EDO UNIVERSITY, IYAMHO
EDO STATE, NIGERIA
DEPARTMENT OF ECONOMICS

ECO 412: APPLIED ECONOMETRICS

INSTRUCTOR: David Umoru, Email: david.umoru@edouniversity.edu.ng
Alternative email: david.umoru@yahoo.com

LECTURES: Tuesday, 10am – 12pm, Lecture Classroom 5 (LC5),
Mobile Line: (+234)8033888414

OFFICE HOURS: Wednesdays, 1pm to 3.30pm, Office: 1st Floor, MH Administrative Block

GENERAL OVERVIEW OF LECTURE: This course is designed primarily for students at the postgraduate level. The aim of the course is to build upon the students' existing knowledge of econometrics and essentially, to help the students develop a practical knowledge of econometrics and its applications to real-world economic data.

PREREQUISITES: Students should be familiar with the concepts of introductory econometrics (ECO 313 & ECO 323) with specific knowledge of econometrics methods, Ordinary Least Squares(OLS) method, single equation modelling, hypothesis testing, random variables and their distributions, limit theorems, finding moment functions, and working with characteristic functions and also be able to solve matric algebra, etc.

LEARNING OUTCOMES:
The students at the end of this course should be able to:

i. Demonstrate a sound understanding of the econometric modeling and estimation as well as exhibit evidence-based policy making
ii. Prove the Gauss Markov theorem and detect, and find solutions to econometric problems in the context of estimated regression models
iii. Enter policy dialogues at the national and international levels, and engage in related policy research to provide new solutions to existing problems in a changing environment
iv. Obtain data that are relevant to the stated economic problem and present a model that is suited to deal with the phenomena under study
v. Demonstrate competence in the use of econometric packages (like E-VIEWS) as may be needed to perform the analysis.
vi. Use the econometric model for analysis and prediction. This involves exploring the economic implications of the empirical results.
vii. Carry out good quality applied economic research with confidence

ASSIGNMENTS: Classroom test and an Econometrics term paper will also be given to facilitate learning of the more challenging areas of the course. This will make up the continuous assessment of 30% of the final grade of every student.

GRADING: We will assign 10% of this class grade to homeworks, 10% for the programming projects, 10% for the mid-term test and 70% for the final exam. The Final exam is comprehensive. The grading for this course is a combination of continuous assessment and final examinations. A final examination will be written at the end of the course and this will cover 70%.

REFERENCE TEXTS

The recommended textbooks for this class are as stated:
Title: *Applied Econometrics.*
Authors: Asteriou, D. and Hall, S.G.
Publisher: Palgrave Macmillan, New York, 2nd Edition
Year: 2007

Title: *Econometric Methods*
Author(s): Johnston, J. and J. DiNardo
Publisher: McGraw Hill International Editions, 4th Edition
Year: 2008

Title: *Applied Time Series Modelling and Forecasting*
Author: Harris, R.I.D and Sollis, R.
Publisher: John Wiley & Sons, Inc., 2nd Edition,
Year: 2003

MAIN LECTURE

LECTURE 1: AUTO-CORRELATION

CONTENTS

INTRODUCTION

Violation of the basic assumptions of OLS estimator leads to econometrics problems such as autocorrelation, heteroscedasticity and multicollinearity.

OBJECTIVES

At the end of this lecture, students will be able to understand the meaning of autocorrelation, types, causes, consequences of autocorrelation and correction of the problem of autocorrelation.

AUTO-CORRELATION

Autocorrelation is the serial dependence of the successive values of the stochastic error term. In other words, the stochastic disturbance in the current period depends on its immediate past values. It is indeed, a violation of the assumption of the "absence of autocorrelation" or "serial independence" of the OLS technique for estimating the classical linear regression model. In effect, the value taken on by the stochastic disturbance in one period depends on the value it takes on in the previous period such that:

$$Cov(u_t, u_{t-1}) \neq 0 \ \forall \ t = 1, 2, ..., n$$

Thus, $u_t = f(u_{t-1}) + e_t$

Autocorrelation has significant occurrence in time-series economic data and hence in time series econometrics.

Forms of Autocorrelation

There are different forms of autocorrelation. These include:

(a) First order autocorrelation
(b) Second-order autocorrelation
(c) Third-order autocorrelation
(d) Kth-order autorrelation

The first-order autoregressive scheme is specified as:

$$U_t = \rho U_{t-1} + e_t \quad |\rho| < 1$$

Where ρ is the coefficient of first-order autocorrelation, e_t is the stochastic error term which satisfies the usual OLS assumptions, that is,

$$E(e_t) = 0$$
$$E(e^2) = \sigma_e^2$$
$$E(e_t e_{t-1}) = 0$$

The second-order autoregressive scheme is specified as:

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + e_t$$

The third-order autoregressive scheme is specified as:

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + \rho_3 U_{t-3} + e_t$$

The kth-order autoregressive scheme is specified as:

$$U_t = \rho_1 U_{t-1} + \rho_2 U_{t-2} + ... + \rho_k U_{t-k} + e_t$$

Causes of Auto-correlation

Incidence of Inertia

Inertia is the fluctuations in time series variables such as employment. So, in regressions involving time series data, successive observations are likely to be interdependent.

Non-Stationarity

A time series is stationary if its mean, covariance and variance do not vary with time. This means that a non-stationary time series is a series that changes with time. Non-stationarity causes autocorrelation because the mean, variance and covariance of a non-stationary series are time variant, that is, they change over time

Difference-Transformation

Autocorrelation is easily induced by first-difference transformation of the time series variables. In short, serial correlation most often characterized models regressed on the successive differences of the values of variables. Consider the following level and difference models:

$$Y_t = \alpha + \beta Z_t + u_t$$

$$\Delta Y_t = \beta \Delta Z_t + \Delta \varepsilon_t$$

$$\text{where } \Delta Y_t = Y_t - Y_{t-1}$$

$$\Delta Z_t = Z_t - Z_{t-1}$$

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

Note that the error term in the level equation is not autocorrelated but it can be shown that the error term in the first difference form is autocorrelated.

Specification Bias

Specification bias is in two folds, the

(a) Omitted variable bias and
(b) Incorrect functional or mathematical form bias.

Omitted Variable Bias

Autocorrelation will occur when there is an omission of an important explanatory variable from the regression model. In other words, the omission bias takes place when variables that are germane to the phenomenon being studied are excluded from the model.

Functional Bias

Wrong functional form of a model induces autocorrelation. For example, when a linear mathematical form of a model is specified instead of a quadratic form as in the following equations, autocorrelation is induced.

$$C_t = \alpha_0 + \alpha_1 Q_t + \varepsilon_{1t}$$

$$C_t = \alpha_0 + \alpha_1 Q_t^2 + \varepsilon_{2t}$$

Thus, when a researcher fits an empirical model with a wrong functional form to the available data, such as, fitting a linear model when a non-linear model is actually the most appropriate, autocorrelation is provoked automatically.

Lag Structures in Economics

The use of lags in economics is a major cause of autocorrelation. Thus, autoregressive models are mostly auto correlated because the error term is a reflection of the systematic pattern due to the influence of the lagged term in the model. Consider the usual autoregressive consumption function:

$$C_t = \alpha_0 + \alpha_1 Y^d + \alpha_2 C_{t-1} + \upsilon_t$$

Is this not a reflection of $u_t = f(u_{t-1}) + e_t$ .The underlying assumption is that consumers' expenditures in the current period depend for most times on their previous level of expenditures. As it were, economic agents most often do not significantly change their consumption habits. Thus, due to the presence of C_{t-1}, that is, the effect of C_{t-1} on C_t , autocorrelation is provoked. In

this very instance, the error term will reflect the pattern of serial dependence given the similar pattern of consumption.

Data Mining

Data mining is the interpolation and extrapolation of data. It is a smoothening process of data manipulation which dampens the fluctuation in the original data set. Such data manipulation is known as data massaging which leads to systematic pattern in the disturbances thereby initiating serial correlation. A good example of data mining is averaging of quarterly data.

Consequences of Autocorrelation on the OLS Estimator

(a) The unbiasedness property of the OLS estimator is not affected even in the presence of auto correlated errors. Thus, the OLS estimator, β, is still unbiased and highly consistent.
(b) The efficiency property of the OLS estimator is destroyed. The variances and the standard errors of the OLS estimator are upwardly biased.
(c) Consequently, the t-values of the coefficient estimates are distorted and rendered unreliable.
(d) Erroneous statistical decisions are made. This is because the use of the conventional t and F statistical tests to evaluate the statistical significance of the estimated coefficients of a model are no longer valid.
(e) Confidence intervals of the OLS estimates are flawed and at best outsized.
(f) Also, the residual variance $(\hat{S}^2)$ will under estimate the true parameter σ^2. As a result, the coefficient of determination, R^2 will be overestimated.
(g) Autocorrelation can be subjugated for predictions. The reason is that an autocorrelated time series is probabilistically predictable because future values depend on current and past values.

Statistical Tests for Autocorrelation

Graphic Method

The graphical method entails plotting the values of the error terms against time. Three graphical tools for assessing the autocorrelation of a time series are the time series plot, the lagged scatter plot, and the autocorrelation function.

Durbin-Watson d Test Statistic

Due to Durbin and Watson (1970),in their article titled, "Testing for Serial Correlation in Least Squares Regression", the d test for autocorrelation is defined as the ratio of the sum of squared differences in successive residuals to the sum of squared residuals.

$$d = \frac{\sum_{t=2}^{n}(u_t - u_{t-1})^2}{\sum_{t=1}^{n} u_t^2}$$

Given that $-1 \leq \rho \leq 1$, the values of the D-W statistic lies between 0 and 4, that is, $0 \leq d \leq 4$. Thus,

When $d = 2$,

$$\rho = 1 - \frac{1}{2}(2)$$

$$= 0$$

signifying absence of autocorrelation

When $d = 0$,

$$\rho = 1 - \frac{1}{2}(0)$$

$$= 1$$

signifying perfect positive autocorrelation

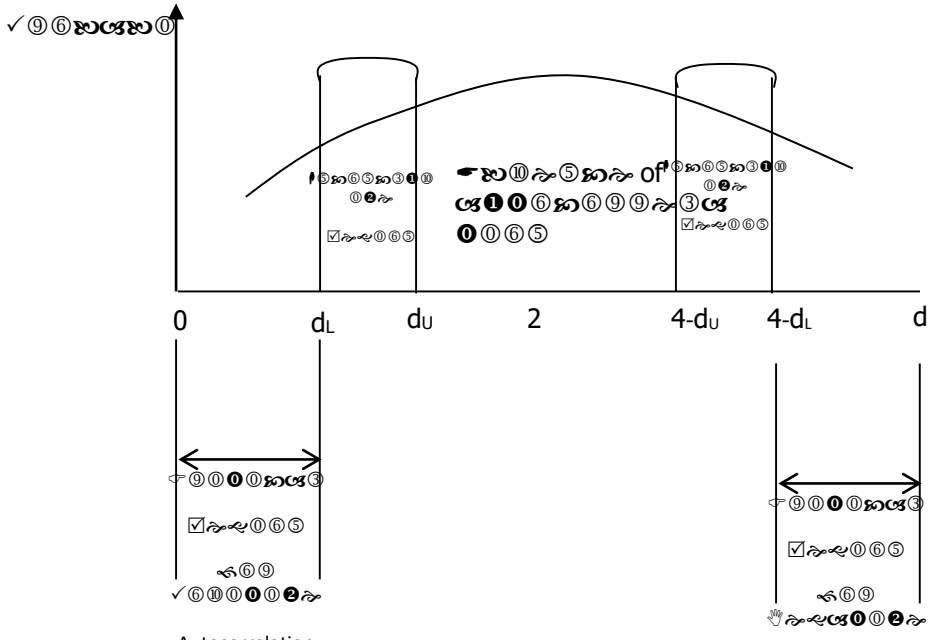
When $d = 4$,

$$\rho = 1 - \frac{1}{2}(4)$$

$$= -1$$

signifying perfect negative autocorrelation

Graphical Representation of Durbin-Watson d Test Statistic



The Durbin-Watson empirical value d_w^* must be compared with the critical d values denoted by d_U , the upper limit for the significance level of d_w^* and d_L , the lower limit for the significance level of d_w^*. Given these conditions; the d_w^* test is applied on the basis of the following hypothesis:

$$H_0 : \hat{\vartheta}_t = 0 \ \ (non-autocorrelated\ \ residuals)$$

$$H_1 : \hat{\vartheta}_t \neq 0 \ \ (autocorrelated\ \ residuals)$$

The application of the d_w^* statistic depends on the following conditions

(a) An intercept term must be included in the regression model. This is because if the model is fitted without an intercept, residual sum of squares [RSS] even if computed might not sum to zero thereby surmounting the notion of zero sum of squares. Consequently,R^2 might be negative [see Johnston and Dinardo (1997)]
(b) The regressors of the model must be truly exogenous i.e. non stochastic,
(c) The error process must be generated by the fist-order autoregressive scheme and not by higher orders. As it were, the d_w^* is only applicable for testing the presence or otherwise of first-order autoregressive scheme,
(d) The error term must be normally distributed with zero mean and constant variance,
(e) No missing data point in the entire time series or series of observations
(f) The regression model must not include lagged endogenous variable as explanatory variable. Thus, if the model to be fitted is of the money demand type.

$$M_t^D = \varphi_0 + \varphi_1 Y_t + \varphi_2 M_{t-1}^D + \upsilon_t$$

The d_w^* statistic becomes inappropriate and often erroneously settling for the absence of auto-correlation even if there is. The only way out of this empirical problem is to utilize the Durbin-h statistic.

