition for invertibility it must be that: $\left|\frac{1}{\theta_1}\right| > 1 \Leftrightarrow |\theta_1| < 1$,

$E(X_t) = \mu$, $\gamma_0 = Var(X_t) = (1+\theta_1^2)\sigma^2$, $\begin{matrix} \gamma_k = -\theta_1\sigma^2, \; k=1 \\ = 0, \; k>1 \end{matrix}$ and $\begin{matrix} \rho_k = \dfrac{-\theta_1}{1+\theta_1^2}, \; k=1 \\ = 0, \; k>1 \end{matrix}$. As far the partial autocorrelation is concerned, it can be shown that is given by $\varphi_{\text{kk.}} = -\dfrac{\theta_1^k(1-\theta_1^2)}{1-\theta_1^{2(k+1)}}$.

A few Remarks

If $\theta_1 > 0$, then $\rho_1 < 0$, and the partial autocorrelation decays exponentially while all of them are negative.

If $\theta_1 < 0$, then $\rho_1 > 0$, and partial autocorrelation alternates sign.

If an MA(1) process is invertible, then it can be expressed as an AR(∞) process. This holds for MA processes of any order.

A2.3 Mixed Autoregressive Moving Average Models

Many stationary random processes cannot be modelled as purely MA or as AR, since they have the qualities of both types of processes. The logical extension of the models presented in the last two sections is the **Mixed Autoregressive Moving Average** process, which we denote as *ARMA(p,q)*. The mathematical expression of such a process is: $X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \ldots + \varphi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$ (3.17). Equation (3.17) can be rewritten as: $\varphi_p(B)X_t = c + \theta_q(B) \varepsilon_t$ (3.17.1)

Obviously, the ARMA(p,q) model can be thought as a generalisation of the previous models since the family of AR and MA models are nested in it. For instance, an AR(p) model can be seen as an ARMA(p,0) model and similarly a MA(q) model can be viewed as an ARMA(0,q) model. It can be shown that the necessary and sufficient condition for the stationarity of the ARMA model is that the roots of the polynomial $\varphi_p(B)$ lie outside the unit root circle. By symmetry, the necessary and sufficient condition for the invertibility of the ARMA model is that the roots of the polynomial $\theta_q(B)$ lie outside the unit root circle.

A2.3.1 First-order Autoregressive - First-order Moving Average Process

For $p = 1$ and $q = 1$ model (3.17) becomes: $X_t = c + \varphi_1 X_{t-1} + \varepsilon_t - \theta_1\ \varepsilon_{t-1}$ (3.17.2) The stationarity and invertibility conditions are as previous shown. The moments of the process are given as follows: $\gamma_0 = \frac{1+\theta_1^2 - 2\varphi_1\theta_1}{1-\varphi_1^2}\sigma^2$ and $\gamma_1 = \frac{(\varphi_1 - \theta_1)(1-\varphi_1\theta_1)}{1-\varphi_1^2}$. From the above it can be shown that $\rho_1 = \frac{(\varphi_1 - \theta_1)(1-\varphi_1\theta_1)}{1+\theta_1^2 - 2\varphi_1\theta_1}$.

As far as the partial autocorrelation is concerned, the first order one is equal to the autocorrelation coefficient and the subsequent behave similarly to those of a MA(1) model. Since the polynomials characterising the model are of first order they have a unique real root. As a result, the correlogram and the graph of the partial autocorrelation will decay exponentially after the first lag.

A2.4 The Box-Jenkins Methodology

The **Box-Jenkins** (BJ hereafter) approach to time-series model building is a method of finding, for a given data set, an ARMA model that adequately represents the data-generating process. This is an iterative approach, where the researcher's aim is to relate a model to the data, and basically consists of the following three steps:

- **Identification**, where by the use of data, and of any information about how the series was generated, to suggest a subclass of parsimonious models to be entertained.
- **Estimation**, where by efficient use of the data the researcher aims at making inferences about parameters conditional on the adequacy of the model entertained.
- **Diagnostic Checking**, where we check the fitted model in its relation to the data with intent to reveal inadequacies and so achieve model improvement.

It should also be borne in mind that the researcher is aiming at a **parsimonious** model. Parsimony means that the model should have as few parameters as possible, consistent with the aim of capturing the major features of the data. In the words of Box and Jenkins "*…we employ the smallest number of parameters for adequate representations*"

A2.4.1 Identification and its Objectives

It should first be said that identification and estimation necessarily overlap. Thus, we may estimate the parameters in a model, which is more elaborate than that which we expect to find, so as to decide at what point simplification is possible.

At the identification stage no precise formulation of the problem is available, statistically 'inefficient' methods must necessarily be used. It is a stage at which graphical methods are particularly useful and judgement must be exercised. However, it should be borne in mind that preliminary identification commits us to nothing except tentative consideration of a class of models that will later be efficiently fitted and checked.

A2.4.2 Identification Techniques

The objective is to select p and q in the ARMA(p,q) model to be fitted to the data. In principle, one attempts to match the theoretical autocorrelation and partial autocorrelation patterns with observed sample counterparts. In practice, the autocorrelations of the underlying process, the **population** autocorrelations, are not known. Therefore, one must rely on **estimates** based on realisations of a given time series. These estimates are called **sample** autocorrelations. Here are a few tips, which are useful in identifying:

- Pure AR models of order p are indicated when sample partial autocorrelations cut off after lag *p*. The autocorrelation of such models do not cut off, but decay toward zero. So, the basic characteristic that identifies an AR model is the behaviour of its sample partial autocorrelation, which should be 'negligible' after the pth term.
- Pure MA models of order q are indicated when sample autocorrelations cut off after lag *q*. The partial autocorrelation of such models do not cut off, but decay toward zero.
- If both the autocorrelation and partial autocorrelation tail off, a mixed process is suggested. Furthermore, the autocorrelation function of a mixed process containing a *p-th* order autoregressive component and a *q-th* order moving average component, is a mixture of exponentials and damped sine waves after the first *q-p* lags.
- Conversely, a mixture of exponential and damped sine waves dominates the partial autocorrelation function after the first *p-q* lags. So in plain English, mixed ARMA models do not yield a cut off in either the autocorrelation or partial autocorrelation patterns. Rather, the autocorrelation function decays toward zero in a complicated pattern for lag larger than q.

It is a difficult task to successfully identify p and q when both are non zero, but experience suggests that it is rare for either of these to be larger than 2. In fact some researchers suggest it is rare for *(p+q)* to be larger than 2. As mentioned above, the researcher should bear in mind that a parsimonious model should always be preferred over an elaborate one. However, in practice more than one parsimonious model may be identified and carried forward to the next stage.

In general, autoregressive (moving average) behaviour, as measured by the autocorrelation function, tends to mimic moving average (autoregressive) behaviour as measured by the partial autocorrelation function. For example, the autocorrelation function of a first-order AR process decays exponentially, while the partial autocorrelation function cuts off after the first lag. Correspondingly, for a first-order MA process, the autocorrelation function cuts off after the first lag. Although not precisely exponential, the partial autocorrelation function is dominated by exponential terms and has the general appearance of an exponential.

A2.4.3 Hypothesis testing for sample autocorrelation and partial autocorrelation

As mentioned above assessing whether sample autocorrelations and sample partial autocorrelations are ‘negligible’ lies at the heart of the identification stage. However, one has to define the term 'negligible' in this context, and furthermore quantify it so as to avoid any subjective criteria. In other words, we are interested in the statistical significance of the sample moments, we want to be able to decide whether a point estimate of the (partial) autocorrelation is significantly different from zero or can be treated as zero for statistical purposes. Thus, in our context ‘negligible’ is synonymous to 'statistically insignificant from zero'.

Estimated autocorrelations can have rather large variances and can be highly correlated with each other. For this reason, as emphasised by Kendall, detailed adherence to the theoretical autocorrelation function cannot be expected in the estimated function. In particular, moderately large estimated autocorrelations can occur after the theoretical function has damped out, and apparent ripples and trends can occur in the estimated function, which have no basis in the theoretical function. In employing the estimated function as a tool for identification, it is usually possible to be fairly sure about broad characteristics, but more subtle indications may or may not represent real effects, and two or more related models may need to be entertained and investigated further. Thus, we need some means for judging whether the (partial) autocorrelation are effectively zero after some specific lag q or p respectively. So, skipping the maths and distributional assumptions, it turns out that if the point estimate (sample) of an autocorrelation does not lie in the interval: $(\frac{-2}{\sqrt{T}}, \frac{2}{\sqrt{T}})$, where T stands for the sample size, then the corresponding population autocorrelations are statistically different from zero. In other words, we reject the null hypothesis that the population (partial) autocorrelation is equal to zero.

