Developing Research Toolkit for Economists (551376)

Dr. Keshab Bhattarai
University of Hull Business School, Hull, England, UK*

December 30, 2022

Abstract

This workbook on Research Toolkits for economists aims to present basic econometric methods that economists have developed over years for testing various propositions of standard economic theories. It focuses on basics of OLS estimators, their applications, problems including multicollinearity, heteroskedasiticity and autocorrelation and remedial measures, unit root and cointegration, simultaneous equations and panel data analysis. Tutorial problems and assignments are integral parts of the module and models discussed here could be applied to analyse economic issues and to write research reports or journal articles.

JEL Classification: C

Keywords: Research toolkits, Econometrics, Empirical Economics

Contents

*Hull University, HU6 7RX, Yorkshire, Hull, UK. email: K.R.Bhattarai@hull.ac.uk.

List of Tables

1 What is research toolkit for economists?

The major objective of research in economics is to find out the truth about economic questions that is bothering individual households, communities, policy makers in the local and national governments or the international community as a whole. Some questions are quantitative by nature such as the distribution of income, employment level by sectors, prices and costs of commodities, demand and supply of various goods and services in the economy, international trade, growth rates of output employment, capital stock, investment, rate of returns on financial assets, primary, secondary or the university level education. Others are qualitative and philosophical. Effective analyses of economic issues such as the efficiency of firms, the welfare of individuals or households require both quantitative and qualitative techniques. The major aim of these is to understand the behavioural and psychological factors that underpin the decision making process of individuals and firms or the policy makers and be able to predict the course of variables in coming years.

Economists have developed many theories regarding how the various markets function or should function. How the various pieces of economic activities make the national or international economy. Economic research therefore is divided into two main groups 1) theoretical research 2) applied research. Theoretical research often involves derivation of demand, supply and equilibrium conditions using some sort of optimising process. Diagrams, equations or simply the logical statements are often used for theoretical deduction. Standard micro or macroeconomic models (or extensions of those models in various fields of economics) are applied to study how consumers and producers optimise and how those determine prices for goods and services or factors of production in related markets. The general equilibrium models quantify the entire economy. Intertemporal models show the process of accumulation, investment and growth. Statistical inferences based on marginal or cumulative distributions of populations, samples with law of large numbers are used to test claims of these theories. Abstract models require algebra, calculus, matrix, econometrics, real analysis or stochastic probability theory to represent and test these theoretical ideas.

Theories need to be applied in practice to make them useful for improvement in the welfare of human society. The application involves systematic collection of information on variables identified by the relevant theory. Empirical research tests the claims made by those theories stated in linear or non-linear functions using various estimation or computation techniques. As the amount of information has grown so has the need to process the information. The applied research is basically about processing information consistently, coherently, systematically using inductive methods. Applied research can also vary according to the nature of method used in analysis. There are mainly four categories of applied research: 1) statistical and econometric analysis 2) calibration and computations of system of equations 3) strategic analysis 4) experimental analysis. Statistical analysis involves designing, implementing and collecting data on economic variables scientifically in an unbiased manner. This also involves determining the properties of distribution of those variables, collecting information on central tendencies, finding correlations and the pattern of causality among variables. Econometric analysis involves techniques and applications to process data for testing various economic theories based on cross sections and time series data. Calibration and computation of system of equations involves solving N number of equations on the basis of certain assumption about their behaviour, such as market demand and supply functions, or input-output analysis or a general equilibrium system. Linear, non-linear or dynamic programming is often used to determine such a system. Game theory is becoming increasingly popular tool to analyse inter-dependence among economic agents where the action to be taken by one is determined by the beliefs or perception of that individual about the action taken by other people in the economy. They are applied to analyse the process and outcome of bargaining, strategic contingency planning or just in describing the behaviours of economic agents. Experimental analysis has the concept of using control groups for testing economic theories, such as impacts of certain policy in economic stability, such as the adoption of the euro, effect of certain drugs, or certain measures on productivity, health or educational attainment.

The broader aim of the research toolkit module is to raise level of confidence of students in economics to engage themselves in analysis of these challenging problems around the world and urge them to develop skills that enable them to exchange ideas efficiently for the benefit of humankind by maintaining steady progress of our economies and to contribute to improve the living standards of millions of people around the globe. Given the tight time framework this module will focus on the basics of econometric techniques. This workbookd ncludes aims and objectives of the module, teaching programme and learning outcomes, assignments, essential readings, contact addresses of relevant staff and other information essential for students. Essentially this complements various other modules.

Introduction to the empirical economics by an example:

Consider a problem of a consumer who maximises utility subject to the budget constraint as:

$$\max U = C_1^{\alpha} C_2^{1-\alpha} \quad \text{s.t. } P_1 C_1 + P_2 C_2 = I \tag{1}$$

Where U denotes utility, C_1 and C_2 amount of goods 1 and 2 respectively with P_1 and P_2 as their prices,α is the share of income spent in good 1 and $(1-\alpha)$ is in good 2, I is income. Following the optimisation (i.e. after solving the first order conditions of Lagrangian constrained optimisation) this results in separate demand functions for two goods as:

$$C_1 = \frac{\alpha I}{P_1} \tag{2}$$

$$C_2 = \frac{(1-\alpha) I}{P_2} \tag{3}$$

Now consider estimating only the demand for C_1. Taking log both sides of the first demand function one gets an estimable demand function:

$$\ln C_1 = \ln \alpha + \ln I - \ln P_1 \tag{4}$$

For a market demand function assume that there are $i = 1...n$ consumers in a community. Define variables for a standard regression model as:

$\ln C_{1,i} = Y_i$; $\beta_1 = \ln \alpha + \ln I$; $\ln P_{1,i} = \beta_2 X_i$. Then a simple linear regression model of standard demand function based economic theory of consumer optimisation is:

$$Y_i = \beta_1 + \beta_2 X_i + e_i \tag{5}$$

Here β_1 measures income effect as well as preference for good one (α) by consumers, slope β_2 measures the impact of price on demand; e_i includes all other elements (omitted variables, misspecification bias). Expected signs $\beta_1 > 0$ and $\beta_2 < 0$. Income (I_i) can be explicitly included and it is expected to have positive impact on demand ($\beta_3 > 0$)

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 \ln I_i + e_i \tag{6}$$

This is an example on the theoretical basis of a linear regression model[1]. For many products (i.e. car, house, food, clothes, sports) demand is determined by their prices of products and income of consumers. Empirical economics discusses techniques to test propositions of economic theory from the evidence available in the real data on variables Y_i, X_i and I_i by estimating (or calibrating) parameters.

1.0.1 Guidelines for Empirical Research

1. Construction of hypotheses: specify one or more equations of a model based on economic theory.
   (a) Macroeconomic theory:
      i. determinants of growth across regions or countries
      ii. total factor productivity
      iii. growth, inequality and environment
      iv. models of fluctuation or business cycle (interest rate rule)
      v. determinants of consumption, saving, investment, exports, imports, output, employment
      vi. inflation and unemployment
      vii. interest rate, exchange rate, inflation, trade balance,
      viii. deposit expansion, credit flows
   (b) Microeconomic theory
      i. Demand for a commodity (necessary, luxury or normal goods; durable or perishable items)
      ii. Cost of production of certain commodity (car, plane, computer, TV, electronic goods, machines)
      iii. Market price for a certain commodity (rice, wheat, maize, millet; meat and other livestock products)
      iv. Profits, revenue of a certain company or industry (Barclays, British Airways, BT, Train/Bus, Low cost airlines)
      v. Wage rates, rental rate of capital, salary of executives (job satisfaction and performance)
      vi. structure of markets (competition, monopoly, oligopoly or monopolistic compitition)
      vii. foreign direct investment (joint ventures, franchising, licensing)
      viii. merger and acquisition; economies of scale (concentration ratios)
      ix. research and development; new management practices; business models
      x. Efficiency and welfare

[1] Parameters of the systems of equations are estimated together using a simultaneous equation model or a VAR model. This is discussed later.

(c) Trade theory: trade, prices, wages, capital flows; revealed comparative advantage, gravity of trade ; terms of trade, real and nominal exchange rates

(d) Public finance

   i. Determinants of revenue and spending
   ii. Budget deficit, public debt
   iii. Impact of taxes on demand and supply of goods and products; income distribution and welfare
   iv. Impacts of taxes, spending and debt on long run growth and short run fluctuations

(e) Environment

   i. Determinants of pollution and costs of abatement
   ii. Global warming and carbon footprints

(f) Employment and labour markets

   i. Demand and supply of labour
   ii. Employment/labour force, population
   iii. skill, qualifications and performance
   iv. Unemployment and inflation
   v. Earnings, wages and work hours, pensions
   vi. Migration, remittances

(g) Finance

   i. Savings and accumulation of wealth by households and investment by firms
   ii. Deposits, loans and interest rate spreads
   iii. Cost and benefit and networth of projects
   iv. Optimal portfolio, allocation, risk free return, return on financial assets
   v. Prices of stocks, bonds; spot, option and future prices
   vi. Liquidity and efficiency of the financial system and growth
   vii. Foreign exchange market, commodity markets

(h) Economic development

   i. Education and manpower development
   ii. Investment and growth
   iii. Human capital and productivity
   iv. Structural transformation
   v. Estimation of surplus labour in dual labour market
   vi. Gini coefficient

Theoretical derivation requires using first and second order conditions and solving a system of equations of a model; survey experiments.

2. Represent theory using a set of interlinked diagrams. Indicate optimum conditions in the diagram.

1. Preparation of data

Once clear about the hypothesis, required data can be obtained by primary survey or downloaded from standard secondary sources like

1. (a) Economic and social data (ESDS) www.esds.ac.uk/international
   (b) datastream
   (c) http://finance.yahoo.com/
   (d) http://www.unctad.org; www.wto.org
   (e) Eurostat: http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes
   (f) BHPS: http://www.esds.ac.uk/government/; http://www.esds.ac.uk/government/surveys/
   (g) http://www.data-archive.ac.uk/findingData/bhpsTitles.asp
   (h) http://www.statistics.gov.uk
   (i) http://www.economicsnetwork.ac.uk/links/data_free#uk
   (j) see links at http://www.hull.ac.uk/php/ecskrb/Confer/research.html
3. Data files from http://www.data-archive.ac.uk/
   (a) Excel or CSV format for PcGive.
   (b) Large scale data are available on SPSS or
   (c) STATA format (*.dat).

You can get lost at this stage if you are not precise on the research question or the hypothesis. Be focused and get only data you require. Ask for help if confused.

4. Estimations and interpretation of results
   (a) Specify equations according to the hypothesis set in stage 1
   (b) Derive equations for the OLS estimators
   (c) Examine the properties of those estimators - do they have expected signs, are they significant? what do they mean?
   (d) Mean and variance of those estimators; reliability and confidence interval for estimated parameters.
   (e) Estimate parameters using the data collected above.
5. Analyse variance by R^2, t, F and χ^2 tests as appropriate
   (a) Write analytical forms for t, F and χ^2 tests.

(b) Determine the degrees of freedom and critical values from the theoretical tables.

(c) Compare empirical results with those theoretical values and determine the significance.

(d) Take a decision regarding significance of the model or coefficient based on these tests.

(e) Think of improving the model based on tests.

(f) see http://www.medcalc.org/manual/t-distribution.php.

6. Tests of multicollinearity

(a) Write estimators at least with two explanatory variables.

(b) Show how the estimator breaks down with perfect multicollinearity.

(c) Determine variables that are correlated to each other based on analysis of correlation among explanatory variables.

(d) Determine the variance inflation factor.

(e) Drops correlated variables and re-estimate the model until getting the sensible results.

7. Restrictions and dummy variables

(a) Consider theoretically appropriate restrictions in the model.

(b) Write analytical forms of F-test that can be used to test restrictions.

(c) Determine the validity of restrictions.

(d) Introduce dummy variables to capture structural changes in time series or individual effects in the cross section Analysis.

8. Heteroskedasticity

(a) Write the analytical form of the heteroskedasticity problem.

(b) Show how the properties of the OLS estimators are affected by presence of heteroskedasticity.

(c) Write test statistics to detect heteroskedasticity.

(d) Transform the model to remove heteroskedasticity.

(e) Construct a cross section data appropriate for heteroskedacity analysis.

(f) Interpreted meaning of the heteroskedasticity consistent standard errors.

9. Autocorrelation

(a) Find causes why there is autocorrelation and consequences of it.

(b) Write analytical form of the Durbin-Watson Statistics.

(c) Show how the properties of the OLS estimators are affected by presence of autocorrelation.

(d) Estimate the model with AR(1) autocorrelation.

(e) Transform the model to remove autocorrelation using iterative procedure.

10. Stationarity

(a) Explain when a variable is stationary and when non-stationary.

(b) Show the impact of non-Stationarity in the variance of the variable in an AR(1) model.

(c) What is Dickey-Fuller test and Augmented Dickey-Fuller test? Phillip-Peron tests.

(d) Determine whether your series are stationary based on DF and ADF statistics.

11. Causality and cointegration

(a) Show the procedure for Granger causality test.

(b) What is order of integration and what is cointegration?

(c) Show analytical forms to test cointegration.

(d) Determine cointegration in a single equation model.

13. Write an essay or article based on research experience gained in following steps 1 to 11

2 Linear Regression Model

- Consider a linear regression model:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \qquad i = 1...N \tag{7}$$

Error term (ε_i) represents all missing elements from this relationship; a positive error term indicates that true observation is above the fitted line and a negative error term indicates that actual observation is below the fitted line. On average these pluses and minuse errors cancel out resulting in a mean zero value for each error. Therefore the Gauss-Markov thoerem assumes that these errors (ε_i) are normally distributed random variables with a zero mean and a constant variance, σ^2 , represented as follows:

$$\varepsilon_i \sim N\left(0, \sigma^2\right) \tag{8}$$

- Normal equations of above regression (derivation is given in the next section):

$$\sum Y_i = \widehat{\beta}_1 N + \widehat{\beta}_2 \sum X_i \tag{9}$$

$$\sum Y_i X_i = \widehat{\beta}_1 \sum X_i + \widehat{\beta}_2 \sum X_i^2 \tag{10}$$

Each dot in the above graph represents an observation. Some observations lie above the least square $\widehat{Y}_i$ line and other observations lie below it. These errors represent all sorts of elements missing from this relationship. Some of them might be due to the missing variables, others might be due to measurement errors, still other may be from the mis-specification of the relationship. The least square line is the line of best fit; line that fits the data set with minimum sum of square of errors. Differences between each observation and the line $\widehat{Y}_i$ is represented by error terms e_i. As some of them are above the line and others below the line, positive errors cancel out with the negative errors. Note that the least square line passes through the average values of variables X and Y (prove it).

A system method of estimation is required when variables are endogenously related to each other (see sections of simultaneous equation and VAR for such models).

2.0.2 Ordinary Least Square (OLS): Assumptions

List the OLS assumptions on error terms e_i .

1) Normality of Errors

$$E(\varepsilon_i) = 0 \tag{11}$$

2) Homoskedasticity

$$var(\varepsilon_i) = \sigma^2 \quad for \ \forall \ i \tag{12}$$

3) No autocorrelation

$$covar(\varepsilon_i \varepsilon_j) = 0 \tag{13}$$

4) Independence of errors from dependent variables

$$covar(\varepsilon_i X_i) = 0 \tag{14}$$

Details on what happens if above assumptions are violated is explained in sections of heteroskedasticity, autocorrelation and specification bias.

2.0.3 Derivation of normal equations for the OLS estimators

Standard problem of the ordinary least square (OLS) method is to choose parameters $\widehat{\beta}_1$ and $\widehat{\beta}_2$ to minimise sum of square errors as:

$$\underset{\widehat{\beta}_1\widehat{\beta}_2}{Min}\ S = \sum \varepsilon_i^2 = \sum \left(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{1,i}\right)^2 \tag{15}$$

First order conditions of minimisation are[2]:

$$\frac{\partial S}{\partial \widehat{\beta}_1} = 0; \frac{\partial S}{\partial \widehat{\beta}_2} = 0; \tag{16}$$

$$\sum \left(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_i\right)(-1) = 0 \tag{17}$$

$$\sum \left(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_i\right)(-X_i) = 0 \tag{18}$$

Normal equations are obtained by re-ordering these first order conditions as:

$$\sum Y_i = \widehat{\beta}_1 N + \widehat{\beta}_2 \sum X_i \tag{19}$$

$$\sum Y_i X_i = \widehat{\beta}_1 \sum X_i + \widehat{\beta}_2 \sum X_i^2 \tag{20}$$

There are two unknown $\widehat{\beta}_1$ and $\widehat{\beta}_2$ and two equations. One way to find $\widehat{\beta}_1$and $\widehat{\beta}_2$ is to use the substitution and reduced form method.

Determine the slope estimator by the reduced form equation method as follows:

Multiply the second equation by N and first by $\sum X_i$

$$\sum X_i \sum Y_i = \widehat{\beta}_1 N \sum X_i + \widehat{\beta}_2 \left(\sum X_i\right)^2 \tag{21}$$

$$N \sum Y_i X_i = \widehat{\beta}_1 N \sum X_i + \widehat{\beta}_2 N \sum X_i^2 \tag{22}$$

By subtraction this reduces to

$$\sum X_i \sum Y_i - N \sum Y_i X_i = \widehat{\beta}_2 \left(\sum X_i\right)^2 - \widehat{\beta}_2 N \sum X_i^2 \tag{23}$$

$$\widehat{\beta}_2 = \frac{\sum X_i \sum Y_i - N \sum Y_i X_i}{\left(\sum X_i\right)^2 - N \sum X_i^2} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{24}$$

This is the OLS Estimator of $\widehat{\beta}_2$, the slope parameter; $\frac{\partial Y}{\partial X} = \widehat{\beta}_2$. It measures how much change in Y occurs after a change in one unit of X.

[2] Notice that there are as many first order conditions as many parameters to be estimated.

When $\widehat{\beta}_2$ is known, it is easy to find the intercept estimator $\widehat{\beta}_1$ by averaging out the regression $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ as:

$$\widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2 \overline{X} \tag{25}$$

Proof for deviation method:
$\frac{\sum X_i \sum Y_i - N \sum Y_i X_i}{(\sum X_i)^2 - N \sum X_i^2} = \frac{\sum x_i y_i}{\sum x_i^2}$;

$$\begin{aligned} LHS &= \frac{\sum X_i \sum Y_i - N \sum Y_i X_i}{\left(\sum X_i\right)^2 - N \sum X_i^2} = \frac{N\overline{X} N\overline{Y} - N \sum Y_i X_i}{\left(N\overline{X}\right)^2 - N \sum X_i^2} \\ &= \frac{N\overline{X} N\overline{Y} - N \sum Y_i X_i}{\left(N\overline{X}\right)^2 - N \sum X_i^2} = \frac{N\overline{XY} - \sum Y_i X_i}{N\overline{X}^2 - \sum X_i^2} = \frac{\sum Y_i X_i \ - \ N\overline{XY}}{\sum X_i^2 \ - \ N\overline{X}^2} \\ &= \frac{\left(\sum Y_i \text{ - } \overline{Y}\right)\left(\sum X_i \text{ - } \overline{X}\right)}{\left(\sum X_i \text{ - } \overline{X}\right)^2} = \frac{\sum x_i y_i}{\sum x_i^2} = RHS; \quad QED. \end{aligned} \tag{26}$$

Matrix method of finding $\widehat{\beta}_1$ and $\widehat{\beta}_2$ is more general and convenient; particularly useful for a multiple regression model (review determinant and inverse of a matrix in the matrix section).

2.0.4 Normal equations in matrix form

Let Y be $N \times 1$ vector of a dependent variable (regressor); X be $N \times K$ matrix of independent variables (regressands); β be $K \times 1$ vector of unknown paramerters e be $N \times 1$ vectro of errors. That means data and parameters in matrix are:

$$Y = \begin{bmatrix} y_1 \\ y_1 \\ \\ y_1 \end{bmatrix}; \; X = \begin{bmatrix} 1 & x_{1,1} & .. & x_{k,1} \\ 1 & x_{1,2} & .. & x_{k,2} \\ 0 & & & \\ 1 & x_{1,N} & .. & x_{k,N} \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ \beta_k \end{bmatrix}; \quad e = \begin{bmatrix} e_1 \\ e_1 \\ \\ e_1 \end{bmatrix}.$$

Then a regression model can be written as:

$$Y = X\beta + e \tag{27}$$

$$\widehat{\beta} = \left(X'X\right)^{-1} X'Y \tag{28}$$

For a case of simple regression model with $k = 2$:

$$y_1 = \beta_0 + \beta_1 x_1 + e_1 \tag{29}$$

$$.. \tag{30}$$

$$y_{_N} = \beta_0 + \beta_1 x_{_N} + e_{_N} \tag{31}$$

Sum and cross products are required to evaluate the normal equations:

$$\sum Y_i = \widehat{\beta}_1 N + \widehat{\beta}_2 \sum X_i \tag{32}$$

$$\sum Y_i X_i = \widehat{\beta}_1 \sum X_i + \widehat{\beta}_2 \sum X_i^2 \tag{33}$$

Represtantion of above normal equations in matrix form:

$$\begin{bmatrix} \sum Y_i \\ \sum Y_i X_i \end{bmatrix} = \begin{bmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}; \;\; \widehat{\beta} = (X'X)^{-1} X'Y \tag{34}$$

Estimators in matrix form:

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum Y_i X_i \end{bmatrix} \tag{35}$$

Consider a small example with only 8 data observations.

2.0.5 Data Table: An Example

DATA		
y	Contant	x
4	1	5
6	1	8
7	1	10
8	1	12
11	1	14
15	1	17
18	1	20
22	1	25

$\Rrightarrow$

$$\begin{bmatrix} \text{Y} \\ \\ 4 \\ 6 \\ 7 \\ 8 \\ 11 \\ 15 \\ 18 \\ 22 \end{bmatrix}_{8\times 1} = \begin{bmatrix} \text{X} & \\ & \\ 1 & 5 \\ 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \\ 1 & 17 \\ 1 & 20 \\ 1 & 25 \end{bmatrix}_{8\times 2} \begin{pmatrix} \beta \\ \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix}_{2\times 1} + \begin{bmatrix} e \\ \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix}_{8\times 1}$$

Derivation of OLS Estimators:
Matrix multiplication: $\left[\left(X^{'}X\right) = \begin{array}{cc} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{array}\right]$

$$\left(X^{'}X\right) = \underset{2\times 8}{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 5 & 8 & 10 & 12 & 14 & 17 & 20 & 25 \end{bmatrix}} \underset{8\times 2}{\begin{bmatrix} 1 & 5 \\ 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \\ 1 & 17 \\ 1 & 20 \\ 1 & 25 \end{bmatrix}} = \underset{2\times 2}{\begin{bmatrix} 8 & 111 \\ 111 & 1843 \end{bmatrix}} \tag{36}$$

OLS in Matrix

$$\left(X^{'}Y\right) = \underset{2\times 8}{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 5 & 8 & 10 & 12 & 14 & 17 & 20 & 25 \end{bmatrix}} \underset{8\times 1}{\begin{bmatrix} 4 \\ 6 \\ 7 \\ 8 \\ 11 \\ 15 \\ 18 \\ 22 \end{bmatrix}} = \underset{2\times 1}{\begin{bmatrix} 91 \\ 1553 \end{bmatrix}} \tag{37}$$

2.0.6 Summary of Data

$$\begin{bmatrix} \sum Y_i = 91 \\ \sum Y_i X_i = 1553 \end{bmatrix} = \begin{bmatrix} N = 8 & \sum X_i = 111 \\ \sum X_i = 111 & \sum X_i^2 = 1843 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} \tag{38}$$

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 8 & 111 \\ 111 & 1843 \end{bmatrix}^{-1} \begin{bmatrix} 91 \\ 1553 \end{bmatrix} \tag{39}$$

Solving by the Cramer's rule
Determinant (cross-product)

$$|X'X| = \begin{vmatrix} 8 & 111 \\ 111 & 1843 \end{vmatrix} = (8 \times 1843) - (111 \times 111) = 2423 \tag{40}$$

2.0.7 Estimates

$$\widehat{\beta}_1 = \frac{1}{2423}\begin{vmatrix} 91 & 111 \\ 1553 & 1843 \end{vmatrix} = \frac{167713 - 172383}{2423} = \frac{-4670}{2423} = -1.9274 \tag{41}$$

$$\widehat{\beta}_2 = \frac{1}{2423}\begin{vmatrix} 8 & 91 \\ 111 & 1553 \end{vmatrix} = \frac{12424 - 10101}{2423} = \frac{2323}{2423} = 0.9587 \tag{42}$$

You can evaluate the inverse of a matrix in Excel easily using following steps:

1. select the cell where to put the result and press shift and control continously by two fingers of left hand
2. use mouse by right hand to choose math and trig function
3. choose MINVERSE
4. Select matrix for which to evaluate the determinant
5. press OK and you will see the reslut.

To evaluate a determinant - select a cell where you want to put the result, then choose MDETERM; select the matrix, then press ok.

For matrix multiplication follow conformability of matrix multiplication. This means number of columns in the first matrix should equal the number of rows in the second matrix. When a X matrix of order $N \times K$ is multiplied to a B matrix of order $K \times 1$ then the resutling matrix Y will be of $N \times 1$ order, K cancels out.

2.0.8 Predicted Y

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_i \implies \widehat{Y}_i = -1.9274 + 0.9587 X_i \tag{43}$$

Both slope and intercepts make economic sense. In this sample expenditure on food (Y) is determined by weekly income of an individual (X), people spend 95.6% percent of their weekly income in food expenditure. Poor people who do not have any income, receive a subsidy of 1.93 pence per week.

- Mean prediction

We can use this estimated equation to find the predicted value $\widehat{y}_i$ for each observation of x_i. If the weekly income is 40, predicted food expenditure will be 36.42.

$\widehat{Y}_i = -1.9274 + 0.9587 X_i = -1.9274 + 0.9587\,(40) = 36.42$

Predicted values of Y_i for the entire sample is as follows:

$$\widehat{Y}_1 = -1.9274 + 0.9587\,(5) = 2.866 \tag{44}$$

$$\widehat{Y}_2 = -1.9274 + 0.9587\,(8) = 5.742 \tag{45}$$

$$\widehat{Y}_3 = -1.9274 + 0.9587\,(10) = 7.660 \tag{46}$$

$$\widehat{Y}_4 = -1.9274 + 0.9587\,(12) = 9.577 \tag{47}$$

$$\widehat{Y}_5 = -1.9274 + 0.9587\,(14) = 11.495 \tag{48}$$

$$\widehat{Y}_6 = -1.9274 + 0.9587\,(17) = 14.371 \tag{49}$$

$$\widehat{Y}_7 = -1.9274 + 0.9587\,(20) = 17.247 \tag{50}$$

$$\widehat{Y}_8 = -1.9274 + 0.9587\,(25) = 22.041 \tag{51}$$

Error terms are also estimated using the definition of error as:
$\widehat{e}_i = Y_i - (-1.9274) - 0.9587X_i = Y_i + 1.9274 - 0.9587X_i$

$$\widehat{e}_i = Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_i \;\; = Y_i - (-1.9274 + 0.9587X_i) \tag{52}$$

$$\widehat{e}_1 = 4 + 1.9274 - 0.9587\,(5) = 1.134 \tag{53}$$

$$\widehat{e}_2 = 6 + 1.9274 - 0.9587\,(8) = 0.258 \tag{54}$$

$$\widehat{e}_3 = 7 + 1.9274 - 0.9587\,(10) = -0.660 \tag{55}$$

$$\widehat{e}_4 = 8 + 1.9274 - 0.9587\,(12) = -1.580 \tag{56}$$

$$\widehat{e}_5 = 11 + 1.9274 - 0.9587\,(14) = -0.495 \tag{57}$$

$$\widehat{e}_6 = 15 + 1.9274 - 0.9587\,(17) = 0.629 \tag{58}$$

$$\widehat{e}_7 = 18 + 1.9274 - 0.9587\,(20) = 0.753 \tag{59}$$

$$\widehat{e}_8 = 22 + 1.9274 - 0.9587\,(25) = 0.000 \tag{60}$$

- Use of regression estimates to calculate the elasticities

The definition of elasticity of food expenditure on income evaluated at the mean values of Y and X is given by

$$\eta = \frac{\frac{\partial Y}{Y}}{\frac{\partial X}{X}} = 0.9587 \times \frac{13.857}{11.375} = 1.1683 \tag{61}$$

This suggests that the expenditure on food is elastic around the mean. There will be £1.17 pence more expenditure to every £1 rise in weekly income.

$$\begin{aligned} TSS &= \sum \left[Y_i - \overline{Y}_i\right]^2 = \sum \left[\widehat{Y}_i - \overline{Y}_i + \;\widehat{e}_i\right]^2 \\ &= \sum \left(\widehat{Y}_i - \overline{Y}_i\right)^2 + \sum \widehat{e}_i^2 + 2\sum \left(\widehat{Y}_i - \overline{Y}_i\right)\widehat{e}_i \\ &= \sum \left(\widehat{Y}_i - \overline{Y}_i\right)^2 + \sum \widehat{e}_i^2 \quad \because 2\sum \left(\widehat{Y}_i - \overline{Y}_i\right)\widehat{e}_i = 0 \end{aligned} \tag{62}$$

Consider $2\sum \left(\widehat{Y}_i - \overline{Y}_i\right)\widehat{e}_i = 2\sum \left(X\widehat{\beta} - \overline{Y}_i\right)\widehat{e}_i = 2\left(X\widehat{\beta} - \overline{Y}_i\right)\sum \widehat{e}_i = 0$; this is try by assumption $covar\,(\varepsilon_i X_i) = 0$.