Durbin h Test Statistic

The inapplicability of the d statistic in models with lagged endogenous variables as regressors necessitated the Durbin h-statistic [see Durbin (1974)]. The h-statistic is defined as:

$$h = \hat{\vartheta} \sqrt{\frac{N}{1 - n[Var(\hat{\varphi}_2)]}}$$

Where N is the sample size, $Var(\hat{\varphi}_2)$ is the variance of the estimated coefficient of the lagged endogenous regressor, that is, the variance of the coefficient of M_{t-1}^D in the money demand function and $\hat{\vartheta}$ is the estimated first-order autocorrelated.

Corrective Measures of Autocorrelation

The remedial measures of the autocorrelation problem exist for both when the autocorrelation coefficient is known and when it is unknown.

Generalized Least Squares [GLS] Estimator

The generalized least squares estimator is basically the application of OLS to the transformed model that fulfilled the classical OLS assumptions. However, its application is facilitated when the first-order autocorrelation coefficient is known. Given the underlying model whose error term follows the $AR[1]$ scheme:

$$Y_t = \beta_0 + X_t\beta_1 + \upsilon_t \quad (5.9)$$

$$where \upsilon_t = \vartheta\upsilon_{t-1} + v_t$$

$$v_t \sim [0, \sigma^2_{v_t}], [-1 < \vartheta < 1]$$

To correct the model for autocorrelation, we lag (5.1) by one-period so that we have:

$$Y_{t-1} = \beta_0 + X_{t-1}\beta_1 + \upsilon_{t-1}$$

Multiply the resulting equation by the autocorrelation coefficient to obtain

$$\hat{\vartheta}Y_{t-1} = \hat{\vartheta}\beta_0 + X_{t-1}\hat{\vartheta}\beta_1 + \hat{\vartheta}\upsilon_{t-1} \quad (5.10)$$

Subtract equation (5.10) from equation (5.9) above to obtain:

$$Y_t - \hat{\vartheta}Y_{t-1} = (1-\hat{\vartheta})\beta_0 + (X_t - \hat{\vartheta}X_{t-1})\beta_1 + \upsilon_t - \hat{\vartheta}\upsilon_{t-1} \quad (5.11)$$

$$Y_t^* = \beta_0^* + \beta_1 X_t^* + v_t^*$$

$$where\ Y_t^* = Y_t - \hat{\vartheta}Y_{t-1}$$

$$\beta_0^* = (1-\hat{\vartheta})\beta_0$$

$$X_t^* = (X_t - \hat{\vartheta}X_{t-1})$$

$$v_t^* = \upsilon_t - \hat{\vartheta}\upsilon_{t-1}$$

Cochrane-Orcutt Iterative Method

- The Cochrane-Orcutt iterative method of transformation entails a process of convergence to the autocorrelation coefficient. Accordingly, the applicability of the Cochrane-Orcutt transformation requires the model to be estimated. Thus haven applied the OLS technique; the estimated model can be given as:

$$\hat{y}_t = \hat{\varphi}_0 + X_t\hat{\varphi}_1 + \hat{\upsilon}_t$$

To correct the model for autocorrelation, compute the OLS residual series as follows:

$$\hat{\upsilon}_t = Y_t - \hat{Y}_t$$

$$\hat{\upsilon}_t = Y_t - \hat{\varphi}_0 - X_t\hat{\varphi}_1$$

Using the residual series, run the following auxiliary regression or simply compute the autocorrelation coefficient from the variance-covariance ratio as follows:

$$\hat{\upsilon}_t = \hat{\vartheta}\hat{\upsilon}_{t-1} + v_t$$

$$\hat{\vartheta} = \frac{\sum \hat{\upsilon}_t \hat{\upsilon}_{t-1}}{\sum \hat{\upsilon}^2_{t-1}}$$

Using the estimated $\hat{\vartheta}$, estimate the generalized first-difference model. In other words, use $\hat{\vartheta}$ to transformed model and use OLS to further estimate the model.

$$y_t^* = \beta_0^* + \beta_1 X_t^* + v_t^*$$

Compute the second round OLS residual series, $\hat{\hat{\upsilon}}_t$

$$\hat{\hat{\upsilon}}_t = Y_t - \hat{\hat{Y}}_t$$

$$\hat{\hat{\upsilon}}_t = Y_t - \hat{\hat{\varphi}}_0 - X_t \hat{\hat{\varphi}}_1$$

Compute the second-round autocorrelation coefficient from the variance-covariance ratio as follows:

$$\hat{\hat{\upsilon}}_t = \hat{\hat{\vartheta}} \hat{\hat{\upsilon}}_{t-1} + v_t$$

$$\hat{\hat{\vartheta}} = \frac{\sum \hat{\hat{\upsilon}}_t \hat{\hat{\upsilon}}_{t-1}}{\sum \hat{\hat{\upsilon}}_{t-1}^{\hat{\hat{2}}}}$$

Use $\hat{\hat{\vartheta}}$ to transform the original model and apply OLS estimator to the model:

$$y_t^{**} = \beta_0^{**} + \beta_1 X_t^{**} + v_t^{**}$$

Compute the third-round OLS residual series, $\hat{\hat{\upsilon}}_t$

$$\hat{\hat{\hat{\upsilon}}}_t = Y_t - \hat{\hat{\hat{Y}}}_t$$

$$\hat{\hat{\hat{\upsilon}}}_t = Y_t - \hat{\hat{\hat{\varphi}}}_0 - X_t \hat{\hat{\hat{\varphi}}}_1$$

Compute the third-round autocorrelation coefficient as follows:

$$\hat{\hat{\hat{\vartheta}}} = \frac{\sum \hat{\hat{\hat{\upsilon}}}_t \hat{\hat{\hat{\upsilon}}}_{t-1}}{\sum \hat{\hat{\hat{\upsilon}}}_{t-1}^{\hat{\hat{\hat{2}}}}}$$

Use $\hat{\hat{\hat{\vartheta}}}$ to transform the original model and apply OLS estimator to the model:

$$y_t^{***} = \beta_0^{***} + \beta_1 X_t^{***} + v_t^{***}$$

The iterative process continues until the autocorrelation effect is eliminated from the estimated set of regression results.

Numerical Examples

Example

These are regression results using OLS for 21 observations with standard errors in parentheses:

$$Y_t = 1.3 + 0.97 Y_{t-1} + 2.31 X_t$$

$$(0.3) \quad (0.18) \quad (0.04)$$

$$D-W = 1.21$$

Test for the presence of autocorrelation in the disturbances [Greene (2003): 281, (Exercise 3)]

Solution

Given that the D-W statistic cannot be used to test for autocorrelation in dynamic models, we resolve to using the Durbin h-statistic whose test statistic is given as:

$$h = \hat{\vartheta} \sqrt{\frac{N}{1 - n[Var(\hat{\varphi}_2)]}}$$

The autocorrelation coefficient can be estimated from D-W the statistic

$$d_w^* = 2(1-\hat{\vartheta})$$
$$= 2-2\hat{\vartheta}$$
$$2\hat{\vartheta} = 2-d_w^*$$
$$\hat{\vartheta} = \frac{2-d_w^*}{2}$$
$$\hat{\vartheta} = 1-\frac{d_w^*}{2}$$
$$\hat{\vartheta} = 1-\frac{1.21}{2}$$
$$= 0.395$$

Applying the Durbib h-statistic, we have as follows:

$$h = 0.395\sqrt{\frac{21}{1-21[0.0016]}}$$
$$= 0.395\sqrt{\frac{21}{1-0.0336}}$$
$$= 0.395\sqrt{\frac{21}{0.9664}}$$
$$= 0.395\sqrt{21.73}$$
$$= 0.395[4.66]$$
$$= 1.84$$

ASSIGNMENT

Consider that in a regression analysis that relates the consumption level of electrical appliance to the income level of consumers, the econometric results that follow below were obtained for the sample period, 1980-2010.

$$C_t = 83.55-0.13I_t$$
$$t-ratios\ (0.009)\ (0.005),\ R^2 = 0.99, R^2\ (Adjusted) = 0.89$$
$$\sum \upsilon_t^2 = 1.28624,\ \sum (\upsilon_t - \upsilon_{t-1})^2 = 1.02683$$

(a) Test for autocorrelation
(b) Comment succinctly on the presence or otherwise of autocorrelation
(c) Estimate ϑ, the autocorrelation coefficient

CONCLUSION

Autocorrelation is an econometric problem that is mostly found in time series data

LECTURE 2: MULTICOLLINEARITY

CONTENTS

INTRODUCTION

Violation of the full rank assumption of OLS estimator leads to multicollinearity problem.

OBJECTIVES

At the end of this lecture, students will be able to understand the meaning, types, causes, and consequences of Multicollinearity as well as the corrective measures of the problem of Multicollinearity.

MULTICOLLINEARITY

- Multicollinearity refers to the linear relationship between the explanatory variables of a multiple regression model (MRM). In other words, multicollinearity occurs when the regressors of an MRM $X_1, X_2,...,X_k$ are highly correlated with each other.
- By econometric intuition, linear relationship ought not to exist between the explanatory variables in a MRM. This is actually a desecration of the full rank assumption of the classical linear regression model.

Extreme Cases of Multicollinearity

There are two extreme scenarios of multicollinearity. These include perfect multicollinearity and orthogonal multicollinearity.

Perfect Multicollinearity

- Perfect multicollienarity means that the relationship between the explanatory variables is exact in the sense that the correlation coefficient between the explanatory variables is unity, that is, $r_{x_i} r_{x_j} = 1$. Consider the log-linear regression of LnY_{1t} on LnX_{1t} and LnX_{2t}.

$$LnY_{1t} = \phi_0 + \phi_1 LnX_{1t} + \phi_2 LnX_{2t} + \upsilon_t$$

$$where \ \ r_{x_i} r_{x_j} = 1$$

ϕ_1 is suppose to give the rate of charge in the mean value of Y_{1t} as X_{1t} changes by a percentage point holding X_{2t} fixed. Unfortunately enough, the perfect collinearity between the regressors X_{1t} and X_{2t} means that X_{2t} cannot be kept constant while X_{1t} changes and vice-versa.

Orthogonal Multicollinearity

- An orthogonality means that explanatory variables have no relationship. In effect, the correlation coefficient between the explanatory variables is zero. Thus, $r_{x_i} r_{x_j} = 0$. In particular, the explanatory variables in the multiple regression models are not correlated in any form. Thus, orthogonal variables are the variables whose covariance is zero.

$$LnY_{1t} = \phi_0 + \phi_1 LnX_{1t} + \phi_2 LnX_{2t} + \upsilon_t$$
$$where \ \ r_{x_i} r_{x_j} = 0$$

Imperfect Multicollinearity

In the practice of applied econometrics, neither of the two extremes of orthogonality and perfect multicollinearity exist. The multicollinearity problem that exists in practice lies in between the two extremes and it is called "imperfect multicollinearity" which means high but less than perfect multicollinearity. In this case, the correlation coefficient lies between zero and unity, that is, $0 < r_{x_i} r_{x_j} < 1$.

$$LnY_{1t} = \phi_0 + \phi_1 LnX_{1t} + \phi_2 LnX_{2t} + \upsilon_t$$
$$where \ \ 0 < r_{x_i} r_{x_j} < 1$$

Causes of Multicollinearity

Distributed Lag Models

Multicollinearity can be caused by the use of lagged variables in a multiple regression model. This is because it is innate for successive values of a particular variable to be highly intercorrellated. Consider the following system of equations:

$$LnM_t^D = \mu_0 + \mu_1 Y_t + \mu_2 Y_{t-1} + \mu_3 r_t + LnM_t^D + \upsilon_{1t}$$
$$LnC_t = \varepsilon_0 + \varepsilon_1 LnY_t + \varepsilon_2 LnY_{t-1} + \varepsilon_3 LnC_{t-1} + \upsilon_{3t}$$
$$LnI_t = \lambda_0 + \lambda_1 LnY_{1t} + \lambda_2 LnY_{2t-1} + \lambda_3 LnI_{1t-1} + \upsilon_{2t}$$

The inclusion of past and present levels of income in the money demand, consumption and investment equations respectively can induce the problem of multicollinearity as these variables are certainly going to be highly correlated.

Over-determined Models

An over-determined regression model is the model whose explanatory variables are more than the number of observations. In this type of regression, multicollinearity problem is severe.

Consequences of Perfect Multicollinearity

- **Indeterminacy of the OLS Estimator**:
  Coefficient estimates of the OLS estimator are indeterminate

Empirically, with perfect multicollinearity such that $[r_{x_i x_j} = 1]$, it implies that the true relationship between the explanatory variables $x_1 \ and \ x_2$ is exact and as such $x_1 = x_2$.The indeterminacy of the OLS estimator arises because the data matrix of the explanatory variables $(X'X)$ in the OLS estimator cannot be inverted. Consequently, OLS estimator $[(X'X)^{-1}X'Y]$ breaks down.

This shows that in the presence of perfect multicollinearity, the variances and hence the standard errors of the OLS estimators are infinite and consequently indeterminate.

Consequences of Imperfect Multicollinearity

- Accordingly, as collinearity increases between any two regressors, the variances and hence standard errors of the OLS estimators evenly increases. Thus, the VIF measures the extent by which the variance of OLS estimator is inflated due to the presence of multcollinearity.

Empirically, with imperfect multicollinearity $[0 < r_{x_i x_j} < 1]$, the following consequences are on the OLS estimator:

(a) Large Variances and standard errors: The variances and standard errors of the OLS estimator are unduly large, and so the t-ratios are rendered statistically insignificant which leads to a type II error of accepting an incorrect null hypothesis null hypothesis instead of rejecting it.
(b) Erroneous Statistical Inferences: Misleading statistical inferences are drawn from the test of hypothesis
(c) Wide confidence intervals: Confidence intervals are unduly outsized
(d) Unstable coefficients: The OLS parameter estimates become highly unstable. Such instability of coefficients could cause a dramatic change in the coefficient sign as the degree of multicollinearity increases.