A2.4.4 Estimation

The methods for estimating the various *ARMA(p,q)* models are highly significant for the course' s progression, however a detailed technical demonstration lies outside of the course's scope.

A2.4.5 Diagnostic Checking

The third stage of the BJ methodology is to check whether the model fits the data. In ARMA modelling model checking is particularly important not least because the model selection stage of the cycle involves application of a certain amount of skill and judgement in being able to recognise (partial) autocorrelation patterns.

Thus, after the identification and estimation stages still the question remains of deciding whether the model is adequate. If there is good evidence of serious inadequacy, we shall need to know how the model should be modified in the next iterative cycle. What we are doing is described partially by the words 'testing goodness of fit'. **Diagnostic checks** must be such that they place the model at jeopardy. There are three main groups of diagnostic checks (i) Residual analysis, (ii) Fitting Extra parameters-the underspecified model, and (iii) Underfitting-the overspecified model.

A2.4.6 Residual Analysis

Residual Analysis is usually based on the fact that the residuals of an adequate model should be approximately white noise. Recall that for a white-noise process the autocorrelations are zero. Therefore, the significance of the residual autocorrelation is tested. One way of testing the null is to plot the residual autocorrelation and visually inspect whether all sample estimates lie within the two bands discussed above. A more formal way of testing for the absence of autocorrelation from the residuals is to compute the so-called **Portmanteau statistic**. Thus to check the overall acceptability of the residual autocorrelation the following statistic, known as **Ljung-Box statistic**, is used: $Q = T(T+2)\sum_{k=1}^{K}\frac{1}{T-k}r_k^2$, where $Q \sim \chi^2_{K-p-q}$. Thus, if Q is higher than the appropriate critical value, then we reject the null hypothesis that autocorrelations up to order K are zero.

A2.4.7 Fitting extra parameters-the underspecified model

In order to verify that the estimated model contains the appropriate number of parameters to represent the data, one can include an additional parameter to see if the addition results in an improvement over the original model.

A2.4.8 Underfitting-the overspecified model

Another very useful check on the model adequacy is to evaluate whether the current model does not contain redundant parameters. Redundant parameters can be explored simply by employing the standard t-statistic in order to test for individual coefficient significance and the F-test for joint significance of the coefficients.

A2.4.9 Information Criteria for Model Selection

Model selection criteria are based on the estimate of the variance of the residuals. If $\hat{\varepsilon}_t$ are the residuals from the estimated model, then $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2$ is their sample variance. The first criterion considered is the so-called Akaike Information Criterion (AIC), $AIC = \ln\hat{\sigma}^2 + 2\frac{k}{T}$, and the second is the Bayesian Information Criterion (BIC), $BIC = \ln\hat{\sigma}^2 + \frac{k}{T}\ln T$. The selection of the model when at least two candidate models are considered is made by choosing the model, which minimises the selected information criterion.

3. Deviations from OLS assumptions II: Non-constant Variance

3.1 Heteroscedasticity Tests

3.1.1 The Goldfeld-Quandt test

Consider the model: $y = X\beta + u$ with $\operatorname{var}(u_t) = \sigma_t^2 = \delta Z_t,\ \delta > 0$, where Z is a variable taken from the regressor matrix X (possibly a transformation). Therefore this setup assumes that we have previously identified Z as the source of the problem.

<u>Description of the test</u>

Divide the overall sample into two sub-samples, omitting anywhere from 1/6 to 1/3 of the observations *from the middle of the data set*. The sub-samples will have n_1 and n_2 observations (where the first set of observations will have smaller variance).Estimate separate regressions for each sub-sample, and obtain the residual sum of squares (ESS) from these equations. If the errors from the original equation are normally distributed then the quantity: $\frac{ESS_i}{\sigma_i^2} \sim \chi^2_{n_i - k}$

The test statistic for the Goldfeld-Quandt test is obtained by taking the ratio of the two independent chi-square random variables above, dividing each by its degrees of freedom. The ratio:

$$\frac{\left(\frac{ESS_2}{\sigma_2^2}\right)/\left(n_2 - k\right)}{\left(\frac{ESS_1}{\sigma_1^2}\right)/\left(n_1 - k\right)} \sim F_{(n_2-k),(n_1-k)}\,.$$

If the null hypothesis of Homoscedasticity is true then the test statistic reduces to:

$$\frac{\left(ESS_2\right)/\left(n_2 - k\right)}{\left(ESS_1\right)/\left(n_1 - k\right)} = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^1} \sim F_{(n_2-k),(n_1-k)}$$

We then expect the value of this statistic to be approximately equal to one. If the test statistic exceeds the critical F-value then *we reject the null hypothesis of Heteroscedasticity*.

3.1.2 The Breusch-Pagan test

The main difference in the philosophies of the two tests is that the first one utilizes info from the two sub-samples while the second one utilizes info from the variance of the whole sample and uses an *auxiliary regression*. In this case, one is trying to explicitly test whether a particular variable *explains* the variance of the residual (under the null it should not).

Description of the test

Estimate the equation of interest: $y = X\beta + u$, then obtain the residuals, then calculate: $\hat{\sigma}^2 = \sum \hat{u}_t^2 / n$, And then estimate: $\frac{\hat{u}_t^2}{\hat{\sigma}^2} = \alpha_0 + a_1 Z + v$ (*auxiliary regression).*

The statistic is an LM taking the form: *n*R*^{2}, with 1df (in other contexts the degrees of freedom will be equal to the number of parameters set equal to zero under the null hypothesis).

3.2 Time-Varying Volatility: The GARCH family

3.2.1 The ARCH model

There might be certain periods during which the environment within which agents operate is more uncertain (volatile). Such cases would be during wars, after major recessions, different economic regimes etc. In general, there might simply be persistence in uncertainty that would render the assumption of constant error variance inappropriate. Engle (1982) has developed the basic model that may be used in such cases, the so-called ARCH (Auto-Regressive Conditional Heteroscedasticity) model.

$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$, and $\varepsilon_t^2 = \nu_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}$.

Alternatively, $\varepsilon_t \square N\left(0, h_t\right)$ and $h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$.

Originally a 2-step procedure was suggested in order to test for ARCH effects:

- Estimate the basic model and retrieve the squared residuals.
- Inspect the sample autocorrelation function (ACF) of the squared residuals, where under the absence of ARCH effects the ACF should be insignificantly different from zero at all lags.

In case, ARCH effects are present then this new information regarding the distributional characteristics of the error term should be taken into account. Use the maximum likelihood method to maximise the following function:

$$LLF = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln h_t - \frac{1}{2}h_t \sum \left(y_t - \beta x_t\right)^2 \text{, where } h_t = \alpha_0 + \alpha_1\left(y_{t-1} - \beta x_{t-1}\right)^2 .$$

We expect $\alpha_0, \alpha_1 > 0$ and $0 < \alpha_1 < 1$. In case there is a structural reason which suggests that the volatility of the error term is affected by a certain variable then we may extent the model to the ARCH-in-Mean (ARCH-M) where we enter this variable as a determinant of the h_t function.

3.2.2 The GARCH model

This is simply an extension of the ARCH model allowing, apart from past values of the squared error term, for past levels of h_t itself to have an impact on current volatility. Hence the model reads:

$$\varepsilon_t^2 = \alpha_0 + \sum \gamma_i \varepsilon_{t-i}^2 + \sum \beta_j h_{t-j}$$

A GARCH-M extension is also possible.

4. Trends and Seasonality in Time Series

4.1.1 Stationary vs. Non-Stationary Processes

Shocks to a stationary time series are temporary; over time the effects of shocks will dissipate, and the series will revert to its long-run level. We already know that a weakly stationary series will:

- Exhibit ***mean reversion*** in that it fluctuates around a constant long-run mean.
- Has a finite variance that is time-invariant
- Has a theoretical correlogram (autocorrelation function) that diminishes as lag length increases.

4.1.2 Non-stationary series

Shocks to a non-stationary time series are permanent. In addition, non-stationary series have the following properties:

- The mean and the variance of a non-stationary series are time-dependent
- There is no long-run mean to which the series returns
- The variance goes to infinity as time approaches infinity
- Theoretical autocorrelations do not decay but, in finite samples, the sample correlogram dies out slowly.