[Total variation] = [Explained variation] + [Residual variation]

[Total variation] = [Regression Sum Square] + [Residual sum square]

$$TSS = RSS + ESS \tag{63}$$

2.0.9 Degrees of freedom (df)

Degrees of freedom for N observations and K explanatory variables:

[Total variation] = [Explained variation] + [Residual variation]

[Total variation] = [Regression Sum Square] + [Residual sum square]

df : N-1 K-1 N-K

2.0.10 Variances

$\sum \widehat{e}_i^2 = \widehat{e}_1^2 + \widehat{e}_2^2 + \widehat{e}_3^2 + \widehat{e}_4^2 + \widehat{e}_5^2 + \widehat{e}_6^2 + \widehat{e}_7^2 + \widehat{e}_8^2 = (1.134)^2 + (0.258)^2 + (-0.660)^2 + (-1.580)^2$

$+ (-0.495)^2 + (0.629)^2 + (0.753)^2 + (0.000)^2 = 5.484$

$$var\,(\,\widehat{e}_i) = E\left(\widehat{\varepsilon}_i^2\right) = \frac{\sum \widehat{e}_i^2}{N-k} = \widehat{\sigma}^2 \tag{64}$$

Where k is the number of parameters in the regression; N is the number of observations.

$$\frac{\sum \widehat{e}_i^2}{N-k} = \frac{5.4841}{8-2} = 0.914 \tag{65}$$

Easy way to calculate total sum squares:

$$\sum y_i^2 = \sum \left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - N\overline{Y}^2 = 1319 - 8 \times 11.375^2 = 283.875 \tag{66}$$

$$\sum x_i^2 = \sum \left(X_i - \overline{X}\right)^2 = \sum X_i^2 - N\overline{X}^2 = 1843 - 8 \times 13.875^2 = 302.875 \tag{67}$$

2.0.11 R-square and F Statistic

Regression Y on X is the line of the best fit for the data; R^2 indicates how well the model fits to the data. It is called the coefficient of determination. It is given by $R^2 = \frac{\sum \widehat{y}_i^2}{\sum y_i^2} = \frac{\text{Regression sum square}}{\text{Total sum square}}$. Its value is between zero and one, $0 \leq R^2 \leq 1$. That $R^2 = 1$ means that the model explains everything and $R^2 = 0$ means that it does not explain anything. Most often it is in between these two extremes; the higher the value of R^2 better is the fit of the model.

$$\sum \widehat{y}_i^2 = \widehat{\beta}_2^2 \sum x_i^2 = 0.9587^2 \times 302.875 = 278.390 \tag{68}$$

Coefficient of determination is a measure in the regression analysis that shows the explanatory power of independent variables (regressors) in explaining the variation on dependent variable (regressand). The total variation on the dependent variable can be decomposed as following:

$$R^2 = \frac{\sum \widehat{y}_i^2}{\sum y_i^2} = \frac{278.390}{283.875} = 0.981 \tag{69}$$

For N observations and K explanatory variables

[Total variation (TSS)] = [Explained variation (RSS)] + [Residual variation (ESS)]

df = N-1 K-1 N-K

$$F = \frac{RSS/(K-1)}{ESS/(N-k)} = \frac{\frac{278.390}{1}}{\frac{5.4841}{6}} = \frac{278.390}{0.9140} = 304.579 \tag{70}$$

$$\overline{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K} = 1 - (1 - 0.981)\frac{8-1}{8-2} = 0.978 \tag{71}$$

$R^2 > \overline{R}^2$

Prove that two forms $\overline{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K}$ or $\overline{R}^2 = R^2\frac{N-1}{N-K} - \frac{K-1}{N-K}$ are equivalent.

Relation between Rsquare and Rbarsquare Prove that two forms $\overline{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K}$ or $\overline{R}^2 = R^2\frac{N-1}{N-K} - \frac{K-1}{N-K}$ are equivalent.

Proof

$$\begin{aligned}
LHS &= \overline{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K} = R^2 + \left(1 - R^2\right) - \left(1 - R^2\right)\frac{N-1}{N-K} \\
&= R^2 - \left(1 - R^2\right)\left[\frac{N-1}{N-K} - 1\right] = R^2 + \left(1 - R^2\right)\left[\frac{N-1-N+K}{N-K}\right] \\
&= R^2 - \left(1 - R^2\right)\left[\frac{K-1}{N-K}\right] = R^2 + R^2\frac{K-1}{N-K} - \frac{K-1}{N-K} \\
&= R^2\left(1 + \frac{K-1}{N-K}\right) - \frac{K-1}{N-K} \\
&= R^2\left(\frac{N-K+K-1}{N-K}\right) - \frac{K-1}{N-K} = R^2\left(\frac{N-1}{N-K}\right) - \frac{K-1}{N-K} \\
&\quad RHS; \quad QED
\end{aligned} \tag{72}$$

2.0.12 Variance, Standard Error and t-value of Slope Parameter

$$Var\left(\widehat{\beta}_2\right) = var\left[\frac{\sum\left(X_i - \overline{X}\right)}{\sum\left(X_i - \overline{X}\right)^2}\right] var\left(y_i\right) = \frac{1}{\sum x_i^2}\widehat{\sigma}^2 \tag{73}$$

Using 67 and 65 above:

$$var\left(\widehat{\beta}_2\right) = \frac{0.914}{302.875} = 0.0030 \tag{74}$$

Standard error of $\widehat{\beta}_2$:

$$SE\left(\widehat{\beta}_2\right) = \sqrt{0.0030} = 0.0548 \tag{75}$$

t-value of $\widehat{\beta}_2$ from the sample :

$$t_{\widehat{\beta}_2} = \frac{\widehat{\beta}_2 - \beta_2}{SE\left(\widehat{\beta}_2\right)} = \frac{0.9587 - 0}{0.0548} = 17.495 \tag{76}$$

Compare it to the the table of value of t for degrer of freedom N-K = 8-2=6 for chosen level of significance (α); usually α is either 1 percent (more accurate) or 5 percent (acceptable); but α really depends on the degree of precision required in accepting or rejecting the hypothesis.

Variance, Standard Error and T value of Intercept Parameter

$$var\left(\widehat{\beta}_1\right) = \left[\frac{1}{N} + \frac{\overline{X}^2}{\sum x_i^2}\right]\widehat{\sigma}^2 \tag{77}$$

$$var\left(\widehat{\beta}_1\right) = \left[\frac{1}{8} + \frac{13.875^2}{302.875}\right] \times 0.914 \tag{78}$$

$$var\left(\widehat{\beta}_1\right) = [0.125 + 0.634] \times 0.914 = 0.6937 \tag{79}$$

$$SE\left(\widehat{\beta}_1\right) = \sqrt{0.6937} = 0.833 \tag{80}$$

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)} = \frac{-1.9774 - 0}{0.833} = -2.374 \tag{81}$$

Matrix method is easier to calcuate variance of a parameter:

$$cov\left(\widehat{\beta}\right) = \left(X'X\right)^{-1}\widehat{\sigma}^2 = \begin{bmatrix} 8 & 111 \\ 111 & 1843 \end{bmatrix}^{-1} \times 0.914 = \frac{1}{2423}\begin{bmatrix} 1843 & -111 \\ -111 & 8 \end{bmatrix} \times 0.914 \tag{82}$$

$$cov\left(\widehat{\beta}\right) = \begin{bmatrix} \frac{1843}{2423} \times 0.914 & -\frac{111}{2423} \times 0.914 \\ -\frac{111}{2423} \times 0.914 & \frac{8}{2423} \times 0.914 \end{bmatrix} = \begin{bmatrix} 0.6952 & -0.0419 \\ -0.0419 & 0.0030 \end{bmatrix} \tag{83}$$

Diagonal elements of this matrix gives variances of parameters; $var\left(\widehat{\beta}_1\right) = 0.6952$ and $var\left(\widehat{\beta}_2\right) = 0.0030$. These are almost the same as above but the matrix method more precise than the hand-calculation.

2.1 Review of Matrix Algebra

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}; \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}; \quad C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix};$$

Addition:

$$A + B = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix} \tag{84}$$

Subtraction:

$$A - B = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} \\ a_{21} - b_{21} & a_{22} - b_{22} \end{bmatrix} \tag{85}$$

Multiplication:

$$AB = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \tag{86}$$

Algebra

2.1.1 Determinant and Transpose of a Matrix

Determinant of A (difference of cross products)

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = (a_{11}a_{22} - a_{21}a_{12}); \tag{87}$$

Determinant of B $|B| = \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = (b_{11}b_{22} - b_{21}b_{12})$

Determinant of C $|C| = \begin{vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{vmatrix} = (c_{11}c_{22} - c_{21}c_{12})$

Transposes of A, B and C (interchange of rows to columns and columns to rows)

$$A' = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix}; \; B' = \begin{bmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \end{bmatrix}; \; C' = \begin{bmatrix} c_{11} & c_{21} \\ c_{12} & c_{22} \end{bmatrix} \tag{88}$$

Singular matrix $|D| = 0$. non-singular matrix $|D| \neq 0$.

2.1.2 Inverse of A

$$A^{-1} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{|A|} adj\,(A) \tag{89}$$

$$adj\,(A) = C' \tag{90}$$

For C cofactor matrix. For this cross the row and column corresponding to an element and multiply by $(-1)^{i+j}$

$$C = \begin{bmatrix} |a_{22}| & -|a_{21}| \\ -|a_{12}| & |a_{11}| \end{bmatrix} = \begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix} \tag{91}$$

$$C' = \begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}' = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{92}$$

Inverse of A

$$\begin{aligned} A^{-1} &= \frac{1}{(a_{11}a_{22} - a_{21}a_{12})} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \\ &= \begin{bmatrix} \frac{a_{22}}{(a_{11}a_{22}-a_{21}a_{12})} & -\frac{a_{12}}{(a_{11}a_{22}-a_{21}a_{12})} \\ -\frac{a_{21}}{(a_{11}a_{22}-a_{21}a_{12})} & \frac{a_{11}}{(a_{11}a_{22}-a_{21}a_{12})} \end{bmatrix} \end{aligned} \tag{93}$$

2.1.3 Exercise on matrix manipulations

1) Find B^{-1}.

2) Some examples for addition, subtraction, determinant and inverse of matrices.

An electronic store has branches both in Hull and York and sells computers and TV before and after the Christmas. Quantities and prices are as given below.

Table 1: Hypothetical Data on Quantities and Prices

	Hull				York			
	Computer		TV		Computer		TV	
	Before	After	Before	After	Before	After	Before	After
Quantities (Y_i)	300	500	600	400	300	500	600	800
Prices (X_i)	500	400	100	60	525	400	120	80

Represent quantities and prices in the matrix form

1. (a) Total quantities and prices sold in both markets before and after, i.e. (Q_B) and (Q_A) and (P_B) and (P_A).
   (b) Difference in sales in these two places $(Q_B - Q_A)$.
   (c) Total sales revenue in both places before and after the Christmas ($R_B = Q_B P_B$, $R_A = Q_A P_A$). Remember in Hull store can sell its product in Hull prices or price of York and York can sell its products in Hull pirces (diagonal element represent revenue from saling products in local price and the off-diagonal elements show revenue in price of another city; clue for covariance term).

(d) If total revenue (R_B, R_A) and quantities (Q_B, Q_A) are known, show formula to find prices (P_B, P_A).[$P = Q^{-1}R$; here all P, R and Q^{-1} matrices are 2×2).

3) Portfolio of stocks in A, B, C, and D companies is $P = \begin{bmatrix} 200 & 300 & -1100 & 600 \end{bmatrix}$ and their prices in good and bad economic states are $S' = \begin{bmatrix} G & 1.3 & 1.2 & 1.0 & 1.5 \\ B & 1.5 & 0.83 & 0.95 & 1.2 \end{bmatrix}$ respectively. Find the expected values of above portfolio in good and bad states (PS).

2.1.4 Exercise 1

Regress demand for a product (Y_i) on its own prices (X_i) as following

$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad i = 1 \ ...N$$

where e_i is a randomly distributed error term for observation i.

1. (a) List the OLS assumptions on error terms e_i .
   (b) Derive the normal equations and the OLS estimators of $\widehat{\beta}_1$ and $\widehat{\beta}_2$.
   (c) A shopkeeper observed the data quantities and prices as given in Table 2 below. What are the OLS estimates of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ implied by these data? Is this a normal good?
   (d) What are the variances of e_i and Y_i?
   (e) What are R^2 and $\overline{R}^2$?
   (f) Determine the overall significance of this model by F-test at 5 percent level of significance. [Critical value of F for df(1,4) =7.71]
   (g) What are the variances and standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$?
   (h) Compute t-statistics and determine whether parameters $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are statistically significant at 5 percent level of significance [Critical value of t for five percent significance for 4 degrees of freedom is 2.776(i.e $t_{crit,0.05,4} = 2.777$)].
   (i) What is the prediction of Y when X is 0.5?
   (j) What is the elasticity of demand evaluated at the mean values of Y_i and X_i?
   (k) Reformulate the model to include price of a substitute product in the model. What will happen to this estimation if these two prices are exactly correlated?
   (l) How would you decide whether the demand for this product varies by gender?

Table 2: Data on Quantities and Prices

Quantities (Y_i)	5	10	15	20	25	30
Prices (X_i)	10	8	6	4	2	1

Hints: $\left[\sum X_i = 31 \quad \sum X_i^2 = 221 \quad \sum Y_i^2 = 2275; \sum Y_i = 105 \quad \sum Y_i X_i = 380\right]; (X'X)^{-1} = \begin{bmatrix} 0.605 & -0.085 \\ -0.085 & 0.0164 \end{bmatrix}$

Test whether work-hours depend on weekly or annual pay among UK counties using data Unempl_pay-couties.csv.

After estimating slopes and intercepts of demand and supply functions in two interdependent markets, the equilibrium prices and quantities could be found by solving the simultaneous equation system as:

Market 1:

$$X_1^d = 10 - 2p_1 + p_2 \tag{94}$$

$$X_1^S = -2 + 3p_1 \tag{95}$$

Market 2:

$$X_2^d = 15 + p_1 - p_2 \tag{96}$$

$$X_2^S = -1 + 2p_2 \tag{97}$$

Equilibrium in both markets implies:

$X_1^d = X_1^S$ implies $10 - 2p_1 + p_2 = -2 + 3p_1$
$X_1^d = X_1^S$ implies $15 + p_1 - p_2 = -1 + 2p_2$

This in matrix notation:

$$\begin{bmatrix} 5 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 12 \\ 16 \end{bmatrix} \tag{98}$$

Application of Matrix in solving equations

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 5 & -1 \\ -1 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 12 \\ 16 \end{bmatrix} \tag{99}$$

Determinant
$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} 5 & -1 \\ -1 & 3 \end{vmatrix} = (5 \times 3 - (-1)(-1)) = 15 - 1 = 14;$
Cofactor transpose:
$C' = \begin{bmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{bmatrix}' = \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$
Solution by matrix inversion:

$$\begin{aligned}\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} &= \frac{1}{14}\begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}\begin{bmatrix} 12 \\ 16 \end{bmatrix} \\ &= \frac{1}{14}\begin{pmatrix} (3\times 12)+(1\times 16) \\ (1\times 12)+(5\times 16) \end{pmatrix} = \begin{pmatrix} \frac{52}{14} \\ \frac{92}{14} \end{pmatrix} = \begin{pmatrix} \frac{26}{7} \\ \frac{46}{7} \end{pmatrix} \end{aligned} \quad (100)$$

Cramer's Rule is easier

$$p1 = \frac{\begin{vmatrix} 12 & -1 \\ 16 & 3 \end{vmatrix}}{\begin{vmatrix} 5 & -1 \\ -1 & 3 \end{vmatrix}} = \frac{36+16}{15-1} = \frac{26}{7}; \quad p2 = \frac{\begin{vmatrix} 5 & 12 \\ -1 & 16 \end{vmatrix}}{\begin{vmatrix} 5 & -1 \\ -1 & 3 \end{vmatrix}} = \frac{80+12}{15-1} = \frac{46}{7} \quad (101)$$

Market 1:

$$LHS = 10 - 2p_1 + p_2 = 10 - 2\left(\frac{26}{7}\right) + \left(\frac{46}{7}\right) = \frac{64}{7} = -2 + 3p_1 = \frac{64}{7} = RHS \quad (102)$$

Market 2:

$$LHS = 15 + p_1 - p_2 = 15 + \frac{26}{7} - \frac{46}{7} = \frac{85}{7} = -1 + 2p_2 = \frac{85}{7} = RHS \quad (103)$$

QED.

Extension to N-markets is obvious. Matrix makes solving large models much easier.

3 Statistical inference

What is the statistical inference?

- Inference is statement about population based on sample information.
- Economic theory provides basic relations among variables. Statistical inference is about empirically testing whether those relations are true based on available cross section, time series or panel data.
- Hypotheses are set up according to the economic theory, parameters are estimated using OLS (similar other) estimators.
- Consider a linear regression

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \quad i = 1 \ldots N \quad (104)$$

Here the true values of β_1 and β_2 are unknown parameters. Their values can be estimated using the OLS technique. $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are such estimates. Validity of these estimates are tested using statistical distributions. Two most important tests for a linear regression are:

1. Significance of an individual coefficient: **t-test**
2. Overall significance of the model: **F -test**
3. **Overall fit of the data to the model is indicated by** R^2. (χ^2, Durbin-Watson, Unit root tests to be discussed later).

Suppose z is a normally distributed variable. The ordinate at z (probability of z) is given by the standard normal density function

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

The probability of gettin $x \leq z$ is given by the area under standard normal distribution function:

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt.$$

See normal.xls file to feel about the normal distributions.

3.1 Hypothesis Tests: t-test, F-test

A hypothesis is a statement about the relationship between economic variables based on the economic theory. For example in normal circumstances, the marginal propensity to conusume is positive but less than one. Given a regression model of the form $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ there are two major hypotheses. First one is about the significance of individual coefficients, t-test of β_1 and β_2 , and second one is is about the validity of the overal model.

t-test Null hypothesis: value of intercept and slope coefficients are zero;

e.g. there is no meaningful relationship between Y_i and X_i. This is stated as:

$$H_0 : \beta_1 = 0$$
$$H_0 : \beta_2 = 0$$

Alternative hypotheses: intercept and slope coefficients are non -zero; there is a meaningful relationship between Y_i and X_i.

$$H_A : \beta_1 \neq 0$$
$$H_A : \beta_2 \neq 0$$

Parameter β_2 is slope, $\frac{\partial Y}{\partial X}$; it measures how much Y will change when X changes by one unit. Parameter β_1 is intercept. It shows amount of Y when X is zero.

Economic theory: a normal demand function should have $\beta_1 > 0$ and $\beta_2 < 0$; a normal supply function should have $\beta_1 \neq 0$ $\beta_2 > 0$. This is the hypothesis to be tested empirically.

F-test Overal validity of the model model is determined by the F-test. It states whether a model is significant as a whole. An individual coefficient can be insignificant but the model can still be meaningful.

Null hypothesis: both intercept and slope coefficients are zero; model is meaningless and irrelevant:

$$H_0 : \beta_1 = \beta_2 = 0$$

Alternative hypotheses: at least one of the parameters is non -zero, model is relevant:

$$H_A : \text{either } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both } \beta_1 \neq 0, \beta_2 \neq 0$$

As is often seen, some of the coefficients in a regression model may be insignificant but F-statistics is significant and model is valid.

An Exmaple : An Example of regression on deviations from the mean

Table 3: Data Table:Price and Quantity

X (price)	1	2	3	4	5	6
Y (demand)	6	3	4	3	2	1

A standard linear demand is given by a regression where demand for a product (Y) depends on its price (P) as:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \quad i = 1 \ldots N$$

Hypotheis about the coefficients of the model:

$$H_0 : \beta_1 \geq 0$$
$$H_0 : \beta_2 \leq 0$$

What are the estimates of $\widehat{\beta}_1$ and $\widehat{\beta}_2$?

Here $\sum X_i = 21$; $\sum Y_i = 19$; $\sum Y_i X_i$ =52 $\sum X_i^2$ =91 $\sum Y_i^2$ =75 $\overline{Y} = 3.17$ $\overline{X} = 3.5$

OLS estimators

$$\widehat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2}; \qquad \widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2 \overline{X} \tag{105}$$

3.1.1 Normal equations and its deviation form

- Normal equations of above regression

$$\sum Y_i = \widehat{\beta}_1 N + \widehat{\beta}_2 \sum X_i$$
$$\sum Y_i X_i = \widehat{\beta}_1 \sum X_i + \widehat{\beta}_2 \sum X_i^2$$

Define deviations as

$$x_i = \left(X_i - \overline{X}\right) \tag{106}$$

$$y_i = (Y_i - \overline{y}) \tag{107}$$

$$\sum \left(X_i - \overline{X}\right) = 0; \sum (Y_i - \overline{y}) = 0 \tag{108}$$

Putting these in the Normal equations

$$\sum (Y_i - \overline{y}) = \widehat{\beta}_1 N + \widehat{\beta}_2 \sum \left(X_i - \overline{X}\right) \tag{109}$$

$$\sum \left(X_i - \overline{X}\right)(Y_i - \overline{y}) = \widehat{\beta}_1 \sum \left(X_i - \overline{X}\right) + \widehat{\beta}_2 \sum \left(X_i - \overline{X}\right)^2 \tag{110}$$

Since terms $\sum \left(X_i - \overline{X}\right) = 0; \sum (Y_i - \overline{y}) = 0$ the first equation drop out. From the second equation $\sum \left(X_i - \overline{X}\right)(Y_i - \overline{y}) = \sum x_i y_i$ and $\sum \left(X_i - \overline{X}\right)^2 = \sum x_i^2$

This is a regression through origin. Therefore estimator of slope ceofficient with deviation

$$\widehat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{111}$$

$$\widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2 \overline{X} \tag{112}$$

- The reliability of $\widehat{\beta}_2$ and $\widehat{\beta}_1$ depends on their variances; t-test is used to determine their significance.

3.1.2 Deviations from the mean

Useful short-cuts (though matrix method is more accurate, sometimes quick short cuts like this can be handy)

$$\sum x_i^2 = \sum \left(X_i - \overline{X}\right)^2 = \sum X_i^2 - N\overline{X}^2 = 91 - 6(3.5)^2 = 17.5 \tag{113}$$

$$\sum y_i^2 = \sum \left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - N\overline{Y}^2 = 91 - 6(3.17)^2 = 14.7 \tag{114}$$

$$\begin{aligned}\sum y_i x_i &= \sum \left(Y_i - \overline{Y}\right) \sum \left(X_i - \overline{X}\right) \\ &= \sum Y_i X_i - \overline{Y} \sum X_i - \overline{X} \sum Y_i + N\overline{YX} = \\ &\quad \sum Y_i X_i - \overline{Y} N \overline{X} - \overline{X} N \overline{Y} + N\overline{YX} \\ &= \sum Y_i X_i - \overline{Y} N \overline{X} = 52 - (3.5)\,(6)\,(3.17) = -14.57 \end{aligned} \quad (115)$$

3.1.3 OLS estimates by the deviation method

Estimate of the slope coefficient:

$$\widehat{\beta}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{-14.57}{17.5} = -0.833 \quad (116)$$

This is negative as expected.

Estimate of the intercept coefficient.

$$\widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2 \overline{X} = 3.17 - (-0.833)\,(3.5) = 6.09 \quad (117)$$

It is positive as expected.

Thus the **regression line fitted from the data**

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_i = 6.09 - 0.833 X_i \quad (118)$$

How reliable is this line? Answer to this should be based on the analysis of variance and statistical tests.

3.1.4 Variation of Y, predicted Y and error

Total variation to be explained:

$$\sum y_i^2 = \sum \left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - N\overline{Y}^2 = 75 - 6(3.17)^2 = 14.707 \quad (119)$$

Total sum of square (TSS): **Variation explained by regression:**

$$\begin{aligned}\sum \widehat{y}_i^2 &= \sum (\widehat{\beta}_2 x_i)^2 = \widehat{\beta}_2^2 \sum {x_i}^2 = \left(\frac{\sum y_i x_i}{\sum x_i^2}\right)^2 \sum {x_i}^2 \\ &= \frac{\left(\sum y_i x_i\right)^2}{\sum x_i^2} = \frac{(-14.57)^2}{17.5} = \frac{212.28}{17.5} = 12.143 \end{aligned} \quad (120)$$

Note that in deviation form: $\sum \widehat{y}_i = \sum \widehat{\beta}_2 x_i$.

Unexplained variation (accounted by various errors):

$$\sum \widehat{e}_i^2 = \sum y_i^2 - \sum \widehat{y}_i^2 = 14.707 - 12.143 = 2.564 \quad (121)$$

3.1.5 Measure of Fit: R-square and Rbar-square

The **measure of fit** R^2 is ratio of total variation explained by regression $\left(\sum \widehat{y}_i^2\right)$ to total variation that need to be explained $\left(\sum y_i^2\right)$

$$R^2 = \frac{\sum \widehat{y}_i^2}{\sum y_i^2} = \frac{12.143}{14.707} = 0.826 \tag{122}$$

This regression model **explains about 83 percent** of variation in y.

$$\overline{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K} = 1 - (1 - 0.826)\frac{5}{4} = 0.78 \tag{123}$$

Variance of error indicates the unexplained variation

$$var\left(\widehat{e}_i\right) = \widehat{\sigma}^2 = \frac{\sum \widehat{e}_i^2}{N-K} = \frac{2.564}{4} = 0.641 \tag{124}$$

$$var\left(y_i\right) = \frac{\sum y_i^2}{N-1} = \frac{14.7}{5} = 2.94 \tag{125}$$

3.1.6 Variance of parameters

Reliability of estimated parameters depends on their variances, standard errors and t-values

$$var\left(\widehat{\beta}_2\right) = \frac{1}{\sum {x_i}^2}\widehat{\sigma}^2 = \frac{0.641}{17.5} = 0.037 \tag{126}$$

$$var\left(\widehat{\beta}_1\right) = \left[\frac{1}{N} + \frac{\overline{X}^2}{\sum {x_i}^2}\right]\widehat{\sigma}^2 = \left[\frac{1}{6} + \frac{3.5^2}{17.5}\right]0.641 = (0.867)\,0.641 = 0.556 \tag{127}$$

Prove these formula (see later on).