Statistical Tests for Multicollinearity

There are several tests for detecting the problem of multicollinearity. As it is, we have the formal and informal tests for multicollinearity.
The informal tests for multicollinearity include:

(a) High R^2: Even in the presence of insignificant t-values, the overall measure of goodness-of-fit, R^2 could very high.
(b) Low t-ratios
(c) Wrong coefficient sign
(d) R^2 delete

The formal statistical tests include the:

(a) Variance Inflation Factor [VIF]
(b) Farrar-Glauber test

Variance Inflation Factor [VIF]

- The variance inflation factor (VIF) quantifies the proportion by which the variance of the OLS estimator is inflated. In other words, the VIF quantifies the variance of the OLS estimator due to multicollinearity.

- Computationally, it is defined as the reciprocal of the tolerance index. Applied econometricians most often desire lower values of VIF, as higher values of VIF are known to adversely affect the regression results. For example, a VIF of 8 implies that the standard errors of the OLS estimator are larger by a factor of 8 than would otherwise be the case, if there were no multicollinearity between the regressors in the multiple regression analysis.

$$VIF(\hat{\beta}) = \frac{1}{1-r_{ij}^2}$$

$$Tolerance(\hat{\beta}) = \frac{1}{VIF}$$
$$= 1 - r_{ij}^2$$

Where $r_{x_i x_j}$ is the simple correlation coefficient between any pairs of regressors say, $[x_1, x_2]$ defined by the Karl Pearson's product moment equation below:

$$r = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \sum x_2^2}}$$

or

$$r_{x_i x_j} = \frac{N\left[\sum x_i x_j - \sum (x_i x_j)\right]}{\sqrt{\left[N\sum x_i^2 - (\sum x_i)^2\right]\left[N\sum x_j^2 - (\sum x_j)^2\right]}}$$

Thus, as the correlation coefficient, $r_{x_i x_j}$ measures the degree of collinearity between the regressors, the VIF quantifies the speed at which variances and covariances increases. For this reason, as the collinearity between the regressors increases and tends to unity i.e the case of perfect multicollinearity, the VIF approaches infinity (∞). Also, if the collinearity between the regressors, $[x_1, x_2]$ is zero, the variance- inflating factor will be equal to unity.

Corrective Measures for Multicollinearity

The remedial measures to be adopted if multicollinearity exists in a model depends on the following factors, severity of the multicollinearity problem, availability of data, the importance of the collinear regressors, the purpose of estimation etc. In any case, the remodel measures are discussed as follows:

Christ's Correction

Christ (1966:389) suggested that multicollinearity can be corrected by increasing the sample size. This entails bringing into the sample, more data points. However, the remedial measures are only valid if error of measurement in the explanatory variables is the cause of multicollinearity.

Dropping Variables

Another measure of resolving severe multicollinearity problem is to drop the variable that is highly collinear with the others. However, dropping one of the collinear variables from a model may put the econometrician at the verge of committing a specification error. This is because in line with economic theory, if interest rate and income are the key determinants of money demand and as such must be included in the money demand function, dropping either of interest rate or income would mean committing a specification error. Under this scenario, the cure could be worse than the disease.

Transformation

Transformation of variables to is generally regularly useful as a way out of the multicollinearity problem. Thus, instead of running the regression in the original variables themselves, what becomes desire is to run multiple regression model on the transformed data matrix on the variables under specification.

Pooling Data

A combination of both cross section and time series data can help to resolve the problem of multicollinearity. Pooling has the following limitations:

(a) Pooled observations do suffer from the serial correlation problem,
(b) Pooling observations in different time periods do erroneously assume stability in the casual relationship across time rather than variation across sub-periods
(c) Pooling cannot distinguish between variations across time and across sectionals. For example, an inclusion of a dummy variable takes into care the different slopes or baseline values rather than the different slopes with various periods.

ASSIGNMENT

Test for Multicollinearity I the following data

Y	20.6	12.3	19.8	15.7	16.2
X	2.6	8.2	2.9	6.4	3.8

CONCLUSION

It has been asserted that the problem of multicollinearity is harmless if the objective for estimating a model is to forecast the values of the endogenous variable only [Christ (1966): 390, Greene (2003)]. In this case, the values of the collinear variables can be included in the model while ignoring the consequence. This can only be successful provided the econometrician is certain that the correlation pattern that exists between the explanatory variables will remain the same throughout the prediction period.

LECTURE 3: HETEROSKEDATSICITY

CONTENTS

INTRODUCTION

Violation of the homoscedasticity assumption of OLS estimator leads to heteroscedasticity problem.

OBJECTIVES

At the end of this lecture, students will be able to understand the meaning, types, causes, and consequences of Heteroskedasticity as well as the corrective measures of the problem of Heteroskedasticity.

HETEROSKEDASTICITY

- Heteroskedasticity means that the variance of the stochastic disturbance term *(u_i)* is not constant (the same) for all values of the explanatory variables. This is because the variance of the stochastic disturbance is no longer given by a finite constant and thus would tend to change with an increasing range of values of the explanatory variables thereby making it impossible to be taken out of summation.

Thus, *the homoskedastic variance – covariance matrix is given by*:

$$E(u_i, u_j^{'}) = \begin{bmatrix} \sigma_u^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_u^2 \end{bmatrix}$$

$$= \sigma_u^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \sigma_u^2 I$$

However, given the presence of heteroskedasticity,

$$Var(u_i) = E(u_i^2) \neq \sigma_u^2 = \sigma_{u_i}^2$$

In effect, the heteroskedastic variance-covariance matrix is given by:

$$E(u_i, u_j^{'}) = \begin{bmatrix} \sigma_{u_1}^2 & 0 & 0 \\ 0 & \sigma_{u_2}^2 & 0 \\ 0 & 0 & \sigma_{u_3}^2 \end{bmatrix}$$

$$= \sigma_{u_i}^2 I$$

The subscript i denote the fact that the variances of each stochastic disturbance are all different.

The occurrence of heteroskedasticity is found in both time series and cross section data but more often encountered and severe with cross section data. This is because the assumption of constant variance over the heterogeneous units may be rather unrealistic.

Causes of Heteroskedasticity

(a) Outliers Problem

Outlying observations are the root cause of heteroskedasticity. An outlier is an observation that is either excessively small or excessively large in relation to other observations in the sample. The table below illustrates a scenario of an outlier

Y	X
328	820
22.6	680
1.2	18200

In other words, the outlying observation exhibits huge difference from others in the sample. In effect, the population of the outlying observation is different from the population of the other sample observations

(b) Omitted Variable Bias

The omission of key explanatory variables from a regression model causes heteroskedasticity. For example, consider the following model of consumption expenditure:

$$C_t = \beta_0 + \beta_1 r_t + \beta_2 I_t + \beta_3 T_t + \beta_4 C_{t-1} + \beta_5 F_t + \beta_6 W_t + \beta_7 Y_t^d + u_t$$

where C is consumption exp*enditutes, r is interest rate, I is inflation rate*

T is taste of the consumer, F is fashion, W is weather condition

Y^d *is disposable income*

In line with economic theory, income is the most crucial determinant of consumption expenditures. Thus, if income is omitted from the model, the omitted variable bias would have been induced. This in turn attracts heteroscedasticity.

(c) Error Learning Factors/Models

As people learn everyday from their past mistakes, their errors of behavior become smaller and smaller over time and as such cannot be relatively constant.

(d) Wrong Functional Form

Specification error or incorrect functional form of a regression model causes heteroskedasticity. This occurs when a model is being regressed with a pool of "level" and "log" variables at the same time

$$LnC_t = \beta_0 + \beta_1 r_t + \beta_2 LnY_t^d + u_t$$

where C is consumption exp*enditutes, r is interest rate,* Y^d *is disposable income*

(e) Erroneous data transformation

Incorrect data transformation is another cause of heteroskedasticity. It occurs when a regression model is being regressed with a pool of ratio and first-difference set f data at the same time

(f) Skewness

Skewness in the distribution of the explanatory variables causes heteroskedasticity. For example, the distribution of income and wealth is most often unequal but skewed in such a way that the bulk of income and wealth is owed by a few individuals at the top. Thus, while the spending behavior or the expenditure profile of a cross-section of families with low income may exhibit similar pattern in addition to being relatively stable, such expenditure profile of the cross-section of the rich families with high income could be different and highly volatile

Consequences of Heteroskedasticity

To investigate the effects of heteroskedasticity on the OLS estimator, its variance and standard errors, it becomes desirous that we revert to matrix specification of the classical linear regression model [CLRM].

$$Y_i = X_i'\beta + \upsilon_i$$

$$where\ Var[\upsilon_i / X] \neq \sigma^2 I$$

$$\forall\ i = 1,2,...,N$$

Given that $Var(u_i) \neq \sigma^2 I$, the general form of the heteroskedastic variance-covariance matrix can then be described as:

$$Var[\upsilon_i / X] = \sigma^2 \Omega$$

Where Ω is a positive definite matrix such that $\sigma^2\Omega = Diag\left[\sigma_1^2, \sigma_2^2, \sigma_3^2, ..., \sigma_N^2\right]$. Thus, the variance-covariance [V-C] matrix of the OLS estimator β will be given by:

$$\begin{aligned} Var - Cov(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E\left[[(X'X)^{-1}X'\mu][(X'X)^{-1}X'\mu]'\right] \\ &= [(X'X)^{-1}X'\mu][(X'X)^{-1}X\mu'] \\ &= [(X'X)^{-1}X']E(\mu\mu')[(X'X)^{-1}X] \\ &= (X'X)^{-1}[X'X]\sigma^2\Omega(X'X)^{-1} \end{aligned}$$

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

This portrays the fact that with heteroskedasticity,

(a) The variances and standard errors of the OLS estimator are no longer efficient, not even asymptotically. In other words, not even in large samples. Thus, the minimum variance property of the OLS estimator is lost.
(b) The variances and standard errors are overestimated by the OLS estimator thereby getting the standard errors of the estimated coefficients distorted by being over boosted.
(c) Statistical test of significance are rendered invalid. As it were, the validity of the conventional formulae for t and f test statistics becomes impaired
(d) Statistical inferences are erroneous. Consequently, with heteroskedasticity, there is a higher risk of committing type 1 error which entails rejecting a correct null hypothesis instead of accepting it and also there is the likelihood of committing a type II error which has to do with the acceptance of an incorrect null hypothesis rather than rejecting it.
(e) Confidence interval of the estimated coefficients becomes inordinately wide. In other words, confidence intervals are overly outsized
(f) In general, the hypothesis-testing procedures on the basis of the OLS estimates are contaminated and spurious.
(g) Heteroskedasticity does not destroy the unbiasedness property of the OLS estimator. As a matter of empirical fact, $E[\upsilon_i / X] = 0$ still holds. Consequently, the OLS estimator $\hat{\beta}$ remains unbiased.

Statistical Tests for Heteroskedasticity

There are numerous tests for detecting the presence or otherwise of the problem of heteroskedasticity. These include the informal and formal techniques.

The formal methods for detecting the presence or otherwise of heteroskedatsticity are methods that suggest that the econometrician has some aprori information set about the true pattern of heteroskedasticity. In effect, the econometrician's task is to conduct the regression analysis on the assumed pattern of heteroskedasticity.

Glejser Test

Due to Glejser (1969), the Glejser test is a formal test for heteroskedasticity that regresses the absolute values of the estimated residuals $\hat{\upsilon}$ on various powers of the explanatory variable of the model. The test is based on the following hypothesis.

$$H_0 : \upsilon_i \text{ are homoskedastic}$$

$$H_1 : \upsilon_i \text{ are heteroskedastic}$$

Spearman's Rank Correlation Test [SRCT]

The SRCT statistic is a detective measure of heterokedasticity that ranks the values of the explanatory variable and the estimated regression residuals either in ascending or in descending order of magnitude without regard for the signs of the residuals. Given the following regression model to be estimated:

$$Y = \beta_0 + \beta_1 X + u$$

What follows next is to:

(a) Fit the regression to the data on Y_i and X_i and
(b) Generate the residuals u.

(c) Disregarding the sign of the estimated residual, that is taking only the absolute value of the estimated residual, we rank $|u|$ and X_i either in a descending or in an ascending order.

(d) Next, is the computation of the Spearman rank correlation coefficient (SRCC). The test statistic is given as:

$$r_{X,\upsilon}^{k} = 1 - 6\left[\frac{\sum d_i^2}{n\left(n^2-1\right)}\right]$$

Where d is the difference between the values of corresponding pairs of X and υ, n is the number of observations in the sample, that is, the number of individual units being ranked. The test is based on the following hypothesis.

$$H_0 : \upsilon_i \text{ are homoskedastic}$$

$$H_1 : \upsilon_i \text{ are heteroskedastic}$$

Decision rule:

If $r_{X,\upsilon}^{k}$ is low, accept H_0, the error valances is homoskedastic

If $r_{X,\upsilon}^{k}$ is high, accept H_1, the error variance is heteroskdastic

Alternatively,

$$\text{If } r_{X,\upsilon}^{k} < \left[t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right], \text{ accept } H_0 \text{ and reject } H_1$$

$$\text{If } r_{X,\upsilon}^{k} > \left[t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right], \text{ accept } H_1 \text{ and reject } H_0$$

$$\text{If } r_{X,\upsilon}^{k} < \left[z = \frac{1.96}{\sqrt{n-2}}\right], \text{ accept } H_0 \text{ and reject } H_1$$

$$\text{If } r_{X,\upsilon}^{k} > \left[z = \frac{1.96}{\sqrt{n-2}}\right], \text{ accept } H_1 \text{ and reject } H_0$$

Where r is the computed SRCT statistic, $1.96/\sqrt{n-2}$ is the standard normal z critical value. Here, the standard error of the SRCT statistic is assumed to obey the standard normal distribution in this regard. The critical t-value can be obtained as $t_{\alpha\%,n-2}$ and α is the level of significance and $n-2$ is the degree of freedom.