Notes

(i) The sample correlogram is a very useful tool for detecting the presence of non-stationarity. However, it is not a formal way which can help us in deciding with conviction (within the limits of statistical inference) whether a series is non-stationary or not. For instance, the sample correlogram of an *AR(1)* series with $\phi = 0.99$ will exhibit the type of gradual decay of a non-stationary process.

Non-stationarity seems a natural feature of economic life. Legislative change is one obvious source of non-stationarity, often inducing structural breaks in time series, but it is far from the only one. Economic growth, perhaps resulting from technological progress, ensures secular trends in many time series. Such trends need to be incorporated into statistical analyses, which could be done in many ways, including the venerable linear trend.

Our focus here will be on a type of stochastic non-stationarity induced by persistent accumulation of past effects, called unit-root processes (an explanation for this terminology is provided below). Such processes can be interpreted as allowing a different `trend' at every point in time, so are said to have stochastic trends. There are many plausible reasons why economic data may contain stochastic trends. For example, technology involves the persistence of acquired knowledge, so that the present level of technology is the accumulation of past discoveries and innovations. Economic variables depending closely on technological progress are therefore likely to have a stochastic trend. The impact of structural changes in the world oil market is another example of non-stationarity. Other variables related to the level of any variable with a stochastic trend will `inherit' that non-stationarity, and transmit it to other variables in turn: nominal wealth and exports spring to mind, and therefore income and expenditure, and so employment, wages etc. Similar consequences follow for every source of stochastic trends, so the linkages in economies suggest that the levels of many variables will be non-stationary, sharing a set of common stochastic trends. A non-stationary process is, by definition, one which violates the stationarity requirement, so its means and variances are non-constant over time. For example, a variable exhibiting a shift in its mean is a non-stationary process, as is a variable with a heteroscedastic variance over time. We will focus here on the non-stationarity caused by stochastic trends and discuss its implications for empirical modelling.

4.1.3 A Special Case: The Random Walk model

Consider the special case of an *AR(1)* model taking the following form $y_t = y_{t-1} + \varepsilon_t$ or $\Delta y_t = \varepsilon_t$. Clearly, this is a special case of an *AR(1)* model where $\phi = 1$ and $\phi_0 = 0$ (the intercept). This is the so-called **Random-Walk without drift**. You can think of this as describing your wealth from betting on the outcome of a coin toss, and a head adding £1 to your wealth while a tail costing you £1. Let $\varepsilon_t = £+1$ if a head appears and $\varepsilon_t = £-1$ in the event of a tail. Thus, your current wealth y_t equals last period's wealth y_{t-1} plus the realized value of ε_t. If y_0 is a given initial condition then the following is true: $y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i$.

Taking expected values, we obtain $E(y_t) = E(y_{t-s}) = y_0$, thus the mean of a random walk is a constant. ***However, all stochastic shocks have non-decaying effects.*** Given the first t realizations of the ε_t process, the conditional mean of y_{t+1} is: $E(y_{t+1}) = E_t(y_t + \varepsilon_t) = y_t$, similarly the conditional mean of y_{t+s} (for any s > 0) can be obtained from: $E(y_{t+s}) = E_t(y_t + \varepsilon_t) = y_t$

Basically it is the Markovian property implying that the conditional means for all future dates are equal to the current value. However, an ε_t shock has a non-decaying effect, so the sequence of y_t is permanently influenced by a shock.

Notice that the variance is time-dependent. To see that recall that: $Var(y_t) = Var(\varepsilon_t + \varepsilon_{t-1} + ... + \varepsilon_1) = t\sigma^2$, so $Var(y_{t-s}) = Var(\varepsilon_{t-s} + \varepsilon_{t-s-1} + ... + \varepsilon_1) = (t-s)\sigma^2$

So as it becomes apparent, the variance is not constant and furthermore as $t \to \infty$, the variance of y_t also approaches to infinity. The autocovariance of y_t and y_{t-s} is: $E\left[\left(y_t - y_0\right)\left(y_{t-s} - y_0\right)\right] = E\left[\left(\varepsilon_t + \varepsilon_{t-1} + ... + \varepsilon_1\right)\left(\varepsilon_{t-s} + \varepsilon_{t-s-1} + ... + \varepsilon_1\right)\right] = E\left[\left(\varepsilon_{t-s}^2\right) + \left(\varepsilon_{t-s-1}^2\right) + ... + \left(\varepsilon_1^2\right)\right] = \left(t - s\right)\sigma^2$. The autocorrelation takes the following form: $\rho_s = \left[\left(t - s\right)/t\right]^{0.5}$

4.1.4 Unit Roots and the problems with inference

Suppose we know that a series is generated by an *AR(1)* model: $y_t = \alpha_1 y_{t-1} + \varepsilon_t$. First, suppose that we wish to test the null hypothesis that $\alpha_1 = 0$. Under the maintained null hypothesis, we can estimate the parameters by OLS the fact theta the error term is a white noise and that $|\alpha_1| < 1$ we can obtain an efficient estimate for the parameters of interest. Then, obtain the standard error of the estimate and calculate the standard t-stat in order to carry on with the hypothesis testing.

However, the situation is quite different if we want to test the hypothesis that $\alpha_1 = 1$. Now, under the null hypothesis y_t is non-stationary and we know that the variance becomes infinitely large as t increases. Under the null hypothesis, it is inappropriate to use classical statistical methods to estimate and perform significance tests on the coefficient α_1. It is rather simple to show that the OLS estimate will yield a biased estimate. Recall that $\rho_1 = \left[\left(t - 1\right)/t\right]^{0.5} < 1$. Since the estimate of α_1 is directly related to the value of ρ_1, the estimated value is based to be below its true value of unity. The estimated model will mimic that of a stationary *AR(1)* process with a near unit root. Hence, the usual t-stat cannot be used to test the hypothesis of interest.

4.1.5 Spurious Regression

Consider the following regression model: $y_t = \beta_0 + \beta_1 z_t + u_t$. The assumptions of the CLRM necessitate that both series are stationary and that the errors have a zero mean and a finite variance. In the presence of non-stationary variables, there might be what Granger and Newbold (1974) called a ***spurious regression***. Such a regression typically has a high R-squared, t-stats that appear to be insignificant, but h results are without any economic meaning.

What Granger and Newbold (1974), using **Monte Carlo** analysis, showed was that when using two independent random walks such as: $y_t = y_{t-1} + \varepsilon_{yt}$, and $z_t = z_{t-1} + \varepsilon_{zt}$, then the regression model above is meaningless; any relationship between the two variables is spurious. Surprisingly, at the 5% level of significance, they were able to reject the null hypothesis $\beta_1 = 0$ in approximately 75% of the time. Moreover, the regression had very high R-squared values and the residuals exhibited a high degree of autocorrelation. Effectively what the regression is picking-up is the two stochastic trends which dominate the behaviour of the two series.

4.1.6 Testing for Unit Roots (Augmented Dickey-Fuller Test)

Recall that a time series is said to be (weakly) stationary if the population mean, variance, and (auto)covariances exist and do not change over time. A stationary series is said to be integrated to order 0, or I(0). Nonstationary series can take many forms, including those with deterministic shifts and/or explosive properties, but those that when first-differenced produce an invertible stationary series are called I(1). If a series has to be differenced d times to induce stationarity, it is said to be an I(d) series, or integrated to order d.

Dickey and Fuller (1979) provided a formal way for testing for the presence of unit roots. Starting with an AR(1) model of the form $y_t = \beta y_{t-1} + \varepsilon_t$, and subtract y_{t-1}

from both sides, a new but equivalent equation arises $\Delta y_t = \gamma y_{t-1} + \varepsilon_t$, where $\gamma = \beta - 1$. This model can be used to test for the presence of a unit root. Note that acceptance of the null, $\gamma = 0$, implies that the model can be expressed totally in terms of changes in the variable. Rejection of the null ($\gamma < 0$) implies that $\gamma \neq 0$or equivalently that $\beta < 1$, is appropriate. Thus, the modified model is useful in distinguishing between levels and differences. However, two points are worth making about regressions such as $\Delta y_t = \gamma y_{t-1} + \varepsilon_t$ and a similar relationship exists for higher order models, that is those with autocorrelated residualsε_t, as is noted below. First, a standard t-test or asymptotic normal test should not be used for the test $\gamma = 0$, the test that distinguishes between changes and levels. Instead, the t-ratio for $\hat{\gamma}$ should be compared to a critical value from Fuller's (1976) tables; such a test is known as a Dickey-Fuller test in the first-order case, or an augmented Dickey-Fuller test (ADF) in higher order models discussed below. The distribution of the Dickey-Fuller test critically depends on whether the constant, α, is zero in the equation that generates the data, the so-called dgp (data generating process) but not on the order of the autoregressive process describingε_t . It is worth considering four separate cases.