Standard errors

$$SE\left(\widehat{\beta}_2\right) = \sqrt{var\left(\widehat{\beta}_2\right)} = \sqrt{0.037} = 0.192 \tag{128}$$

$$SE\left(\widehat{\beta}_1\right) = \sqrt{var\left(\widehat{\beta}_1\right)} = \sqrt{0.556} = 0.746 \tag{129}$$

3.1.7 Test of significance of parameters (t-test)

$$t\left(\widehat{\beta}_2\right) = \frac{\widehat{\beta}_2}{SE\left(\widehat{\beta}_2\right)} = \frac{-0.833}{0.192} = -4.339 \tag{130}$$

$$t\left(\widehat{\beta}_1\right) = \frac{\widehat{\beta}_1}{SE\left(\widehat{\beta}_1\right)} = \frac{6.09}{0.746} = 8.16 \tag{131}$$

These calculated t-values need to be compared to t-values from the theoretical t-table.

Test of significance of parameters (t-test)

Theoretical values of t are given in a t Table, often given in the appendix of every econometrics or statistical text (http://mathworld.wolfram.com/Studentst-Distribution.html; http://www.statsoft.com/textbook/distribution-tables/).

Column of t-table have level of significance (α) and rows have degrees of freedom.

Here $t_{\alpha,df}$ is t-table value for degrees of freedom ($df = n - k$) and α level of significance. df = 6-2=4.

Table 4: Relevant t-values (one tail) fron t-Table

(n, α)	0.05	0.025	0.005
1	6.314	12.706	63.657
2	2.920	4.303	9.925
4	2.132	2.776	4.604

$t\left(\widehat{\beta}_1\right) = 8.16 > t_{\alpha,df} = t_{0.05,4} = 2.132$. Thus the intercept is statistically significant; $t\left(\widehat{\beta}_2\right) = |-4.339| > t_{\alpha,df} = t_{0.05,4} = 2.132$. Thus the slope is also statistically significant at 5% and 2.5% level of significance.

Decision rule: (one tail test following economic theory)

- Accept $H_0 : \beta_1 > 0$ if $t\left(\widehat{\beta}_1\right) < t_{\alpha,df}$;
- Reject $H_0 : \beta_1 > 0$ or accept $H_A : \beta_1 \ngeq 0$ if $t\left(\widehat{\beta}_1\right) > t_{\alpha,df}$
- Accept $H_0 : \beta_2 < 0$ if $t\left(\widehat{\beta}_2\right) < t_{\alpha,df}$
- Reject $H_0 : \beta_2 < 0$ or accept $H_A : \beta_2 \nleq 0$ if $t\left(\widehat{\beta}_2\right) > t_{\alpha,df}$

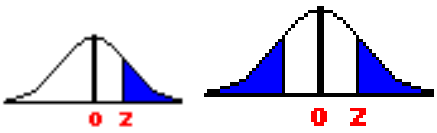
P-value: Probability of test statistics exceeding that of the sample statistics.

3.1.8 Level of significance in a t-test

3.1.9 One- and Two-Tailed Tests

If the area in only one tail of a curve is used in testing a statistical hypothesis, the test is called a **one-tailed test**; if the area of both tails are used, the test is called **two-tailed**.

The decision as to whether a one-tailed or a two-tailed test is to be used depends on the alternative hypothesis.



One-tailed test $X \leq z$ Two-tailed test $-z \leq X \leq z$

3.1.10 Confidence interval on the slope parameter

A researcher may be interested more in knowing the interval in which the true parameter may lie than in the point estimte where α is the level of significance or the probability of error such as 1% or 5%. That means accuracy of the estimate is $(1-\alpha)$ %.

A 95% level confidence interval for β_1 and β_2 is:

$$P\left[\widehat{\beta}_2 - SE\left(\widehat{\beta}_2\right) t_{\alpha,n} < \beta_2 < \widehat{\beta}_2 + SE\left(\widehat{\beta}_2\right) t_{\alpha,n}\right] = (1-\alpha) \tag{132}$$

$$\begin{aligned} & P\left[-0.833 - 0.192\,(2.132.) < \beta_2 < -0.833 + 0.192\,(2.132.)\right] \\ = \;& (1-0.05) = 0.95 \end{aligned} \tag{133}$$

$$P\left[-1.242 < \beta_2 < -0.424\right] = 0.95 \tag{134}$$

There is 95 confidence that the true value of slope β_2 lies between -0.424 and -1.242.

Confidence interval on the intercept parameter

95 % confidence interval on the slope parameter:

$$P\left[\widehat{\beta}_1 - SE\left(\widehat{\beta}_2\right) t_{\alpha,n} < \beta_1 < \widehat{\beta}_1 + SE\left(\widehat{\beta}_2\right) t_{\alpha,n}\right] = (1-\alpha) \tag{135}$$

$$\begin{aligned} & P\left[6.09 - 0.746\,(2.132.) < \beta_1 < 6.09 + 0.746\,(2.132.)\right] \\ = \;& (1-0.05) = 0.95 \end{aligned} \tag{136}$$

$$P\left[4.500 < \beta_2 < 7.680\right] = 0.95 \tag{137}$$

There is 95 confidence that the true value of intercept β_1 lies between 4.500 and 7.680.

3.1.11 F-test

F-value is the ratio of sum of squared normally distributed variables (χ^2) adjusted for relevant degrees of freedom.

$$F = \frac{V_1/n_1}{V_2/n_2} = F\left(n_1, n_2\right) \tag{138}$$

Where V_1 and V_2 are variances of numberator and denomenator and $n_1$and n_2 are degrees of freedom of numberator and denomenator; e.g. $F = \frac{RSS/(K-1)}{ESS/(N-k)}$.

H_0: Variance are the same; H_A: Variance are different. F_{crit} values are obtained from F-distribution table. Accept it if $F_{Calc} < F_{crit}$ and reject if $F_{Calc} > F_{crit}$.

F- is ratio of two χ^2 distributed variables with degrees of freedom $n2$ and $n1$.

$$F_{calc} = \frac{\frac{\sum \widehat{y}_i^2}{K-1}}{\frac{\sum \widehat{e}_i^2}{N-K}} = \frac{\frac{12.143}{1}}{\frac{2.564}{4}} = \frac{12.143}{0.641} = 18.94 \tag{139}$$

Table 5: Relevant F-values from the F-Table

	1% level of significance			5% level of significance		
$(n2, n1)$	1	2	3	1	2	3
1	4042	4999.5	5403	161.4	199.5	215.7
2	98.50	99.00	99.17	18.51	19.00	19.16
4	21.20	18.00	16.69	7.71	6.94	6.59

$n1$ = degrees of freedom of numerator; $n2$ =degrees of freedom of denominator; for 5% level of significance $F_{n1,n2} = F_{1,4} = 7.71$; $F_{calc} > F_{1,4}$;for 1% level of significance $F_{n1,n2} = F_{1,4} = 21.20$; $F_{calc} > F_{1,4} \Longrightarrow$imply that this model is statistically significant at 1% as well as at 5% level of significance. Model is meaningful.

Exercise: Revise data as following and do all above calculations.

This should give a line of perfect fit. What does it impy to $\sum \widehat{e}_i^2$?

Table 6: Data Table:Price and Quantity

X	1	2	3	4	5	6
Y	6	5	4	3	2	1

3.1.12 Exercise 2

A sport centre has a gym. A hypothetical data set on the monthly charges (X) and number of people using the gym (Y) are given in the following table.

Table 7: Monthy charges and number of customers

X_i	10	8	7	6	3	5	9	12	11	10
Y_i	60	75	90	100	150	120	125	100	80	65

1. Represent X and Y in a Scattered diagram.
2. Draw horizontal and vertical lines with the mean of X and Y in that diagram.
3. Draw a line by your hand that best represents all sample observations.
4. Write a classical linear regression model in which X causes Y.
5. Write the assumptions of the error terms.
6. Derive normal equations of the OLS estimator minimising sum of squared errors. Estimate parameters of the model using above information. Use the deviation technique in your estimation.
7. What is your prediction of Y when X is 13?
8. Calculate the sum of variation in Y.
9. Decompose this total variance into explained and residual components.
10. Find the coefficient of determination or the R-square of this model.
11. Find the variance and standard error of the slope parameter.
12. Calculate the t-statistics and determine its level of significance using the T-table.
13. Construct a 95 percent confidence interval for the slope parameter.
14. Find the variance and the standard error of the intercept parameter.

3.1.13 Exercise 3

One major use of an econometric model is prediction. Suppose that a local supermarket wants you to estimate a model that determines expenditure on food in terms of income, and to predict food demand next year. Consider a simple regression model of the following form:

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \qquad t = 1...T \tag{140}$$

where Y_t is expenditure on food, X_t is income and ε_t is independently and identically distributed random error term with a zero mean and a constant variance.

From the sample information on food expenditure and income contained in "food.csv" file find estimated values of β_1 , β_2 and σ^2. You also want to predict the amount of expenditure on food Y_0 next year using information on likely income next year, X_0 . You may safely assume that as before $\varepsilon_0 \sim N(0, \sigma^2)$.

1. Write down your prediction equation. Give an equation for the mean prediction, $E(Y_0)$.
2. What is your prediction of food expenditure if the income is £250? How can you compute your prediction error?
3. What is the variance of prediction error?
4. Construct a 95% confidence interval for your prediction. Explain what this interval means. How would you modify your model if the confidence interval of prediction is very large?
5. Give a graphical explanation of your answers in (a)-(d), labelling your diagrams carefully.

4 Economic theory underlying an empirical estimation

Setting up the labour demand for the firms that sells its product (Q) at price (P) produced employing labour (L) at wage rate (w) and with productivity α.

Firms problem:

$$\max \Pi = PQ - wL \;\text{ s.t. } Q = L^{\alpha} \tag{141}$$

Differentiate it with respect to L. Then based on the first order condition the optimal demand for labour is:

$$L^{\alpha-1} = \frac{w}{\alpha P};\; L = \left(\frac{\alpha P}{w}\right)^{\frac{1}{1-\alpha}} \tag{142}$$

Taking logs:

$$\ln L = \frac{1}{1-\alpha}\ln\alpha + \frac{1}{1-\alpha}\ln P - \frac{1}{1-\alpha}\ln W \tag{143}$$

There are $i = 1...n$ firms. Define $\ln L = Y_i$; $\beta_1 = \frac{1}{1-\alpha}\ln\alpha + \frac{1}{1-\alpha}\ln P$; $\ln W_i = X_i$; $\beta_2 = \frac{1}{1-\alpha}$;

$$Y_i = \beta_1 + \beta_2 X_i \tag{144}$$

Expected signs $\beta_1 > 0$ and $\beta_2 < 0$. Firms will employ more worker if wages are lower but fewer workers when wages are high. This is the theoretical predication. Is it true? Need to test with the data. Cross section or time series data could be constructed on employment and real wages to test this hypothesis.

Application: Estimating Elasticities

Elasticity of Y with respect to X measures the proportionate change in Y due to change in X; $\frac{dy/y}{dx/x}$. Regression coefficient are helpful in measuring these elasticities and the exact value of elasticities depend on the form of regression as illustrated below.

Elasticity in a linear regression model:

- $Y_i = \beta_1 + \beta_2 X_i + e_i$

$\frac{\partial Y_i}{\partial X_i} = \beta_2 \qquad e = \frac{\partial Y_i}{\partial X_i}\frac{\overline{X}}{\overline{Y}} = \beta_2\frac{\overline{X}}{\overline{Y}}$

Elasticity in a log dependent variable linear regression model:

- $ln\left(Y_i\right) = \beta_1 + \beta_2 X_i + e_i$

$\frac{\partial Y_i}{\partial X_i}\frac{1}{Y_i} = \beta_2 \qquad e = \frac{\partial Y_i}{\partial X_i}\frac{X}{Y}Y = \beta_2\overline{X}$

Elasticity in a log explanatory variable linear regression model:

- $Y_i = \beta_1 + \beta_2 \ln\left(X_i\right) + e_i$

$\frac{\partial Y_i}{\partial X_i} = \beta_2\frac{1}{X_i} \qquad e = \frac{\partial Y_i}{\partial X_i}\frac{X}{Y} = \beta_2\frac{1}{X_i}\ \frac{X}{Y} = \beta_2\ \frac{1}{Y}$

Elasticity in a double log linear regression model:

- $ln\left(Y_i\right) = \beta_1 + \beta_2 \ln\left(X_i\right) + e_i$.

$\frac{\partial Y_i}{\partial X_i}\frac{1}{Y_i} = \beta_2\frac{1}{X_i} \qquad e = \frac{\partial Y_i}{\partial X_i}\frac{X}{Y} = \beta_2\frac{Y}{X_i}\ \frac{X}{Y} = \beta_2$

Elasticity in a regression model linear in reciprocal of an explanatory variable:

- $Y_i = \beta_1 + \beta_2\frac{1}{X_i} + e_i$.

$\frac{\partial Y_i}{\partial X_i} = -\beta_2\frac{1}{X_i^2} \qquad e = \frac{\partial Y_i}{\partial X_i}\frac{X}{Y} = -\beta_2\frac{1}{X_i^2}\frac{X}{Y} = \beta_2\frac{1}{X_i}\frac{1}{Y_i}$

Elasticity in a quadratic regression model:

- $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + e_i$

$\frac{\partial Y_i}{\partial X_i} = \beta_2 + 2\beta_3 X_i^2; \quad e = \frac{\partial Y_i}{\partial X_i}\frac{X}{Y} = \left(\beta_2 + 2\beta_3 X_i\right)\frac{X}{Y}$

Use MacKinnon, White and Davison test to choose between linear and log-linear models.

Return in a portfolio Investor's net return from a selection of portfolios R_p is risky assets is R_p is given by return on the asset adjusted for the risk free return (R_f), which is a return in safe asset like the treasury bill.

$$R_p = R_s - hR_f \tag{145}$$

The investor like to choose the hedging factor, $0 < h < 1$, to minimise the variance on portfolio.

$$var\left(R_p\right) = var\left(R_s\right) - 2hcov(hR_f) + h^2 var(R_f) \tag{146}$$

By minimise the variance of the portfolio w.r.t. h

$$\frac{var\left(R_p\right)}{\partial h} = -2hcov(R_s, R_f) + 2hvar(R_f) = 0 \tag{147}$$

$$h = \frac{cov(R_s, R_f)}{var(R_f)} = \frac{\sqrt{var(R_s)}\sqrt{var(R_f)}}{\sqrt{var(R_s)}\sqrt{var(R_f)}}\frac{cov(R_s, R_f)}{var(R_f)} = \rho\frac{\sigma_s}{\sigma_f} \tag{148}$$

Empirical question then would be to estimate the value of h from given data on R_p, R_s and R_f. See the Datastream or http://download.finance.yahoo.com/ have data on stock prices.

4.1 Regression in Matrix Notations

Let Y is $N \times 1$ vector of dependent variables, X is $N \times K$ matrix of explanatory variables, e is $N \times 1$ vector of independently and identically distributed normal random variable with mean equal to zero and a constant variance $e \sim N(0, \sigma^2 I)$; β is a $K \times 1$ vector of unknown coefficients

$$Y = \beta X + e \tag{149}$$

4.1.1 Objective

Objective is to choose β that minimise sum square errors

$$\begin{aligned} \underset{\beta}{Min}\ S\left(\beta\right) &= e'e = \left(Y - \beta X\right)'\left(Y - \beta X\right) \\ &= Y'Y - Y'\left(\beta X\right) - \left(\beta X\right)'Y + \left(\beta X\right)'\left(\beta X\right) \quad (150) \\ &= Y'Y - 2\beta X'Y + \left(\beta X\right)'\left(\beta X\right) \quad (151) \end{aligned}$$

First order condition:

$$\frac{\partial S\left(\beta\right)}{\partial \beta} = -2X'Y + 2\widehat{\beta}X'X = 0 \Longrightarrow \widehat{\beta} = \left(X'X\right)^{-1}X'Y \tag{152}$$

$$(X'X)^{-1} = \begin{pmatrix} N & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}^{-1} = \frac{1}{N\sum X_i^2 - (\sum X_i)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & N \end{bmatrix} \tag{153}$$

$$(X'X)^{-1} = \begin{bmatrix} \frac{\sum X_i^2}{N\sum X_i^2 - (\sum X_i)^2} & -\frac{\sum X_i}{N\sum X_i^2 - (\sum X_i)^2} \\ -\frac{\sum X_i}{N\sum X_i^2 - (\sum X_i)^2} & \frac{N}{N\sum X_i^2 - (\sum X_i)^2} \end{bmatrix} \tag{154}$$

$$\widehat{\beta} = (X'X)^{-1} X'Y; \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \frac{\sum X_i^2}{N\sum X_i^2 - (\sum X_i)^2} & -\frac{\sum X_i}{N\sum X_i^2 - (\sum X_i)^2} \\ -\frac{\sum X_i}{N\sum X_i^2 - (\sum X_i)^2} & \frac{N}{N\sum X_i^2 - (\sum X_i)^2} \end{bmatrix} \begin{bmatrix} \sum Y_i \\ \sum X_i' Y_i \end{bmatrix} \tag{155}$$

Derivation of Parameters (with Matrix Inverse)

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i' Y_i}{N\sum X_i^2 - (\sum X_i)^2} \\ \frac{N\sum X_i' Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - (\sum X_i)^2} \end{bmatrix} = \begin{bmatrix} \frac{\sum X_i \sum X_i' Y_i - \sum X_i^2 \sum Y_i}{N\sum X_i^2 - (\sum X_i)^2} \\ \frac{\sum X_i \sum Y_i - N\sum X_i' Y_i}{N\sum X_i^2 - (\sum X_i)^2} \end{bmatrix} \tag{156}$$

Compare to what we had earlier:

$$\widehat{\beta}_2 = \frac{\sum X_i \sum Y_i - N\sum Y_i X_i}{(\sum X_i)^2 - N\sum X_i^2} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{157}$$

$$\widehat{e} = \left(Y - \widehat{\beta} X\right) \tag{158}$$

$$\widehat{\sigma}^2 = \frac{\sum \widehat{e}_i^2}{N-k} = \frac{e'e}{N-k} \tag{159}$$

$$\sum \widehat{e}_i^2 = \sum y_i^2 - \sum \widehat{y}_i^2 \tag{160}$$

$$\sum \widehat{y}^2 = \sum (x\beta)' (\beta x) \qquad x = X - \overline{X} \tag{161}$$

$$\sum \widehat{y}_i^2 = \sum (\widehat{\beta}_2 x_i)^2 = \widehat{\beta}_2^2 \sum {x_i}^2 \tag{162}$$

$$R^2 = \frac{\sum \widehat{y}_i^2}{\sum y_i^2} \text{ and } F_{calc} = \frac{\frac{\sum \widehat{y}_i^2}{K-1}}{\frac{\sum \widehat{e}_i^2}{N-K}}; \; F_{calc} = \frac{R^2}{K-1} \frac{N-K}{(1-R^2)} \tag{163}$$

4.2 Variance in matrix notation

$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$

$$Y = \widehat{Y} + e = \widehat{\beta} X + e \tag{164}$$

$$Var(Y) = \sum y_i^2 = Y'Y - N\overline{Y}^2 \tag{165}$$

When there are two explantory variables in deviation from the mean form:

$$\widehat{y} = \widehat{\beta}_1 x_1 + \widehat{\beta}_1 x_2 \tag{166}$$

$$\begin{aligned}
\sum \widehat{y}^2 &= \left(\widehat{\beta}_1 \sum x_1 + \widehat{\beta}_2 \sum x_2\right)^2 = \\
&\quad \widehat{\beta}_1^2 \sum x_1^2 + \widehat{\beta}_1 \widehat{\beta}_2 \sum x_1 x_2 + \widehat{\beta}_1 \widehat{\beta}_2 \sum x_1 x_2 + \widehat{\beta}_2^2 \sum x_2^2 \\
&= \widehat{\beta}_1 \left(\widehat{\beta}_1 \sum x_1^2 + \widehat{\beta}_2 \sum x_1 x_2\right) + \widehat{\beta}_2 \left(\widehat{\beta}_1 \sum x_1 x_2 + \widehat{\beta}_2 \sum x_1^2\right) \\
&= \widehat{\beta}_1 \sum x_1 y + \widehat{\beta}_2 \sum x_2 y
\end{aligned} \tag{167}$$

$$\sum \widehat{e}_i^2 = \sum y_i^2 - \sum \widehat{y}_i^2 \tag{168}$$

In matrix notation:

$$\sum \widehat{y}^2 = \begin{bmatrix} \widehat{\beta}_1 & \widehat{\beta}_2 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & . & x_{1N} \\ x_{21} & x_{22} & . & x_{2N} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_N \end{bmatrix} = \widehat{\beta}' x' y \tag{169}$$

$$e'e = Y'Y - \widehat{\beta}' x' y \tag{170}$$

$$R^2 = \frac{\widehat{\beta}' x' y}{Y'Y} \text{ and } F_{calc} = \frac{\frac{\sum \widehat{y}_i^2}{K-1}}{\frac{e'e}{N-K}};\ F_{calc} = \frac{R^2}{K-1} \frac{N-K}{(1-R^2)} \tag{171}$$

4.2.1 Blue Property in Matrix: Linearity and Unbiasedness

$$\widehat{\beta} = (X'X)^{-1} X'Y \tag{172}$$

$$\widehat{\beta} = aY; \quad a = (X'X)^{-1} X' \tag{173}$$

Linearity proved.

$$E\left(\widehat{\beta}\right) = E\left[(X'X)^{-1} X' (X\beta + e)\right] \tag{174}$$

$$E\left(\widehat{\beta}\right) = E\left[\left(X'X\right)^{-1}X'X\beta\right] + E\left[\left(X'X\right)^{-1}X'e\right] \tag{175}$$

$$E\left(\widehat{\beta}\right) = \beta + E\left[\left(X'X\right)^{-1}X'e\right] \tag{176}$$

$$E\left(\widehat{\beta}\right) = \beta \tag{177}$$

Unbiasedness is proved.
Blue Property in Matrix: Minimum Variance

$$E\left(\widehat{\beta}\right) - \beta = E\left[\left(X'X\right)^{-1}X'e\right] \tag{178}$$

$$\begin{aligned} E\left[E\left(\widehat{\beta}\right) - \beta\right]^2 &= E\left[\left(X'X\right)^{-1}X'e\right]'\left[\left(X'X\right)^{-1}X'e\right] && (179)\\ &= \left(X'X\right)^{-1}X'XE\left(e'e\right)\left(X'X\right)^{-1} = \widehat{\sigma}^2\left(X'X\right)^{-1} && (180)\end{aligned}$$

Take an alternative estimator b

$$b = \left[\left(X'X\right)^{-1}X' + c\right]Y \tag{181}$$

$$b = \left[\left(X'X\right)^{-1}X' + c\right](X\beta + e) \tag{182}$$

$$b - \beta = E\left[\left(X'X\right)^{-1}X'e + ce\right] \tag{183}$$

4.2.2 Blue Property in Matrix: Minimum Variance

Now it need to be shown that

$$cov\left(b\right) > cov\left(\widehat{\beta}\right) \tag{184}$$

Take an alternative estimator b

$$b - \beta = E\left[\left(X'X\right)^{-1}X'e + ce\right] \tag{185}$$

$$\begin{aligned} cov\left(b\right) &= E\left[\left(b-\beta\right)\left(b-\beta\right)'\right]\\ &= E\left[\left(X'X\right)^{-1}X'e + ce\right]\left[\left(X'X\right)^{-1}X'e + ce\right]\\ &= \sigma^2\left(X'X\right)^{-1} + \sigma^2c^2 && (186)\end{aligned}$$

$$cov\left(b\right) > cov\left(\widehat{\beta}\right) \tag{187}$$

Proved.
Thus the OLS is BLUE = Best, Linear, Unbiased Estimator.

5 Multiple Regression Model in Matrix

Consider a multiple linear regression model in which there are more than one eplanatory variables as:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + ... + \beta_k X_{k,i} + \varepsilon_i \quad i = 1 \ ...N \tag{188}$$

Basic assumptions are the same as in the single explanatory variable model discussed so far as:

$$E(\varepsilon_i) \ = 0 \tag{189}$$

$$E(\varepsilon_i x_{j,i}) \ = 0; \ var(\varepsilon_i) \ = \sigma^2 \quad for \ \ \forall \ \ i; \ \ \varepsilon_i \tilde{} N\left(0, \sigma^2\right) \tag{190}$$

$$covar(\varepsilon_i \varepsilon_j) \ = 0 \tag{191}$$

One more important assumption in the multiple linear regression model is that the explanatory variables are uncorrelated.

$$E(X_{1,i} X_{1,j}) = 0 \tag{192}$$

Objective of the estimation is to choose parameters that minimise the sum of squared errors as:

$$\underset{\widehat{\beta}_0 \widehat{\beta}_1 \widehat{\beta}_2 ... \widehat{\beta}_k}{Min} S = \sum \varepsilon_i^2 = \sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1,i} - \widehat{\beta}_2 X_{2,i} - \widehat{\beta}_3 X_{3,i} - .... - \widehat{\beta}_k X_{k,i}\right)^2 \tag{193}$$

Minimisartion results in following first order conditions:

$$\frac{\partial S}{\partial \widehat{\beta}_0} = 0; \frac{\partial S}{\partial \widehat{\beta}_1} = 0; \frac{\partial S}{\partial \widehat{\beta}_2} = 0; \frac{\partial S}{\partial \widehat{\beta}_3} = 0; .....\frac{\partial S}{\partial \widehat{\beta}_k} = 0 \tag{194}$$

This equations result in the normal equations.