Goldfeld-Quandt Test

This is a formal test for heteroskedasticity due to Goldfeld and Quandt (1972). By definition, the G-Q test is a fundamental *F-test statistic* that entails the ordering of the set of observation in accordance to the magnitude of the values of the explanatory variable and thereafter divides the set of observations into three parts such that the first and third halves are equal. The middle half which is made up of one-quarter of the total number of observations in the sample $(n/4)$ is excluded from the test.

Thus, if $n = 16$, it implies that the 4 middle observations in the ordered set must be omitted or deleted and the balance 12 observations divided into two equal halves of 6 observations each. Having ascertained this division, separate error variances or residual variances are estimated

from the OLS regression of the two equal halves of 6 observations each. Worthy of note is the fact that the ordering of the data set is determined on the basis of an ascending order of the values of the explanatory variable.

In sum, the G-Q method requires the following steps:

Step 1: Ranking the data points on the regressor, X_i starting with the least observation

Step 2: Divide the ordered set of observations into equal groups each of $[n-f]/2$ observations haven omitted the central group of observations $[f]$. The central group of observations to be excluded from the heteroskedasticity test should be about one quarter of the total number of observations $[n]$. For example, if $n = 32$, about $32/4 = 8$ data points must be excluded and the remaining 22 divided into two equal halves of n_1 *and* n_2 observations, where $n_1 = 11$ *and* $n_2 = 11$.

Step 3: Fitting the OLS regression on separate basis to the two groups of 11 observations each, and obtain the residual sums of squares RSS_1, $\sum \hat{\upsilon}_1^2$ and RSS_2 $\sum \hat{\upsilon}_2^2$ with the degree of freedom given by $[n_1 - k \text{ and } n_2 - k]$ respectively, and k is the number of parameters to be estimated.

Step 4: Compute the F- ratio as follows:

The test statistic is given as:

$$F_{statistic} = \frac{\sigma_{2\upsilon}^2}{\sigma_{1\upsilon}^2}$$

$$\sigma_u^2 = \frac{SSE}{n_2 - k}$$

$$\sigma_u^2 = \frac{\sum u^2}{n_2 - k}$$

$$\sigma_{2u}^2 = \frac{\sum (Y - \hat{Y})^2}{n_2 - k}$$

$$\sigma_{2u}^2 = \frac{u'u}{n_2 - k}$$

$$\sigma_{1u}^2 = \frac{SSE}{n_1 - k}$$

$$\sigma_{1u}^2 = \frac{\sum u^2}{n_1 - k}$$

$$\sigma_{1u}^2 = \frac{\sum (Y - \hat{Y})^2}{n_1 - k}$$

$$\sigma_{1u}^2 = \frac{u'u}{n_1 - k}$$

$$F_{statistic} = \frac{\sum (Y - \hat{Y})^2 / n_2 - k}{\sum (Y - \hat{Y})^2 / n_1 - k}$$

Where σ_{υ}^2 is the error variance, otherwise known as the residual variance, $\sum \hat{\upsilon}^2$ is the sum of squared residuals, n is the number of observations in the sample [sample size] and k is the number of estimated regression coefficients.

The test is based on the hypothesis:

$$H_0 : \upsilon_i \text{ are homoskedastic}$$

$$H_1 : \upsilon_i \text{ are heteroskedastic}$$

Decision Rule:

If $S_{1\upsilon}^2 = S_{2\upsilon}^2$, accept H_0, the error variance is homoskedastic

If $S_{1\upsilon}^2 \neq S_{2\upsilon}^2$, accept H_1, the error variance is heteroskedastic

The G-Q test ratio therefore obeys the F-distribution with $n_1 - k$ *and* $n_2 - k$ degress of freedom. Significant F indicates presence of heteroscedasticity and vice versa.

If $F_{calculated} < F_{critical}$, accept H_0 and reject H_1

If $F_{calculated} > F_{critical}$, reject H_0 and accept H_1

The associated critical F-value is obtained as $F_{\alpha\%, n_1 - k, n_2 - k}$. The decision rule is to accept the homoskedasticity assumption if $F_{cal} < F_{crit}$ and reject it if $F_{cal} > F_{crit}$. In which case, if the both the numerator and the denominator are equal we accept H_0.

Corrective Measures for Heteroskedasticity

Transformation based on the Pattern of Heteroskedasticity

In econometric literature, different assumptions about the error term have been made and this warrants the type of data transformation in order to eliminate heteroskedasticity from an empirical model.

Logarithmic Transformation [LT]

Given the unknown nature of heteroskedatsicity, a logarithmic transformation [estimating the original model in log] is also applicable in resolving the problem of heteroskedatsicity. In this case, the transformation of the original model becomes:

$$LnY_i = \varphi_1 + \varphi_2 LnX_i + \upsilon_i$$

Merits of Logarithmic Transformation

In general, logarithmic transformation helps to reduce if not total elimination of the problem of heteroskedatsicity. This is evident in the following facts. Logarithmic transformation:

- Compresses the scales in which the variables are measured, thereby reducing a tenfold difference between two values a two-fold difference. For example, the number 120 is ten times larger than 12, but $Ln(120)$ gives 4.787 which is just about twice as large as $Ln(12)$ which is equal to 2.485
- Yields direct elasticities. For example, the slope coefficient of a logarithmic transformed model measure the elasticity of the dependent variable with respect to the regressor in question. In particular, it measures the percentage change in the dependent variable due to a percentage change in the explanatory variable.

Logarithmic Transformation Problems

- Log transformation is not applicable if some of the observations [data points] for both the dependent and the explanatory variables are zero or negative
- The problem of spurious correlation will be encountered between the ratios of the transformed variables even when the original variables are uncorrelated or random.
- The conventional F and t tests for model robustness are only valid in large samples given that the variances are unknown and are estimated from any of the transformation procedures.
- In the multiple regression model [MRM], model with more than one regressors, it is difficult to ascertain on apriori basic which of the regressors to be chosen for data transformation

Example

Consider the data below:

Test for the presence or otherwise of heteroskedasticity using the spearman rank coefficient test statistics.

Y	2	4	6	3	6
X	20	12	16	14	18

Solution

First: We state the hypothesis:

$$H_0 : e_i \text{ are homoskedatsic}$$

$$H_1 : e_i \text{ are heteroskedatsic}$$

Secondly: We would estimate an SRM which is of the following specification

$$Y = \ell_0 + \ell_1 X + e$$

Y	X	X^2	YX	ŷ	y-ŷ
2	10	100	20	2.2	-0.2
4	12	144	48	3.2	0.8
6	16	256	96	5.2	0.8
3	14	196	42	4.2	-1.2
6	18	324	108	6.2	-0.2
21	**70**	**1,020**	**314**	**21**	**0**

$$\hat{\ell}_0 = \frac{\sum Y \sum X^2 - \sum X \sum YX}{N\sum X^2 - \left[\sum X\right]^2}$$

$$= \frac{2(1020) - 70(314)}{5(1020) - (70)^2}$$

$$\hat{\ell}_0 = -2.8$$

$$\hat{\ell}_1 = \frac{N\sum YX - \sum Y \sum X}{N\sum X^2 - \left[\sum X\right]^2}$$

$$= \frac{5(314) - 21(70)}{5(1020) - (70)^2}$$

$$\hat{\ell}_1 = 0.5$$

Estimated Regression line: $\hat{Y}_i = -2.8 + 0.5X_i$

$$when\ X = 10,\ \hat{Y}_i = -2.8 + 0.5(10) = 2.2$$

$$when\ X = 12,\ \hat{Y}_i = -2.8 + 0.5(12) = 3.2$$

$$when\ X = 16,\ \hat{Y}_i = -2.8 + 0.5(16) = 5.2$$

$$when\ X = 14,\ \hat{Y}_i = -2.8 + 0.5(14) = 4.2$$

$$when\ X = 18,\ \hat{Y}_i = -2.8 + 0.5(18) = 6.2$$

Y	$\hat{Y}_i$	$e = Y - \hat{Y}_i$		X	r_X	r_e	$d = r_X - r_e$	d^2
2	2.2	-0.2	1	10	1	1.5	-0.5	0.25
4	3.2	0.8	3	12	2	3.5	-1.5	2.25
6	5.2	0.8	4	16	4	3.5	0.5	0.25
3	4.2	-1.2	5	14	3	5	-2	4
6	6.2	-0.2	2	18	5	1.5	3.5	12.25
								19.0

Now, we can apply the spearman rank coefficient test statistic

$$r_{x.e} = 1 - \left[\frac{6\sum d^2}{n\left[n^2 - 1\right]}\right]$$

$$r_{x.e} = 1 - \left[\frac{6(19)}{5(24)}\right]$$

$$= 1 - 0.95$$

$$= 0.05$$

where d is the difference between the values of corresponding pairs of X *and* e observations, n is the number of observations in the sample. Solving ties: $\frac{1+2}{2} = 1.5$

$$\frac{3+4}{2} = 3.5$$

Decision Rule: We can evaluate the z and t critical values as follows:

$$r_{X,e} < \left(\begin{array}{l} z = \dfrac{1.96}{\sqrt{5-2}} \\ = \dfrac{1.96}{\sqrt{4}} \\ = \dfrac{1.96}{2} \\ = 0.98 \end{array} \right)$$

$$0.05 < 0.98$$

$$r_{x.e} < \left(t = \frac{0.05\sqrt{5-2}}{\sqrt{1-(0.05)^2}} \right)$$

$$0.05 <$$

Given that Since $r_{x.e}$ is on the low side, we conclude that the error variances are homoskedastic. So we would accept H_0 and reject H_1

Example 4.5

Given:

Y	X
10	20
12	22
14	26
16	28
18	28
10	30
12	42
18	26
10	30
16	20
22	26
18	22
16	28
28	28
26	62

Test for heteroskedasticity using the Goldfeld-Quandt test statistic at both the 5% and 15 significance levels.

Solution 4.5

Date Ordering

Y	X	
10	20	
16	20	
12	22	
18	22	
14	26	
18	26	
22	26	
16	28	
18	28	
18	28	
16	28	
28	28	
10	30	
10	30	
12	42	
26	62	

$16 - 4 = 12 = 6$, 2

Delete $\frac{1}{4}$ x $16 = 4$

$16 - 4 = 12 = 6$, 2

Hypothesis:

H_0 : *residual variance is homoskedatsic*

H_1 : *residual variance is heteroskedatsic*

$$F - statistic = \frac{\sigma_{2e}^2}{\sigma_{1e}^2}$$

$$where\ \sigma_{1\upsilon}^2 = \frac{\sum \hat{\upsilon}^2}{n_1 - 2}$$

$$= \frac{\sum \left[Y - \hat{Y} \right]^2}{n_1 - 2}$$

Solving first half

Y	X	X^2	YX	ŷ	$(y - ŷ)^2$
10	20	400	200	13.34	11.16
16	20	400	320	13.34	7.08
12	22	484	264	14.26	5.11
18	22	484	396	14.26	13.99
14	26	676	364	16.1	4.41
18	26	676	468	16.1	3.61
88	136	3120	2,012		45.36

The underlying model can be specifies thus: $Y_{11} = b_{01} + b_{11}X_{11} + \upsilon_{11}$

$$where \quad \hat{b}_{01} = \frac{\sum Y \sum X^2 - \sum X \sum YX}{N\sum X^2 - \left[\sum X\right]^2}$$

$$= \frac{88(3120) - 136(2012)}{6(3120) - (136)^2}$$

$$\hat{b}_{01} = 4.14$$

$$\hat{b}_{11} = \frac{N\sum YX - \sum Y \sum X}{N\sum X^2 - \left[\sum X\right]^2}$$

$$= \frac{6(2012) - 88(136)}{6(3120) - (136)^2}$$

$$\hat{b}_{11} = 0.46$$

Estimated Regression line: $\hat{Y}_i = 4.14 + 0.46X_i$

To get $\hat{Y}_i$ *we procede as follows*:

$$when \ X = 20, \ \hat{Y}_i = 4.14 + 0.46[20] = 13.34$$

$$when \ X = 20, \ \hat{Y}_i = 4.14 + 0.46[20] = 13.34$$

$$when \ X = 22, \ \hat{Y}_i = 4.14 + 0.46[22] = 14.26$$

$$when \ X = 22, \ \hat{Y}_i = 4.14 + 0.46[22] = 14.26$$

$$when \ X = 26, \ \hat{Y}_i = 4.14 + 0.46[26] = 16.10$$

$$when \ X = 26, \ \hat{Y}_i = 4.14 + 0.46[26] = 16.10$$

$$\sigma^2_{1\upsilon} = \frac{\sum \hat{\upsilon}^2}{n_1 - 2}$$

$$= \frac{45.36}{4}$$

$$= 11.34$$

$$Also, \ \sigma^2_{2\upsilon} = \frac{\sum \hat{\upsilon}^2}{n_2 - 2}$$

$$= \frac{\sum \left[Y - \hat{Y}\right]^2}{n_2 - 2}$$

Solving third half

Y	X	X^2	YX	ŷ	Y - ŷ	$(y - \hat{y})^2$

16	28	784	448	15.05	0.95	0.90
28	28	784	784	15.05	12.95	167.70
10	30	900	300	15.51	-5.51	30.36
10	30	900	300	15.51	-5.51	30.36
12	42	1764	504	18.27	-6.27	39.31
26	62	3844	1610	22.87	3.13	9.67
102	220	8976	3948		0	278.43

The underlying model can be specifies thus: $Y_{22} = b_{02} + b_{22}X_{22} + \upsilon_{22}$

$$where \quad \hat{b}_{02} = \frac{\sum Y \sum X^2 - \sum X \sum YX}{N\sum X^2 - \left[\sum X\right]^2}$$

$$= \frac{102(8976) - 220(3948)}{6(8976) - (220)^2}$$

$$\hat{b}_{02} = 8.61$$

$$\hat{b}_{22} = \frac{N\sum YX - \sum Y \sum X}{N\sum X^2 - \left[\sum X\right]^2}$$

$$= \frac{6(3948) - 102(220)}{6(8976) - (220)^2}$$

$$\hat{b}_{22} = 0.23$$

Estimated Regression line: $\hat{Y}_i = 8.61 + 0.23X_i$

To get $\hat{Y}_i$ *we procede as follows*:

$$when \ X = 28, \hat{Y}_i = 8.61 + 0.23[28] = 15.05$$

$$when \ X = 28, \hat{Y}_i = 4.14 + 0.46[28] = 15.05$$

$$when \ X = 30, \hat{Y}_i = 4.14 + 0.46[30] = 15.51$$

$$when \ X = 30, \hat{Y}_i = 4.14 + 0.46[30] = 15.51$$

$$when \ X = 42, \hat{Y}_i = 4.14 + 0.46[42] = 18.27$$

$$when \ X = 62, \hat{Y}_i = 4.14 + 0.46[62] = 22.87$$

$$\sigma^2_{2\upsilon} = \frac{\sum \hat{\upsilon}^2}{n_1 - 2}$$

$$= \frac{278.43}{4}$$

$$= 69.61$$

$$F-statistic=\frac{\sigma_{2e}^2}{\sigma_{1e}^2}$$
$$=\frac{69.61}{11.34}$$
$$=6.14$$

Decision Rule: At $F_{\alpha\%,[n_1-k,n_2-k]}=F_{5\%,[4,4]}=6.39$. Since $\sigma_{1v}^2\neq\sigma_{2v}^2$ we accept the null hypothesis and reject the alternative. Alternatively, $F-computed\ (6.14)<F-critical\ (6.39)$, we accept H_0 and conclude that the error variance are homoskedastic.