i. The true model is a random walk without drift, $y_t = \beta y_{t-1} + \varepsilon_t$ and one estimates $\Delta y_t = \gamma y_{t-1} + \varepsilon_t$ that is a regression without a constant. The 5% critical value (one-sided) is -1.95 for all reasonable sample sizes compared to -1.64 for a large sample normal test.

ii. The true model is a random walk without drift, $y_t = \beta y_{t-1} + \varepsilon_t$ and $\Delta y_t = \alpha + \gamma y_{t-1} + \varepsilon_t$ is estimated, that is a regression with a constant. The 5% critical value (one-sided) varies between -2.93 (50 observations) and -2.86 (large samples).

iii. The true model is a random walk with drift, $y_t = \alpha + \beta y_{t-1} + \varepsilon_t$ with $\alpha \neq 0$ and $\Delta y_t = \alpha + \gamma y_{t-1} + \varepsilon_t$ is estimated, that is a regression with a constant. The t-test on $\hat{\gamma}$ is asymptotically normal because the time trend that results from a non-zero constant dominates the lagged dependent variable. However, the variance of the estimate depends upon unknown parameters and, therefore, this test is not feasible.

iv. The true model is a random walk with drift, $y_t = \alpha + \beta y_{t-1} + \varepsilon_t$, with $\alpha \neq 0$, and $\Delta y_t = \alpha + \delta t + \gamma y_{t-1} + \varepsilon_t$ is estimated, that is a regression with a constant and a time trend (t). The 5% critical value (one-sided) varies between -3.50 (50 observations) and -3.41 (large samples).

Decision rule: If the pseudo t-stat is lower (higher) than the DF critical value (obtained from the appropriate tables), then reject (do not reject) the null of non-stationarity. The choice of case depends on what the data generating process might be believed to be. Most practitioners use case (ii) when a series is not thought to drift and case (iv) when drift is apparent. One argument in favour of these two cases is that the model should be ‘reasonable’ under both the null (nonstationarity) and the alternative (stationarity -possibly about a linear time trend). However, as in all hypothesis tests, rejection of the null does not imply the alternative is correct nor does failure to reject imply that the null is correct.

The DF test assumes that the error term is white noise. However, sometimes this is not a realistic assumption. Effectively, one account for this possibility by using the so-called Augmented DF test where lags of the dependent variable are added in the following fashion:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=2}^{p} \Delta y_{t-i+1} + \varepsilon_t$$

Naturally, the outcome of the test depends upon the choice of p. If p is too short, the residuals are autocorrelated and the test is biased. On the other hand, if p is too large, the equation is over-parameterised and a lack of power will result. The choice of p can be made on the basis of a sequence of t-tests or some other criterion such as AIC or BIC.

4.1.7 Unit Roots and the Order of Integration

A unit root implies that a series is non-stationary; note though that a series may have more than one unit roots. If y_t has a single unit root then Δy_t is stationary. In other words, y_t needs to be differenced once to achieve stationarity. It is useful to remember that: Order of integration = number of times a series needs to be differenced in order to achieve stationarity = number of unit roots. For instance, if $y_t \sim I(n)$ then y_t is integrated of order n. Typically, $n = 0, 1,$ or 2.

Although `classical' econometric theory generally assumed stationary data, particularly constant means and variances across time periods, empirical evidence is strongly against the validity of that assumption. Nevertheless, stationarity is an important basis for empirical modelling, and inference when the stationarity assumption is incorrect can induce serious mistakes. To develop a more relevant basis, we considered recent developments in modelling non-stationary data, focusing on autoregressive processes with unit roots. We showed that these processes were non-stationary but could be transformed back to stationarity by differencing and cointegration transformations, where the latter comprised linear combinations of the variables that did not have unit roots. We investigated the comparative properties of stationary and non-stationary processes, reviewed the historical development of modelling non-stationarity and presented a re-run of a famous Monte Carlo simulation study of the dangers of ignoring non-stationarity in static regression analysis. Next, we described how to test for unit roots in scalar autoregressions, then extended the approach to tests for cointegration. Finally, an extensive empirical illustration using two gasoline prices implemented the tools described in the preceding analysis. Unit-root non-stationarity seems widespread in economic time series, and some theoretical models entail unit roots. Links between variables will then `spread' such non-stationarities throughout the economy. Thus, we believe it is sensible empirical practice to assume unit roots in (log) levels until that is rejected by well-based evidence. Cointegrated relations and differenced data both help model unit roots, and can be related in equilibrium-correction equations, as we illustrated. For modelling purposes, a unit-root process may also be considered as a statistical approximation when serial correlation is high. Monte Carlo studies have demonstrated that treating near-unit roots as unit roots in situations where the unit-root hypothesis is only approximately correct makes statistical inference more reliable than otherwise. Unfortunately, other sources of non-stationarity may remain, such as changes in parameters (particularly shifts in the means of equilibrium errors and growth rates) or data distributions, so careful empirical evaluation of fitted equations remains essential. We reiterate the importance of having white-noise residuals, preferably homoscedastic, to avoid misleading inferences. This emphasizes the advantages of accounting for the dynamic properties of the data in equilibrium-correction equations, which not only results in improved precision from lower residual variances, but delivers empirical estimates of adjustment parameters. Later we will attempt to explain cointegration analysis will address system methods. Since cointegration inherently links several variables, multivariate analysis is natural, and recent developments have focused on this approach.

5. Dynamic Specification, Error Correction Model, Cointegration in Single Equation setting

5.1 A Taxonomy of Dynamic models

As discussed earlier the presence of serial correlation implies that the dependent variable (*Y*) is not only affected by *X* at time t but also by past values of *X*. In other words, serial correlation might not be of pure form but simply an indication that the model in misspecified. Therefore, it could be the case that the appropriate strategy to deal with autocorrelation is not to model it directly but to incorporate time lags explicitly instead. This would be the case when the dependent variable does not respond immediately to a specific change in the independent variables but does so with some delay. Think of a general model where we allow Y to depend on its own history, on the current and past level of X. (Assuming that all series are stationary):

$$Y_t = \delta + \theta Y_{t-1} + \phi_0 X_t + \phi_1 X_{t-1} + \varepsilon_t$$

The above model is usually called Autoregressive Distributed Lag (ADL). An interesting element is that it describes the dynamic effects of a change in X_t on current and future values of Y. Taking partial derivatives, we can obtain the **immediate response**, given by: $\frac{\partial Y_t}{\partial X_t} = \phi_0$

Sometimes also called **impact multiplier**. The effect after one period is:

$$\frac{\partial Y_{t+1}}{\partial X_t} = \frac{\partial Y_t}{\partial X_t} + \phi_1 = \theta\phi_0 + \phi_1$$

After two periods is: $\frac{\partial Y_{t+2}}{\partial X_t} = \theta \frac{\partial Y_{t+1}}{\partial X_t} = \theta\left(\theta\phi_0 + \phi_1\right)$

It is obvious that after the first period the effect is decreasing as long as $|\theta| < 1$ (we know what this condition implies!). One can now derive the **long-run multiplier** (or also called **equilibrium multiplier**)

$$\phi_0 + (\theta\phi_0 + \phi_1) + \theta(\theta\phi_0 + \phi_1) + ... = \phi_0 + \left(1 + \theta + \theta^2 + ...\right)(\theta\phi_0 + \phi_1) = \frac{\phi_0 + \phi_1}{1-\theta}$$

This captures the fact that the unit increase in X_t will have a cumulative effect on the current and future levels of Y_t. To arrive at a similar conclusion return to the ADL model and derive the equilibrium relationship between the variables by effectively imposing that in the long–run the following will hold:

$E(X_t) = E(X_{t-1}), E(Y_t) = E(Y_{t-1})$. Hence: $E(Y_t) = \delta + \theta E(Y_t) + \phi_0 E(X_t) + \phi_1 E(X_t)$

or $E(Y_t) = \frac{\delta}{1-\theta} + \frac{\phi_0 + \phi_1}{1-\theta} E(X_t)$.