5.0.3 Normal equations in a multiple regression model

Normal equations for two explanatory variable case:

$$\sum Y_i = \widehat{\beta}_0 N + \widehat{\beta}_1 \sum X_{1,i} + \widehat{\beta}_2 \sum X_{2,i} \tag{195}$$

$$\sum X_{1,i} Y_i = \widehat{\beta}_0 \sum X_{1,i} + \widehat{\beta}_1 \sum X_{1,i}^2 + \widehat{\beta}_2 \sum X_{1,i} X_{2,i} \tag{196}$$

$$\sum X_{2,i} Y_i = \widehat{\beta}_0 \sum X_{2,i} + \widehat{\beta}_1 \sum X_{1,i} X_{2,i} + \widehat{\beta}_2 \sum X_{2,i}^2 \tag{197}$$

$$\begin{bmatrix} \sum Y_i \\ \sum X_{1,i} Y_i \\ \sum X_{2,i} Y_i \end{bmatrix} = \begin{bmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i} X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i} X_{2,i} & \sum X_{2,i}^2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} \tag{198}$$

5.0.4 Normal equations in matrix form:

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum Y_i X_{1,i} \\ \sum Y_i X_{2,i} \end{bmatrix} \tag{199}$$

$$\beta = (X'X)^{-1} X'Y \tag{200}$$

$$\widehat{\beta}_0 = \frac{\begin{vmatrix} \sum Y_i & \sum X_{1,i} & \sum X_{2,i} \\ \sum Y_i X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum Y_i X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}}{\begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}} \tag{201}$$

5.0.5 Cramer Rule to find estimators of model paramers:

$$\widehat{\beta}_1 = \frac{\begin{vmatrix} N & \sum Y_i & \sum X_{2,i} \\ \sum X_{1,i} & \sum Y_i X_{1,i} & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum Y_i X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}}{\begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}} \tag{202}$$

$$\widehat{\beta}_2 = \frac{\begin{vmatrix} N & \sum X_{1,i} & \sum Y_i \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum Y_i X_{1,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum Y_i X_{2,i} \end{vmatrix}}{\begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}} \tag{203}$$

Matrix must be non-singular to get a meaningful estimator:

$$(X'X)^{-1} = \begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix} \neq 0 \tag{204}$$

Covariance of Parameters

$$cov\left(\widehat{\beta}\right) = \begin{pmatrix} var(\widehat{\beta}_1) & cov(\widehat{\beta}_1\widehat{\beta}_2) & cov(\widehat{\beta}_1\widehat{\beta}_3) \\ cov(\widehat{\beta}_1\widehat{\beta}_2) & var(\widehat{\beta}_2) & cov(\widehat{\beta}_2\widehat{\beta}_3) \\ cov(\widehat{\beta}_1\widehat{\beta}_3) & cov(\widehat{\beta}_2\widehat{\beta}_3) & var(\widehat{\beta}_3) \end{pmatrix} \tag{205}$$

$$cov\left(\widehat{\beta}\right) = (X'X)^{-1}\sigma^2 \tag{206}$$

$$cov\left(\widehat{\beta}\right) = \left[\begin{array}{ccc} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{array} \right]^{-1} \widehat{\sigma}^2 \tag{207}$$

It is easier to consider normal equations in the deviation form:

$$\left[\begin{array}{c} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \right] = \left[\begin{array}{cc} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{array} \right]^{-1} \left[\begin{array}{c} \sum y_i x_{1,i} \\ \sum y_i x_{2,i} \end{array} \right] \tag{208}$$

$$\beta = \left(X'X\right)^{-1} X'Y \tag{209}$$

$$\widehat{\beta}_1 = \frac{\left| \begin{array}{cc} \sum y_i x_{1,i} & \sum x_{1,i}x_{2,i} \\ \sum y_i x_{2,i} & \sum x_{2,i}^2 \end{array} \right|}{\left| \begin{array}{cc} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{array} \right|} \tag{210}$$

$$\widehat{\beta}_2 = \frac{\left| \begin{array}{cc} \sum x_{1,i}^2 & \sum y_i x_{1,i} \\ \sum x_{1,i}x_{2,i} & \sum y_i x_{2,i} \end{array} \right|}{\left| \begin{array}{cc} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{array} \right|} \tag{211}$$

$$\begin{aligned} & \left[\begin{array}{cc} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{array} \right]^{-1} \\ = \quad & \frac{1}{\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2} \left[\begin{array}{cc} \sum x_{2,i}^2 & -\sum x_{1,i}x_{2,i} \\ -\sum x_{1,i}x_{2,i} & \sum x_{1,i}^2 \end{array} \right] \end{aligned} \tag{212}$$

Total sum of squared errors:

$$\sum \widehat{e}_i^2 = \sum y_i^2 - \sum \widehat{y}_i^2 \tag{213}$$

$$E\left(\widehat{\varepsilon}_i\right)^2 = \frac{\sum \widehat{e}_i^2}{N-k} = \widehat{\sigma}^2 \tag{214}$$

$$var\left(\widehat{\beta}_1\right) = \frac{\sum x_{2,i}^2}{\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2} \sigma^2 \tag{215}$$

$$var\left(\widehat{\beta}_2\right) = \frac{\sum x_{1,i}^2}{\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2} \sigma^2 \tag{216}$$

All done by a calculator:

Table 8: Price, Income and Sales

$X_{1,i}$	11	7	6	5	3	2	1
$X_{2,i}$	2	2	4	5	6	5	4
Y_i	1	2	3	4	5	6	7

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 7 & 35 & 28 \\ 35 & 245 & 117 \\ 28 & 117 & 126 \end{bmatrix}^{-1} \begin{bmatrix} 28 \\ 97 \\ 126 \end{bmatrix} \tag{217}$$

Inverse: Determinants and Cofactors

j. It can be solved by the Cramer rule but it is easier to derive variance when solved by the inverse method:

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \frac{1}{\begin{vmatrix} 7 & 35 & 28 \\ 35 & 245 & 117 \\ 28 & 117 & 126 \end{vmatrix}} \tag{218}$$

$$\begin{bmatrix} \begin{vmatrix} 245 & 117 \\ 117 & 126 \end{vmatrix} & -\begin{vmatrix} 35 & 117 \\ 28 & 126 \end{vmatrix} & \begin{vmatrix} 35 & 245 \\ 28 & 117 \end{vmatrix} \\ -\begin{vmatrix} 35 & 28 \\ 117 & 126 \end{vmatrix} & \begin{vmatrix} 7 & 28 \\ 28 & 126 \end{vmatrix} & -\begin{vmatrix} 7 & 35 \\ 28 & 117 \end{vmatrix} \\ \begin{vmatrix} 35 & 28 \\ 245 & 117 \end{vmatrix} & -\begin{vmatrix} 7 & 28 \\ 35 & 117 \end{vmatrix} & \begin{vmatrix} 7 & 35 \\ 35 & 245 \end{vmatrix} \end{bmatrix}^T \begin{bmatrix} 28 \\ 97 \\ 126 \end{bmatrix} \tag{219}$$

Evaluating Inverse

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \frac{1}{\begin{pmatrix} 216090 + 114660 + 114660 \\ -192080 - 95823 - 154350 \end{pmatrix}} \tag{220}$$

$$\begin{bmatrix} 17181 & -1134 & -2765 \\ -1134 & 98 & 161 \\ -2765 & 161 & 490 \end{bmatrix}^T \begin{bmatrix} 28 \\ 97 \\ 126 \end{bmatrix} \tag{221}$$

OLS Estimates

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \frac{1}{(3157)} \begin{bmatrix} 17181 & -1134 & -2765 \\ -1134 & 98 & 161 \\ -2765 & 161 & 490 \end{bmatrix} \begin{bmatrix} 28 \\ 97 \\ 126 \end{bmatrix} \tag{222}$$

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 5.44 & -0.360 & -0.876 \\ -0.360 & 0.031 & 0.051 \\ -0.876 & 0.051 & 0.155 \end{bmatrix} \begin{bmatrix} 28 \\ 97 \\ 126 \end{bmatrix} \tag{223}$$

OLS Estimates

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} (5.44 \times 28) + (-0.360 \times 97) + (-0.876 \times 126) \\ (-0.360 \times 28) + (0.031 \times 97) + 0.051 \times 126 \\ (-0.876 \times 28) + (0.051 \times 97) + 0.155 \times 126 \end{bmatrix} \quad (224)$$

$$= \begin{pmatrix} 8.158 \\ -0.647 \\ -0.051 \end{pmatrix} \quad (225)$$

Using OLS estimates for prediction

These hand-calculations are very close to the Excel routines. Discrepancy must be due the rounding errors as excel has precision of 12 decimal points.

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{pmatrix} 7.18 \\ -0.621 \\ -0.020 \end{pmatrix} \quad (226)$$

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1,i} + \widehat{\beta}_2 X_{2,i} = 7.18 - 0.621 X_{1,i} - 0.020 X_{2,i} \quad (227)$$

k. Prediction when $X_{1,i} = 5$ and $X_{2,i} = 4$.

$$\widehat{Y}_i = 7.18 - 0.621\,(5) - 0.020\,(4) = 3.995 \quad (228)$$

5.1 Testing for Restrictions

When a research has some prior information about the value and sign of coefficient in a regression model, such infromation could be put in the estimation process as a restriction. There can be one or several restrictions in a model but the validity of these restrictions could be tested using F-statistics.

- Consider a linear regression

$$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i \quad i = 1 \ldots N \quad (229)$$

and assumptions

$$E(\varepsilon_i) = 0 \quad (230)$$

$$E(\varepsilon_i x_{j,i}) = 0 \quad (231)$$

$$var(\varepsilon_i) = \sigma^2 \quad for \ \forall \ i \quad (232)$$

$$covar(\varepsilon_i \varepsilon_j) = 0 \quad (233)$$

$$\varepsilon_i \tilde{} N\left(0, \sigma^2\right) \quad (234)$$

- Objective is to choose parameters that minimise the sum of squared errors

$$\underset{\widehat{\beta}_1,\widehat{\beta}_2,\widehat{\beta}_3}{Min}\, S = \sum \varepsilon_i^2 = \left(Y_i - \widehat{\beta}_1 X_{1,i} - \widehat{\beta}_2 X_{2,i} - \widehat{\beta}_3 X_{3,i}\right)^2 \tag{235}$$

Derivation of Normal Equations

$$\frac{\partial S}{\partial \widehat{\beta}_1} = 0; \frac{\partial S}{\partial \widehat{\beta}_2} = 0; \frac{\partial S}{\partial \widehat{\beta}_3} = 0; \tag{236}$$

- Normal equations for three explanatory variable case

$$\sum X_{1,i}Y_i = \widehat{\beta}_1 \sum X_{1,i}^2 + \widehat{\beta}_2 \sum X_{1,i}X_{2,i} + \widehat{\beta}_3 \sum X_{1,i}X_{3,i} \tag{237}$$

$$\sum X_{2,i}Y_i = \widehat{\beta}_1 \sum X_{1,i}X_{2,i} + \widehat{\beta}_2 \sum X_{2,i}^2 + \widehat{\beta}_3 \sum X_{2,i}X_{3,i} \tag{238}$$

$$\sum X_{3,i}Y_i = \widehat{\beta}_1 \sum X_{1,i}X_{3,i} + \widehat{\beta}_2 \sum X_{2,i}X_{3,i} + \widehat{\beta}_3 \sum X_{3,i}^2 \tag{239}$$

$$\begin{bmatrix} \sum X_{1,i}Y_i \\ \sum X_{2,i}Y_i \\ \sum X_{3,i}Y_i \end{bmatrix} = \begin{bmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} \tag{240}$$

Normal equations in matrix form

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum X_{1,i}Y_i \\ \sum X_{2,i}Y_i \\ \sum X_{3,i}Y_i \end{bmatrix} \tag{241}$$

$$\beta = \left(X'X\right)^{-1} X'Y \tag{242}$$

$$\widehat{\beta}_1 = \frac{\begin{vmatrix} \sum X_{1,i}Y_i & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{2,i}Y_i & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{3,i}Y_i & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{vmatrix}}{\begin{vmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{vmatrix}} \tag{243}$$

Use Cramer Rule to solve for paramers

$$\widehat{\beta}_2 = \frac{\begin{vmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}Y_i \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}Y_i \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}Y_i \end{vmatrix}}{\begin{vmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{vmatrix}} \tag{244}$$

$$\widehat{\beta}_2 = \frac{\begin{vmatrix} \sum X_{1,i}^2 & \sum X_{1,i}Y_i & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}Y_i & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{3,i}Y_i & \sum X_{3,i}^2 \end{vmatrix}}{\begin{vmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{vmatrix}} \tag{245}$$

Covariance of Parameters

5.1.1 Matrix must be non-singular

$$(X'X)^{-1} \neq 0 \tag{246}$$

$$cov\left(\widehat{\beta}\right) = \begin{pmatrix} var(\widehat{\beta}_1) & var(\widehat{\beta}_1\widehat{\beta}_2) & var(\widehat{\beta}_1\widehat{\beta}_3) \\ var(\widehat{\beta}_1\widehat{\beta}_2) & var(\widehat{\beta}_2) & var(\widehat{\beta}_2\widehat{\beta}_3) \\ var(\widehat{\beta}_1\widehat{\beta}_3) & var(\widehat{\beta}_2\widehat{\beta}_3) & var(\widehat{\beta}_3) \end{pmatrix} \tag{247}$$

$$cov\left(\widehat{\beta}\right) = (X'X)^{-1}\sigma^2 \tag{248}$$

$$cov\left(\widehat{\beta}\right) = \begin{bmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{bmatrix}^{-1} \widehat{\sigma}^2 \tag{249}$$

Data (text book example Carter, Griffith and Hill)

Table 9: Data for a multiple regression

y	1	-1	2	0	4	2	2	0	2
x1	1	-1	1	0	1	0	0	1	0
x2	0	1	0	1	2	3	0	-1	0
x3	-1	0	0	0	0	0	1	1	1

$$\begin{bmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 16 & -1 \\ 0 & -1 & 4 \end{bmatrix} \text{ and}$$

$$\begin{bmatrix} \sum X_{1,i}Y_i \\ \sum X_{2,i}Y_i \\ \sum X_{3,i}Y_i \end{bmatrix} = \begin{bmatrix} 8 \\ 13 \\ 3 \end{bmatrix}$$

Estimation of Parameters

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 16 & -1 \\ 0 & -1 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 13 \\ 3 \end{bmatrix} \tag{250}$$

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.063 & 0.016 \\ 0 & 0.016 & 0.254 \end{bmatrix} \begin{bmatrix} 8 \\ 13 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.6 \\ 0.873 \\ 0.968 \end{bmatrix} \tag{251}$$

Estimated equation

$$\widehat{Y}_i = 1.6X_{1,i} + 0.873X_{2,i} + 0.968X_{3,i} \tag{252}$$

Estimation of Errors

$$\widehat{e}_i = Y_i - 1.6X_{1,i} + 0.873X_{2,i} + 0.968X_{3,i} \tag{253}$$

$$\widehat{e}_1 = 1 - 1.6\,(1) + 0.873\,(0) + 0.968\,(-1) = 0.368 \tag{254}$$

$$\widehat{e}_2 = -1 - 1.6\,(-1) + 0.873\,(1) + 0.968\,(0) = -0.273 \tag{255}$$

$$\widehat{e}_3 = 2 - 1.6\,(1) + 0.873\,(0) + 0.968\,(0) = 0.4 \tag{256}$$
$$\widehat{e}_4 = 0 - 1.6\,(0) + 0.873\,(1) + 0.968\,(0) = -0.873 \tag{257}$$
$$\widehat{e}_5 = 4 - 1.6\,(1) + 0.873\,(2) + 0.968\,(0) = 0.654 \tag{258}$$
$$\widehat{e}_6 = 2 - 1.6\,(0) + 0.873\,(3) + 0.968\,(0) = -0.619 \tag{259}$$
$$\widehat{e}_7 = 2 - 1.6\,(0) + 0.873\,(0) + 0.968\,(1) = 1.032 \tag{260}$$
$$\widehat{e}_8 = 0 - 1.6\,(1) + 0.873\,(-1) + 0.968\,(1) = -1.695 \tag{261}$$
$$\widehat{e}_9 = 2 - 1.6\,(0) + 0.873\,(0) + 0.968\,(1) = 1.032 \tag{262}$$

Sum of Error square, variance and covariance of Beta

$$\begin{aligned} \sum \widehat{e}_i^2 &= 0.368^2 + (-0.273)^2 + 0.4^2 \\ &\quad + (-0.873)^2 + (0.654)^2 + (-0.619)^2 + 1.032^2 \\ + (-1.695)^2 + 1.032^2 &= 6.9460 \end{aligned} \tag{263}$$

Variance of errors

$$var(e) = E\left(\widehat{\varepsilon}_i\right)^2 = \frac{\sum \widehat{e}_i^2}{N-k} = \frac{6.9460}{9-3} = 1.1577 = \widehat{\sigma}^2 \tag{264}$$

$$\begin{aligned} cov\left(\widehat{\beta}\right) &= \begin{bmatrix} \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i}X_{3,i} \\ \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i}X_{3,i} \\ \sum X_{1,i}X_{3,i} & \sum X_{2,i}X_{3,i} & \sum X_{3,i}^2 \end{bmatrix}^{-1} \widehat{\sigma}^2 \\ &= \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 0.063 & 0.016 \\ 0 & 0.016 & 0.254 \end{bmatrix} (1.1577) = \begin{bmatrix} 0.232 & 0 & 0 \\ 0 & 0.074 & 0.018 \\ 0 & 0.018 & 0.294 \end{bmatrix} \end{aligned} \tag{265}$$

Test of Restrictions

$$var\left(\widehat{\beta}_1\right) = 0.232; var\left(\widehat{\beta}_2\right) = 0.074; var\left(\widehat{\beta}_1\right) = 0.294; \tag{266}$$

$$cov\left(\widehat{\beta}_1\widehat{\beta}_2\right) = cov\left(\widehat{\beta}_1\widehat{\beta}_3\right) = 0; cov\left(\widehat{\beta}_2\widehat{\beta}_3\right) = cov\left(\widehat{\beta}_3\widehat{\beta}_2\right) = 0; \tag{267}$$

F-test

$$F = \frac{(Rb-r)'\left[Rcov\left(b\right)R'\right]^{-1}(Rb-r)}{J} \tag{268}$$

Hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ against H_A: $\beta_1 \neq 0$; $\beta_2 \neq 0$; or $\beta_3 \neq 0$

Here $J = 3$ is the number of restrictions

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; b = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix}; r = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{269}$$

Test of Restrictions

$$F = \frac{\begin{array}{c}\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right)' \\ \left[\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 0.232 & 0 & 0 \\ 0 & 0.074 & 0.018 \\ 0 & 0.018 & 0.294 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}'\right]^{-1} \\ \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right)\end{array}}{J=3} \tag{270}$$

See matrix_restrictions.xls for calculations.

Test of Restrictions

$$F = \frac{\begin{pmatrix} 1.6 & 0.873 & 0.968 \end{pmatrix}\begin{bmatrix} 4.3190 & 0 & 0 \\ 0 & 13.821 & -0.8638 \\ 0 & -0.8638 & 3.455 \end{bmatrix}\begin{bmatrix} 1.6 \\ 0.873 \\ 0.968 \end{bmatrix}}{3} \tag{271}$$

$$F = \frac{\begin{pmatrix} 1.6 & 0.873 & 0.968 \end{pmatrix}\begin{bmatrix} 6.91042 \\ 11.22943 \\ 2.59141 \end{bmatrix}}{3} = \frac{23.37}{3} = 7.79 \tag{272}$$

$F_{(m1,m2),\alpha} = F_{(3,6),5\%} = 4.76$.

Critical value for F at degrees of freedom of (3,6) at 5% confidence interval is 4.76.

F calculated is bigger than F critical => Reject null hypothesis, which says

$$H_0\text{: } \beta_1 = \beta_2 = \beta_3 = 0$$

At least one of these parameters is significant and explains variation in y, in other words accept

$$H_A\text{: } \beta_1 \neq 0;\ \ \beta_2 \neq 0; \text{ or } \beta_3 \neq 0$$

Instructions for testing linear restrictions in PcGive for cross section data like this:

a. regress Y on $X_{1,i}$ $X_{2,i}$, $X_{3,i}$ and $X_{4,i}$.

b. click on test/linear restriction, put the restrictions in the matrix box. one line for each restriction. For instance if β_0+β_1+β_2+β_3+ $\beta_4 = 0$. to be tested then type 1 1 1 1 1 0 , then click ok , it will test validity of that restriction. If there are two restrictions

β_0+β_1+β_2+β_3+ $\beta_4 = 0$ and β_3- $\beta_4 = 0$ then

1 1 1 1 1 0

0 0 0 1 -1 0

put this input in the matrix box, then click OK. This will test for both restrictions.

6 Dummy Variables in a Regression Model

Qualitative data can be incorporated in a regression model using a set of dummy variables. Consider some examples:

- It represents qualitative aspect or characteristic in the data

  Quality : good, bad; Location: south/north/east/west; characterisitcs: fat/thin or tall/short

  Time: Annual 1970s/ 1990s.; seasonal: Summer,Autumn, Winter, Spring;

- Industry or sectors of production: agriculture, mining, transportation, manufacturing, tourism, education, public services;

- Gender: male/female; Education: GCSE/UG/PD/PhD

  Subjects: Math/English/Science/Economics

- Ethnic backgrounds: Black, White, Asian, Cacasian, European, American, Latinos, Mangols, Ausis.

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \gamma D_i X_i + \varepsilon_i \quad i = 1 \ldots N \tag{273}$$

$$\varepsilon_i \sim N\left(0, \sigma^2\right) \tag{274}$$

- Here D_i is special type of variable

$$D_i = \int \begin{array}{l} 1 = \text{if the certain quality exists} \\ 0 = \text{otherwise} \end{array} \tag{275}$$

The term $\gamma D_i X_i$ picks up the interaction effect between D_i and X_i .

- Three types of dummy

  1. Intercept dummy
  2. Slope dummy
  3. Interaction between slope and intercept

Draw diagrams for this.
Examples:

- 1.
  - Earnding differences by gender, region, ethnicity or religion, occupation, education level.
  - Unemployment duration by gender, region, ethnicity or religion, occupation, education level.
  - Demand for a product by by weather, season, gender, region, ethnicity or religion, occupation, education level.
  - Test scores by gender, previous background, ethnic origin
  - Growth rates by decades, countries, exchange rate regimes

Dummy Variables Trap: Consider seasonal dummies as

$$Y_i = \beta_1 + \beta_2 X_i + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \beta_5 D_4 + \varepsilon_i \tag{276}$$

where

$$D_1 = \int \begin{array}{l} 1 = \text{if summer} \\ 0 \text{ otherwise} \end{array} \tag{277}$$

$$D_2 = \int \begin{array}{l} 1 = \text{if autumn} \\ 0 \text{ otherwise} \end{array} \tag{278}$$

$$D_3 = \int \begin{array}{l} 1 = \text{if winter} \\ 0 \text{ otherwise} \end{array} \tag{279}$$

$$D_4 = \int \begin{array}{l} 1 = \text{if spring} \\ 0 \text{ otherwise} \end{array} \tag{280}$$

- Since $\sum D_i = 1$, it will cause multicollinearity as:

$$D_1 + D_2 + D_3 + D_4 = 1 \tag{281}$$

  drop on of D_i to avoid the dummy variable trap.

Dummy Variables in a piecewise linear regression models

- Threshold effects in sales
- tariff charges by volume of transaction -mobile phones
- Panel regression: time and individual dummies
- Pay according to hierarchy in an organisation
- profit from whole sale and retail sales
- age dependent earnings -Scholarship for students, pensions and allowances for elderly
- tax allowances by level of income or business
- Investment credit by size of investment
- prices, employemnts, profits or sales for small, medium and large scale corporations
- requirements according to weight or hight of body

6.0.2 Test of Structural Change

Chow Test for stability of parameters or structural change based on sum of squared residuals.

- Use n1 and n2 observations to estimate overall and separate regressions with (n1+n2-k, n1-k, and n2-k) degrees of freedoms;
- obtain SSR1(with n1+n2-k dfs),
- SSR2 (with n1-k dfs),
- SSR3 (with n2-k dfs) and
- SSR4 = SSR1+ SSR2 (with n1+n2-2k dfs),
- obtain S5 = SSR1-SSR4;
- do F-test

$$F = \frac{\frac{SSR1}{k}}{\frac{S_5}{(n1+n2-2k)}} \tag{282}$$

The advantage of this approach to the Chow test is that it does not require the construction of the dummy and interaction variables.

6.1 Exercise 4

Suppose that you are interested in estimating the demand for beer in Yorkshire pubs and consider the following multiple regression model:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_{1,i}) + \beta_2 \ln(X_{2,i}) + \beta_3 \ln(X_{3,i}) + \beta_4 \ln(X_{4,i}) + \varepsilon_i \qquad i = 1 \ldots N \tag{283}$$

where Y_i is the demand for beer, $X_{1,i}$ is the price of beer, $X_{2,i}$ is the price of other liquor products, $X_{3,i}$ is the price of food and other services, $X_{4,i}$ is consumer income. Coefficients β_0,β_1,β_2,β_3,and β_4 are the set of unknown elasticity coefficients you would like to estimate. Again assume that errors ε_i are independently normally distributed, $\varepsilon_i \sim N(0, \sigma^2)$.

1. (a) Estimate the unknown parameters of this model using data in Beer1.csv.
   (b) How would you determine the overall significance of this model? Write down your test criterion. Compare that test statistic with another test statistic that you would use to test whether a particular coefficient, such as β_3, is statistically significant or not.
   (c) How would you establish whether a particular variable is helping to explain the variation in beer consumption?
   (d) Further suppose that you have some non-sample information on the relation between the price and income coefficients as following:
      i. sum of the elasticities equals zero: $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$.
      ii. two cross elasticities are equal: $\beta_3 = \beta_4 = 0$ or $\beta_3 - \beta_4 = 0$
      iii. income elasticity is equal to unity: $\beta_5 = 1$
   (e) How do you test whether these restrictions are valid or not ?
   (f) In addition to the variables listed in the above model you suspect that gender and level of education of individuals are important determinants of beer consumption. Explain how you could incorporate these variables in this model.
   (g) The income of an individual also depends upon his/her age. Income in turn determines the consumption of beer. Thus age interacts with income. How would you introduce this age-income interaction effect in the above model?

Instructions for testing linear restrictions in PcGive for cross section data like this:

a. regress Y on $X_{1,i}$, $X_{2,i}$, $X_{3,i}$ and $X_{4,i}$.

b. click on test/linear restriction, put the restrictions in the matrix box. one line for each restriction. For instance if $\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$ to be tested then type 1 1 1 1 1 0 , then click ok , it will test validity of that restriction. If there are two restriction

β_0+β_1+β_2+β_3+ $\beta_4 = 0$ and β_3- $\beta_4 = 0$ then
1 1 1 1 1 0
0 0 0 1 -1 0
put this input in the matrix box, then click OK. This will test for thoth restrictions.