ASSIGNMENT

Consider the data below:

Test for the presence or otherwise of heteroskedasticity using the spearman rank coefficient test statistics.

Y	228	478	625	398	625
X	120	112	116	114	118

CONCLUSION

In the multiple regression model [MRM], model with more than one regressors, it is difficult to ascertain on apriori basic which of the regressors to be chosen for data transformation

LECTURE 4: IDENTIFICATION PROBLEM

CONTENTS

- Introduction
- Objectives
- Definition of Identification Problem

- Types of Identification
- Identification Restrictions
- Formal Rules for Identification
- Implications for Identification
- Conclusion
- Assignment

INTRODUCTION

Identification in econometrics has to do with being able to solve for unique values of the parameters of the structural model from the coefficients of the reduced-form of the model.

OBJECTIVES

At the end of this lecture, students will be able to understand the meaning and types of identification, identifying restrictions, formal rules and implications for identification.

IDENTIFICATION PROBLEM

- Identification is concerned with the possibility of obtaining meaningful estimates of the structural parameters from the reduced form coefficients such that there must be no other equation in the model that can be formed by algebraic manipulation of some other equations within the model which contains the same variables as the function in question.

- The identification problem thus occurs because different sets of structural coefficients are computed from the same sample data. In other words, a given reduced form equation is found to be compatible with different structural equations thereby making it difficult to disentangle the particular hypothesis that is being tested empirically.

- As it were, the identification problem is a mathematical problem associated with simultaneous equation systems. It is therefore a problem of model specification and not of model estimation.

Types of Identification

In econometric modeling, two types of identification are discernible. These are:

(a) Under-identified equation
(b) Identified equation
  (b.1) Exactly (just) identified equation
  (b.2) Over identified equation

- **Under-identification**

An under-identified equation is an equation whose coefficients cannot be estimated. Indeed, an equation is under-identified if its statistical form is not unique.

- **Identified Equation**

A system is identified if all of its equations are identified. An identified equation could either be exactly identified or over-identified.

- **Exactly (Just) Identification**

An equation is exactly or just identified if only one set of structural coefficient estimates can be computed from the reduced-form coefficients.

- **Over Identification**

An equation is over identified if more than one set of structural coefficient can be computed from the coefficients of the reduced form equation. In sum, a model (system of equations) is identified if all the equations in the model are identified.

Identification Restrictions

The identifying restriction entails the placement of restrictions on the variables of a simultaneous equations model using economic theory and extraneous information to solve the identification problem of the simultaneous equations. These restrictions can take a variety of forms such as:

(a) Use of extraneous estimates of parameters,
(b) Knowledge of exact relationship among parameters,
(c) Knowledge of the relative variances of disturbances,
(d) Knowledge of zero correlation between disturbances in different equations,
(e) Zero restrictions, taking the form of specification that certain structural parameters are zero, i.e., that certain endogenous variables and exogenous variables do not appear in certain equations.

Formal Rules for Identification

(a) Order condition for identification
(b) Rank condition for identification

Order Condition

The order condition states that for an equation to be identified, the total number of variables excluded from it but included in other equation of the model must be at least as great (must be equal to or greater than) as the number of equation of the model less one. Mathematically, the order condition is given by:

$$Q - Q^* \geq E - 1$$

Where Q is the total number of variables in the model
Q* is the total number of variables in the particular equation that is being identified
E is the total number of endogenous variables (number of equations) in the model

Illustration 1:

Consider the following simple version of the Keynesian income determination model:

$$C_t = \alpha_0 + \alpha_1 (Y_t - T_t) + \varepsilon_{1t}$$
$$I_t = \beta_0 + \beta_1 I_{t-1} + \beta_2 r_t + \varepsilon_{2t}$$
$$T_t = \delta_0 + \delta_1 Y_t + \varepsilon_{3t}$$
$$Y_t = C_t + I_t + G_t$$

In commenting on the identification status of the above system of equations, we note the following:

(a) There are four (4) endogenous variables, namely C_t, I_t, T_t and Y_t

(b) There are three (3) predictive variables namely, r_t, I_{t-1} and G_t. Applying the order condition to the first and second equations (the consumption and investment equations), we have;

$\underset{\sim}{Q} = 6,$
$\underset{\sim}{Q}^* = 3$
$\underset{\sim}{E} = 4$
$6 - 3 = 4 - 1$
$3 = 3$

We therefore conclude that the consumption and investment equations are exactly identification (just identified). Applying the same order condition to the third equation (the tax equation), we have:

$\underset{\sim}{Q}^* = 2,$
$\underset{\sim}{Q} = 6$
$\underset{\sim}{E} = 4,$
$6 - 2 > 4 - 1$
$3 = 3$

The tax equation is over identified. Therefore, it is possible to fruitfully estimate the structural parameters of the model from the reduced-form equation. In short, the structural parameters can be retrieved from the reduced from coefficients.

Illustration:
Consider the model:

$$C_t = \beta_0 + \beta_1 Y_t + \mu_{1t}$$
$$I_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 I_{t-1} + \mu_{2t}$$
$$Y_t = C_t + I_t + G_t$$

Using the order condition:

Equation (1)

$\underset{\sim}{Q} = 5$, $\underset{\sim}{Q}^* = 2$, $\underset{\sim}{E} = 3$, $5 - 2 > 3 - 1 \rightarrow 3 > 2$ over identified

$\underset{\sim}{Q} = 5$
$\underset{\sim}{Q} = 2$
$\underset{\sim}{E} = 3$

Therefore, $5 - 2 > 3 - 1$
$3 > 2$

The consumption function is over identified

Equation (2)

$\underset{\sim}{Q} = 5$
$\underset{\sim}{Q} = 3$
$\underset{\sim}{E} = 3$

Therefore, $5 - 3 = 3 - 1$

The investment function is exactly (just) identified.

Equation	Degrees of over identification
1	1
2	0
	L = 1

Given that L L6, use ILS estimator to estimate model.

Rank Condition

- The rank condition states that in a system of K equations, a particular equation is identified if and only if it is possible to construct at least one non-zero determinant of order $(K-1)$ from the coefficients of the variables excluded from that particular equation but contained in the other equations of the model.

- This condition is called the rank condition because it refers to the rank of the matrix of parameters of excluded variables and the rank of a matrix is the order of the largest non-zero determinant which can be formed from the matrix.

- In econometric analysis, the relevant matrix is the sub matrix of the coefficient of the excluded variables. It is a "SUFFICIENT" criterion for the identification. When an equation is sufficiently identified, it is necessarily identified but the converse is not the case.

- In effect, the common use of the "order" condition for identification is not justified because it is only a necessary condition for identification. Thus an equation might be necessarily identified but not sufficiently identified. That is, even if the "order" condition is satisfied for a particular equation, it may happen that very equation is not identified.

Illustration:

Considers the structural Keynesian model given below:

$$C_t = l_0 + l_1 Y_t + \mu_{1t}$$
$$I_t = m_0 + m_1 Y_t + \alpha_2 m_{t-1} + \mu_{2t}$$
$$Y_t = C_t + I_t + G_t$$

This model could be re-written in the form

$$0 = -Y_t + l_o + l_1 Y_t + \mu_{1t}$$
$$0 = -I_t + m_o + m_1 Y_t + m_2 I_{t-1} + \mu_{2t}$$
$$0 = -Y_t + C_t + I_t + G_t$$

Ignoring the random disturbance the table of Ps of the model becomes variables.

Equations	C_t	I_t	Y_t	I_{t-1}	G_t
Equation C	-1	0	l_1	0	0
Equation I	0	-1	m_1	m_2	0
Equation Y	1	1	-1	0	1

Since we are identifying equation (1), the consumption function, we strike out the first row in the table of structural parameters as follows.

Table of structural parameters

Equations	C_t	I_t	Y_t	I_{t-1}	G_t
Equation C	-1	0	l_1	0	0
Equation I	0	-1	m_1	m_2	0
Equation Y	1	1	-1	0	1

Tables of parameters of Excluded Variables

I_t	γ_{t-1}	G_t
-1	α_2	0
1	0	1

Forming the determinant(s) of order (M – 1) x (M – 1) that is (3 – 1) by (3 – 1) = 2 x 2, we have that:

$$\Delta_1 = \begin{vmatrix} -1 & m_2 \\ 1 & 0 \end{vmatrix}$$

$$0 - m_2 = -m_2 \neq 0$$

$$\Delta_2 = \begin{vmatrix} -1 & 0 \\ 1 & 1 \end{vmatrix}$$

$$-1 - 0 = -1 \neq 0$$

$$\Delta_3 = \begin{vmatrix} m_2 & 0 \\ 0 & 1 \end{vmatrix}$$

$$m_2 - 0 = m_2 \neq 0$$

We are able to form 3 non zero determinants of order 2, the consumption function of the model is identified.

Implications of Identification

(a) If an equation (model) is under-identified it is impossible to estimate its parameters with any econometric technique.
(b) If an equation (model), its coefficients can be estimated. The suitable estimation technique is ascertained by identification status, i.e. exactly identified or over-identified.
(c) If an equation is exactly identified, the appropriate econometric technique to be used for its estimation is the ILS
(d) If an equation is over identified, the appropriate econometric technique to be used for estimating it is the 2SLS, 3SLS, ML etc.

ASSIGNMENT

Consider the following simple version of the Keynesian income determination model:

$$C_t = \alpha_0 + \alpha_1 Y + \varepsilon_{1t}$$
$$I_t = d_0 + d_1 I_{t-1} + \varepsilon_{2t}$$
$$Y_t = C_t + I_t + G_t$$

Determine the identification status of the model and suggest an estimator for the model.

CONLUSION

The order condition is which is a "NECESSARY" condition for identification is indeed based on the counting rule of the variables included and excluded from the particular equation that is being identified

LECTURE 5: GAUSS-MARKOV THEOREM

CONTENTS

- Introduction
- Objectives
- Gauss-Markov Theorem
- Proof of Gauss-Markov Theorem

- Numerical Application
- Conclusion
- Assignment

INTRODUCTION

The Gauss Markov Theorem [GMT] has to do with unbiasedness and efficiency properties of the OLS estimator.

OBJECTIVES

At the end of this lecture, students will be able to understand the meaning and know how to give the mathematical proof of Gauss-Markov theorem.

GAUSS-MARKOV THEOREM

The Gauss Markov Theorem [GMT] states that the OLS estimator provides the best, linear and unbiased [BLU] estimator. In other words, in the class of linear and unbiased estimators, the OLS estimator is the most efficient estimator.