5.1.1 Finite Distributed Lag model

A Restricted version of the ADL with the autoregressive coefficient being zero is given below: $Y_t = a + \sum_{i=0}^{k} \beta_i X_{t-i} + \varepsilon_t$

As before, the impact of X on Y occurs (is distributed) over a finite number of periods. The first beta (β_0) coefficient denotes the impact of a unit change in X on the mean of Y in the same period, given the lagged values of X. It is called the **impact (short-run) multiplier**. The beta coefficients of the lagged values of X, which relate changes in X from previous periods to the mean of Y are called **interim multipliers**. For instance, β_1 denotes the effect of a unit change in X last period on the mean of Y in the current period. It is called **interim multiplier of order 1**. If this one unit of change in X is maintained indefinitely, the mean of Y changes by β_0 in the initial period and by (β_0 + β_1) after two periods. The latter sum denotes the **two-period interim multiplier**. The equilibrium multiplier is given by the sum of betas

A useful metric often derived from the DL model is the average or **Mean Lag**, defined as: $ML = \frac{\sum_{i=0}^{k} i\beta_i}{\beta}$.

The mean lag is simply a weighted average of the betas, where these values do not necessarily sum to 1. The mean lag denotes the average number of periods during which a sustained change in X influences Y.

Technically, one should have no problem in estimating model as it stands. However, in practice it is unusual to have prior information for the order of the distributed lag model, which means that the selection of the lag order is part of the estimation process. Why do we care about specifying the correct lag order? Because if too few lags are included specification error occurs leading to biased and inconsistent parameter estimates. If too many lags are included, the 'irrelevant' X's will potentially contribute to a multicollinearity problem adversely affecting the magnitudes of lagged X's as well as their t-statistics. Notice that even if we knew the 'true' lag order, if it is too high then we would have a severe loss of degrees of freedom since each lag consumes one degree of freedom. In general, when we usually do not know the 'true' lag order we run both the dual risks of both multicollinearity and specification error. A number of methods for estimating DL models have been proposed; in this course we will review the **Koyck's model (KDL)**.

5.1.2 Koyck Model

Assuming that the coefficients continuously diminish, so that the influence of successively distant values of *X* become smaller, and that these weights decrease geometrically, the coefficient for the ith period lag of *X*, β_i can be expressed as:

$$\beta_i = \beta_0 \lambda^i$$

Where λ is a constant. If $\lambda = 1$ the assumption of declining lag coefficients is violated. When $\lambda = 0$, *Y* is not affected by *X*. Therefore, for geometrically declining weights this constant must satisfy the following restriction: $0 < \lambda < 1$. The constant λ indicates the **rate of decay** for the distributed lag. The closer it is to one, the slower the decrease in successive lag coefficients, while for small values of *lambda,* the coefficients fall more rapidly. The quantity $1-\lambda$ is called the **speed of adjustment**, which indicates the rate at which successive lag coefficients decrease.

5.1.3 Error Correction Model (ECM)

Starting from the ADL subtract Y_{t-1} from both sides to obtain: $\Delta Y_t = \delta - (1-\theta)Y_{t-1} + \phi_0 \Delta X_t + (\phi_0 + \phi_1) X_{t-1} + \varepsilon_t$ or

$\Delta Y_t = \phi_0 \Delta X_t - (1-\theta)[Y_{t-1} - a - \beta X_{t-1}] + \varepsilon_t$.

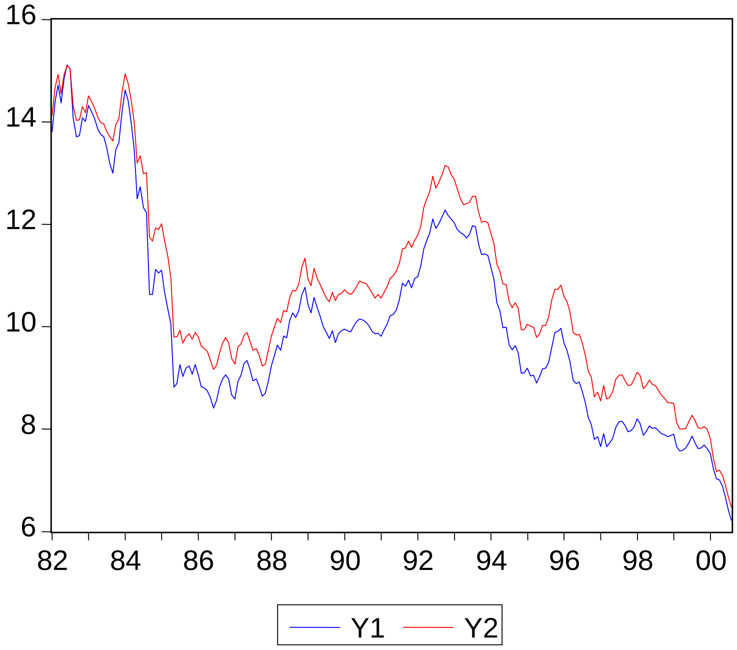
Intuition: the change in Y is due to the current change in X (the first differences) plus an **error correction mechanism**. If Y_{t-1} is above the equilibrium value that corresponds to X_{t-1} (the underlying equilibrium economic relationship), that is if the ‘equilibrium error’ in square brackets is positive, an additional negative adjustment in Y_t is generated. Very important is the **speed of adjustment**, determined by $(1-\theta)$ and is positive since $|\theta| < 1$ (stability).

5.1.4 Partial Adjustment Model

This is an alternative model with an intuitive economic interpretation. Let Y_t^* denote the optimal or desired level of Y and assume that: $Y_t^* = a + \beta X_t + \eta_t$. The actual value of Y_t will differ from Y_t^* because an adjustment to its optimal level is not immediate. Suppose that the adjustment is only partial in the sense that: $Y_t - Y_{t-1} = (1-\theta)(Y_t^* - Y_{t-1})$

5.2 Dynamic Comovement: Cointegration

An example of dynamic comovement (US interest rates of maturities 1 and 2 years, monthly data from 1981 to 2000) is given in the graph below:



A necessary, but not sufficient condition for cointegration is that the two series should be integrated of the same order. Even if a pair of series is individually non-stationary, certain linear combinations of contemporaneous observations seem to be stationary in the sense that they do not require further differencing to exhibit limited dependence (Stock, 1987). Suppose that both *X* and *Y* are I (1), as the case at hand, so that their changes are I (0). Then typically any linear combination of *X* and *Y* will be I (1). However, if there exists a linear combination such that $Z=X-\alpha Y$ is I (0), then *X* and *Y* are said to be co-integrated. Put more formally, following Engle and Granger (1987) the components of a vector *Z* are said to be co-integrated of order d, b, denoted as $Z \sim CI\ (d\text{-}b)$ if:

(i) all components of *Z* are I(d);

(ii) there exists a vector $\beta\,(\neq 0\,)$ such that $\beta' Z\sim I\,(\,d\text{-}b\,),\, b > 0$. The vector β is called the co-integrating vector.

In the case of $d = b = 1$, co-integration would mean that the so-called equilibrium error would be I (0) and therefore *Z* will rarely drift far from zero (in the case of a zero mean) and will often cross the zero line. In other words the equilibrium relationship described by *Z* will occur occasionally, so it will not be meaningless. The basic idea is that at least in the long run, the two series will move together, despite their individual non-stationarity, so there exist(s) linear combination (s) of them which is (are) stationary. In a sense, the line $X\text{-}\alpha Y\text{=}0$ can be considered to be an ‘equilibrium’ or ‘attractor’ of the system in the phase-space, so that *Z* can be interpreted as the extent to which the system is out of equilibrium.

5.2.1 Testing for Cointegration in single equation framework: The Engle-Granger approach

An equivalent question as to whether the relationship between *X* and *Y* is spurious is whether they are cointegrated. Granger (1981) introduced the case $y_t = \alpha + \beta x_t + \varepsilon_t$, where the individual time series are I(1) but the error term is I(0). That is, the error term might be autocorrelated but, because it is stationary, the relationship will keep returning to the equilibrium or long-run equation $y_t = \alpha + \beta x_t$. More formally, if a vector of time series is I(d) but a linear combination is integrated to a lower order, the time series are said to be ***Cointegrated***. However, it is instructive to return to an I(1) world to put the cointegrated model in perspective. Granger (1981) and Engle and Granger (1987) demonstrated that, if a vector of time series is cointegrated, the long-run parameters can be estimated directly without specifying the dynamics because, in statistical terms, the estimated long-run parameter estimates converge to their true values more quickly than those operating on stationary variables. That is, they are ‘superconsistent’ and a two-step procedure of first estimating the long-run relationship and estimating the dynamics, conditional on the long run becomes possible. As a result, simple static models came back in vogue in the late 1980’s but it rapidly became apparent that small sample biases can indeed be large (Banerjee et al, 1986).