7 Multicollinearity

Multiple Regression Model in Matrix

- Consider a linear regression

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + ... + \beta_k X_{k,i} + \varepsilon_i \quad i = 1 \ ...N \tag{284}$$

and assumptions

$$E(\varepsilon_i) = 0 \tag{285}$$

$$E(\varepsilon_i x_{j,i}) = 0; \ var(\varepsilon_i) = \sigma^2 \quad for \ \forall \ i; \ \varepsilon_i \sim N(0, \sigma^2) \tag{286}$$

$$covar(\varepsilon_i \varepsilon_j) = 0 \tag{287}$$

Explanatory variables are uncorrelated.

$$E(X_{1,i} X_{1,j}) = 0 \tag{288}$$

- Objective is to choose parameters that minimise the sum of squared errors

$$\underset{\widehat{\beta}_0 \widehat{\beta}_1 \widehat{\beta}_2 ... \widehat{\beta}_k}{Min \ S} = \sum \varepsilon_i^2 = \sum \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1,i} - \widehat{\beta}_2 X_{2,i} - \widehat{\beta}_3 X_{3,i} - .... - \widehat{\beta}_k X_{k,i} \right)^2 \tag{289}$$

Derivation of Normal Equations

$$\frac{\partial S}{\partial \widehat{\beta}_0} = 0; \frac{\partial S}{\partial \widehat{\beta}_1} = 0; \frac{\partial S}{\partial \widehat{\beta}_2} = 0; \frac{\partial S}{\partial \widehat{\beta}_3} = 0; .....\frac{\partial S}{\partial \widehat{\beta}_k} = 0 \tag{290}$$

- Normal equations for two explanatory variable case

$$\sum Y_i = \widehat{\beta}_0 N + \widehat{\beta}_1 \sum X_{1,i} + \widehat{\beta}_2 \sum X_{2,i} \tag{291}$$

$$\sum X_{1,i} Y_i = \widehat{\beta}_0 \sum X_{1,i} + \widehat{\beta}_1 \sum X_{1,i}^2 + \widehat{\beta}_2 \sum X_{1,i} X_{2,i} \tag{292}$$

$$\sum X_{2,i} Y_i = \widehat{\beta}_0 \sum X_{2,i} + \widehat{\beta}_1 \sum X_{1,i} X_{2,i} + \widehat{\beta}_2 \sum X_{2,i}^2 \tag{293}$$

$$\begin{bmatrix} \sum Y_i \\ \sum X_{1,i} Y_i \\ \sum X_{2,i} Y_i \end{bmatrix} = \begin{bmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i} X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i} X_{2,i} & \sum X_{2,i}^2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} \tag{294}$$

Normal equations in matrix form

$$\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum Y_i X_{1,i} \\ \sum Y_i X_{2,i} \end{bmatrix} \tag{295}$$

$$\beta = (X'X)^{-1} X'Y \tag{296}$$

Use Cramer Rule to solve for paramers:

$$\widehat{\beta}_0 = \frac{\begin{vmatrix} \sum Y_i & \sum X_{1,i} & \sum X_{2,i} \\ \sum Y_i X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum Y_i X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}}{\begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}} \tag{297}$$

$$\widehat{\beta}_1 = \frac{\begin{vmatrix} N & \sum Y_i & \sum X_{2,i} \\ \sum X_{1,i} & \sum Y_i X_{1,i} & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum Y_i X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}}{\begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}} \tag{298}$$

$$\widehat{\beta}_2 = \frac{\begin{vmatrix} N & \sum X_{1,i} & \sum Y_i \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum Y_i X_{1,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum Y_i X_{2,i} \end{vmatrix}}{\begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}} \tag{299}$$

Evaluate the determinant:

$$|X'X| = \begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix} \tag{300}$$

For this calculation, repeate first two columns as:

$$|X'X| = \begin{vmatrix} N & \sum X_{1,i} & \sum X_{2,i} & N & \sum X_{1,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} & \sum X_{1,i} & \sum X_{1,i}^2 \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 & \sum X_{2,i} & \sum X_{1,i}X_{2,i} \end{vmatrix} \tag{301}$$

Dederminant = (sum of cross product from to left to right - sum of cross product from bottom left to right) as:

$$|X'X| = N\sum X_{1,i}^2 \sum X_{2,i}^2 + \sum X_{1,i} \sum X_{1,i}X_{2,i} \sum X_{2,i} + \sum X_{2,i} \sum X_{1,i}X_{2,i} \sum X_{1,i}$$
$$- \sum X_{2,i} \sum X_{2,i} \sum X_{1,i}^2 - N\sum X_{1,i}X_{2,i} \sum X_{1,i}X_{2,i} - \sum X_{2,i}^2 \sum X_{1,i} \sum X_{1,i}$$

7.0.1 Exact multicollinearity: Singularity

In existence of exact multicollinearity $X'X$ is singular, i.e. $|X'X| = 0$

If $X_{1,i} = \lambda X_{2,i}$ then
$$|X'X| = N\sum X_{1,i}^2 \sum X_{2,i}^2 + \sum X_{1,i} \sum X_{1,i}X_{2,i} \sum X_{2,i} + \sum X_{2,i} \sum X_{1,i}X_{2,i} \sum X_{1,i}$$
$$- \sum X_{2,i} \sum X_{2,i} \sum X_{1,i}^2 - N\sum X_{1,i}X_{2,i} \sum X_{1,i}X_{2,i} - \sum X_{2,i}^2 \sum X_{1,i} \sum X_{1,i}$$

Substituting out $X_{1,i}$

$$|X'X| = N\lambda^2 \sum X_{2,i}^2 \sum X_{2,i}^2 + \lambda^2 \sum X_{2,i} \sum X_{2,i}^2 \sum X_{2,i} + \lambda^2 \sum X_{2,i} \sum X_{2,i}^2 \sum X_{2,i}$$
$$-\lambda^2 \sum X_{2,i} \sum X_{2,i} \sum X_{2,i}^2 - N\lambda^2 \sum X_{2,i}^2 \sum X_{2,i}^2 - \lambda^2 \sum X_{2,i}^2 \sum X_{2,i} \sum X_{2,i}$$
$$= 0$$

$$|X'X| = \begin{vmatrix} N & \lambda\sum X_{2,i} & \sum X_{2,i} \\ \lambda\sum X_{2,i} & \lambda^2\sum X_{2,i}^2 & \lambda\sum X_{2,i}X_{2,i} \\ \sum X_{2,i} & \lambda\sum X_{2,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix} = 0 \tag{302}$$

Parameters are indeterminate in model with exact multicollinearity

$$\widehat{\beta}_0 = \frac{\begin{vmatrix} \sum Y_i & \sum X_{1,i} & \sum X_{2,i} \\ \sum Y_i X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum Y_i X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}}{0} = \infty \tag{303}$$

$$\widehat{\beta}_1 = \frac{\begin{vmatrix} N & \sum Y_i & \sum X_{2,i} \\ \sum X_{1,i} & \sum Y_i X_{1,i} & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum Y_i X_{2,i} & \sum X_{2,i}^2 \end{vmatrix}}{0} = \infty \tag{304}$$

$$\widehat{\beta}_2 = \frac{\begin{vmatrix} N & \sum X_{1,i} & \sum Y_i \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum Y_i X_{1,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum Y_i X_{2,i} \end{vmatrix}}{0} = \infty \tag{305}$$

Covariance of parameters cannot be estimated in model with exact multicollinearity

$$(X'X)^{-1} = \infty \tag{306}$$

$$cov\left(\widehat{\beta}\right)=\left(\begin{array}{ccc} var(\widehat{\beta}_1) & cov(\widehat{\beta}_1\widehat{\beta}_2) & cov(\widehat{\beta}_1\widehat{\beta}_3) \\ cov(\widehat{\beta}_1\widehat{\beta}_2) & var(\widehat{\beta}_2) & cov(\widehat{\beta}_2\widehat{\beta}_3) \\ cov(\widehat{\beta}_1\widehat{\beta}_3) & cov(\widehat{\beta}_2\widehat{\beta}_3) & var(\widehat{\beta}_3) \end{array}\right)=\infty \tag{307}$$

$$cov\left(\widehat{\beta}\right)=\left(X'X\right)^{-1}\sigma^2=\infty \tag{308}$$

$$cov\left(\widehat{\beta}\right)=\left[\begin{array}{ccc} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{array}\right]^{-1}\widehat{\sigma}^2=\infty \tag{309}$$

Table 10: Data for testing multicollinearity

y	3	5	7	6	9	6	7
x1	1	2	3	4	5	6	7
x2	5	10	15	20	25	30	35

Numerical example of exact multicollinearity Evaluate the determinant

$$|X'X|=\left|\begin{array}{ccc} N & \sum X_{1,i} & \sum X_{2,i} \\ \sum X_{1,i} & \sum X_{1,i}^2 & \sum X_{1,i}X_{2,i} \\ \sum X_{2,i} & \sum X_{1,i}X_{2,i} & \sum X_{2,i}^2 \end{array}\right|=\left|\begin{array}{ccc} 7 & 28 & 140 \\ 28 & 140 & 700 \\ 140 & 700 & 3500 \end{array}\right|; \tag{310}$$

$$\left[\begin{array}{c} \sum Y_i \\ \sum Y_iX_{1,i} \\ \sum Y_iX_{2,i} \end{array}\right]=\left[\begin{array}{c} 43 \\ 188 \\ 980 \end{array}\right]$$

Numerical example of exact multicollinearity

$|X'X|=N\sum X_{1,i}^2\sum X_{2,i}^2+\sum X_{1,i}\sum X_{1,i}X_{2,i}\sum X_{2,i}+\sum X_{2,i}\sum X_{1,i}X_{2,i}\sum X_{1,i}$
$-\sum X_{2,i}\sum X_{2,i}\sum X_{1,i}^2-N\sum X_{1,i}X_{2,i}\sum X_{1,i}X_{2,i}-\sum X_{2,i}^2\sum X_{1,i}\sum X_{1,i}$
$=(7\times 140\times 3500+28\times 700\times 140+140\times 700\times 28$
$-140\times 140\times 140-7\times 700\times 700-28\times 28\times 3500)=0$

Evaluate determinants easily in excel using following steps:

1. select the cell where to put the result.and press shift and control continously by two fingers of left hand
2. use mouse by right hand to choose math and trig function
3. choose MDETERM
4. Select matrix for which to evaluate the determinant
5. press OK and you will see the reslut.

Normal equations of a multiple regression in deviation form:

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i x_{1,i} \\ \sum y_i x_{2,i} \end{bmatrix} \tag{311}$$

$$\beta = (X'X)^{-1} X'Y \tag{312}$$

$$\widehat{\beta}_1 = \frac{\begin{vmatrix} \sum y_i x_{1,i} & \sum x_{1,i}x_{2,i} \\ \sum y_i x_{2,i} & \sum x_{2,i}^2 \end{vmatrix}}{\begin{vmatrix} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{vmatrix}} \tag{313}$$

$$\widehat{\beta}_2 = \frac{\begin{vmatrix} \sum x_{1,i}^2 & \sum y_i x_{1,i} \\ \sum x_{1,i}x_{2,i} & \sum y_i x_{2,i} \end{vmatrix}}{\begin{vmatrix} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{vmatrix}} \tag{314}$$

Variances of parameters:

$$\begin{aligned} & \begin{bmatrix} \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{bmatrix}^{-1} \\ = \quad & \frac{1}{\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2} \begin{bmatrix} \sum x_{2,i}^2 & -\sum x_{1,i}x_{2,i} \\ -\sum x_{1,i}x_{2,i} & \sum x_{1,i}^2 \end{bmatrix} \end{aligned} \tag{315}$$

$$var\left(\widehat{\beta}_1\right) = \frac{\sum x_{2,i}^2}{\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2}\sigma^2 \tag{316}$$

$$var\left(\widehat{\beta}_2\right) = \frac{\sum x_{1,i}^2}{\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2}\sigma^2 \tag{317}$$

Variance Inflation Factor (VIF) in inexact multicollinearity Significant R^2 but insignificant t-ratios. why?

When Variance is high the standard errors are high and that makes t-statistics very small and insignificant

$$\begin{aligned} SE\left(\widehat{\beta}_2\right) &= \sqrt{var\left(\widehat{\beta}_2\right)}\ ;\ SE\left(\widehat{\beta}_1\right) = \sqrt{var\left(\widehat{\beta}_1\right)}\ ; \\ t_{\widehat{\beta}_1} &= \frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)};\ t_{\widehat{\beta}_2} = \frac{\widehat{\beta}_2 - \beta_2}{SE\left(\widehat{\beta}_2\right)} \end{aligned} \tag{318}$$

.since $0 < r_{12} < 1$ it raises the variance and hence stancard errors and lowers t-values.

1. First detect the pairwise correlations between explalantory variables such $X_{1,i}$ and $X_{3,i}$ be given by r_{12}.

2. Drop highly correlated variables.

3. Adopts Klein's rule of thumb:

4. Compare R_y^2 from overall regression to R_x^2 from auxiliary regression. Determine multicollinearity if $R_x^2 > R_y^2$. Drop highly correlated variables.

Let correlations between $X_{1,i}$ and $X_{2,i}$ be given by r_{12}. Then Variance inflation factor is $\frac{1}{(1-r_{12}^2)}$

$$\begin{aligned} var\left(\widehat{\beta}_2\right) &= \frac{\sum x_{1,i}^2}{\left[\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2\right]}\sigma^2 \\ &= \frac{1}{\left[\frac{\sum x_{1,i}^2 \sum x_{2,i}^2}{\sum x_{1,i}^2} - \frac{\left(\sum x_{1,i}x_{2,i}\right)^2}{\sum x_{1,i}^2}\right]}\sigma^2 \\ &= \frac{1}{\sum x_{2,i}^2 \left[\frac{\sum x_{1,i}^2}{\sum x_{1,i}^2} - \frac{\left(\sum x_{1,i}x_{2,i}\right)^2}{\sum x_{2,i}^2 \sum x_{1,i}^2}\right]}\sigma^2 \\ &= \frac{1}{\sum x_{2,i}^2 \left[1 - r_{12}^2\right]}\sigma^2 \\ &= \frac{1}{(1-r_{12}^2)}\frac{1}{\sum x_{2,i}^2}\sigma^2 \end{aligned} \tag{319}$$

Let correlations between $X_{1,i}$ and $X_{2,i}$ be given by r_{12}. Then Variance inflation factor is $\frac{1}{(1-r_{12}^2)}$

$$\begin{aligned} var\left(\widehat{\beta}_1\right) &= \frac{\sum x_{2,i}^2}{\sum x_{1,i}^2 \sum x_{2,i}^2 - \left(\sum x_{1,i}x_{2,i}\right)^2}\sigma^2 \\ & \frac{1}{\left[\frac{\sum x_{1,i}^2 \sum x_{2,i}^2}{\sum x_{2,i}^2} - \frac{\left(\sum x_{1,i}x_{2,i}\right)^2}{\sum x_{2,i}^2}\right]}\sigma^2 \\ &= \frac{1}{\sum x_{1,i}^2 \left[\frac{\sum x_{2,i}^2}{\sum x_{2,i}^2} - \frac{\left(\sum x_{1,i}x_{2,i}\right)^2}{\sum x_{1,i}^2 \sum x_{2,i}^2}\right]}\sigma^2 \\ &= \frac{1}{\sum x_{1,i}^2 \left[1 - r_{12}^2\right]}\sigma^2 \end{aligned} \tag{320}$$

Table 11: Data on income, performace and quality of work

y	3	5	7	6	9	6	7
x_1	1	2	3	4	5	6	7
x_2	5	10	15	20	25	30	35

7.0.2 Exercise 5

1. Data on income (y), performance indicator (x_1) and quality of workers (x_2) in a certatin reputable company is given as following.

Fit a regression model $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$ for this company. If any problem suggest remedial measures.

8 Heteroskedasticity

Heteroskedasticity occurs when variances of errors are not constant, $var(\varepsilon_i) \neq \sigma_i^2$ variance of errors vary for each i. This is mainly a cross section problem. OLS estimator is still unbiased as:

$$E\left(\widehat{\beta}_2\right) = \beta_2 \tag{321}$$

But it is not consistency as its variance tends to infinity

$$Var\left(\widehat{\beta}_2\right) = \frac{1}{\sum x_i^2}\widehat{\sigma}^2 \tag{322}$$

OLS estimators are still unbiased but they are no long efficient:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \quad i = 1 \ ...N \tag{323}$$

and assumptions

$$E(\varepsilon_i) \ = 0 \tag{324}$$

$$E(\varepsilon_i x_i) \ = 0 \tag{325}$$

$$var(\varepsilon_i) \ = \sigma^2 \quad for \ \ \forall \ \ i \tag{326}$$

$$covar(\varepsilon_i \varepsilon_j) \ = 0 \tag{327}$$

Then the OLS Regression coefficients are:

$$\widehat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}; \ \ \widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2 \overline{X} \tag{328}$$

Main reason for this are

- Learning reduces errors;

- driving practice, driving errors and accidents
- typing practice and typing errors,
- defects in productions and improved machines

- Improved data collection: better formulas and goods software
- More heteroscedasticity exists in cross section than in time series data.

Nature of Heteroskedasticity

$$E\left(\varepsilon_i\right)^2 = \sigma_i^2 \tag{329}$$

$$\widehat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{330}$$

$$E\left(\widehat{\beta}_2\right) = \sum w_i y_i \tag{331}$$

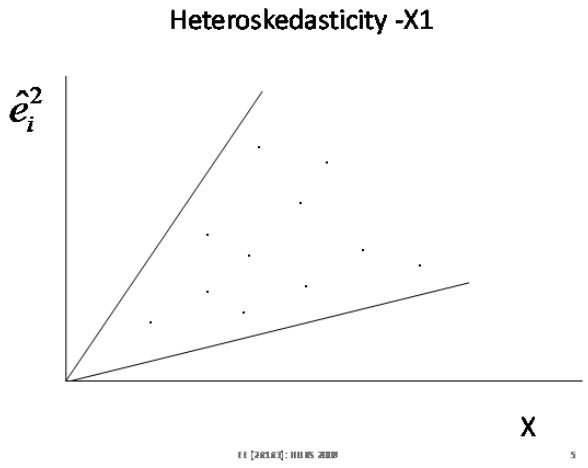
where

$$w_i \quad = \frac{x_i}{\sum x_i^2} = \frac{\left(X_i - \overline{X}\right)}{\sum \left(X_i - \overline{X}\right)^2} \tag{332}$$

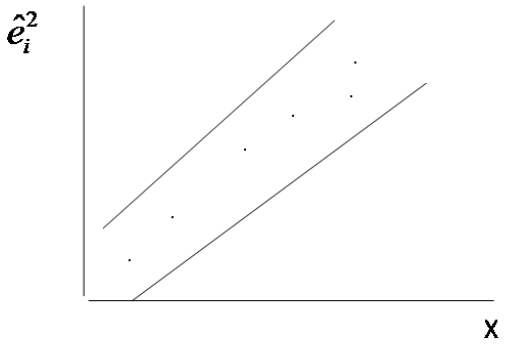
$$Var\left(\widehat{\beta}_2\right) = var\left[\frac{\sum \left(X_i - \overline{X}\right)}{\sum \left(X_i - \overline{X}\right)^2}\right] var\left(y_i\right) = \frac{\sum x_i^2 \sigma_i^2}{\left[\sum x_i^2\right]^2} \tag{333}$$

8.1 Graphical detection of the Heteroskedasticity

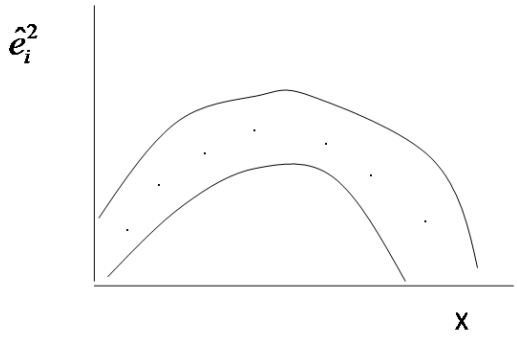
Under heteroskedasticity there is systematic pattern of errors with explanatory variable:



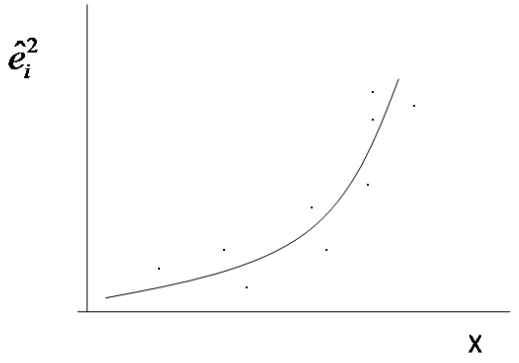
Hetroskedasticity -X2

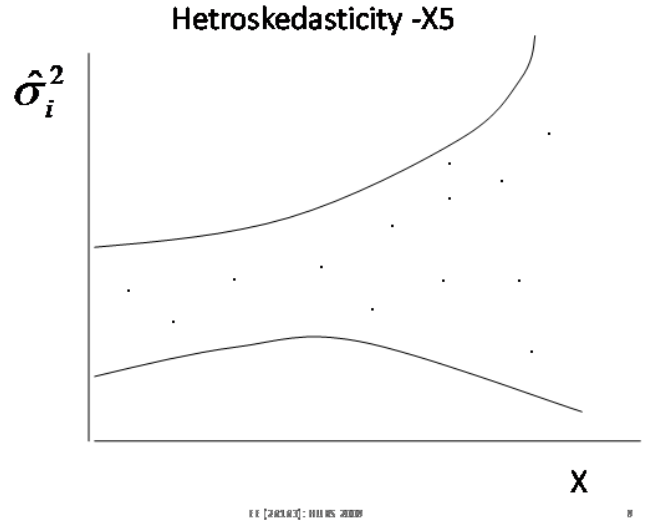


Hetroskedasticity -X3



Heteroskedasticity -X4





As mentioned before when errors are homoskedastic parameters are sunbiased $E\left(\widehat{\beta}_2\right) = \beta_2$ and efficient, $Var\left(\widehat{\beta}_2\right) = \frac{1}{\sum x_i^2}\widehat{\sigma}^2$ and $\lim_{N \to \infty} Var\left(\widehat{\beta}_2\right) = \lim_{N \to \infty} \frac{\frac{\widehat{\sigma}^2}{N}}{\frac{\sum x_i^2}{N}} = 0$
.

OLS Estimator is still unbiased

$$\widehat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \sum w_i y_i \tag{334}$$

$$E\left(\widehat{\beta}_2\right) = E\left(\sum w_i y_i\right) = E\sum w_i\left(\beta_1 + \beta_2 X_i + \varepsilon_i\right) \tag{335}$$

$$E\left(\widehat{\beta}_2\right) = \beta_1 E\left(\sum w_i\right) + \beta_2 E\left(\sum w_i x_i\right) + E\left(\sum w_i \varepsilon_i\right) \tag{336}$$

But the OLS parameter is inconsistent in the presence of heteroskedasticity

$$E\left(\widehat{\beta}_2\right) = \sum w_i y_i \tag{337}$$

$$E\left(\widehat{\beta}_2\right) = E\left(\sum w_i y_i\right) = E\sum w_i\left(\beta_1 + \beta_2 X_i + \varepsilon_i\right) \tag{338}$$

$$E\left(\widehat{\beta}_2\right) = \beta_1 E\left(\sum w_i\right) + \beta_2 E\left(\sum w_i x_i\right) + E\left(\sum w_i \varepsilon_i\right) \tag{339}$$

$$E\left(\widehat{\beta}_2\right) = \beta_2 + E\left(\sum w_i \varepsilon_i\right) \tag{340}$$

$$Var\left(\widehat{\beta}_2\right) = E\left[E\left(\widehat{\beta}_2\right) - \beta_2\right]^2 = E\left(\sum w_i\varepsilon_i\ \right)^2 \tag{341}$$

$$Var\left(\widehat{\beta}_2\right) = E\left(\sum\sum w_i^2\varepsilon_i^2\ \right) + \sum\sum cov\left(\varepsilon_i\varepsilon_j\right)^2 \tag{342}$$

$$Var\left(\widehat{\beta}_2\right) = \frac{\sum x_i^2\sigma_i^2}{\left[\sum x_i^2\right]^2} \tag{343}$$

OLS Estimator is inconsistent assymptotically becasue $Var\left(\widehat{\beta}_2\right) \Rightarrow \infty$ and the sample size becomes larger, $N \to \infty$.

$$Var\left(\widehat{\beta}_2\right) = \frac{\sum x_i^2\sigma_i^2}{\left[\sum x_i^2\right]^2} \tag{344}$$

$$\lim_{N\to\infty} Var\left(\widehat{\beta}_2\right) = \lim_{N\to\infty} \frac{\sum x_i^2\sigma_i^2}{\left[\sum x_i^2\right]^2} \Rightarrow \infty \tag{345}$$

Various tests of heteroskedasticity

- Spearman Rank Test
- Park Test
- Goldfeld-Quandt Test
- Glesjer Test
- Breusch-Pagan,Godfrey test
- White Test
- ARCH test

  (See food_hetro.xls excel spreadsheet for some exmaples on how to compute these. Gujarati (2003) Basic Econometrics,McGraw Hill is a good text for Heteroskedasticity; x-hetro test in PcGive). STATA and Eview have many options to test for the heterskedsticity.

Spearman rank test of heteroskedactity

$$r_s = 1 - 6 \times \frac{\sum_i d_i^2}{n\left(n^2-1\right)} \tag{346}$$

- steps:
- run OLS of y on x.
- obtain errors e
- rank e and y or x
- find the difference of the rank
- use t-statistics if ranks are significantly different assuming n>8 and rank correlation coefficient ρ=0.

$$t = 1 - 6 \times \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \quad with \quad df \quad (n-2) \tag{347}$$

- If $t_{cal} > t_{crit}$ there is heteroskedasticity.

Glesjer Test of heteroskedasticity

- Model

$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad i = 1 \ldots N \tag{348}$$

- There are a number of versions of it:

$$|e_i| = \beta_1 + \beta_2 X_i + v_i \tag{349}$$

$$|e_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i \tag{350}$$

$$|e_i| = \beta_1 + \beta_2 \frac{1}{X_i} + v_i \tag{351}$$

$$|e_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i \tag{352}$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i \tag{353}$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i \tag{354}$$

- In each case dot-test H_0: $\beta_i = 0$ against H_A: $\beta_i \neq 0$. If is significant then that is the evidence of heteroskedasticity.

White test **White test of heteroskedasticity is more general test**
$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i \quad i = 1 \ldots N$

- run OLS and obtain error squares $\widehat{e}_i^2$
- regress $\widehat{e}_i^2 = \alpha_0 + \alpha_1 X_{1,i} + \alpha_2 X_{2,i} + + \alpha_3 X_{3,i} + \alpha_4 X_{1,i}^2 + \alpha_5 X_{2,i}^2 + \alpha_6 X_{3,i}^2 + \alpha_7 X_{1,i} X_{2,i} + \alpha_8 X_{2,i} X_{3,i} + v_i$
- Compute test statistics $n.R^2 = \chi_{df}^2$
- If the calculated χ_{df}^2 value is greater that the χ_{df}^2 table value then, there is evidence of heteroskedasticity.

Park test of heteroskedasticity

- Model
$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad i = 1 \ldots N \tag{355}$$
- Error square:
$$\sigma_i^2 = \sigma^2 X_i^{\beta} e_i^{v_i} \tag{356}$$
- Or taking log
$$ln\sigma_i^2 = \ln \sigma^2 + \beta_2 X_i + v_i \tag{357}$$
- steps : run the OLS regression for (Y_i) and get the estimates of error terms (e_i) .
- Square e_i , and then run a regression of lne_i^2 with x variable. Do t-test H0: $\beta_2 = 0$ against HA: $\beta_2 \neq 0$. If is significant then that is the evidence of heteroskedasticity.

Goldfeld-Quandt test of heteroskedasticity

- Model
$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad i = 1 \ldots N \tag{358}$$
- Steps:
  - Rank observations in ascending order of one of the x variable
  - Omit c numbers of central observations leaving two groups $\frac{N-C}{2}$ with number of osbervations
  - Fit OLS to the first $\frac{N-C}{2}$ and the last $\frac{N-C}{2}$ observations and find sum of the squared errors from both of them.
  - Set hypothesis $\sigma_1^2 = \sigma_2^2$ against $\sigma_1^2 \neq \sigma_2^2$.
  - compute $\lambda = \frac{ERSS_2/df2}{ERSS_1/df1}$.
  - It follows F distribution.

Breusch-Pagan,Godfrey test of heteroskedasticity $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + ... + \beta_k X_{k,i} + \varepsilon_i \quad i = 1 \ ...N$

- run OLS and obtain error squares
- Obtain average error square $\widehat{\sigma}^2 = \frac{\sum_i e_i^2}{n}$and $p_i = \frac{e_i^2}{\widehat{\sigma}^2}$
- regress p_i on a set of explanatory variables
- $p_i = \alpha_0 + \alpha_1 X_{1,i} + \alpha_2 X_{2,i} + \alpha_3 X_{3,i} + ... + \alpha_k X_{k,i} + \varepsilon_i$
- obtain squares of explained sum (EXSS)
- $\theta = \frac{1}{2}(EXSS)$
- $\theta = \frac{1}{m-1}(EXSS) \sim \chi^2_{m-1}$
- $H_0 : \alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = .. = \alpha_k = 0$
- No heteroskedasticity and $\sigma_i^2 = \alpha_1$ a constant. If calculated χ^2_{m-1} is greater than table value there is an evidence of heteroskedasticity.

ARCH test of heteroskedasticity Engle (1987) autoregressive conditional heteroskedasticy (ARCH): more useful for time series data

Model has mean and variance equations.

$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \beta_3 X_{3,t} + ... + \beta_k X_{k,t} + e_t$

$\varepsilon_t \sim N\left(0, \left(\alpha_0 + \alpha_2 e_{t-1}^2\right)\right)$

Variance equation is:

$$\sigma_t^2 = \alpha_0 + \alpha_2 e_{t-1}^2 \tag{359}$$

Here σ_t^2 not observed. Simple way is to run OLS of Y_t and get $\widehat{e}_t^2$

- ARCH (1)

$\widehat{e}_t^2 = \alpha_0 + \alpha_2 \widehat{e}_{t-1}^2 + v_t$

- ARCH (p)

$\widehat{e}_t^2 = \alpha_0 + \alpha_2 \widehat{e}_{t-1}^2 + \alpha_3 \widehat{e}_{1-1}^2 + \alpha_4 \widehat{e}_{1-1}^2 + .. + \alpha_p \widehat{e}_{1-p}^2 + v_t$

- Compute the test statistics

$n.R^2 \sim \chi^2_{df}$

Again if the calculated χ^2_{df} is greater than table value there is an evidence of ARCH effect and heteroskedasticity.

Both ARCH and GARRCH modesl are estimated using iterative Maximum Likelihood procedure.

GARCH tests of heteroskedasticity Bollerslev's generalised autoregressive conditional heteroskedasticy (GARCH) process is more general

- GARCH (1)

$$\sigma_t^2 = \alpha_0 + \alpha_2 \widehat{e}_{t-1}^2 + \beta \sigma_{t-1}^2 + v_t \tag{360}$$

Here α terms measure the impact of unknown elements, and β measure conditional elements.