PROOF OF GAUSS-MARKOV THEOREM

$$Y = Xa + \mu \qquad (1)$$

$$\hat{Y} = X\hat{a} \qquad (2)$$

$$\mu = Y - \hat{Y}$$

$$\Rightarrow Y - X\hat{a}$$

$$\hat{\mu}'\hat{\mu} = (Y - X\hat{a})'(Y - X\hat{a})$$

$$= Y'Y - \hat{a}Y'X - \hat{a}'X'Y + \hat{a}'X'X\hat{a}$$

Observe that $\hat{a}Y'X$ *and* $\hat{a}'X'Y$ *are both scalars and thus equal to their transpose*

$$\hat{\mu}'\hat{\mu} = Y'Y - \hat{a}Y'X - \hat{a}'X'Y + \hat{a}'X'X\hat{a}$$

$$\frac{\partial(\hat{\mu}'\hat{\mu})}{\partial\hat{a}} = -2X'Y + 2\hat{a}X'X$$

Setting $\frac{\partial(\hat{\mu}'\hat{\mu})}{\partial\hat{a}} = 0$

$$-2X'Y + 2\hat{\beta}X'X = 0$$

$$X'Y = \hat{\beta}X'X$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

where Y *is an* $(n \times 1)$ *column vector of endogeneous variables*

X *is an* $(n \times k)$ *matrix of exogeneous variables*

β *is a* $(k \times 1)$ *column vector of population parameter*

μ *is an* $(n \times 1)$ *column vector of stochastic disturbances*

Combining the twin assumptions of homoskedasticity and absence of autocorrelation, that is,

$$Var(\mu_i) = E(\mu_i^2) = \sigma_\mu^2 = \sigma^2 \quad \forall_i = (1, 2, ..., n)$$

$$Cov(\mu_i, \mu_j') = E(\mu_i, \mu_j') = 0 \quad \forall_i = (1, 2, ..., n)$$

We now derive the variance – covariance matrix as follows:

$$E(\mu_i, \mu_j') = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \sigma^2 I$$

Proof of Unbiasedness

$$\hat{a} = (X'X)^{-1}X'Y \qquad (1)$$

Recall that $Y = Xa + \mu$ (2)

Substituing equation (2) *into* (1)

$$\hat{a} = (X'X)^{-1}X'[Xa + \mu] \qquad (3)$$

$$= (X'X)^{-1}X'Xa + (X'X)^{-1}X'\mu \qquad (4)$$

$$= \frac{(X'X)}{(X'X)}a + (X'X)^{-1}X'\mu \qquad (5)$$

$$\hat{\beta} = \beta + (X'X)^{-1}X'\mu \qquad (6)$$

$$E(\hat{a}) = E[a + (X'X)^{-1}X'\mu] \quad (7)$$

$$= a + (X'X)^{-1}X'E(\mu) \quad (8)$$

$$E(\hat{a}) = a \quad (9)$$

Variance – Covariance

$$Var - Cov(\hat{a}) = E[(\hat{a} - a)(\hat{a} - a)'] \quad (10)$$

$$= E\left[[(X'X)^{-1}X'\mu][(X'X)^{-1}X'\mu]'\right] \quad (11)$$

$$= [(X'X)^{-1}X'\mu][(X'X)^{-1}X\mu'] \quad (12)$$

$$= [(X'X)^{-1}X']E(\mu\mu')[(X'X)^{-1}X] \quad (13)$$

$$= (X'X)^{-1}[X'X]\sigma^2 I(X'X)^{-1} \quad (14)$$

$$Var - Cov(\hat{a}) = \sigma^2(X'X)^{-1} \quad (15)$$

Consider another linear estimator a^*

$$a^* = \hat{a} + DY \quad (16)$$

Substituting for $\hat{a}$ *and* Y

$$a^* = a + (X'X)^{-1}X'\mu + DXa + D\mu \quad (17)$$

$$= a + DXa + (X'X)^{-1}X'\mu + D\mu \quad (18)$$

$$a^* = a + DXa + [(X'X)^{-1}X' + D]\mu \quad (19)$$

Taking the expected value of equation (19)

$$E(a^*) = E\left[a + DXa + [(X'X)^{-1}X' + D]\mu\right] \quad (20)$$

$$= a + DXa + [(X'X)^{-1}X' + D]E(\mu) \quad (21)$$

$$E(a^*) = a + DXa \quad (22)$$

For a^* *to be unbiased* $DX = 0$

$$E(a^*) = a \quad (23)$$

$Variance - Covariance\ \ (d^*)$

$$Var - Cov(d^*) = E[(a^* - a)(a^* - a)'] \quad (24)$$

$$= E\left[\left\{[(X'X)^{-1}X' + D]\mu\right\}\left\{[(X'X)^{-1}X' + D]\mu\right\}'\right] \quad (25)$$

$$= E\left[\left\{[(X'X)^{-1}X' + D]\mu\right\}\left\{[(X'X)^{-1}X' + D']\mu'\right\}\right] \quad (26)$$

$$= [(X'X)^{-1}X' + D]E(\mu\mu')[(X'X)^{-1}X + D'] \quad (27)$$

$$= [(X'X)^{-1}X' + D]\sigma^2 I[(X'X)^{-1}X + D'] \quad (28)$$

$$= [(X'X)^{-1}X' + D][\sigma^2(X'X)^{-1}X + \sigma^2 D'] \quad (29)$$

$$= (X'X)^{-1}X'\sigma^2(X'X)^{-1}X + (X'X)^{-1}X'\sigma^2 D' + (X'X)^{-1}X\sigma^2 D + \sigma^2 DD' \quad (30)$$

$$Var - Cov(\hat{a}) = \sigma^2(X'X)^{-1} + \sigma^2 DD' \quad (32)$$

$$Var - Cov(\hat{a}) < Var - Cov(a^*)$$

Thus, the OLS estimator, $\hat{a}$ *is best, linear and unbiased* (BLU)

In sum, the OLS estimator formulae in matrix are given by:

$$\begin{bmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \end{bmatrix} = (X'X)^{-1}X'Y$$

$$X'Y = \begin{bmatrix} \sum Y \\ \sum YX_1 \\ \sum YX_2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} n & \sum X_1 & \sum X_2 \\ \sum X_1 & \sum X_1^2 & \sum X_1X_2 \\ \sum X_2 & \sum X_1X_2 & \sum X_2^2 \end{bmatrix}$$

$$\sigma^2 = \frac{\mu'\mu}{n-k}$$

$$where\ \mu'\mu = Y'Y - \hat{a}'X'Y$$

$$R^2 = \frac{\hat{a}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

$$F = \frac{\hat{a}'X'Y - n\bar{Y}^2/(k-1)}{Y'Y - \hat{a}'X'Y/(n-k)}$$

$$Var - Cov(\hat{a}) = \sigma^2(X'X)^{-1} = \begin{bmatrix} \sigma^2_{a_0} & 0 & 0 \\ 0 & \sigma^2_{a_1} & 0 \\ 0 & 0 & \sigma^2_{a_2} \end{bmatrix}$$

$$Var(\hat{a}_0) = \sigma^2_{a_0}$$

$$S_{a_0} = \sqrt{\sigma^2_{a_0}}$$

$$Var(\hat{a}_1) = \sigma^2_{a_1}$$

$$S_{a_1} = \sqrt{\sigma^2_{a_1}}$$

$$Var(\hat{a}_2) = \sigma^2_{a_2}$$

$$S_{a_2} = \sqrt{\sigma^2_{a_2}}$$

Numerical Example

Years	Life-Expectancy (Y)	Heath-care Spending (Z_1)	National Income (Z_2)
	"000	"000	"000
2000	45	10	14
2001	46	12	22
2002	49	13	20
2003	50	15	36
2004	52	18	48
2005	53	22	56
2006	58	25	87

Formulate the basic model in matrix format. Estimate the determinants of life expectancy in Nigeria. Interpret the results of the model.

Solution

Basic Model [Matrix formulation]: $Y = Z\beta + \mu$

OLS Estimator: $\beta = (Z'Z)^{-1} Z^1 Y$

$$Z^1Z = \begin{bmatrix} n & \Sigma Z_1 & \Sigma Z_2 \\ \Sigma Z_1 & \Sigma Z_1^2 & \Sigma Z_1 Z_2 \\ \Sigma Z_2 & \Sigma Z_2 Z_1 & \Sigma Z_2^2 \end{bmatrix} = \begin{bmatrix} 7 & 115 & 283 \\ 115 & 2{,}071 & 5{,}475 \\ 283 & 5{,}475 & 15{,}385 \end{bmatrix}$$

$$Z'Y = \begin{bmatrix} \Sigma Y \\ \Sigma Z_1 Y \\ \Sigma Z_2 Y \end{bmatrix} = \begin{bmatrix} 353 \\ 5941 \\ 14932 \end{bmatrix}$$

$$Y'Y = 17919$$

$$\left(Z'Z\right)^{-1} = \frac{Adj\left(Z'Z\right)}{\left|Z'Z\right|}$$

$$\left|Z^1Z\right| = 243{,}776$$

$$Adj\left(Z'Z\right) = \left[cofactor\ matrix\ of\ \left(Z'Z\right)\right]^T$$

$$c_{11} = (+)\begin{vmatrix} 2{,}071 & 5{,}475 \\ 5{,}475 & 15{,}385 \end{vmatrix} = 1{,}886{,}710$$

$$c_{12} = (-)\begin{vmatrix} 115 & 5{,}475 \\ 283 & 15{,}385 \end{vmatrix} = -219{,}850$$

$$c_{13} = (+)\begin{vmatrix} 115 & 2{,}071 \\ 283 & 5{,}475 \end{vmatrix} = 43{,}532$$

$$c_{21} = (-)\begin{vmatrix} 115 & 283 \\ 5{,}475 & 15{,}385 \end{vmatrix} = -219{,}850$$

$$c_{22} = (+)\begin{vmatrix} 7 & 283 \\ 283 & 15{,}385 \end{vmatrix} = 27{,}606$$

$$c_{23} = (-)\begin{vmatrix} 7 & 115 \\ 283 & 5{,}475 \end{vmatrix} = -5{,}780$$

$$c_{31} = (+)\begin{vmatrix} 115 & 283 \\ 2{,}071 & 5{,}475 \end{vmatrix} = 43{,}532$$

$$c_{32} = (-)\begin{vmatrix} 7 & 283 \\ 115 & 5{,}475 \end{vmatrix} = -5{,}780$$

$$c_{33} = (+)\begin{vmatrix} 7 & 115 \\ 115 & 2{,}071 \end{vmatrix} = 1{,}272$$

$$\text{Cofactor Matrix, } C_{[Z'Z]} = \begin{bmatrix} 1{,}886{,}710 & -219{,}850 & 43{,}532 \\ -219{,}850 & 27{,}606 & -5{,}780 \\ 43{,}532 & -5{,}780 & 1{,}272 \end{bmatrix}$$

$$Adj_{[Z'Z]} = C^{T}_{[Z'Z]} = \begin{bmatrix} 1{,}886{,}710 & -219{,}850 & 43{,}532 \\ -219{,}850 & 27{,}606 & -5{,}780 \\ 43{,}532 & -5{,}780 & 1{,}272 \end{bmatrix}$$

$$(Z'Z)^{-1} = \frac{1}{243{,}776}\begin{bmatrix} 1{,}886{,}710 & -219{,}850 & 43{,}532 \\ -219{,}850 & 27{,}606 & -5{,}780 \\ 43{,}532 & -5{,}780 & 1{,}272 \end{bmatrix}$$

$$= \begin{bmatrix} 7.739 & -0.902 & 0.179 \\ -0.902 & 0.113 & -0.024 \\ 0.179 & -0.024 & 0.005 \end{bmatrix}$$

$$\beta = \begin{bmatrix} 7.739 & -0.902 & 0.179 \\ -0.902 & 0.113 & -0.024 \\ 0.179 & -0.024 & 0.005 \end{bmatrix}\begin{bmatrix} 353 \\ 5{,}941 \\ 14{,}932 \end{bmatrix} = \begin{bmatrix} 46.09 \\ -5.44 \\ 285.26 \end{bmatrix}$$

Estimated Model: $\hat{Y} = 46.09 - 5.44Z_1 + 285.26Z_2$

ASSIGNMENT

Consider the following data

Y	141	121	101	151	121	81
X_1	21	30	34	51	52	64
X_2	12	10	10	12	10	14

(a) Estimate the model $Y = X\beta + \mu$
(b) Estimate the unadjusted coefficient of determination
(c) Determine the statistical significance of the coefficients
(d) Determine the statistical significance of the regression model

CONCLUSION

The BLU property of the OLS estimator can be explained under the following: Unbiasedness, Efficiency, Consistency, Linearity and Sufficiency.

LECTURE 6: SIMULTANEOUS EQUATIONS MODELLING

CONTENTS

INTRODUCTION

Simultaneous equations model (SEM) is a system of equations representing a set of relationships among variables and thereby relating the joint dependence of variables

OBJECTIVES

At the end of this lecture, students will be able to understand the meaning of Simultaneous-Equation Bias: Endogeneity, explain cause of Simultaneity Bias/Endogeneity and give a proof of Omitted Variable Bias.

SIMULTANEOUS EQUATIONS MODELLING

- A simultaneous equations model (SEM) is a system of equations in which the dependent variables in some equations are explanatory variables in other equations and thereby feeding-off shocks to each other. Thus, a SEM is a system of equations representing a set of relationships among variables and thereby relating the joint dependence of variables. The feedback effect of the SEM can be demonstrated using the structural model:

$$Y_{1t} = d_0 + d_1 Y_{2t} + d_2 X_{1t} + u_{1t} \qquad (1)$$

$$Y_{2t} = \phi_0 + \phi_1 Y_{1t} + \phi_2 X_{2t} + u_{2t} \qquad (2)$$

  where Y_{1t} *and* Y_{2t} are mutually dependent variables, X_{1t} *and* X_{2t} are the exogenous variable u_{1t} *and* u_{2t} are the stochastic disturbance terms. If u_1 increases by a given proportion, it will automatically increase Y_1.

- The increase in Y_1 will in turn cause Y_2 to increase. This feedback effect between the two structural equations is contemporaneous and indeed continuous, an indication that the endogenous variables Y_{1t} *and* Y_{2t} are jointly dependent. The correlation is that an increase in u_1 increases Y_1 which in turn increases Y_2 . So u_{1t} *and* Y_{2t} are positively correlated.

Single Equation Model	Simultaneous Equation Model
The SEM represents only one relationship among variables	The SEM represents more than one relationship among variables
The SEM has only one equation.	The SEM has more than one equation.

The estimation method is mainly OLS	The OLS estimation technique cannot be applied to estimate the SEM
In the SEM, only the parameters of the single equation can be estimated	In the SEM , more than one parameters can be estimated simultaneously
In the SEM, there is a single dependent variable and one or more explanatory variables	In the SEM, there are more than one dependent variables and more than one explanatory variables

Specification:

A SEM has these specifications:

(a) Reduced form specification

(b) Structural form specification

The structural model is a complete system of equation, which describes the structure of the relationship between economic variables such that the endogenous variables are expressed as function of other endogenous variables, predetermined variables and stochastic disturbances. The regressors of structural equations correlated with stochastic disturnaces. The structural specification of a simultaneous equations model can be given as:

$$Y_{1t} = d_0 + d_1 Y_{2t} + d_2 X_t + u_{1t}$$
$$Y_{2t} = \phi_0 + \phi_1 Y_{1t} + \phi_2 X_t + u_{2t}$$

Where Y_{1t} *and* Y_{2t} are the mutually dependent variables, X_t is an exogenous variable and u_{1t} and *and* u_{2t} are the stochastic disturbance terms

The reduced-form model (RFM) is that model in which the endogenous variables are expressed as an explicit function of only the exogenous and predetermined variables. In other words, the RFM expresses an endogenous variable solely in terms of the predetermined variable and the stochastic disturbances. The RFM that corresponds to the above structural model is given by:

$$Y_{2t} = \pi_{20} + \pi_{22} X_t + e_2$$
$$Y_{1t} = \pi_{10} + \pi_{12} X_t + e_1$$

Simultaneous-Equation Bias: Endogeneity

- Simultaneous-equation bias is an endogeneity problem which entails reverse causation between the explanatory and the dependent variables of a model. Thus, simultaneity bias is a loop of causality between the dependent variables and the regressors of a model. It occurs when a variable on the right-hand side of the causal inferential model and the variable on the left-hand side of the same model influence each other at the same time.
- In effect, both the endogenous and the explanatory variables are related to each other. Accordingly, by sequencing the causality between the dependent and independent variables of a model, endogeneity is induced. Endogeneity refers to the correlation between the endogenous explanatory variable, that is, the endogenous regressor and the random error term.