Two major problems typically arise in a regression such as (6.2.1). First, it is not always clear whether one should regress y_t on x_t or vice versa. Endogeneity is not an issue asymptotically because the simultaneous equations bias is of a lower order of importance and, indeed, is dominated by the nonstationarity of the regressor. However, least squares is affected by the chosen normalisation and the estimate of one regression is not the inverse of that in the alternative ordering unless $R^2 = 1$. Secondly, the coefficient $\hat{\beta}$ is not asymptotically normal when x_t is I(1) without drift, even if ε_t is iid. Of course, autocorrelation in the residuals produces a bias in the least squares standard errors, even when the regressor is nonstationary, and this effect is in addition to that caused by nonstationarity. The preceding discussion is based on the assumption that the disturbances are stationary. In practice, it is necessary to pre-test this assumption. Engle and Granger suggested a number of alternative tests but that which emerged as the ‘popular’ method is the use of an ADF test on the residuals without including a constant or a time trend. Naturally, this test depends upon the normalisation rule and, hence, conflict can, and often does arise. This led some researchers to conduct the test in both directions, but such an approach would cause severe size distortions. The ADF critical values are inappropriate because, as the least squares procedure is designed to minimise the residual variance, disturbances which are, in fact, nonstationary will produce estimated residuals that are biased towards a finite variance (stationary) time series. For example, the bivariate critical value is approximately -3.2.

6. VAR, Dynamic Systems (atheoretical econometrics)

6.1 The VAR Representation

Employing a Vector Autoregression (VAR) is a way to avoid classifying variables between exogenous and endogenous; similarly, no *a priori* restrictions are used. Hence, sometimes called *atheoretical econometrics* (Sims). In essence all variables are treated symmetrically. Consider the two-variable case:

$$\begin{aligned} y_t &= b_{10} - b_{12} z_t + \gamma_{11} y_{t-1} + \gamma_{12} z_{t-1} + \varepsilon_{yt} \\ z_t &= b_{20} - b_{21} y_t + \gamma_{21} y_{t-1} + \gamma_{22} z_{t-1} + \varepsilon_{zt} \end{aligned} \qquad (1)$$

Where the error terms are white noises and are independent across equations. Such a model is called a *first-order vector autoregression, VAR(1).* We can re-write the model in a *reduced form*:

$$\begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix}$$

or (2)

$$\mathbf{Bx_t} = \mathbf{\Gamma_0} + \mathbf{\Gamma_1 x_{t-1}} + \boldsymbol{\varepsilon}_\mathbf{t}$$

the standard VAR form is obtained as follows:

$$\begin{aligned} &\mathbf{B}^{-1}\mathbf{Bx_t} = \mathbf{B}^{-1}\mathbf{\Gamma_0} + \mathbf{B}^{-1}\mathbf{\Gamma_1 x_{t-1}} + \mathbf{B}^{-1}\boldsymbol{\varepsilon}_\mathbf{t} \\ &\mathbf{x_t} = \mathbf{A_0} + \mathbf{A_1 x_{t-1}} + \mathbf{e_t} \\ &\text{where} \\ &\mathbf{A_0} = \mathbf{B}^{-1}\mathbf{\Gamma_0}, \mathbf{A_1} = \mathbf{B}^{-1}\mathbf{\Gamma_1}, \mathbf{e_t} = \mathbf{B}^{-1}\boldsymbol{\varepsilon}_\mathbf{t} \end{aligned} \qquad (3)$$

6.1.1 Stability and Stationarity

Solving the system backwards we obtain:

$$\mathbf{x_t} = \mathbf{A_0} + \mathbf{A_1}\left(\mathbf{A_0} + \mathbf{A_1 x_{t-2}} + \mathbf{e_{t-1}} + \mathbf{e_t}\right) = \left(\mathbf{I} + \mathbf{A_1}\right)\mathbf{A_0} + \mathbf{A_1^2 x_{t-2}} + \mathbf{A_1 e_{t-1}} + \mathbf{e_t} \qquad (4)$$

and after n iterations, this yields:

$$\mathbf{x_t} = \left(\mathbf{I} + \mathbf{A_1} + ...\mathbf{A_1^n}\right)\mathbf{A_0} + \sum_{\mathbf{i=0}}^{\mathbf{n}} \mathbf{A_1^i e_{t-i}} + \mathbf{A_1^{n+1} x_{t-n-1}} \qquad (5)$$

As we continue to iterate backward, it is clear that convergence requires the expression $\mathbf{A}_\mathbf{1}^\mathbf{n}$ vanish as n approaches infinity. So stability requires that the roots of $(1-a_{11}L)(1-a_{22}L)-(a_{11}a_{22}L^2)$ lie outside the unit root circle.

6.1.2 Estimation and Identification

Consider the following multivariate generalization:

$$\mathbf{x_t = A_0 + A_1x_{t\text{-}1} + A_2x_{t\text{-}2} + ... + A_px_{t\text{-}p+} + e_t} \qquad (6)$$

The variables to be included in the model are selected according to the relevant economic theory. Apart from that there is a whole set of other things to consider, such as: (i) Lag length and (ii) Identification.

6.1.3 Choice of the appropriate Lag length

It is of particular importance because we need to avoid over-parameterization or under-parameterization and furthermore the lag length affects the degrees of freedom. If *p* is the lag length, then each of the *n* equations contains *np* coefficients plus the intercept term. Incorrect choice of lag order results in: (i) if too small then the model is misspecified, (ii) if too large then degrees of freedom are wasted.

Let Σ_u, Σ_r be the variance-covariance matrices of the unrestricted and restricted systems respectively, and let *c* denote the maximum number of regressors contained in the equation with highest lag order. Asymptotically the test statistic: $\mathbf{LR = (T\text{-}c)\left[log|\Sigma_u|\text{-}log|\Sigma_r|\right]}$ (7) has a chi-square distribution with degrees of freedom equal to the restrictions in the system. Additionally one could resort to the standard criteria (AIC, SBC).

6.1.4 Identification

Model 1, in the structural form, cannot be estimated due to the feedback inherent in the system, so z_t is correlated with the error term. Note that there is no such problem with versions (3) or (6) for that matter. The issue is whether one can recover all the information present in system (1). More formally, is model (1) identifiable given the OLS estimates of the VAR model in the form of (3)? The answer to that is NO, unless we restrict system (1) in an appropriate way. To show you why, consider this: the structural model has 10 parameters, while the VAR model has 9 parameters. Hence, one of the parameters has to be restricted. Following Sims (1980) imposing the restriction on system (1) that b_{21} equals zero, yields:

$$y_t = b_{10} - b_{12}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \varepsilon_{yt}$$
$$z_t = b_{20} + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \varepsilon_{zt}$$

so now $B^{-1} = \begin{bmatrix} 1 & -b_{12} \\ 0 & 1 \end{bmatrix}$

and the system in a VAR form now looks like:

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & -b_{12} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} 1 & -b_{12} \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}\begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix} \text{or}$$

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} b_{10} - b_{12}b_{20} \\ b_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{11} - b_{12}\gamma_{21} & \gamma_{12} - b_{12}\gamma_{22} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}\begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{yt} - b_{12}\varepsilon_{zt} \\ \varepsilon_{zt} \end{bmatrix}$$

Estimating the following system:

$$y_t = a_{10} + a_{11}y_{t-1} + a_{12}z_{t-1} + e_{yt}$$
$$z_t = a_{20} + a_{21}y_{t-1} + a_{22}z_{t-1} + e_{zt}$$

yields the following estimates:

$$a_{10} = b_{10} - b_{12}b_{20}$$
$$a_{11} = \gamma_{11} - b_{12}\gamma_{21}$$
$$a_{12} = \gamma_{12} - b_{12}\gamma_{22}$$
$$a_{20} = b_{20}$$
$$a_{21} = \gamma_{21}$$
$$a_{22} = \gamma_{22}$$

we can also calculate the parameters of the variance-covariance matrix:

$$Var\left(e_1\right) = \sigma_y^2 + b_{12}^2\sigma_z^2$$
$$Var\left(e_2\right) = \sigma_z^2$$
$$Cov\left(e_1, e_2\right) = -b_{12}\sigma_z^2$$

Now we have 9 equations which can be solved to solve for the 9 parameters. Rethinking the restriction imposed: we have restricted y not to have any contemporaneous effect on z. Hence, contemporaneous shocks from both equations affect y, but only shocks in z affect its contemporaneous value. Such decomposition, in a triangular fashion, is called a **Choleski Decomposition**.