- GARCH (p,q)
- $\sigma_t^2 = \alpha_0 + \alpha_2 \widehat{e}_{t-1}^2 + \alpha_3 \widehat{e}_{t-2}^2 + \alpha_4 \widehat{e}_{t-3}^2 + .. + \alpha_p \widehat{e}_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + ..\beta_q \sigma_{t-q}^2 + .. + v_t$
- Compute the test statistics $n.R^2 \tilde{} \chi_{df}^2$
- Sometimes written as
- $h_t = \alpha_0 + \alpha_2 \widehat{e}_{t-1}^2 + \alpha_3 \widehat{e}_{t-2}^2 + \alpha_4 \widehat{e}_{t-3}^2 + .. + \alpha_p \widehat{e}_{t-p}^2 + \beta_1 h_{t-1} + \beta_2 h_{t-2} + ..\beta_q h_{t-q} + .. + v_t$
- $where \quad h_t = \sigma_t^2$
- Various functional forms of h_t
- $h_t = \alpha_0 + \alpha_2 \widehat{e}_{t-1}^2 + \beta_1 \sqrt{h_{t-1}} + v_i$ or $h_t = \alpha_0 + \alpha_2 \widehat{e}_{t-1}^2 + \sqrt{\beta_1 h_{t-1} + \beta_2 h_{t-2}} + v_i$
- Both ARCH and GARCH modesl are estimated using iterative Maximum Likelihood procedure. Volatility package in PcGive estimates ARCH-GARCH models; Eviews, STATA or RATS also have these routines.

8.1.1 Relation between a GARCH and ARCH process

GARCH (1,1) process is infinite ARCH(p) process. It is proved as following:
Let error variance as $h_t = \sigma_t^2$ and write GARCH (1,1) for $0 < \delta < 1$ as:

$$h_t = \gamma_0 + \gamma_1 \widehat{e}_{t-1}^2 + \delta h_{t-1} \tag{361}$$

Now contineously substitute out h_{t-q} tersm as:

$$\begin{aligned}
h_t &= \gamma_0 + \gamma_1 \widehat{e}_{t-1}^2 + \delta \left[\gamma_0 + \gamma_1 \widehat{e}_{t-2}^2 + \delta h_{t-2}\right] \\
&= \gamma_0 + \gamma_1 \widehat{e}_{t-1}^2 + \delta\gamma_0 + \delta\gamma_1 \widehat{e}_{t-2}^2 + \delta^2 h_{t-2} && (362) \\
&= \gamma_0 + \gamma_1 \widehat{e}_{t-1}^2 + \delta\gamma_0 + \delta\gamma_1 \widehat{e}_{t-2}^2 + \delta^2 \left[\gamma_0 + \gamma_1 \widehat{e}_{t-3}^2 + \delta h_{t-3}\right] && (363) \\
&= \gamma_0 + \gamma_1 \widehat{e}_{t-1}^2 + \delta\gamma_0 + \delta\gamma_1 \widehat{e}_{t-2}^2 + \delta^2\gamma_0 + \delta^2\gamma_1 \widehat{e}_{t-3}^2 + \delta^3 h_{t-3} && (364) \\
&= .. && (365) \\
&= \gamma_0 + \delta\gamma_0 + \delta^2\gamma_0 + ... + \gamma_1 \widehat{e}_{t-1}^2 + \delta\gamma_1 \widehat{e}_{t-2}^2 + \delta^2\gamma_1 \widehat{e}_{t-3}^2 + .. + \delta^J h_t && (366) \\
&\simeq \frac{\gamma_0}{1-\delta} + \gamma_1 \left[\widehat{e}_{t-1}^2 + \delta \widehat{e}_{t-2}^2 + \delta^2 \widehat{e}_{t-3}^2\right] = \frac{\gamma_0}{1-\delta} + \gamma_1 \sum_{J=1}^{\infty} \delta^{j-1} \widehat{e}_{t-j}^2 && (367)
\end{aligned}$$

Now $h_t = \frac{\gamma_0}{1-\delta} + \gamma_1 \sum_{J=1}^{\infty} \delta^{j-1} \widehat{e}_{t-j}^2$ is just ARCH(∞) process.

STATA and Eviews are very handy in estimating ARCH/GARCH models. Eviews gives mean and variance estimates simultaneouly. With a date file in *.csv format, use File/option/foreign data as worksheet file/quick/estimation/ARCH sequnce of command to estimate an ARCH model. From view command check all diagnostic. Do forecsts and study the conditional volatility as given by the model.

8.1.2 Weighted Least Square Method

GLS Solution of the heteroskedasticity problem when the variance is known:

$$\frac{Y_i}{\sigma_i} = \frac{\beta_1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i} \quad i = 1 \ldots N \tag{368}$$

Variance with this tranformation equals 1. $var\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{\sigma_i^2}{\sigma_i^2} = 1$
if

$$\sigma_i^2 = \sigma^2 X_i \tag{369}$$

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{\varepsilon_i}{X_i} \; ; \quad var\left(\frac{\varepsilon_i}{x_i}\right) = \frac{\sigma^2 x_i^2}{x_i^2} = \sigma^2 \tag{370}$$

In matrix notation

$$\beta_{OLS} = \left(X'X\right)^{-1}\left(X'Y\right) \tag{371}$$

$$\beta_{GLS} = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right) \tag{372}$$

Ω^{-1} is inverse of variance covariance matrix.

9 Autocorrelation

Autocorrelation occurs when covariances of errors are not zero, $covar\left(\varepsilon_t \varepsilon_{t-1}\right)$ $\neq 0$ covariance of errors are nonnegative This is mainly a problem observed in time series data.

Consider a linear regression

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \quad t = 1 \ldots T \tag{373}$$

Classical assumptions

$$E\left(\varepsilon_t\right) \; = 0 \tag{374}$$

$$E\left(\varepsilon_t x_{_t}\right) \; = 0 \tag{375}$$

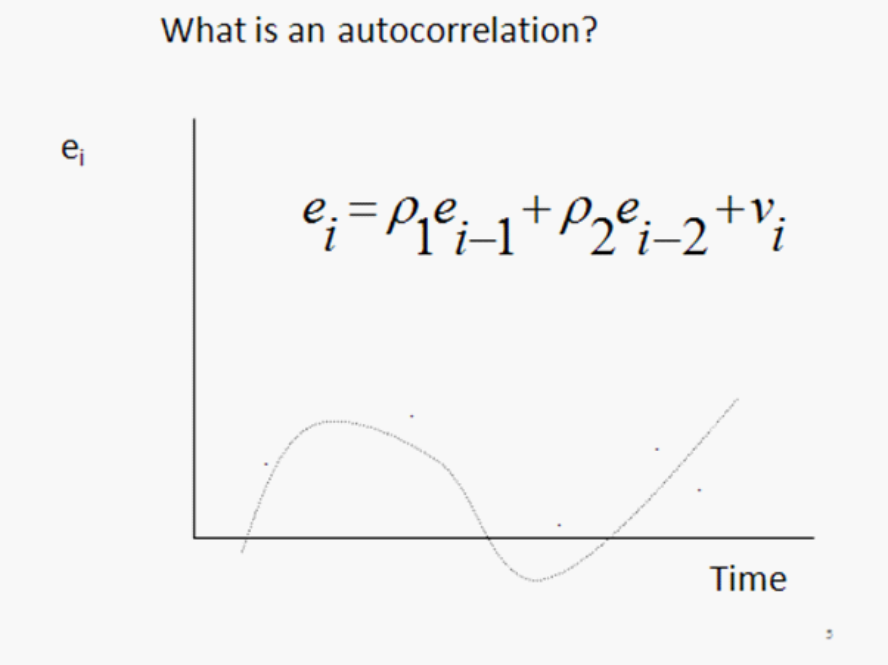
$$var\left(\varepsilon_t\right) = \sigma^2 \quad for\ \forall\ t \quad covar\left(\varepsilon_t \varepsilon_{t-1}\right) = 0 \tag{376}$$

In presence of autocorrelation (first order)

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t \tag{377}$$

Then the OLS Regression coefficients are:

$$\widehat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2};\ \ \widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2 \overline{X}\ \ ; \widehat{\rho} = \frac{\sum e_t e_{t-1}}{\sum e_t^2} \tag{378}$$



Causes of autocorrelation

- inertia , specification bias, cobweb phenomena
- manipulation of data

Consequences of autocorrelation

1. (a) Estimators are still linear and unbiased, but
   (b) they there not the best, they are inefficient.

Remedial measures

1. (a) When ρ is known - transform the model
   (b) When ρ is unknown estimate it and transform the model

9.0.3 Nature of Autocorrelation

$$\widehat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} \tag{379}$$

$$E\left(\widehat{\beta}_2\right) = \sum w_t y_t \tag{380}$$

where

$$E\left(\varepsilon_t\right)^2 = \sigma^2 \tag{381}$$

$$E\left(\widehat{\beta}_2\right) = \beta_2 + E\left(\sum w_t \varepsilon_t\right) \tag{382}$$

$$E\left(\widehat{\beta}_2\right) = \beta_1 E\left(\sum w_t\right) + \beta_2 E\left(\sum w_t x_t\right) + E\left(\sum w_t \varepsilon_t\right) \tag{383}$$

$$Var\left(\widehat{\beta}_2\right) = E\left[E\left(\widehat{\beta}_2\right) - \beta_2\right]^2 = E\left(\sum w_t \varepsilon_t\right)^2 \tag{384}$$

$$Var\left(\widehat{\beta}_2\right) = \frac{1}{\sum x_t^2}\sigma^2 + \sum \sum cov\left(\varepsilon_t \varepsilon_{t-1}\right) \tag{385}$$

OLS Estimator is still unbiased

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t \tag{386}$$

$$\widehat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} = \sum w_t y_t \tag{387}$$

$$E\left(\widehat{\beta}_2\right) = E\left(\sum w_t y_t\right) = E\sum w_t\left(\beta_1 + \beta_2 X_t + \varepsilon_t\right) \tag{388}$$

$$E\left(\widehat{\beta}_2\right) = \beta_1 E\left(\sum w_t\right) + \beta_2 E\left(\sum w_t x_t\right) + E\left(\sum w_t \varepsilon_t\right) \tag{389}$$

$$E\left(\widehat{\beta}_2\right) = \beta_2 \tag{390}$$

9.0.4 OLS Parameters are inefficient with Autocorrelation

$$E\left(\widehat{\beta}_2\right) = \sum w_t y_t \tag{391}$$

$$E\left(\widehat{\beta}_2\right) = E\left(\sum w_t y_t\right) = E\sum w_t\left(\beta_1 + \beta_2 X_t + \varepsilon_t\right) \tag{392}$$

$$E\left(\widehat{\beta}_2\right) = \beta_1 E\left(\sum w_t\right) + \beta_2 E\left(\sum w_t x_t\right) + E\left(\sum w_t \varepsilon_t\right) \tag{393}$$

$$E\left(\widehat{\beta}_2\right) = \beta_2 + E\left(\sum w_t \varepsilon_t\right) \tag{394}$$

$$Var\left(\widehat{\beta}_2\right) = E\left[E\left(\widehat{\beta}_2\right) - \beta_2\right]^2 = E\left(\sum w_t\varepsilon_t\right)^2 \tag{395}$$

$$Var\left(\widehat{\beta}_2\right) = E\left(\sum\sum w_t^2\varepsilon_t^2\right) + 2\sum\sum w_t w_{t-1} cov\left(\varepsilon_t\varepsilon_{t-1}\right) \tag{396}$$

$$Var\left(\widehat{\beta}_2\right) = \frac{1}{\sum x_t^2}\sigma^2\left[1 + 2\frac{\sum x_t x_{t-1}}{[\sum x_t^2]}\frac{cov\left(\varepsilon_t\varepsilon_{t-1}\right)}{\sqrt{var\left(\varepsilon_t\right)}}\right] \because var\left(\varepsilon_t\right) = var\left(\varepsilon_{t-1}\right) \tag{397}$$

$$Var\left(\widehat{\beta}_2\right) = \frac{1}{\sum x_t^2}\sigma^2\left[\begin{array}{c} 1 + 2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^1 + \\ +2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^2 + .. + 2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^s \end{array}\right] \tag{398}$$

OLS Estimator is inconsistent assymptotically

$$Var\left(\widehat{\beta}_2\right) = \frac{1}{\sum x_t^2}\sigma^2\left[\begin{array}{c} 1 + 2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^1 + \\ +2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^2 + .. + 2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^s \end{array}\right] \tag{399}$$

$$\underset{\lim\ \ N\to\infty}{Var\left(\widehat{\beta}_2\right)} = \frac{1}{\sum x_t^2}\sigma^2\left[\begin{array}{c} 1 + 2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^1 + \\ +2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^2 + .. + 2\frac{\sum(x_t-\overline{x})(x_{t-1}-\overline{x})}{\sum x_t^2}\rho^s \end{array}\right] \Rightarrow \infty \tag{400}$$

9.0.5 Durbin-Watson test of autocorrelation

$$d = \frac{\sum\limits_{t=1}^{T}\left(e_t - e_{t-1}\right)^2}{\sum\limits_{t=1}^{T}e_t^2} \tag{401}$$

$$d = \frac{\sum\limits_{t=1}^{T}\left(e_t^2 - 2e_t e_{t-1} + e_{t-1}^2\right)}{\sum\limits_{t=1}^{T}e_t^2} = 2\left(1-\rho\right);\ \ \because \sum_{t=1}^{T}e_t^2 \simeq \sum_{t=2}^{T}e_{t-1}^2 \tag{402}$$

Autocorrelation coefficient is given by:

$$\rho = \frac{\sum\limits_{t=1}^{T}e_t e_{t-1}}{\sum\limits_{t=1}^{T}e_t^2} \tag{403}$$

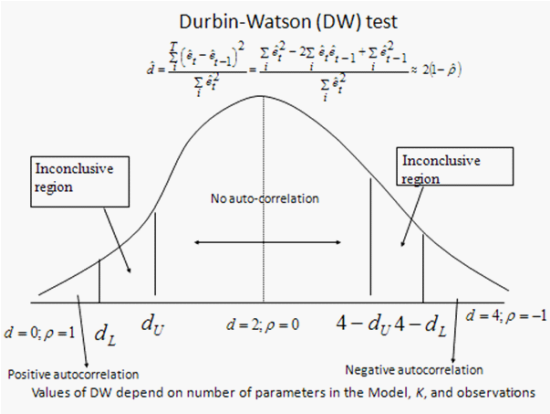
Autocorrelation and Durbin-Watson Statistics

$$d = 2\left(1 - \rho\right) \tag{404}$$

$$\rho = 0 \Longrightarrow d = 2 \tag{405}$$

$$\rho = -1 \Longrightarrow d = 4 \tag{406}$$

Durbin-Watson Distribution



Transformation of the model in the presence of autocorrelation when autocorrelation coefficient is known

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \quad t = 1 \ldots T \tag{407}$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t \tag{408}$$

$$Y_t - \rho Y_{t-1} = \left(\beta_1 - \rho\beta_1\right) + \beta_2\left(X_t - \rho X_{t-1}\right) + \varepsilon_t - \rho\varepsilon_{t-1} \tag{409}$$

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + \varepsilon_t^* \tag{410}$$

Apply OLS in this transformed model β_1^* and β_2 will have BLUE properties.

When autocorrelation coefficient is unknown, this method is similar to the above ones, except that it involves multiple iteration for estimating ρ . Steps are as following:

1. Get estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ from the original model; get error terms $\widehat{e}_i$ and estimate $\widehat{\rho}$

2. Transform the original model multiplying it by $\widehat{\rho}$ and by taking the first difference,

3. Estimate $\widehat{\widehat{\beta}}_1$ and $\widehat{\widehat{\beta}}_2$ from the transformed model and get errors $\widehat{\widehat{e}}_i$ of this transformed model

4. Then again estimate $\widehat{\widehat{\rho}}$ and use those values to transform the original model as

$$Y_t - \widehat{\rho}Y_{t-1} = (\beta_1 - \widehat{\rho}\beta_1) + \beta_2 (X_t - \widehat{\rho}X_{t-1}) + \varepsilon_t - \widehat{\rho}\varepsilon_{t-1} \tag{411}$$

5. Continue this iteration process until $\widehat{\widehat{\rho}}$ converges.

PcGive suggests using differences in variables. Diagnos /ACF options in OLS in Shazam will generate these iterations.

Use Durbin h-statistic when the lagged value of the dependent variable is an explanatory variable. This statistic is derived from the WD statistic as

$$h = \left(1 - \frac{d}{2}\right) \times \sqrt{\frac{n}{1 - n\sigma_y^2}} \tag{412}$$

9.0.6 Breusch-Godfrey LM-test of Serial Correlation

Breusch-Godfrey LM test of serial correlation is another popular test of autocorrelation.

$$LM = (n - p) R^2 \sim \chi^2_{df} \tag{413}$$

The hypothesis is set up for this as follows:

$$Y_t = \beta_1 + \beta_2 X_t + e_t \quad t = 1 \ldots T \tag{414}$$

$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \ldots + \rho_p e_{t-p} + \varepsilon_t \quad p = 1 \ldots p \tag{415}$$

Null hypothesis is that there no autocorrelation:

$$H_0 : \rho_1 = \rho_2 = + \ldots + = \rho_p = 0 \tag{416}$$

Alternative hypothesis is that there is at leat one of ρ_1 is significant. Then compare the statistics with the critical value. If the calculated LM is greater than critical value χ^2_{df} there null of no autocorrelation is rejected. There is an evidence for autocorrelation.

9.1 GLS to solve autocorrelation

In matrix notation

$$\beta_{OLS} = (X'X)^{-1}(X'Y) \tag{417}$$

$$\beta_{GLS} = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right) \tag{418}$$

Ω^{-1} is inverse of variance covariance matrix.
Generalised Least Square
Take a regression

$$Y = X\beta + e \tag{419}$$

Assumption of homoskedasticity and no autocorrelation are violated

$$var(\varepsilon_i) \neq \sigma^2 \quad for \ \forall \ i \tag{420}$$

$$covar(\varepsilon_i\varepsilon_j) \neq 0 \tag{421}$$

The variance covariance of error is given by

$$\Omega = E(ee') = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & .. & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & .. & \sigma_{2n} \\ : & : & : & : \\ \sigma_{n1} & \sigma_{n2} & .. & \sigma_n^2 \end{bmatrix} \tag{422}$$

$$Q'\Omega Q = \Lambda \tag{423}$$

Generalised Least Square

$$\Omega = Q\Lambda Q' = Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q' \tag{424}$$

$$P = Q\Lambda^{\frac{1}{2}} \tag{425}$$

$$P'\Omega P = I \quad ; \quad P'P = \Omega^{-1} \tag{426}$$

Transform the model

$$PY = \beta PX + Pe \tag{427}$$

$$Y^* = \beta X^* + e^* \tag{428}$$

$Y^* = PY \quad X^* = PX \quad and \quad e^* = Pe \qquad \beta_{GLS} = (X'P'PX)^{-1}(X'P'PY)$

$$\beta_{GLS} = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right) \tag{429}$$

10 Time Series

Time series models aim to explain the data generating process for $\{y_t\}_{-\infty}^{\infty} = \{y_{-\infty}.....y_{-1}.y_0.y_1.y_2.....y_T.y_{T+1}.y_{T+1}.....\}$

A Time series consists of trend, cycle, season and irregular component

$$Y = T \times C \times S \times I \tag{430}$$

In a simple method the moving average gives $T \times C$ components and is used to isolate the $S \times I$ components. For instance for a 12 monthly moving average

$$\overline{Y}_i = \frac{1}{12}\left(Y_1 + Y_2 + ... + Y_{12}\right) \tag{431}$$

$$S \times I = \frac{T \times C \times S \times I}{T \times C} = \frac{Y_i}{\overline{Y}_i} = z_t \tag{432}$$

Now to isolate the Irregular component I from $S \times I$ take out the seasonal elements from z_t assuming monthly data for 5 years (60 observations) compute the seasonal indices as following:

$$Month1 : \overline{z}_1 = \frac{1}{5}\left(z_1 + z_{13} + z_{25} + z_{39} + z_{48}\right) \tag{433}$$

$$Month2 : \overline{z}_2 = \frac{1}{5}\left(z_2 + z_{14} + z_{26} + z_{40} + z_{49}\right) \tag{434}$$

$$Month3 : \overline{z}_3 = \frac{1}{5}\left(z_3 + z_{15} + z_{26} + z_{41} + z_{50}\right) \tag{435}$$

.....

$$Month11 : \overline{z}_{11} = \frac{1}{5}\left(z_{11} + z_{23} + z_{35} + z_{47} + z_{59}\right) \tag{436}$$

$$Month12 : \overline{z}_{12} = \frac{1}{5}\left(z_{12} + z_{24} + z_{36} + z_{46} + z_{60}\right) \tag{437}$$

Deseasonalisation of data $Y_i^d = \frac{Y_i}{\overline{z}_i}$ and irregular component should be $i = \frac{z_t}{\overline{z}_i}$.
Trends:
Simple extrapolation

$$Y_t = c_1 + c_2 t \tag{438}$$

Exponential growth

$$Y_t = Ae^{rt} \tag{439}$$

Autoregressive model

$$Y_t = c_1 + c_2 Y_{t-1} \tag{440}$$

Log trend

$$\ln(Y_t) = c_1 + c_2 \ln(Y_{t-1}) \tag{441}$$

Quadratic trends:

$$Y_t = c_1 + c_2 t + c_3 t^2 \tag{442}$$

Logistic trend:

$$Y_t = \frac{1}{k + bt} \quad b > 1 \tag{443}$$

$$Y_t = e^{k_1 - \frac{k_2}{t}} \tag{444}$$

$$ln(Y_t) = k_1 - \frac{k_2}{t} \tag{445}$$

auto lagged with declining weights $\alpha < 1$

$$Y_t = \alpha Y_{t-1} + \alpha(1-\alpha) Y_{t-2} + \alpha(1-\alpha)^2 Y_{t-2} + ... + \alpha(1-\alpha)^n Y_{t-2} \tag{446}$$

Forecasting forward with these models is obvious.

10.1 Time Series Process

Simplest of these is a trend model

$$Y_t = \beta t + \varepsilon_t \tag{447}$$

with mean $E(Y_t) = \beta t$ and variance $E(Y_t - \beta t)^2 = E(\varepsilon_t)^2 = \sigma_\varepsilon^2$

Or it could have been just a constant plus a Gaussian white noise $\varepsilon_t \sim N(0, \sigma^2)$ as:

$$Y_t = \mu + \varepsilon_t \tag{448}$$

with mean $E(Y_t) = \mu$ and variance $E(Y_t - \mu)^2 = E(\varepsilon_t)^2 = \sigma_\varepsilon^2$

Autocovariance of $\{y_t\}_{-\infty}^{\infty}$ for I realisations is

$$\gamma_{tj} = E(Y_t - \mu) E(Y_{t-j} - \mu) = E(\varepsilon_t) E(\varepsilon_{t-j}) = 0 \quad for\ j \neq 0 \tag{449}$$

Stationarity

when neither mean μ nor the autocovariance γ_{ij} depend on time t then the Y_t is covariance stationary or weakly stationary.

$$E(Y_t) = \mu \ for \ \forall \ t \tag{450}$$

$$E\left(Y_t - \mu\right) E\left(Y_{t-j} - \mu\right) = \gamma_j \quad for \text{ any } t \ \ and \ j = \{^{\sigma_\varepsilon^2 \ \ for \ j=0}_{0 \ for \ j \neq 0} \tag{451}$$

For instance 448 is stationary while 447 not covariance stationary because its mean βt is function of time.

If the process is stationary γ_j is the same for any value of t $\gamma_j = \gamma_{-j}$

$$\gamma_j = E\left(Y_{t+j} - \mu\right) E\left(Y_{(t+j)-j} - \mu\right) = E\left(Y_{t+j} - \mu\right) E\left(Y_t - \mu\right) = E\left(Y_t - \mu\right) E\left(Y_{t+j} - \mu\right) \ = \gamma_{-j} \tag{452}$$

10.2 Stationarity

What is a stationary variable?

When its mean and variance are constant.

$$E\left(Y_t\right) = \mu \tag{453}$$

$$var\left(Y_t\right) = \sigma^2 \tag{454}$$

When mean and variances are not constant, that variable is non-starionary, for instance a random walk

$$Y_t = Y_{t-1} + \varepsilon_i \quad t = 1 \ ...T \tag{455}$$

In an autoregressive model

$$Y_t = \rho Y_{t-1} + \varepsilon_i \quad t = 1 \ ...T \tag{456}$$

if the autocorrlation coefficient $\rho = 1$ then it becomes a random walk. This variable is non-stationary.

$$Y_t = \sum_{s=1}^{\infty} \rho^s \varepsilon_{t-s} \tag{457}$$

Current realisations are accumulation of past errors.

Prove that variance of this is given by:

$$var\left(Y_t\right) = t.\sigma^2 \tag{458}$$

Regression among non-stationary variables becomes spurious unless they are cointegrated.

10.2.1 Unit root and order of integration

A Non-Stationary variable can be made stationary by taking first difference as:

$$\Delta Y_t = Y_t - Y_{t-1} \tag{459}$$

If a variable becomes stationary by taking the first difference it is said to be intergrated of order one

$$I(1) \tag{460}$$

If it becomes stationary after differencing d time then it is called $I(d)$ variable.

Dickey-Fuller and Phillip-Perron unit root tests are used to determine stationarity of a variable.

$$Y_t = \rho Y_{t-1} + \varepsilon_i \tag{461}$$

10.2.2 Level, drift, trend and lag terms in unit root test

Dickey-Fuller and Phillip-Perron unit root tests are used to determine stationarity of a variable.

$$Y_t = \rho Y_{t-1} + \varepsilon_i \tag{462}$$

$$\Delta Y_t = (\rho - 1) Y_{t-1} + \varepsilon_i; \qquad \Delta Y_t = \gamma Y_{t-1} + \varepsilon_i; \tag{463}$$

Random walk with drift

$$\Delta Y_t = \alpha_0 + \gamma Y_{t-1} + \varepsilon_i \tag{464}$$

trend stationary

$$\Delta Y_t = \alpha_0 + \alpha_1 t + \gamma Y_{t-1} + \varepsilon_i \tag{465}$$

Augmented Dickey-Fuller test

$$\Delta Y_t = \alpha_0 + \alpha_1 t + \gamma Y_{t-1} + \sum_{i=1}^{m} \rho^s \Delta Y_{t-i} + \varepsilon_i \tag{466}$$

Cointegration in a regression

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \tag{467}$$

First do the regression and then estimate the error as

$$\widehat{\varepsilon}_t = Y_t - \widehat{\beta}_1 - \widehat{\beta}_2 X_t \tag{468}$$

Y_t and X_t are cointegrated if the estimated error is stationary $\widehat{\varepsilon}_t \sim I(0)$

$$\widehat{\varepsilon}_t = \rho\widehat{\varepsilon}_{t-1} + \varepsilon_t \tag{469}$$

if $\rho < 1$ the error $\widehat{\varepsilon}_t$ is stationary and Y_t and X_t are cointegrated. They have a long run relationship.

When variables are cointegrated there is an error correction mechanism.

$$Y_t = \varphi_2 X_t + \epsilon_t \tag{470}$$

$$Y_t = X_t + \epsilon_t \ ; \qquad \varphi_2 = 1 \tag{471}$$

Cointegration: Engle-Granger Representation Theorem

$$\epsilon_t \ = Y_t - X_t \tag{472}$$

For test of cointegration

$$\Delta\epsilon_t \ = \gamma\epsilon_{t-1} \ + u_t \tag{473}$$

$$\Delta\left(Y_t - X_t\right) = \gamma\left(Y_{t-1} - X_{t-1}\right) + u_t \tag{474}$$

$$\Delta Y_t = \Delta X_t + \gamma\left(Y_{t-1} - X_{t-1}\right) + u_t \tag{475}$$

This is an error correction model.Term $\gamma\left(Y_{t-1} - X_{t-1}\right)$ gives the adjustment towards the long run equilibrium and ΔX_t denotes the short run impact.

H_0 : No cointegration; t- statistics can be used instead of DF test in error correction model.