Cause of Simultaneity Bias/Endogeneity

Endogeneity can arise as a result of:

(a) Measurement error

(b) Omitted variable bias

Measurement Error:

Measurement error in the endogenous explanatory variable causes simultaneous equation bias.

Omitted Variable Bias

The omission of key explanatory variables from a regression model causes simultaneous equation bias.

- If the "correct" model that explains the variation in Y, that is, the model to be estimated in mean deviation was:

$$y = a_1x_1 + a_2x_2 + a_3x_3 + u$$

  - If the model is correctly specified, the unbiasedness property would be satisfied in that the regression coefficients would be unbiased estimators of the population parameters and as such

$$E(\hat{a}_1) = a_1,\ E(\hat{a}_2) = a_2 \ \ and \ \ E(\hat{a}_3) = a_3$$

- If by ignorance or carelessness, the econometrician mistakenly omitted x_2 *and* x_3 from the model with the mis-specified equation given as:

$$y = b_1x_1 + \varepsilon$$

Then $E(\hat{b}_1) \neq \beta_1$

Proof: Applying OLS to the mis-specified equation, the slope coefficient of the SRM will be given as:

$$b_1 = \frac{\sum yx_1}{\sum x_1^2}$$

The normal equations for the correctly specified model are given as:

$$\sum yx_1 = a_1\sum x_1^2 + a_2\sum x_1x_2 + a_3\sum x_1x_3 \quad (4)$$

$$\sum yx_2 = a_1\sum x_1x_2 + a_2\sum x_2^2 + a_3\sum x_2x_3 \quad (5)$$

Dividing eqn (4) *by* $\sum x_1^2$,

$$\frac{\sum yx_1}{\sum x_1^2} = a_1 + a_2\frac{\sum x_1x_2}{\sum x_1^2} + a_3\frac{\sum x_1x_3}{\sum x_1^2} \qquad (6)$$

where $\frac{\sum yx_1}{\sum x_1^2} = b_1$, *the slope coefficient of the SRM of Y on X in which* X_2 *is omitted*

$\frac{\sum x_1 x_2}{\sum x_1^2}$ *is the slope coefficient of the SRM of the omitted variable* x_2 *on* x_1 *and is denoted by* b_2, *i.e.* $x_2 = b_2 x_1 + e_{1t}$

$\frac{\sum x_1 x_3}{\sum x_1^2}$ *is the slope coefficient of the SRM of the omitted variable* x_3 *on* x_1 *and is denoted by* b_3, *i.e.* $x_3 = b_3 x_1 + e_{2t}$

Substituing these facts into eqn (6),

$$b_1 = a_1 + a_2 b_2 + a_3 b_3 \qquad (7)$$

Obviouly, the coefficient of the included variable x_1 *in the mis – specified equation has picked up the ceofficient of the omitted variable,* x_2 *that was correlated with* x_1

Taking expectations of eqn (7),

$$E(b_1) = a_1 + a_2 b_2 + a_3 b_3$$

$$Thus, E(b_1) \neq a_1$$

$$Omitted\ \text{variable}\ bias = [E(b_1) - a_1]$$

$$= a_2 b_2 + a_3 b_3$$

In effect, the regression coefficient b_1 in the incorrect model specification differ from the regression coefficient β_1 of the correct model specification. Thus, b_1 is a biased estimator of β_1 and the "bias" is equal to $\beta_2 b_2 + \beta_3 b_3$. Overall, the omission of a key variable from the regression model leads to biased estimates of the parameters of the included variables.

Consequences of Omitted Variable Bias

(a) Estimated coefficients are positively biased
(b) Estimated variances are biased
(c) Estimated standard errors are biased
(d) Hypothesis testing about the significance of parameters is misleading
(e) Confidence intervals are wrongfully estimated
(f) Forecasting is invalid
(g) Estimated coefficients are inconsistent. This is because the "bias" will not disappear even as the sample size gets larger
(h) Residual variance is incorrectly estimated

There are two conditions under which $E(b_1) = a_1$:

(a) $a_2 = 0$

There will be no "bias" if the omitted variable X_2 has no effect on the dependent variable, Y. Of course, if that be the case, it thus means that the model in the first place was not mis-specified

(b) $b_2 = 0$

That is, there will be no "bias" if X_2 and X_1 are not correlated. Thus, if the two

explanatory variables in the correctly specified model are uncorrelated such that $Cov(X_2, X_1) = 0$, then omitting X_2 does not in any way result in biased estimate of the effect of X_1

ASSIGNMENT
Explain omitted variable bias with mathematical proof

CONCLUSION
In SEM, both the endogenous and the explanatory variables are related to each other

LECTURE 7: TIME SERIES PROPERTIES OF VARIABLES

CONTENTS

INTRODUCTION

As a lay down to any estimation process and in view of standard econometrics application, empirical methodology in time series econometrics does proceed in the following steps:

- Testing for stationarity of variables in a model,
- Testing for co-integration and
- Estimating error correction models.

OBJECTIVES

At the end of this lecture, students will be able to understand the meaning of Stationary and Non-stationary Series, Properties of Integrated Processes, Stationarity Tests, Co-integration and Error Correction Modelling.

TIME SERIES PROPERTIES OF VARIABLES

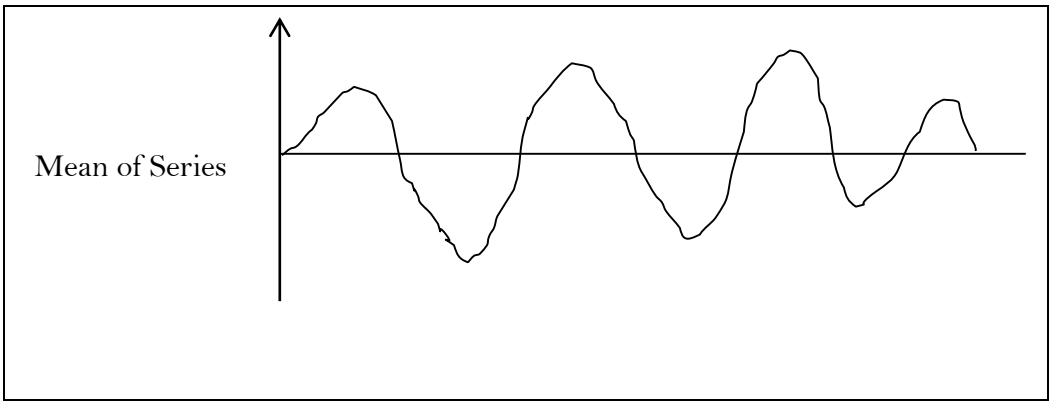
Stationary versus Non-stationary Series

Stationary Series

A time series X_t is stationary if it has no trend whether deterministic or stochastic trend. In other words, there is no systematic change in its mean, variance and other periodic variation. Accordingly, it is a time series with only regular variation. For stationary time series, any shock, that is, perturbation to the series will die-off over- time thereby making it possible for the long-run mean value of the series to be established. This then guarantees the convergence of long-term forecasts based on such series to the constant long-run mean of the series [see Enders (1998)].

Figure 8.1 illustrates the stationary process as it portrays a cumulative convergence towards the initial equilibrium state after an initial shock to equilibrium.

Figure: Stationary Series



Non-stationary Series and the Order of Integration

A non-stationary series is a time series with trend. The trend could be deterministic or stochastic. Only the stochastic trend can be removed by differencing and not the deterministic trend. Thus, a non-stationary series is a series with irregular variation such that its mean, variance and covariance are not relatively constant. Consequently, a non-stationary series is an integrated series which can only be made stationary by differencing.

A series is said to be integrated of order f if it has a stationary representation after differencing the series f time. In other words, the order of integration refers to the number of times a variable has to be differenced to gain stationarity.

Econometrically speaking, if a series becomes stationary after first differencing, it is said to be integrated of order one and it is represented as $M_t^S \sim I(1)$. If the series becomes stationary after differencing twice, it is adjudged integrated of order two and it is represented as $M_t^S \sim I(2)$, while if originally without differencing the series, the time series variable is found to be stationary, then it is said to be integrated of order zero and this is represented as $M_t^S \sim I(0)$. In sum, $M_t^S \sim I(6)$ means that M_t^S is integrated of order six (non-stationary) and must be differenced six times to make it stationary.

Properties of Integrated Processes

The properties of an integrated process can be itemized thus: An integrated process has a

(a) Finite unconditional mean,
(b) Time dependent variance and
(c) Time dependent covariance

Consider a random walk model with drift, a non- stationary model which most often is specified as:

$$Z_t = \delta + Z_{t-1} + \upsilon_t$$
$$where \ \upsilon_t \sim IID(0, \sigma_{Z_t}^2)$$

The first difference of the random walk process

$$\Delta Z_t = \delta + \upsilon_t$$
$$where \ \upsilon_t \sim IID(0, \sigma_{Z_t}^2)$$

is stationary. By intuition, Z_t has a stochastic trend. If a time series is generated from the random walk process with a drift and $\delta > 0$, it means that the series is trending upward and if $\delta < 0$, the series Z_t is trending downward as shown in the figures below:

Figure 7.2: Upward Trend Series

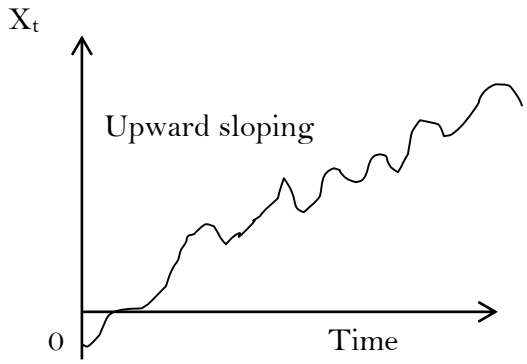
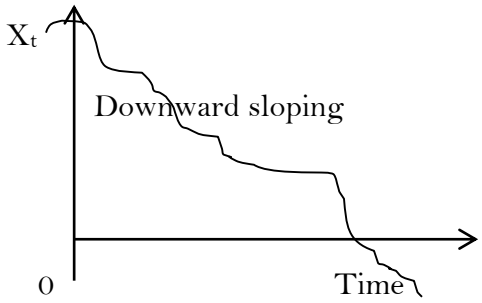


Figure 7.3: Downward Trend Series



A variable that trends upwards or downwards is never stationary. Hence, the presence of trend leads to non-stationarity. In the non-stationary process, there is always a disturbance to the equilibrium that is already established. This results in cumulative divergence from the equilibrium mean value so that it is practically impossible to re-institute equilibrium.

Problems of Non-Stationary Series

There are basically two problems of non-stationary series. These include spurious regression results, and inconsistent regression results.

Inconsistent Regressions

Inconsistent regression estimates are obtained when a stationary series is regressed on a non-stationary series. Given that the non-stationary series will have a time-dependent mean, the value of the coefficients of the regression will not themselves be constant. In such a case, the coefficient estimates are highly sensitive and unstable over different sub-samples. Thus, if we were to draw

valid inference that is not time dependent; then all the variables in the model should be integrated of the same order.

Spurious Regressions

Spurious regression arises when timely unrelated variables are regressed using the OLS and the results are indicative of the fact that the series are correlated even when on aproiri basis the relationship between the variables is not genuine.

In a Monte Carlo experiment, Granger and Newbold (1974) found that when a regression involves non-stationary variables, the OLS estimates become spurious in view of a very high R^2 and low Durbin-Watson (DW) statistic, which tend to cause the OLS estimator to underestimate standard errors and hence overestimate t-values of the regression coefficients. So, one possible way of detecting a spurious regression is the use of autocorrelation statistic, in particular the DW-statistic.

Stationarity Tests

Since most time series variables are often suspected to be strongly trended, that is, to have time-variant mean (a mean that changes over time), central to the stationary test is the determination of the order of integration (the number of times a variable has to be differenced to achieve stationarity) given the pre-notion of unit root that characterized the variables.

The Augmented Dickey-Fuller (ADF) test and the Phillips-Peron (PP) test are mostly employed to experiment for the existence of unit root. The econometric rationale for both tests is that while the PP test makes no distributional assumption about the residuals, the ADF test hypothesized that the residuals from the auxiliary regression are white noise.

In view of the problems that characterized non-stationary series, econometricians often deem it essential to analyze the time-series properties of the variables due to their unknown data generating process (DGP) by performing the unit root test.

Sample Autocorrelations: Correlogram

A correlogram is a plot of the relationship between the sample autocorrelation coefficients, r_k and the time periods. As the time period increases, the autocorrelation coefficients r_k gradually decay to zero. Therefore, for a series to be stationary, the rate of decay of the autocorrelation coefficients as time period increases needs to be rapid. For a non- stationary series, the rate of decay of the autocorrelation coefficient is very slow.