7. Cointegration in a Multiple Equation setting

The Engle-Granger two-step method for testing for cointegration suffers from several drawbacks:

- The problem of normalisation
- Cannot be used to estimate multiple cointegration vectors it does not deal with the endogeneity of the regressors (see discussion on VAR models)

7.1 Three fundamental theorems on cointegrated variables

7.1.1 The rank of the cointegration space

If the vector of variables of interest is of order n there may be up to $n-1$ linearly independent cointegration vectors, Clearly, if there are two variable sin the vector there can be at a maximum one cointegration vector the number of cointegration vectors is called the cointegrating rank.

7.1.2 Cointegration and Common Stochastic Trends

Cointegration of a vector of variables implies that the number of unit roots in the system is less than the number of unit roots in the corresponding univariate series. If two variables are cointegrated then they must share the same stochastic trend. In fact, as a general rule one can move from the number of cointegrating relations (r) and the number of common stochastic trends (q): $r = n - q$.

7.1.3 The Granger Representation Theorem

If two variables are cointegrated then by virtue of the GRT a Vector Error Correction model is in place as follows:

$$\Delta x_t = \gamma_x \left(x_{t-1} - \beta y_{t-1}\right) + \sum_{i=0}^{k} \alpha_i \Delta x_{t-i} + \sum_{i=0}^{k} \delta_i \Delta y_{t-i} + \varepsilon_{xt}$$

$$\Delta y_t = \gamma_y \left(x_{t-1} - \beta y_{t-1}\right) + \sum_{i=0}^{k} \lambda_i \Delta x_{t-i} + \sum_{i=0}^{k} \theta_i \Delta y_{t-i} + \varepsilon_{yt}$$

Where $\left(x_{t-1}-\beta y_{t-1}\right)$ is the error-correction term and γ_x, γ_y are the speed-of-adjustment coefficients. So the VECM is essentially a VAR in first-differences augmented by the level terms capturing the deviation from equilibrium from the last period in matrix form the model is:

$$\Delta X_t = \Gamma_1 \Delta X_{t-1} + \Gamma_2 \Delta X_{t-2} + \ldots + \Gamma_{k-1} \Delta X_{t-k-1} - \Pi X_{t-k} + u_t$$

Where

$$\Gamma_i = -(I - A_1 - \ldots - A_i) \qquad i = 1,2,\ldots,k-1$$

and $\Pi = -(I - A_1 - \ldots - A_k)$

The rank of matrix Π determines whether there are any significant cointegrating vectors between the variables. Clearly if the rank of Π is zero the matrix is null and the model is just a VAR model in first differences. The other extreme case is when Π has full column rank, which is equivalent to the stationarity of the vector process. The intermediate case of reduced column rank implies that there exist stationary linear combinations of the variables, corresponding to the cointegration vectors. $\Pi = \gamma\beta'$ is the matrix of long-run parameters, the first component is the matrix of weights with which each cointegration vector enters each equation.

7.1.4 Digression: Characteristic Roots

Let A be an $\left(n\times n\right)$ square matrix with elements α_{ij}, and x an $\left(n\times 1\right)$ vector. The scalar λ is called a characteristic root of A if: $Ax=\lambda x$, let I be an $\left(n\times n\right)$ identity matrix, then: $Ax-\lambda x=0 \Rightarrow \left(A-\lambda I\right)x=0$. Since x is a vector containing values not identically equal to zero then the previous equation requires that the rows of $\left(A-\lambda I\right)$ be linearly dependent. Equivalently, it requires that: $\left|A-\lambda I\right|=0$ (called the characteristic equation and will always be a polynomial of order (n) in λ, which immediately implies that there will be (n) roots). So the characteristic roots can be found by finding the values of λ that satisfy: $|A-\lambda I|=0$.

Example

$$A=\begin{bmatrix} 0.5 & -0.2 \\ -0.2 & 0.5 \end{bmatrix} \text{ so that } |A-\lambda I|=\begin{vmatrix} 0.5-\lambda & -0.2 \\ -0.2 & 0.5-\lambda \end{vmatrix}$$

Solving for the value of λ such that $|A-\lambda I|=0$ yields a quadratic equation of the form: $\lambda^2-\lambda+0.21=0$, the two values that solve the equation are $\lambda_1=0.7, \lambda_2=0.3$ (the characteristic roots). If the series are not cointegrated the rank is zero and all these characteristic roots will equal zero.

7.2 Formal tests for the rank of the cointegration space

There are two formal tests to help us decide (i) whether a set of series are cointegrated and (ii) the number of cointegration vectors in case of cointegration

These are: $\lambda_{trace}=-T\sum_{i=r+1}^{n}\ln\left(1-\hat{\lambda}_i\right)$ and $\lambda_{\max}=-T\ln\left(1-\hat{\lambda}_{r+1}\right)$.

Where $\hat{\lambda}_i$ is the estimated values of the characteristic roots (eigenvalues) obtained form the estimated ***Π*** matrix**.** T = number of observations (the length of the time series), λ_{trace} tests the null that the number of distinct cointegration vectors is less than or equal to r against a general alternative. It should be clear that λ_{trace} equals zero when all $\hat{\lambda}_i$ are zero. The further the estimated characteristic roots are from zero, the more negative is $\ln\left(1-\hat{\lambda}_i\right)$ and the larger the λ_{trace} statistic. $\lambda_{\max}$ tests the null that the number of cointegration vectors is r against the alternative of r+1 vectors.

8. Causality and Time Series

The question of Cause and Effect is a fundamental one in our attempt to understand our environment. Although Causality is one of the central and most widely discussed concepts in the scientific agenda, scientists of different disciplines disagree about its appropriate definition. Therefore, if one wants to be sharp, has to use an operational definition of Causality. Furthermore, there is a need for invariant and independent of theories criterion. Especially in economics and politics where there is little consensus for the 'laws' governing economic and political systems, a criterion dependent on the theoretical framework adopted, would be undermined by the model's validity. For this reason the use of a purely statistical criterion, independent of economic or political theory is essential. It turns out that such a criterion can be developed if the Granger's definition of causality is used (Granger, 1969).

8.1 Granger Causality

Granger's definition of causality is in terms of predictability. The pivotal idea is that a 'cause' ought to improve our ability to forecast an effect in a stochastic system. In other words, a variable $\boldsymbol{X}$ causes another variable $\boldsymbol{Y}$ if the latter can be more accurately forecasted by using the history of $\boldsymbol{X}$ rather than by not doing so. Thus, Granger's definition of causality is based upon an incremental predictability criterion.

Furthermore, Granger's definition of causality is based on the stochastic nature of the variables and its central feature is the direction of the flow of time. It is purely a statistical criterion relying entirely on the assumption that the future cannot cause the past. To put more formally, the essence of Granger's concept of causality is that X causes Y if the knowledge of X's history leads to improved prediction of Y. Before formally defining causality, the following axiom is assumed to hold:

Axiom: *The past and present may cause the future, but the future cannot cause the past.*

In order to provide an operational definition of causality, **Granger (1980)** will be followed. Suppose that one is interested in the possibility that a series X_t causes another Y_{t+n}. Let J_t be the universal information set available at time t. Define J_t - X_t as the set of elements of J_t without the element X_t.

Then it is,

Definition A If $MSE^2\ (Yt_{+n} \mid J_{t)} < MSE\ (Yt_{+n} \mid J_t - X_{t+n})$

Then X_t *Granger causes* Y_{t+n}, denoted by $X \to Y$.

If a less general information set than the universal is available, as always is the case in economic modelling, then a *prima facie* cause occurs. However, if causality is present in the way defined above, there is no information on whether it holds bilaterally, that is, if *Y* causes *X.* If causality exists in both directions we say that feedback occurs. This is formally defined as:

Definition B

If $MSE\ (Yt_{+n} \mid J_{t)} < MSE\ (Yt_{+n} \mid J_t - X_{t+n})$

Then X_t *Granger causes* Y_{t+n}, denoted by $X \to Y$.

and

If $MSE\ (Xt_{+n} \mid J_{t)} < MSE\ (Xt_{+n} \mid J_t - Y_{t+n})$

Then Y_t *Granger causes* X_{t+n}, denoted by $Y \to X$. Therefore, bidirectional causality is present denoted by: X↔Y

To recap the causal structure of a bivariate system is exhausted by the following four mutually exclusive outcomes: X→Y, unidirectional causality running from X toY, or

[2] MSE stands for Mean Square Error

Y→X, unidirectional causality running from Y to X, or X↔Y, bidirectional causality exists, or X and Y are independent meaning that no causal link exists between the two processes.