Granger Causality Test Estimate the following model where M_t is money Y_t is GDP and test the causality as below:

$$Y_t = \sum_{i=1}^{n}\alpha_i M_{t-i} + \sum_{j=1}^{m}\beta_j Y_{t-j} + u_{1,t} \tag{476}$$

$$M_t = \sum_{i=1}^{n}\lambda_i M_{t-i} + \sum_{j=1}^{m}\delta_j Y_{t-j} + u_{2,t} \tag{477}$$

Unidirection causality from M_t to Y_t requires $\sum_{i=1}^{n}\alpha_i \neq 0$ and $\sum_{j=1}^{m}\delta_j = 0$

Unidirection causality from Y_t to M_t requires $\sum_{j=1}^{m}\delta_j \neq 0$ and $\sum_{i=1}^{n}\alpha_i = 0$

Bilateral causality between Y_t to M_t requires $\sum_{i=1}^{n}\alpha_i \neq 0$ and $\sum_{j=1}^{m}\delta_j \neq 0$

Independence of Y_t to M_t from each other $\sum_{i=1}^{n}\alpha_i = 0$ and $\sum_{j=1}^{m}\delta_j = 0$

10.2.3 Exercise 8

Stationarity, Unit Root and Cointegarion

1. Study the monthly data on unemployment rate and inflation since 1972:1 to 2004:8 as given in "unmnth.xls" file. Use GiveWin PcGive to

- Draw diagrams to represent the rates of unemployment among males and females and the RPI over this period.
- Ascertain whether unit root exists in the overall unemployment rate, URT and RPI at 5% and 1% level of significance in level, in log and in the first difference of these series.
- Detrend the data with Hodrik-Prescott filter and conduct stochastic volatility tests.

2. Regress unemployment rate on inflation rate in levels and in the first differences. Test whether these series are cointegrated using the Engle-Granger procedure. (hint: stationarity of residuals).

3. The time series and represent the underlying data generating processes (DGP) of consumption $\{C_t\}$ and income $\{Y_t\}$. Answer the following questions regarding the properties these series.

   (a) What is meant by saying that$\{C_t\}$ and $\{Y_t\}$ are stationary series? Why is it important that the series are stationary for a robust regression analysis?

   (b) How do you determine whether $\{C_t\}$ and $\{Y_t\}$ are stationary series, or not?

   (c) Analyse the properties of these series when they follow a random walk, or have a unit root.

   (d) What is the meaning of the order of integration in this respect? Discuss any three different methods of checking for stationarity.

   (e) What is the meaning of cointegration between the series and ? How would you decide whether these series$\{C_t\}$ and $\{Y_t\}$ are co-integrated, or not?

   (f) If the original series $\{C_t\}$ and $\{Y_t\}$ are not co-integrated, what transformation can be applied to achieve co-integration? How do you decide the order of co-integration?

   (g) Use time series of consumption and income contained in Quarterly_cons.xls. Determine the order of integration for both consumption and income. Is there an evidence of cointegration between consumption and income in levels or in the first differences?

11 Linear probability, probit and logit models

- Alternative names: dichotomous dependent variables, discrete dependent random variable, binary variable, either or choice variables

$$Y_i = \int \begin{array}{l} Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \quad \text{if the event occurs} \\ 0 = \text{otherwise} \end{array} \tag{478}$$

Examples

- the labour force participation (1 if a person participates in the labour force, 0 otherwise)
- yes or no vote in particular issue ; to marry or not to marry; to study further or to start a job
- to buy or not to buy a particular stock
- choice of transportation mode to work (1 if a person drives to work, 0 otherwise)
- Union membership (1 if one is a member of the union, 0 otherwise)
- Owning a house (1 if one owns 0 otherwise)
- Multinomial choices: work as a teacher, or as a clerk, or as a self employed or professional or as a factory worker
- Multinomial ordered choices: strongly agree, agree, neutral, disagree

Linear Probability Model

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \tag{479}$$

where $Y_i = 1$ if person owns a house, 0 otherwise; X_i is family income.
$E\left[\left(Y_i = 1\right)/X_i\right]$ probability that the event y will occur given x

$$E\left[\left(Y_i = 1\right)/X_i\right] = 0 \times \left[1 - P_i\right] + 1 \times P_i = P_i \tag{480}$$

$$0 \leqslant E\left[\left(Y_i = 1\right)/X_i\right] = P_i = \beta_1 + \beta_2 X_i \leqslant 1 \tag{481}$$

- Problem: Errors are heteroscedastic

$$\varepsilon_i = 1 - \beta_1 - \beta_2 X_i \quad with \quad (1 - P_i) \tag{482}$$

$$\varepsilon_i = -\beta_1 - \beta_2 X_i \quad with \quad P_i \tag{483}$$

Variance of error in a linear probability model

$$var\left(\varepsilon_i\right) = \left(1-\beta_1-\beta_2 X_i\right)^2 \ \left(1-P_i\right) + \left(-\beta_1-\beta_2 X_i\ \right)^2 P_i \tag{484}$$

$$\sigma^2 = \left(1-\beta_1-\beta_2 X_i\right)^2 \ \left(-\beta_1-\beta_2 X_i\right) + \left(-\beta_1-\beta_2 X_i\ \right)^2 \left(1-\beta_1-\beta_2 X_i\right) \tag{485}$$

$$\sigma^2 = \left(1-\beta_1-\beta_2 X_i\right)\left(\beta_1+\beta_2 X_i\right) \ = \left(1-P_i\right) \ P_i \tag{486}$$

Variance depends on X.

Limitations of a linear probability model

It is possible to transform this model to make it homescedastic by dividing the original variables by

$$\sqrt{\left(1-\beta_1-\beta_2 X_i\right)\left(\beta_1+\beta_2 X_i\right)} \ = \sqrt{\left(1-P_i\right) \ P_i} = \sqrt{W_i} \tag{487}$$

$$\frac{Y_i}{\sqrt{W_i}} = \frac{\beta_1}{\sqrt{W_i}} + \beta_2 \frac{X_i}{\sqrt{W_i}} + \frac{\varepsilon_i}{\sqrt{W_i}} \tag{488}$$

- It does not guarantee that the probability lies inside (0,1) bands
- Probability in non-linear phenomenon: at very low level of income a family does not own a house; at very high level of income every one owns a house ; marginal effect of income is very negligible. The linear probability model does not explain this fact well.

Probit Model

-

$$\begin{aligned} \Pr\left(Y_i=1\right) &= \Pr\left(Z_i^* \leq Z_i\right) = F\left(Z_i\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-\frac{t^e}{2}} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_1+\beta_2 X_i+\varepsilon_i} e^{-\frac{t^e}{2}} dt \end{aligned} \tag{489}$$

- Here t is standardised normal variable, $t \tilde{} N\left(0,1\right)$

  probability depends upon unobserved utility index Z_i which depends upon observable variables such as income. There is a thresh-hold of this index when after which family starts owning a house, $Z_i \geqslant Z_i^*$

Logit Model

- variable Y_i which takes value 1 $\left(Y_i=1\right)$ if a student gets a first class mark, value 0 $\left(Y_i=0\right)$otherwise.

- Probability of getting a first class mark in an exam is a function of student effort index denoted by $.Z_i$; where $P_i = \frac{1}{1+e^{-Z_i}}$

  $Z_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ An example of a logit model: what determines that a student gets the first class degree?

$$Z_i = \beta_1 + \beta_2 H_i + \beta_3 E_i + \beta_4 A_i + \beta_2)_i + \varepsilon_i \tag{490}$$

  H = hours of study, E= exercises, A = attendance in lectures and classes; P = papers written for assignment.

- Ratio of odds: $\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i}$;taking log of the odds $ln\left(\frac{P_i}{1-P_i}\right) = Z_i$

Features of a logit Model

- – probability goes from 0 to 1 as the index variable goes from -∞ to $+\infty$. Probability lies between 0 and 1.
  - Log of the odds is linear in x, characteristic variables but probabilities themselves are not linear but non linear function of the parameters. Probabilities are estimated using the maximum likelihood method.
  - Any explanatory variable that determines the value of Z_i , measures how the log of odds of an event (i.e. owning a house) changes as a result of change in explanatory variable such as income.
  - We can calculate P_i for given estimates of $\beta_1$and $\beta_2$or all other $.\beta_i$
  - Limiting case when P_i =1; $ln\left(\frac{P_i}{1-1}\right)$ or when P_i =0 ; $ln\left(\frac{0}{1-0}\right)$OLS cannot be applied in such case but the maximum likelihood method may be used to estimate the parameters.

Logit model on probability of getting married from the dataset constructed from the BHPS (Hours.csv)

Tobit Model

- – It is an extension of the probit model, named after Tobin. We observe variables if the event occurs: ie if some one buys a house. We do not observe explanatory variables for people who have not bought a house. The observed sample is censored, contains observations for only those who buy the house.

$$Y_i = \int \begin{array}{l} Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \quad \text{if the event occurs} \\ 0 = \text{otherwise} \end{array} \tag{491}$$

  - is equal to is the event is observed equal to zero if the event is not observed.

Table 12: Probability of Getting Married

	Coefficient	t-value	t-prob
Intercept	-2.99	-8.44	0.000
Log workhours	0.277	2.13	0.034
Gender	0.269	4.33	0.000
Labour	0.187	2.74	0.006
Liberal	0.330	3.28	0.001
Conservative	0.381	4.60	0.000
Health	0.189	6.56	0.000
Money	-0.036	-2.49	0.013
Children	0.253	23.5	0.000
Job	-0.124	-7.43	0.000
State =2 , AIC = 7244.8 $N = 5790$; LL -3612.4			

- It is unscientific to estimate probability only with observed sample without worrying about the remaining observations in the truncated distribution. The Tobit model tries to correct this bias.
- Inverse Mill's ratio: Example first estimate probability of work then estimate the hourly wage as a function of socio-economic background variables

Summary of Probability Models
The effect of observed variables on probability

- –

$$\frac{\partial P_i}{\partial x_{i,j}} = \begin{cases} \beta_j \\ \beta_j P_j \left(1 - P_j\right) \\ \beta_j \phi \left(Z_i\right) \end{cases} \tag{492}$$

  - where $Z_i = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,j}$ and ϕ is the standard normal density function.

Estimate probability models using data in Hours.csv.

11.0.4 AR, MA, ARMA and ARIMA Forecasting

AR(1) forecast

$$y_t = \delta + \theta_1 y_{t-1} + e_t \tag{493}$$

h =1 ahead Forecast

$$y_{T+1} = \delta + \theta_1 y_T + e_{T+1} \qquad e_{T+1} \tilde{} \; N\left(0, 1\right) \tag{494}$$

Mean forecast:

$$\widehat{y}_{T+1} = E\left(y_{T+1}\right) = \delta + \theta_1 y_T \tag{495}$$

Estimate of Forecast error

$$\widehat{e}_{T+1} = y_{T+1} - \widehat{y}_{T+1} = \delta + \theta_1 y_T + e_{T+1} - \delta - \theta_1 \widehat{y}_T \tag{496}$$

Variance of h =1 Forecast error

$$var\left(\widehat{e}_{T+1}\right) = \sigma_e^2 \tag{497}$$

h =2 ahead Forecast

$$y_{T+2} = \delta + \theta_1 y_{T+1} + e_{T+1} \qquad e_{T+2} \tilde{} \; N(0,1) \tag{498}$$

Mean forecast:

$$\widehat{y}_{T+2} = E\left(y_{T+2}\right) = \delta + \theta_1 y_{T+1} \tag{499}$$

Estimate of Forecast error

$$\begin{aligned} \widehat{e}_{T+2} &= y_{T+2} - \widehat{y}_{T+2} = \delta + \theta_1 y_{T+1} + e_{T+2} - \delta - \theta_1 \widehat{y}_{T+1} \\ &= e_{T+2} + \theta_1 \left(y_{T+1} - \widehat{y}_{T+1}\right) = e_{T+2} + \theta_1 e_{T+1} \end{aligned} \tag{500}$$

Variance of Forecast error

$$var\left(\widehat{e}_{T+2}\right) = \sigma_e^2 \left(1 + \theta_1^2\right) \tag{501}$$

h period ahead Forecast

$$y_{T+h} = \delta + \theta_1 y_{T+h-1} + e_{T+h} \qquad e_{T+h} \tilde{} \; N(0,1) \tag{502}$$

Mean forecast:

$$\widehat{y}_{T+h} = E\left(y_{T+h}\right) = \delta + \theta_1 \widehat{y}_{T+h-1} \tag{503}$$

Estimate of Forecast error

$$\begin{aligned} \widehat{e}_{T+h} &= y_{T+h} - \widehat{y}_{T+2} = \delta + \theta_1 y_{T+h-1} + e_{T+h} - \delta - \theta_1 \widehat{y}_{T+h-1} \\ &= e_{T+h} + \theta_1 \left(y_{T+h-1} - \widehat{y}_{T+h-1}\right) = e_{T+h} + \theta_1 e_{T+h-1} \end{aligned} \tag{504}$$

Variance of Forecast error

$$var\left(\widehat{e}_{T+h}\right) = \sigma_e^2 \left(1 + \theta_1^2 + \theta_1^2 + ... + \theta_1^{2(h-1)}\right) \tag{505}$$

MA(1) forecast Forecast with MA(1)

$$y_t = \mu + e_t + \alpha_1 e_{t-1} \tag{506}$$

h=1 period ahead forecast

$$y_{T+1} = \mu + e_{T+1} + \alpha_1 e_T \tag{507}$$

Mean forecast

$$E\left(y_{T+1}\right) = \widehat{y}_{T+1} = \mu + \alpha_1 e_T \tag{508}$$

Forecast error

$$y_{T+1} - \widehat{y}_{T+1} = \mu + e_{T+1} + \alpha_1 e_T - \mu - \alpha_1 e_{T+1} = e_{T+1} \tag{509}$$

Variance of forecast:

$$var\left(y_{T+1} - \widehat{y}_{T+1}\right) = var\left(e_{T+1}\right) = \sigma_e^2 \tag{510}$$

h=2 period ahead Forecast

$$y_{T+2} = \mu + e_{T+2} + \alpha_1 e_{T+1} \tag{511}$$

Mean forecast

$$E\left(y_{T+2}\right) = \widehat{y}_{T+2} = \mu \tag{512}$$

Forecast error

$$y_{T+2} - \widehat{y}_{T+2} = \mu + e_{T+2} + \alpha_1 e_{T+1} - \mu = e_{T+2} + \alpha_1 e_{T+1} \tag{513}$$

$$var\left(y_{T+2} - \widehat{y}_{T+2}\right) = var\left(e_{T+2}\right) = var\left(e_{T+2} + \alpha_1 e_{T+1}\right) = \sigma_e^2\left(1 + \alpha_1^2\right) \tag{514}$$

Similarly mean and variance of h period ahead forecast:

$$y_{T+h} = \mu + e_{T+h} + \alpha_1 e_{T+h-1} \tag{515}$$

$$E\left(y_{T+h}\right) = \widehat{y}_{T+h} = \mu \tag{516}$$

Forecast error

$$y_{T+h} - \widehat{y}_{T+h} = \mu + e_{T+h} + \alpha_1 e_{T+h-1} - \mu = e_{T+h} + \alpha_1 e_{T+h-1} \tag{517}$$

$$var\left(y_{T+2} - \widehat{y}_{T+2}\right) = var\left(e_{T+h}\right) = var\left(e_{T+h} + \alpha_1 e_{T+h-1}\right) = \sigma_e^2\left(1 + \alpha_1^2\right) \tag{518}$$

ARMA(1,1) forecast Forecasts using ARMA(1,1) process:

$$y_t = \delta + \theta_1 y_{t-1} + e_t + \alpha_1 e_{t-1} \tag{519}$$

h=1 period ahead Forecast

$$y_{T+1} = \delta + \theta_1 y_{t-1} + e_{T+1} + \alpha_1 e_T \tag{520}$$

Mean forecast

$$E\left(y_{T+1}\right) = \widehat{y}_{T+1} = \delta + \theta_1 y_{t-1} + \alpha_1 e_T \tag{521}$$

Forecast error

$$\begin{aligned} \widehat{e}_{T+1} &= \left(y_{T+h} - \widehat{y}_{T+h}\right) = \\ \delta + \theta_1 y_{t-1} + e_{T+1} + e_{T+1} + \alpha_1 e_T - \delta - \theta_1 y_{t-1} - \alpha_1 e_T &= e_{T+1} \end{aligned} \tag{522}$$

Forecast error

$$\begin{aligned} \widehat{e}_{T+1} &= \left(y_{T+h} - \widehat{y}_{T+h}\right) = \delta + \theta_1 y_{t-1} + e_{T+1} + e_{T+1} \\ +\alpha_1 e_T - \delta - \theta_1 y_{t-1} - \alpha_1 e_T &= e_{T+1} \end{aligned} \tag{523}$$

Variance of Forecast error

$$var\left(\widehat{e}_{T+1}\right) = var\left(y_{T+h} - \widehat{y}_{T+h}\right) = var\left(e_{T+1}\right) = \sigma_e^2 \tag{524}$$

$$y_t = \delta + \theta_1 y_{t-1} + e_t + \alpha_1 e_{t-1} \tag{525}$$

h=2 period ahead Forecast

$$y_{T+2} = \delta + \theta_1 y_{t+1} + e_{T+2} + \alpha_1 e_{T+1} \tag{526}$$

Mean forecast and Forecast error

$$E\left(y_{T+2}\right) = \widehat{y}_{T+2} = \delta + \theta_1 y_{t+1} \tag{527}$$

$$\begin{aligned} \widehat{e}_{T+2} &= \left(y_{T+2} - \widehat{y}_{T+2}\right) = \delta + \theta_1 y_{t+1} + e_{T+2} + \alpha_1 e_{T+1} - \delta - \theta_1 \widehat{y}_{T+1} \\ &= \theta_1\left(y_{t+1} - \widehat{y}_{T+1}\right) + e_{T+2} + \alpha_1 e_{T+1} = \left(\theta_1 + \alpha_1\right) e_{T+1} + e_{T+2} \end{aligned} \tag{528}$$

Variance of Forecast error

$$var\left(\widehat{e}_{T+1}\right) = var\left[\left(\theta_1 + \alpha_1\right) e_{T+1} + e_{T+2}\right] = var\left(e_{T+1}\right) = \sigma_e^2\left[\left(\theta_1 + \alpha_1\right)^2 + 1\right] \tag{529}$$

h=3 period ahead Forecast

$$y_{T+2} = \delta + \theta_1 y_{t+2} + e_{T+3} + \alpha_1 e_{T+2} \tag{530}$$

Mean forecast

$$E\left(y_{T+3}\right) = \widehat{y}_{T+3} = \delta + \theta_1 \widehat{y}_{t+2} \tag{531}$$

Forecast error and Variance of Forecast error

$$\begin{aligned} \widehat{e}_{T+3} &= \left(y_{T+3} - \widehat{y}_{T+3}\right) = \delta + \theta_1 y_{t+2} + e_{T+3} + \alpha_1 e_{T+2} - \delta - \theta_1 \widehat{y}_{T+2} \\ &= \theta_1 \left(y_{t+2} - \widehat{y}_{T+2}\right) + e_{T+3} + \alpha_1 e_{T+2} \\ &= e_{T+3} + \alpha_1 e_{T+2} + \left(\theta_1 + \alpha_1\right) e_{T+2} + e_{T+2} \end{aligned} \tag{532}$$

$$\begin{aligned} var\left(\widehat{e}_{T+3}\right) &= var\left[e_{T+3} + \alpha_1 e_{T+2} + \left(\theta_1 + \alpha_1\right) e_{T+2} + e_{T+2}\right] \\ &= \sigma_e^2 \left[1 + \left(1 + \alpha_1\right)^2 + \left(\theta_1 + \alpha_1\right)^2\right] \end{aligned} \tag{533}$$

see: http://www.hull.ac.uk/php/ecskrb/Stochastic_GE_IJTGM%204(2)%20Paper%207.pdf

12 Panel Data Model

Panel Data

for $i = 1,\ldots N$ countries and $t = 1,\ldots,T$ years

Table 13: Structure of Panel Data

Dependent Variable	Explanatory Variable	Random Error
$y_{1,1}$	$x_{1,1}$	$e_{1,1}$
.	.	.
$y_{1,T}$	$x_{1,T}$	$e_{1,T}$
$y_{2,1}$	$x_{2,1}$	$e_{2,1}$
.	.	.
$y_{2,T}$	$x_{2,T}$	$e_{2,T}$
.	.	.
$y_{N,1}$	$x_{N,1}$	$e_{,1}$
.	.	.
$y_{2,T}$	$x_{2,T}$	$e_{2,T}$

12.0.5 Panel Data Model: Fixed Effect Model

$$y_{i,t} = \alpha_i + x_{i,t}\beta + e_{i,t} \qquad e_{i,t} \sim IID\left(0, \sigma_e^2\right) \tag{534}$$

where parameter α_i picks up the fixed effects that differ among individuals,β is the vector of coefficients on explanatory variables. These parameters can be estimated by OLS when N is small but not when that is large but the model need to be transformed to the least square dummy variable method when N is too large.

$$\overline{y}_i = \alpha_i + \overline{x}_i\beta + e_i \qquad \overline{y}_i = T^{-1}\sum_i y_{i,t} \tag{535}$$

$$y_{i,t} - \overline{y}_i = \left(x_{i,t} - \overline{x}_i\right)\beta + \left(e_{i,t} - e_i\right) \tag{536}$$

fixed effect least square dummy variable estimator of β is

$$\beta_{FE} = \left(\sum_t^T\sum_i^N \left(x_{i,t} - \overline{x}_i\right)\left(x_{i,t} - \overline{x}_i\right)'\right)^{-1} \sum_t^T\sum_i^N \left(x_{i,t} - \overline{x}_i\right)\left(y_{i,t} - \overline{y}_i\right)' \tag{537}$$

$$\alpha_i = \overline{y}_i - \overline{x}_i\beta_{FE} \tag{538}$$

fixed effect least square dummy variable estimator of β is

$$\beta_{FE} = \left(\sum_t^T\sum_i^N \left(x_{i,t} - \overline{x}_i\right)\left(x_{i,t} - \overline{x}_i\right)'\right)^{-1} \sum_t^T\sum_i^N \left(x_{i,t} - \overline{x}_i\right)\left(y_{i,t} - \overline{y}_i\right)' \tag{539}$$

$$\alpha_i = \overline{y}_i - \overline{x}_i\beta_{FE} \tag{540}$$

These estimators are unbiased, consistent and efficient with corresponding covariance matrix given by:

$$cov\left(\beta_{FE}\right) = \sigma_e^2\left(\sum_t^T\sum_i^N \left(x_{i,t} - \overline{x}_i\right)\left(x_{i,t} - \overline{x}_i\right)'\right)^{-1} \tag{541}$$

$$\sigma_e^2 = \frac{1}{N\left(T-1\right)}\sum_t^T\sum_i^N \left(y_{i,t} - \alpha_i - x_{i,t}\beta_{FE}\right) \tag{542}$$

Static Panel Data Model of House Price in England :

Datafile: HousePrice_regional.csv: (Coefficients of regional dummies are not as expected; let us look at dynamic panel then)

Exercise: Do this exercise for the UK including London, Wales, Scotland and Northern Ireland.

Table 14: Static Panel Data Model of House Price in England

	Coefficient	t-value	t_Prob
Real income	4.64	46.0	0.000
Population	1.25	0.692	0.490
Mortgage rate	-11.51	-0.050	0.960
Mortgate House Price ratio	-237240	-10.0	0.000
Current deposit	1.94	2.52	0.000
Saving deposit	1.10	2.32	0.012
Constant	124143	6.30	0.021
North East	base region		
North West	-14780.4	-1.68	0.094
York_Humber	-8229.6	-1.57	0.117
Soth West	-12905.4	-2.76	0.006
England	-170937.9	-1.74	0.083
East Midland	-9977.2	-2.54	0.011
West Midland	-10441.1	-2.02	0.044
East Englia	-6245.6	-0.83	0.409
Galaway	-17390.6	-2.07	0.039
South East	-22845.3	-2.13	0.034
R2 = 0.97; N =9; T = 48; Chi2 =12250 [0.000] **			

12.0.6 Panel Data Model: Random Effect

Random effect models are more appropriate for analysing determinants of growth as

$$y_{i,t} = \mu + x_{i,t}\beta + \alpha_i + e_{i,t} \tag{543}$$

where $\alpha_i \tilde{} IID\left(0, \sigma_\alpha^2\right)$ are individual specific random errors and $e_{i,t} \tilde{} IID\left(0, \sigma_e^2\right)$ are remaining random errors.

$$\alpha_i \iota_T + e_i \qquad where\ \iota_T = (1, 1, ...1) \tag{544}$$

$$var\left(\alpha_i \iota_T + e_i\right) = \Omega = \sigma_\alpha^2 \iota_T \iota_T' + \sigma_e^2 I_T \tag{545}$$

Errors are correlated therefore this requires estimation by the Generalised Least Square estimator. Transform the model by pre-multiplying by Ω^{-1} where

$$\Omega^{-1} = \sigma_e^2 \left[I_T - \frac{\sigma_\alpha^2}{\sigma_e^2 + T\sigma_\alpha^2} \iota_T \iota_T' \right] \tag{546}$$

Table 15: Dynamic Panel Data Model of House Price in England)l

	Coefficient	t-value	t_Prob
House price (-1)	0.729	38.6	0.000
Real income	1.297	16.0	0.000
Population	-0.938	-1.56	0.120
Mortgage rate	110.387	1.03	0.305
Mortgate House Price ratio	-89015.9	-19.5	0.000
Current deposit	0.0429	0.34	0.734
Saving deposit	0.594	3.95	0.000
Constant	54443	17.0	0.000
North East	base region		
North West	1943.2	0.876	0.382
York_Humber	1301.2	1.03	0.305
Soth West	-1233.1	-1.65	0.100
England	19171.9	0.78	0.438
East Midland	-598.1	-0.85	0.398
West Midland	335.6	0.27	0.786
East Englia	4402.0	2.86	0.005
Galaway	2068.9	1.07	0.286
South East	715.5	0.316	0.752
R2 = 0.99; N =10; T = 47; Chi2 =190200 [0.000] **			

$$\beta_{GLS} = \left(\sum_{t}^{T}\sum_{i}^{N}\left(x_{i,t}-\overline{x}_i\right)\left(x_{i,t}-\overline{x}_i\right)' + \psi T\sum_{i}^{N}\left(x_{i,t}-\overline{x}_i\right)\left(x_{i,t}-\overline{x}_i\right)'\right)^{-1} \left(\sum_{t}^{T}\sum_{i}^{N}\left(x_{i,t}-\overline{x}_i\right)\left(y_{i,t}-\overline{y}_i\right)' + \psi T\sum_{i}^{N}\left(x_{i,t}-\overline{x}_i\right)\left(y_{i,t}-\overline{y}_i\right)'\right) \quad (547)$$

$$\Omega = \begin{bmatrix} \sigma_\alpha^2+\sigma_e^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & . & . & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2+\sigma_e^2 & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & . & . & \sigma_\alpha^2+\sigma_e^2 \end{bmatrix} \quad (548)$$

$$\Omega^{-\frac{1}{2}} = \frac{1}{\sigma_e}\left[I_T - 1 - \frac{\sigma_e}{\sqrt{\sigma_e^2+T\sigma_\alpha^2}}\right] \quad (549)$$

$$\beta_{GLS} = \sum_i \left(X'\Omega^{-1}X\right)^{-1}\sum_i\left(X'\Omega^{-1}Y\right) \quad (550)$$

Panel Data Model: GMM Estimator

generalised method of moments (GMM) as proposed by Hansen (1982).

$$y_{i,t} = \gamma y_{i,t-1}\beta + \alpha_i + e_{i,t} \qquad \gamma < 1 \tag{551}$$

which generates the following estimator

$$\gamma_{FE} = \frac{\sum_t^T \sum_i^N \left(y_{i,t} - \overline{y}_i\right)\left(y_{i,t} - \overline{y}_{i,t-1}\right)}{\sum_t^T \sum_i^N \left(y_{i,t} - \overline{y}_{i,t-1}\right)^2}; \qquad \overline{y}_i = T^{-1}\sum_i y_{i,t}; and\ \overline{y}_{i,-1} = T^{-1}\sum_i y_{i,t-1} \tag{552}$$

This is not asymptotically unbiased estimator:

$$\gamma_{FE} = \gamma + \frac{\left(\frac{1}{NT}\right)\sum_t^T \sum_i^N \left(e_{i,t} - \overline{e}_i\right)\left(y_{i,t} - \overline{y}_{i,t-1}\right)}{\left(\frac{1}{NT}\right)\sum_t^T \sum_i^N \left(y_{i,t} - \overline{y}_{i,-1}\right)^2} \tag{553}$$

$$\underset{N\to\infty}{p\lim} \left(\frac{1}{NT}\right)\sum_t^T \sum_i^N \left(e_{i,t} - \overline{e}_i\right)\left(y_{i,t} - \overline{y}_{i,t-1}\right) = -\frac{\sigma_e^2}{T^2}\frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2} \neq 0 \tag{554}$$

Panel Data Model: Instrumental Variables for GMM

Instrumental variable methods have been suggested to solve this inconsistency

$$\widehat{\gamma}_{IV} = \frac{\sum_t^T \sum_i^N y_{i,t-2}\left(y_{i,t-1} - \overline{y}_{i,t-2}\right)}{\sum_t^T \sum_i^N y_{i,t-2}\left(y_{i,t-1} - y_{i,t-2}\right)} \tag{555}$$

where $y_{i,t-2}$ is used as instrument of $(y_{i,t-1} - y_{i,t-2})$
It is asymptotically

$$\underset{N\to\infty}{p\lim} \left(\frac{1}{NT}\right)\sum_t^T \sum_i^N \left(e_{i,t} - \overline{e}_i\right) y_{i,t-2} \tag{556}$$

13 VAR Analysis

Consider a vector autoregressive model of order 2, VAR(2) given below.

$$y_t = a_{10} + a_{11}y_{t-1} + a_{12}y_{t-2} + b_{11}x_{t-1} + b_{12}x_{t-2} + e_{1,t} \tag{557}$$

$$x_t = a_{20} + a_{21}y_{t-1} + a_{22}y_{t-2} + b_{21}x_{t-1} + b_{22}x_{t-2} + e_{2,t} \tag{558}$$

where and are two variables for time t range from 1 ... T periods. Errors of each equation, $e_1$and e_2, are identically and independently distributed with zero mean and constant variance and covariance between and is assumed zero.

a. Evaluate the relationship between and in the long run.