- The sample autocorrelation test statistic is given by:

$$r_k = \frac{\sum (X_t - \overline{X}_t)(X_{t-k} - \overline{X}_{t-k})}{\sqrt{\sum (X_t - \overline{X}_t)^2 (X_{t-k} - \overline{X}_{t-k})^2}}$$

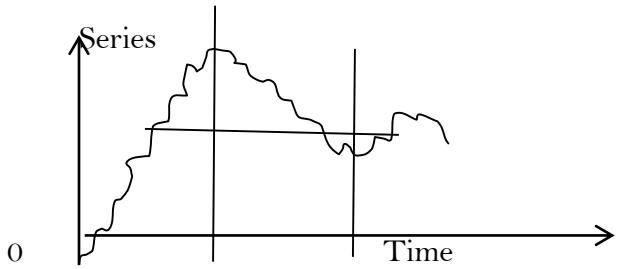
Time Path: Graphical Method of Test

This is a graphical test for stationarity. Accordingly, it basically entails testing for stationarity with the aid of a graph. A consideration can thus be given to the following graphs:

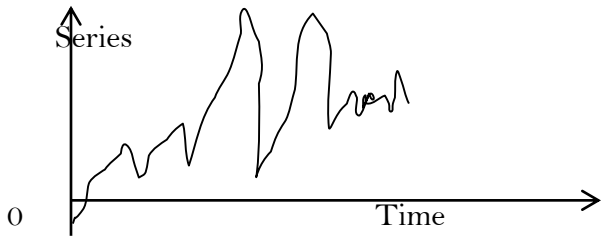
Figure: Time Paths and the Graphical Test for

Autocorrelation

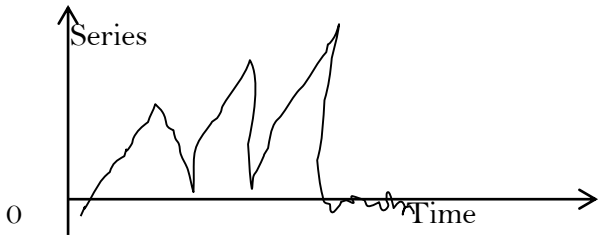
Time-Path Panel [7a]



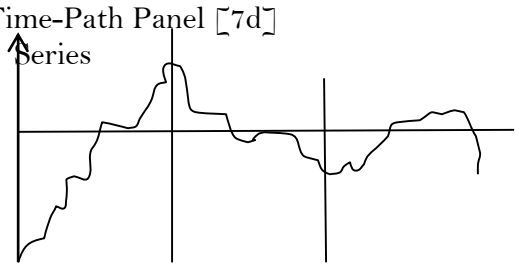
Time-Path Panel [7b]



Time-Path Panel [7c]



Time-Path Panel [7d]



Series Mean of

An informal inspection of four panels reveals that only time-path of panel [10d] is stationary since it can be visualized that the initial disturbance as portrayed by the wave-like trend gradually reconverge to the equilibrium mean value. Hence, the mean and the variance of the series in question are relatively constant.

Unit Root Test

The unit root test is a formal statistical test for stationarity. Given the test equation:

$$Y_t = \delta + \varphi Y_{t-1} + \upsilon_t$$
$$where\ \ \upsilon_t \sim IID(0, \sigma^2) \qquad (8.1)$$

Where Y_t is the variable (series) tested for stationarity, $\in$ is the stochastic error term which should satisfy the zero mean and unit variance conditions. Thus, the test equation is often estimated by the Ordinary Least Squares [OLS] technique under the null hypothesis of a unit root. If $\varphi = 1$, the series Y_t is non-stationary i.e. unit root exists because if $\varphi > 1$, the series Y_t is also non-stationary and if If $\varphi < 1$, the series Y_t is stationary. If $\varphi = 1$, the test equation degenerates to:

$$Y_t = \delta + Y_{t-1} + \upsilon_t$$

which is random walk model with drift and is non stationary. Therefore, the condition for stationarity is that $|\varphi| < 1$. To test for stationarity, the t-statistic is utilized to test whether or not $\varphi = 1$ since the series is not stationary when $|\varphi| \geq 1$. The relevant hypothesis for the test becomes:

$$H_0 : \varphi = 1$$
$$H_1 : \varphi < 1$$

The test statistic is:

$$t_{\varphi} = \frac{\hat{\varphi} - 1}{S_{\hat{\varphi}}}$$

Where $\hat{\varphi}$ is the estimated value of φ and $S_{\hat{\varphi}}$ is the standard error of $\hat{\varphi}$. Decision to accept H_0 is made if $t_{calculated} < t_{critical}$. Thus, H_0 is rejected if $t_{calculated} > t_{critical}$ with the conclusion that the series Y_t is stationary. The limitations of the preceding unit root test procedure include the fact that:

(a) The OLS estimator of φ has a downward bias in small sample because of the presence of lagged dependent variable acting as an explanatory variable,

(b) The distribution of the test-statistic is not normal [non-standard] even in large samples. As it were, the test ratio does not obey the standard t distribution neither is it asymptotically distributed with $N \sim [0,1]$. This is because stationarity was required in the derivation of the standard distribution.

Dickey and Fuller (1979) were the first to tackle these problems. They re-worked equation (8.1) by subtracting Y_{t-1} from both sides of equation (8.1) yields:

$$Y_t - Y_{t-1} = \delta + \varphi Y_{t-1} - Y_{t-1} + \upsilon_t$$
$$\Delta Y_t = \delta + (\varphi - 1) Y_{t-1} + \upsilon_t$$
$$Let\ \theta = (\varphi - 1)$$
$$\Delta Y_t = \delta + \theta Y_{t-1} + \upsilon_t$$

Accordingly, testing for $\varphi = 1$ in equation (8.1) is equivalent to testing for $\theta = 0$ in equation (8.3). In what follows, the new hypothesis to be tested becomes:

$$H_0 : \theta = 0$$
$$H_1 : \theta < 0$$

The test statistic is:

$$t_\theta = \frac{\hat{\theta}}{S_{\hat{\theta}}}$$

In the final analysis, Dickey and Fuller (1979) generated a limiting distribution based on Monte-Carlo experiments with empirical approximation. The term "unit root" is often the name for the test because it involves testing whether or not $\varphi = 1$. The preceding empirical implementation of the Dickey-Fuller unit root test for stationarity incorporates an $AR[1]$ process.

DF test equations are augmented with p-lagged values of the endogenous variable for the drifted and the deterministic trend auxiliary regression as an augmented Dickey Fuller [ADF] test equation given by:

$$\Delta Y_t = \delta + \gamma_t + \theta Y_{t-1} + \sum_{i=1}^{P} \varphi_i \Delta Y_{t-1} + \upsilon_{3t}$$

Where Δ is the difference operator, t is the time trend, $\in$ is the white noise error term which is independently and identically distributed with zero mean and constant variance. Depending on the number of lags the econometrician add to the test equation, utilizing 1 lag order, the one-lag $ADF[1]$ model will be obtained, utilizing 2 lag order, the two-lag $ADF[2]$ model is obtained etc. The lags are continuously added until serial correlation in the residuals of the test equations is eliminated.

Sarghan-Bhargava Method of Test

The Sarghan-Bhargava test is a test for stationarity based on the residual series from a regression. The following steps are involved in the Sarghan-Bhargava test for stationarity.

- Estimate the model:

$$Y_t = \delta + \varphi Z_t + \upsilon_t$$
$$where\ \ \upsilon_t \sim IID(0, \sigma^2)$$

- Save the residuals from the regression of Y_t on Z_t
- Regress the residuals series on its past values as in the auxiliary model:

$$\upsilon_t = c_0 + c_1 \upsilon_{t-1} + v_t$$

  Given that υ_t is acting as a lag regressor in the model, we difference the model for estimation as in the case below:

$$\Delta \upsilon_t = c_0 + (c_1 - 1)\upsilon_{t-1} + v_t$$

- Obtain the DW-statistic from the estimated model.
- Take statistical decision: If the computed DW is close to zero, there is no stationarity and if $\zeta = 0$, *where* $\zeta = (c_1 - 1)$, the series is not stationary co-integration as the model collapses to a random walk.

Co-integration

The need to test for the existence of a long-run relationship between the endogenous variable and its regressors informed the theory of co-integration as propounded by Granger (1981), Granger (1986) and Hendry (1986).

"Theorem" [see Engle and Granger (1987)]

Co-integration theorem holds in general that two variables are co-integrated of order (f,g) if they are both integrated of order f and there exist some linear combinations of them that are integrated of order $(f-g)$ where $(f>0)$. To illustrate ideas, we consider variables Y_t *and* Z_t for co-integration such that:

$$Y_t, Z_t \sim CI[f,g]$$

Suppose $Y_t \sim I(1)$ *and* $Z_t \sim I(1)$ and their linear combination is $W_t \sim I(0)$, then:

$$f-g=0$$
$$1-g=0$$
$$g>0$$

Given that $g>0$, the series are co-integrated and the order of co-integration is $Y_t, Z_t \sim CI[1,1]$. The implication is that two non-stationary series could be co-integrated if their linear combination is stationary. Co-integratiion is therefore a special case within the analysis of the order of integration. Thus the linear combination of an $I(0)$ with another $I(0)$ series will give an $I(0)$ series.

Also the combination of two $I(1)$ series with yield an $I(1)$ series. However a combination of an $I(1)$ series with an $I(0)$ series will give an $I(1)$ series meaning that a higher order series will dominate. In terms of regression analysis the regression of an $I(0)$ series on an $I(1)$ series will be non-stationary and the results will be spurious and statistically inconsistent.

Co-integration refers to long-run relationship between variables. It is therefore a method of avoiding both the spurious and inconsistent regression problems which otherwise occur with the regression of non-stationary series.

Statistical Tests for Co-integration

Two Variable Case of Co-integration Test:

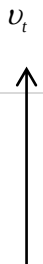
Engle-Granger Two-Step [EGTS] Modus Operandi

The Engle-Granger co-integration methodology is based on testing the OLS residual series for stationarity. According to Engle and Granger (1987), if the long-run relationship between two variables exists, the disequilibrium error should not drift far apart from the zero line. Let us consider the following bivriate co-integrating regression model:

$$M_t = \beta_0 + \beta_1 Z_t + \upsilon_t$$
$$where \ \upsilon_t \sim IID(0,\sigma^2)$$

It is expected that the mean of υ_t should be zero i.e. $E(\upsilon_t)=0$. This is illustrated graphically below:

Figure: Engle-Granger Theorem



Since the disequilibrium error, υ_t fluctuates very much around the mean, the variables in the static equation M_t *and* Z_t are stationary and as such, they are co-integrated. The linear combination of the two variables under the Engle-Granger methodology ought to be unique. As a result, when the coefficient of one of the variables is normalized to unity, a unique stationary linear combination is obtained.

- The following steps are involved in Engle-Granger test procedure:
  - Estimate the bivariate model using the OLS estimator
$$M_t = \beta_0 + \beta_1 Z_t + \upsilon_t$$
  - Save residual series, that is obtain
$$\hat{\upsilon}_t = M_t - \hat{\beta}_0 - \hat{\beta}_1 Z_t$$
  - Test residual series $\hat{\upsilon}_t$ for stationarity using the Dickey-Fuller [AD],

Augmented Dickey-Fuller [ADF], Phillips-Peron test statistics. If the residual series are stationary, there is co-integration. Given that OLS residuals have zero mean and we do not expect them to have deterministic trend. Therefore, both the intercept term and time trend are excluded in testing the residual series for stationarity.

Multivariate Case of Co-integration Test:

Johansen's Maximum Likelihood [JML] Approach

- The Johansen's Maximum Likelihood technique is a multivariate test for co-integration. This entails testing for co-integration between two or more variables. In essence, there could be more than one linear combinations that is stationary and hence, more than one co-integrating vector.

- In the Johansen's procedure, a test for the optimal leg length of the related vector auto-regression [VAR] has to be conducted. This is often necessitated because the JML is preceded by an estimation of a VAR model which in all respect should acquire the appropriate lag length. Indeed, the Johannes's co-integration test is highly sensitive to the appropriate lag length. In this regard, the AIC, FPE, LR, SIC, and HQ are often utilized in selecting the appropriate lag length. In most cases, the lag order supported by more of the five criteria for each equation is chosen as the appropriate lag length.

To save degrees of freedom, the highest lag in the testing down process of the lag length test is in most cases taken.

Error Correction Modelling and Estimation

- According to the Engle-Granger (1987) theorem, the short-run adjustment dynamics can be usefully described by the error correction model. This requires using the one-period lagged residual to correct for deviations of actual values from the long-run equilibrium values. The procedure is to use the residual series generated in the OLS regression that was used to test for stationarity to further reparametrize the dynamic short-run specification.

Over-parameterized ECM

An over-parameterized ECM model in log is often of the form:

$$\Delta LnM_t = \psi_0 + \sum_{i=1}^{k} \psi_i \Delta LnM_{t-1} + \sum_{i=0}^{k} \varpi_i \Delta LnZ_{t-1} \Phi ECM_{t-1} + \upsilon_t$$

$$\upsilon_t \sim IID(0, \sigma_{\upsilon_t}^2)$$

The ECM_{t-1} is one period lagged value of the error term, (Φ) is the adjustment coefficient which gives the percentage of disequilibrium between the long-run and the short-run values that is corrected for in a period, say in a year, a month etc.

Usefulness of the Error Correction Coefficient

- The error correction coefficient integrates both the short-run and the long-run dynamics thereby taking into cognizance the information lost during the time of differencing the variables. This is the Box-Jekins' methodology.
- The error correction coefficient has the advantage of a linear feed-back between the conditional mean and the conditional variance of changes in policy variables. Accordingly, it allows for the possibility of dynamic adjustment transmitted through the one-period lagged error term to the next period.

ASSIGNMENT

Explain Johansen's Maximum Likelihood [JML] Approach for testing for co-integration

CONCLUSION

The Engle-Granger (1987) representation theorem formally established the theoretical basis for error correction modelling.