8.1.1 Testing for causality

The traditional way for testing causality among two variables was developed in the Box-Jenkins framework. Basically, after determining the ARMA model for each of the series, attention was turned to their residuals. Simply, the researcher was looking for evidence of cross correlation among the two residual series in order to rule out independence. Among other problems of this way of testing is that there are questions about the 'statistical power' of this procedure. Furthermore, inferences regarding one-way causality are problematic. However, to do justice to the procedure it should be mentioned that it is less sensitive to the choice of the lag length.

Currently, the dominant procedure for assessing Granger Causality among stationary series[3] assumes that the information relevant to the prediction of the respective variables is contained solely in the time series data on these variables. The test involves estimating the following regressions:

$$Y_t = \sum_{i=0}^{n} \alpha_i X_{t-i} + \sum_{j=0}^{n} \beta_i Y_{t-j} + \varepsilon_{1t} \quad (8.1)$$

$$X_t = \sum_{i=0}^{m} \gamma_i X_{t-i} + \sum_{j=0}^{m} \delta_i Y_{t-j} + \varepsilon_{2t} \quad (8.2)$$

Where it is assumed that the disturbance terms are uncorrelated across equations. The idea behind the above formulation is that each of the variables is related to its past values as well as of the other variable.

The causal structure of a bivariate system, as discussed earlier, is exhausted by four mutually exclusive outcomes. In order to determine which of the four is a better

[3] A generalisation of the procedure exists to accommodate nonstationarity.

description of the data at hand we can formulate a number of hypotheses in terms of the parameters appearing in the above models. We now distinguish four cases:

If the estimated coefficients on the lagged X in (8.1) are **jointly** statistically different from zero **and** the set of estimated coefficients on the lagged Y in (8.2) are not **jointly** different from zero, then **Unidirectional Causality** exists from X to Y. Conversely, if the set of coefficients of lagged X is **jointly** zero and the set of coefficients on lagged Y are **jointly** different from zero, then **Unidirectional Causality** exists from Y to X. Evidence for **Feedback** is indicated when **both** sets of coefficients are different from zero in **both equations**, and finally **Independence** is suggested when the sets of X and Y coefficients are not statistically different from zero in **both the regressions**. The actual testing of the above hypotheses is based on a set of F-tests in the form of restricted and unrestricted regressions models.

A final remark

Before proceeding to applications of the Granger test, keep in mind that the number of lagged terms to be included in the regressions is an important practical question. Also bear in mind that inference on the direction of causality may depend critically on the lag structure adopted.

References

Bollerslev, T. (1986). “Generalised Autoregressive Conditional Heteroscedasticity”, *Journal of Econometrics*, 31, 307-327.

Bollerslev, T., Engle, R. and Wooldridge, J. (1988). “A Capital Asset Pricing Model with Time-Varying Covariances”, *Journal of Political Economy*, 96, 116-172.

Bollerslev, T. and Wooldridge, J. (1992). “Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time Varying Covariances” *Econometric Reviews*, 11, 143–172.

Box, G.E.P. and G.M. Jenkins, (1970). “Time Series Analysis: Forecasting and Control”, Holden Day, San Francisco

Breusch, T.S. and L.G. Godfrey, (1981). “A Review of Recent Work on Testing for Autocorrelation in Dynamic Simultaneous Models”, in Macroeconomic Analysis: Essays in Macroeconomics and Econometrics, eds D. Currie, R. Nobay, and D. Peel, Croom Helm, London.

Chatfield, C., (1989). “The Analysis of Time Series: An Introduction”, Chapman and Hall, 4th edition.

Cohrane, D. and G.H. Orcutt, (1949). “Application of Least Squares Regression to Relationship Containing Autocorrelated Error Terms”, *Journal of the American Statistical Association*, 44, 32-61.

Dickey, D. A., and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**, 427{431.

Dickey, D. A., and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, **49**, 1057{1072.

Drakos, K. (2002). “A Daily View of the Term Structure Dynamics: Some International Evidence”, *De Economist*, 150(1), 1-12.

Drakos, K. and A. Kutan, (2003). “Regional Effects of Terrorism on Tourism: Evidence from Three Mediterranean Countries”, *Journal of Conflict Resolution*, 47(5), 621-641.

Drakos, K. (2004). “Terrorism-Induced Structural Shifts in Financial Risk: The Case of Airline Stocks in the Aftermath of the September 11th Terrorist Attacks”, *European Journal of Political Economy*, 20, 435-446.

Drakos, K. (2004) “Expectations Hypothesis or Market Segmentation? A Stochastic Trends-based Approach”, *European Review of Economics and Finance*, 3(2), 65-81.

Durbin, J. and G.S. Watson, (1950). “Testing for Serial Correlation in Least Squares Regression II”, *Biometrica*, 38, 159-178.

Durbin, J. and G.S. Watson, (1950). “Testing for Serial Correlation in Least Squares Regression I”, *Biometrica*, 37, 409-428.

Durbin, J., (1970). “Testing for Serial Correlation in Least Squares Regression when Some of the Regressors are Lagged Dependent Variables”, *Econometrica*, 38, 410-421.

Engle, R. (1982). “Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation”, *Econometrica*, 50, 987-1007.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, **51**, 277{304. Reprinted in Hendry, D. F., *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers, 1993; and in Ericsson, N. R. and Irons, J. S. (eds.) *Testing Exogeneity*, Oxford: Oxford University Press, 1994.

Engle, R., Lillien, D. and Robins, R. (1987). “Estimating time-varying risk premia in the term structure: The ARCH-M model”, *Econometrica*, 55, 391-408.

Engle, R. F., and Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica*, **55**, 251{276.

Engle, R. and Ng, V. (1993). “Time varying volatility and the dynamic behaviour of the term structure”, *Journal of Money, Credit and Banking*, 35, 336-349.

Engle, R. and Rothschild, M. (1990). “Asset pricing with factor ARCH covariance structure”, *Journal of Econometrics*, 45, 213-238.

Feige, E.L. and D.K. Pearce, (1979). “The Causal Causal Relationship between Money and Income: Some Caveats for the Time Series Analysis”, *Review of Economic and Statistics*, 61, 521-533.

Granger, C.W.J., (1969). “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods”, *Econometrica*, 37, 424-438.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model speci_cation. *Journal of Econometrics*, **16**, 121{130.

Granger, C. W. J., and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics,* **2**, 111{120.

Harvery, A., (1993). “Time Series Models”, Harvester Wheatsheaf, 2nd edition

Kendall, M. and Ord, K., (1990). “Time Series”, Edward Arnold, 3rd edition.

Johansen, S. (1988) Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics and Control*, 12, 2, 231-254.

Johansen, S. (1991). “Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models”, *Econometrica*, 59, 1551-1580.

Johansen, S. (1992). “Testing Weak Exogeneity and the Order of Cointegration in UK Money Demand Data”, *Journal of Policy Modelling*, 14, 313-334.

Johansen, S. (1995). “Likelihood-Based Inference in Cointegrated Vector Autoregressive Models”, Oxford University Press, Oxford.

Johansen, S. and Juselius, K. (1990). “Maximum Likelihood Estimation and Inference on Cointegration-With Application to the Demand for Money”, O*xford Bulletin of Economics and Statistics*, 52, 169-210.

Ljung, G.M. and G.E.P. Box, (1978). “On a Measure of Lack of Fit in Time Series Models”, *Biometrica*, 65, 297-303.

Osterwald-Lenum, M. (1992). “A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics”, *Oxford Bulletin of Economics and Statistics*, 54, 461-472.

Ostrom, C., (1990). “Time Series Analysis: Regression Techniques”, Sage, 2nd edition.

Pennings, P., Keman, H., and Kleinnijenhuis, J., (1999). “Doing Research in Political Science: An Introduction to Comparative Methods and Statistics”, Sage.

Schwarz, G., (1978). “Estimating the Dimension of a Model”, *Annals of Statistics*, 6, 461-464.

Spanos, A., (1999). “Probability Theory an Statistical Inference: Econometric Modeling with Observational Data”, Cambridge University Press.

Stock, J. H. (1987). Asymptotic properties of least squares estimators of cointegrating vectors.*Econometrica*, **55**, 1035{1056.

White, H. (1980). “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity:, *Econometrica*, 48(4), 817-838.

Zellner, A., (1979). “Causality and Econometrics”, In Karl Brunner and Allan Meltzer, eds, Three aspects of policy and policy making: Knowledge, data and institutions, Carnegie-Rochester Conference Series on Public Policy, vol. 10, 9-53, North-Holland, New York.