Answer: Long run relationship is obtained by imposing the steady state relations:

$$\overline{y} = a_{10} + a_{11}\overline{y} + a_{12}\overline{y} + b_{11}\overline{x} + b_{12}\overline{x} \tag{559}$$

$$\overline{y} = \frac{a_{10}}{1 - a_{11} - a_{12}} + \frac{(b_{11} + b_{12})}{1 - a_{11} - a_{12}}\overline{x} \tag{560}$$

$$\overline{x} = a_{20} + a_{21}\overline{y} + a_{22}\overline{y} + b_{21}\overline{x} + b_{22}\overline{x} \tag{561}$$

$$\overline{x} = \frac{a_{20}}{1 - b_{21} - b_{22}} + \frac{(a_{21} + a_{22})}{1 - b_{21} - b_{22}}\overline{y} \tag{562}$$

b. Provide impulse response analysis for and of a unit shock in $e_{1,t}$ and $e_{2,t}$.

Use lag operator $y_{t-1} = Ly_t; y_{t-2} = Ly_{t-1} = L^2y_t$;Then the system changes to

$$y_t = a_{10} + a_{11}Ly_t + a_{12}L^2y_t + b_{11}Lx_t + b_{12}L^2x_t + e_{1,t} \tag{563}$$

$$x_t = a_{20} + a_{21}Ly_t + a_{22}L^2y_t + b_{21}Lx_t + b_{22}L^2x_t + e_{2,t} \tag{564}$$

$$y_t = \frac{a_{10}}{1 - a_{11}L - a_{12}L^2} + \frac{(b_{11} + b_{12})}{1 - a_{11}L - a_{12}L^2}x_t + \frac{1}{1 - a_{11}L - a_{12}L^2}e_{1,t} \tag{565}$$

$$x_t = \frac{a_{10}}{1 - b_{11}L - b_{12}L^2} + \frac{(a_{11} + a_{12})}{1 - b_{11}L - b_{12}L^2}y_t + \frac{1}{1 - b_{11}L - b_{12}L^2}e_{2,t} \tag{566}$$

Terms $\frac{1}{1-a_{11}L-a_{12}L^2}e_{1,t}$ and $\frac{1}{1-b_{11}L-b_{12}L^2}e_{2,t}$ give the impulse response of the first and second equations respectively.

c. Indicate and explain criteria to determine the order of a VAR model like this:

It is wise to use from general to specific approach of David Hendry to determine the order of VAR . First start the model with a large number of lags and then keep reducing the number of lags until the significant relation is found. Likelihood ratio tests are suggested for this.

d. What extra information is needed to make a h period ahead forecast using the above model? VAR is a time series model. Given the past values of time series, it requires distribution of the error terms for h period ahead forecasts.

e. A diagram can show how the variance of the forecast error and the confidence interval of a forecast are sensitive to the number of periods in the forecast horizon. The confidence level of forecast increases with the larger horizon of the forecasts.

Relevant web pages

http://www.khanacademy.org/
http://www.econometricsociety.org/; http://www.aeaweb.org/aer/index.php; http://www.res.org.uk/economic/ejbrowse.asp
http://www.imf.org/external/pubs/ft/weo/2010/01/weodata/index.aspx;
http://www.ifs.org.uk/publications/789
http://www.esds.ac.uk/international/; http://www.bankofengland.co.uk/;
http://www.hm-treasury.gov.uk/
http://www.eea-esem.com/EEA/2010/Prog/ - look at fiscal policy sessions.

13.0.7 Texts in Econometrics

References

[1] Asteriou D. and Stephen Hall (2011) Applied Econometrics, Palgrave, macmillan, 2nd edition.

[2] Amemiya T. (1985) Advanced Econometrics, Harvard University Press.

[3] Baltagi H. (2008) Econometric analysis of panel data, Fourth Ediction, Blackwell.

[4] Bauwens L., M. Lubrano and J. F. Richard (1999) Bayesian Inference in Dynamic Econometric Models, Oxford.

[5] Bhattarai K. (2010) Handbook of Econometric Analysis, University of Hull Business School.

[6] Blundell R. W.K. Newey and T. Persson (2006) Advances in Economics and Econometrics, Econometric Society Monograph, Vol. II and III, Cambridge University Press.

[7] Burke S. P. and J. Hunter (2005) Modelling non-stationary economic time series, Palgrave.

[8] Campbell J. Y., A. W. Lo and A C MacKindlay (1997) The Econometrics of Financial Markets, Princeton.

[9] Carnot N, V. Koen and B. Tissot (2005) Economic Forecasting, Palgrave, macmillan, New York.

[10] Chipman J. S. (2011) Advanced Econometric Theory, Routledge.

[11] Davidson J. (2000) Econometric Theory, Basil Blackwell Publishers.

[12] Davidson R and MacKinnon J. G. (2004) Econometric Theory and Methods, Oxford.

[13] Doornik J A and D.F. Hendry ((2003) PC-Give Volume I-III, GiveWin Timberlake Consultants Limited, London.

[14] Dougherty C. (2002) Introduction to Econometrics (Second Edition), Oxford University Press, 2002

[15] Engle R.F and D. L. McFadden (1997) Handbook of Econometrics, V.4, North-Holland, Amsterdam

[16] Enders W. (1995) Applied Econometric Time Series, John Wiley.

[17] GAMS Users Manual, GAMS Development Corporation, Washington DC. www.gams.com.

[18] Gauss C.F. (1823) Theory of Combination of Observations Least Subject to Errors, SIAM (1995), Philadelphia.

[19] Greene W. (2003) Econometric Analysis, Fifth Edition, Prentice Hall.

[20] George G. J, W.E. Griffiths, R.C. Hill, H. Lutkepohl T.C. Lee (1985) Theory and Practice of Econometrics, JohnWiley.

[21] Gujarati, D. (2011), Econometrics by Example, Palgrave, macmillan.

[22] Gujarati, D. N. (2005), Basic Econometrics, 4rd edition, McGraw Hill.

[23] Griffiths W. E., R.C. Hill and G. G. Judge, (1993), Learning and Practising Econometrics, J Wiley and Sons.

[24] Griliches Z. and M. D. Intriligator eds. (1984) Handbook of Econometrics volms I,II,II, North-Holland, Amsterdam.

[25] Harvey A.C. (1990) The Econometric Analysis of Time Series, Phillip Allan, 2nd edition.

[26] Heckman J.J. and E. Leamer (2001) Handbook of EconometrTics V.5,6 North-Holland, London.

[27] Hendry D.F. (1995) Dynamic Econometric Theory, Oxford.

[28] Hamilton J. D. (1994) Time Series Analysis, Princeton.

[29] Harris R. and R. Sollis (2003) Applied Time Series Modelling and Forecasting, John Wiley and Sons.

[30] Hill, C.R., W.E.Griffiths and G.C. Lim (2008), Principles of Econometrics, 3nd edition, John Wiley and Sons Inc. New York.

[31] Hsiao C. (1993) Analysis of panel data, Cambridge.

[32] Johnston J. (1960) Econometric methods, 7th edition, McGraw Hill.

[33] Koop G. (2009) Analysis of Economic Data, Wiley, 2009.

[34] Kreps D. M. and K. F. Wallis (1997) Advances in Economics and Econometrics, Econometric Society Monograph, Vol. I Cambridge University Press.

[35] Kuthberson K. S. Hall and M. Taylor (1992) Applied time series analysis , Michigan University Press.

[36] Lancaster T. (1990) Econometric Analysis of Transition Data, Blackwell

[37] Lancaster T. (2004) An Introduction to Modern Bayesian Econometrics, Blackwell

[38] Maddala G.S. (2003) Introduction to Econometrics, Cambridge.

[39] Maddala G.S. (1983) Limited Dependent and Qualitative Variables in Econometrics, Cambridge.

[40] Markov A. A. (1900) Calculation of Probability, State Publications, St. Petersburgh.

[41] Patterson K (2000) An Introduction to Applied Econometrics: A time series approach, McMillan

[42] Pyndick R. S. and D. L. Rubinfeld (1998) Econometric Models and Economic Forecasts, 4th edition, McGraw Hill

[43] Ruud P. A. (2000) An Introduction to Classical Econometric Theory, Oxford.

[44] Stock J. H. and M.W. Watson (2007) Introduction to Econometrics, 2nd edition, Addison Wesley.

[45] Syddeater K.P. Hammond P. and A. Seierstad and A. Strom (2008) Further Mathematics for Economic Analysis, second edition, Prentice Hall.

[46] Verbeek M. (2004) A Guide to Modern Econometrics, Wiley.

[47] K.F. Wallis et al. (1987) Models of the UK economy : a fourth review by the ESRC Macroeconomic Modelling Bureau , Oxford.

[48] Wang Peijie (2003) Financial Econometrics, Routledge Advanced Texts.

[49] Wooldridge J. M. (2002) Econometric Analysis of Cross Section and Panel Data, MIT Press.

[50] Wooldridge J.M. (2009) Introductory Econometrics: A Modern Approach, 4th ed., South Western.

13.0.8 Professional articles in econometrics

References

[1] Anderson T.W. (1996) R.A. Fisher and Multivariate Analysis, Statistical Science, 11:1:20-34.

[2] Bartlett M.S. (1947) Multivariate Analysis, Supplement to the Journal of the Royal Statistical Society, 9:2:176-197.

[3] Beyer A, REA Farmer, J Henry and M. Marcellino (2005) Factor Analysis in a New Keynesian Model, European Central Bank, Working Paper Series, no. 510, August

[4] Baltagi Badi H. (1992) Sampling Distributions and Efficiency Comparisons of OLS and GLS in the Presence of Both Serial Correlation and Heteroskedasticity Econometric Theory, Vol. 8, No. 2 (Jun., 1992), pp. 304-305

[5] Blomquist N. Soren (1980) A Note on the Use of the Coefficient of Determination The Scandinavian Journal of Economics, Vol. 82, No. 3 (1980), pp. 409-412

[6] Blundell Richard and Ian Preston (1998) Consumption Inequality and Income Uncertainty The Quarterly Journal of Economics, Vol. 113, No. 2 (May, 1998), pp. 603-640

[7] Blundell Richard and Costas Meghir (1986) Selection Criteria for a Microeconometric Model of Labour Supply Journal of Applied Econometrics, Vol. 1, No. 1 (Jan., 1986), pp. 55-80

[8] Box G and G Jenkins (1976) Time Series Analysis, Forecasting and Conctrol, Holden Day.

[9] Bauwens L , M. Lubrano and J. F. Richard (1999) Bayesian Inference in Dynamic Econometric Models, Oxford University Press.

[10] Chesher A (1984) Improving the efficiency of Probit estimators, Review of Economic Studies,66:3:523-527.

[11] Cochrane D and G. H. Orcutt (1949) Application of Least Squares Regression to Relationships Containing Auto- Correlated Error Terms Journal of the American Statistical Association, Vol. 44, No. 245 (Mar., 1949), pp. 32-61

[12] Davidson Russell, James G. Mackinnon, Russel Davidson (1985) Heteroskedasticity-Robust Tests in Regressions Directions: Annales de l'inséé, No. 59/60, Économétrie non linéaire asymptotique (Jul. - Dec., 1985), pp. 183-218

[13] Dickey D.A. and W.A. Fuller (1979) Distribution of the Estimator for Autoregressive Time Series with a Unit Root, Journal of the American Statistical Association, 74:366:427-437,June.

[14] Durbin J. and G. S. Watson (1950) Testing for Serial Correlation in Least Squares Regression: I, Biometrika, Vol. 37, No. 3/4 (Dec., 1950), pp. 409-428

[15] Durbin J. (1987) Statistics and Statistical Science Journal of the Royal Statistical Society. Series A (General), Vol. 150, No. 3 (1987), pp. 177-191

[16] Durbin J. and M. G. Kendall (1951) The Geometry of Estimation Biometrika, Vol. 38, No. 1/2 (Jun., 1951), pp. 150-158

[17] Durbin J. and G. S. Watson (1951) Testing for Serial Correlation in Least Squares Regression. II Author(s): Source: Biometrika, Vol. 38, No. 1/2 (Jun., 1951), pp. 159-177

[18] Durbin J. (1970) Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors are Lagged Dependent Variables Econometrica, Vol. 38, No. 3 (May, 1970), pp. 410-421

[19] Durbin J. and G. S. Watson (1971) Testing for Serial Correlation in Least Squares Regression. III Biometrika, Vol. 58, No. 1 (Apr., 1971), pp. 1-19

[20] Engle R E and C.W.J. Granger (1987) Co-integration and Error Correction: Representation, Estimation and Testing. Econometrica, vol. 55, No. 2, pp. 251-276.

[21] Farrar Donald E. and Robert R. Glauber (1967) Multicollinearity in Regression Analysis: The Problem Revisited The Review of Economics and Statistics, Vol. 49, No. 1 (Feb., 1967), pp. 92-107

[22] Fisher R. A. (1923) Statistical Tests of Agreement between Observation and Hypothesis Economica, No. 8 (Jun., 1923), pp. 139-147

[23] Fisher Jonas D.M. and Ryan Peters (2010) Using stock returns to identify Government spending shocks, The Economic Journal, 120 (May), 414–436.

[24] Garratt A., K. Lee, M.H. Pesaran and Y. Shin (2003) A Structural Cointegration VAR Approach to Macroeconometric Modelling, Economic Journal,113:487:412-455.

[25] Glejser H. (1969) A New Test for Heteroskedasticity Journal of the American Statistical Association, Vol. 64, No. 325 (Mar., 1969), pp. 316- 323

[26] Godfrey L. G. and M. R. Wickens (1981) Testing Linear and Log-Linear Regressions for Functional Form The Review of Economic Studies, Vol. 48, No. 3 (Jul., 1981), pp. 487-496

[27] Griffiths W. E. and R. C. Hill and G. G. Judge (1993) Learning and Practicing Econometrics, John Willey

[28] Grubb David and Lonnie Magee (1988) A Variance Comparison of OLS and Feasible GLS Estimators Econometric Theory, Vol. 4, No. 2 (Aug., 1988), pp. 329-335

[29] Hall Alastair R. (2000) Covariance Matrix Estimation and the Power of the Overidentifying Restrictions Test Econometrica, Vol. 68, No. 6 (Nov., 2000), pp. 1517-1527

[30] Hall Alastair (1987) The Information Matrix Test for the Linear Model,The Review of Economic Studies, Vol. 54, No. 2 (Apr., 1987), pp. 257-263

[31] Hansen P. L. (1982) Large sample properties of generalised methods of moments estimators, Econometrica, 50:4:1029-1054.

[32] Harvey Andrew (1997) Trends, Cycles and Autoregressions, The Economic Journal, Vol. 107, No. 440 (Jan., 1997), pp. 192-201

[33] Harvey A.C. (1990) The Econometric Analysis of Time Series, Phillip Allan, 2nd edition.

[34] Hansen L.P. (1982) Large sample properties of generalized method of moment estimators, Econometrica, 50:4:1029-1054.

[35] Hamilton James D. (1994) Time Series Analysis, Princeton.

[36] Hausman J.A., (1978), Specification Tests in Econometrics, Econometrica, Vol. 46, No. 6, pp.1251-1271.

[37] Hausman Jerry A. (1975) An Instrumental Variable Approach to Full Information Estimators for Linear and Certain Nonlinear Econometric Models Econometrica, Vol. 43, No. 4 (Jul., 1975), pp. 727-738

[38] Hausman Daniel M. (1989) Economic Methodology in a Nutshell The Journal of Economic Perspectives, Vol. 3, No. 2 (Spring, 1989), pp. 115-127

[39] Heckman James J. (1978) Dummy Endogenous Variables in a Simultaneous Equation System Econometrica, Vol. 46, No. 4 (Jul., , pp. 931-959

[40] Heckman J. J., (1979) Sample Selection Bias as a Specification Error, Econometrica, Vol. 47, No. 1, pp153-161.

[41] Hendry D.F. (1997) The Econometrics of Macroeconomic Forecasting , *Economic Journal*, 107, 444., 1330-1357

[42] Hendry (1995) Dynamic Econometric Theory, Oxford.

[43] Hendry David F. (1974) Stochastic Specification in an Aggregate Demand Model of the United Kingdom Econometrica, Vol. 42, No. 3 (May, 1974), pp. 559-578

[44] Imbens G. W. and T Lancaster (1994) Combining Micro and Macro Data in Microeconometric Models, Review of Economic Studies, 61:4:655-680.

[45] Jarque Carlos M. and Anil K. Bera A Test for Normality of Observations and Regression Residuals: International Statistical Review Vol. 55, No. 2 (Aug., 1987), pp. 163-172

[46] Johansen Soren (1988) Statistical analysis of cointegration vectors, Journal of Economic Dynamics and Control, 12:231-254, North Holland.

[47] Johansen Soren (1988) Estimation and Hypothesis Testing of Cointegration Verctors in Gaussian Vector Autoregressive Models, Econometrica, 59:6, 1551-1580.

[48] Kadane Joseph B. and T. W. Anderson (1977) A Comment on the Test of Overidentifying Restrictions Econometrica, Vol. 45, No. 4 (May, 1977), pp. 1027-1031

[49] Keifer N (1988) Economic duration data and hazard functions, Journal of Economic Literature, 26:647-679.

[50] Klein L. R. and Mitsugu Nakamura (1962) Singularity in the Equation Systems of Econometrics: Some Aspects of the Problem of Multicollinearity International Economic Review, Vol. 3, No. 3 (Sep., 1962), pp. 274-299

[51] King Maxwell L. (1981) The Durbin-Watson Test for Serial Correlation: Bounds for Regressions with Trend and/or Seasonal Dummy Variables Econometrica, Vol. 49, No. 6 (Nov., 1981), pp. 1571-1581

[52] Kiviet Jan F. (1986) On the Rigour of Some Misspecification Tests for Modelling Dynamic Relationships The Review of Economic Studies, Vol. 53, No. 2 (Apr., 1986), pp. 241-261

[53] Knight John L. (1986) Non-Normal Errors and the Distribution of OLS and 2SLS Structural Estimators Econometric Theory, Vol. 2, No. 1 (Apr., 1986), pp. 75-106

[54] Kumar T. Krishna (1975) Multicollinearity in Regression Analysis The Review of Economics and Statistics, Vol. 57, No. 3 (Aug., 1975), pp. 365-366

[55] Lancaster T (1979) Econometric Methods for Duration of Unemployment, Econometrica, 47:4:939-56.

[56] Lancaster T and A Chesher (1983) The Estimation of Models of Labour Market Behviour Review of Economic Studies, 50:4:609-624.

[57] Lancaster T (2004) An Introduction to Modern Bayesian Econometrics, Blackwell

[58] Lovell Michael C. (1986) Tests of the Rational Expectations Hypothesis The American Economic Review, Vol. 76, No. 1 (Mar., 1986), pp. 110-124

[59] Maddala G. S. (1971) Generalized Least Squares with an Estimated Variance Covariance Matrix Econometrica, Vol. 39, No. 1 (Jan., 1971), pp. 23-33

[60] McFadden (1963) Daniel Constant Elasticity of Substitution Production Functions The Review of Economic Studies, Vol. 30, No. 2 (Jun., 1963), pp. 73-83

[61] McFadden Daniel and Paul A. Ruud (1994) Estimation by Simulation The Review of Economics and Statistics, Vol. 76, No. 4 (Nov., 1994), pp. 591-608

[62] Mills Terence C., Gianluigi Pelloni, Athina Zervoyianni (1995) Unemployment Fluctuations in the United States: Further Tests of the Sectoral-Shifts Hypothesis The Review of Economics and Statistics, Vol. 77, No. 2 (May, 1995), pp. 294-304

[63] Nabeya Seiji (2001) Approximation to the Limiting Distribution of t- and F-Statistics in Testing for Seasonal Unit Roots Econometric Theory, Vol. 17, No. 4 (Aug., 2001), pp. 711-737

[64] Nelson Charles R. (1987) A Reappraisal of Recent Tests of the Permanent Income Hypothesis The Journal of Political Economy, Vol. 95, No. 3 (Jun., 1987), pp. 641-646

[65] Patterson K (2000) A Primar to Unit Root Test, Approach,MacMillan.

[66] Pedroni, P. (1999): "Critical values for cointegration tests in heterogeneous panels with multiple regressors", Oxford Bulletin of Economics and Statistics, 61, p.653-670.

[67] Pesaran M. H. (1982) A Critique of the Proposed Tests of the Natural Rate-Rational Expectations Hypothesis The Economic Journal, Vol. 92, No. 367 (Sep., 1982), pp. 529-554

[68] Pesaran, M.H. and R. Smith (1995): "Estimating long-run relationships from dynamic heterogeneous panels", Journal of Econometrics, 68, p.79-113

[69] Phillips Peter C. B. (2003) Laws and Limits of Econometrics The Economic Journal, Vol. 113, No. 486, Conference Papers (Mar., 2003), pp. C26-C52

[70] Phillips P.C.B. (1987) Time Series Regression with an Unit Root, Econometrica, vol. 55, No. 2, 277-301.

[71] Ramsey J. B. (1969) Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis Journal of the Royal Statistical Society. Series B (Methodological), Vol. 31, No. 2 (1969), pp. 350-371

[72] Rao Radhakrishna C (1992) R. A. Fisher: The Founder of Modern Statistics Statistical Science, Vol. 7, No. 1 (Feb., 1992), pp. 34-48

[73] Rao C R (1972) Recent Trends of Research Work in Multivariate Analysis, Biometrics- Special Multivariate Issue, 28:1:3-22.

[74] Rayner A C (1970) The Use of Multivariate Analysis in Development Theory: A Critique of the Approach Adopted by Adelman and Morris, Quarterly Journal of Economics, 84:4:639-647.

[75] Rider P R (1936) Annual Survey of Statistical Technique: Developments in the Analysis of Multivariate Data - Part I, Econometrica, 4:3:264-268.

[76] Schervish MJ (1987) A Review of Multivariate Analysis, Statistical Science, 2:4:396-413.

[77] Sim C.A (1980) Macroeconomics and Reality, Econometrica, 48:1 Jan.1 -45._

[78] Simon Julian L. and Dennis J. Aigner (1970) Cross-Section and Time-Series Tests of the Permanent-Income Hypothesis The American Economic Review, Vol. 60, No. 3 (Jun., 1970), pp. 341-351

[79] Sheshinski Eytan (1967) Tests of the "Learning by Doing" Hypothesis The Review of Economics and Statistics, Vol. 49, No. 4 (Nov., 1967), pp. 568-578

[80] Staigler D., Stock J. H., (1997), "Instrumental Variables Regression with Weak Instruments", Econometrica, Vol. 65, No. 3, pp.557-586.

[81] Stock, J.H., and M.W. Watson (2001). Vector Autoregressions, The Journal of Economic Perspectives, 15 (4): 101-115.

[82] Sorensen CK and JMP Gutierrez (2006) Euro Area Banking Sector Integration Using Hierarchical Cluster Analysis Techniques, European Central Bank, Working Paper Series, no. 626, May.

[83] Stevens DL (1973) Financial Characteristics of Merged Firms: A Multivariate Analysis, Journal of Financial and Quantitative Analysis, 8:2:149-158.

[84] Thomas BAM (1961) Some Industrial Applications of Multivariate Analysis, Applied Statistics, 10:1:1-8.

[85] Suits Daniel B. (1984) Dummy Variables: Mechanics V. Interpretation The Review of Economics and Statistics, Vol. 66, No. 1 (Feb., 1984), pp. 177-180

[86] Theil H. (1956) On the Theory of Economic Policy The American Economic Review, Vol. 46, No. 2, Papers and Proceedings of the Sixty-eighth Annual Meeting of the American Economic Association (May, 1956), pp. 360-366

[87] Theil H. (1965) The Analysis of Disturbances in Regression Analysis Journal of the American Statistical Association, Vol. 60, No. 312 (Dec., 1965), pp. 1067- 1079

[88] Wallace T. D. and C. E. Toro-Vizcarrondo (1969) Tables for the Mean Square Error Test for Exact Linear Restrictions in Regression Journal of the American Statistical Association, Vol. 64, No. 328 (Dec., 1969), pp. 1649-1663

[89] White Betsy Buttrill (1978) Empirical Tests of the Life Cycle Hypothesis The American Economic Review, Vol. 68, No. 4 (Sep., 1978), pp. 547-560

[90] White Halbert (9187) Consistency of OLS Econometric Theory, Vol. 3, No. 1 (Apr., 1987), pp. 159-160

[91] Wooldridge Jeffrey M.(1994) Efficient Estimation under Heteroskedasticity Econometric Theory, Vol. 10, No. 1 (Mar., 1994), p. 223

[92] Zellner A. (1985) Bayesian Econometrics, Econometrica, 53:2:253-270

13.0.9 Some Articles Applied Econometrics

References

[1] McGuinness Tony (1980) Econometric Analysis of Total Demand For Alcoholic Beverages in the U.K., 1956-75 The Journal of Industrial Economics, Vol. 29, No. 1 (Sep., 1980), pp. 85-109

[2] Shafik Nemat (1994) Economic Development and Environmental Quality: An Econometric Analysis Oxford Economic Papers, New Series, Vol. 46, Special Issue on Environmental Economics (Oct., 1994), pp. 757-773

[3] Lee Ray-Shine and Nirvikar Singh (1994) Patterns in Residential Gas and Electricity Consumption: An Econometric Analysis Journal of Business & Economic Statistics, Vol. 12, No. 2 (Apr., 1994), pp. 233-241

[4] Geweke John and William McCausland (2001) Bayesian Specification Analysis in Econometrics American Journal of Agricultural Economics, Vol. 83, No. 5, Proceedings Issue (Dec., 2001), pp. 1181-1186

[5] Carey Kathleen (2000) Hospital Cost Containment and Length of Stay: An Econometric Analysis Southern Economic Journal, Vol. 67, No. 2 (Oct., 2000), pp. 363-380

[6] Deolalikar Anil B. and Robert E. Evenson (1989) Technology Production and Technology Purchase in Indian Industry: An Econometric Analysis The Review of Economics and Statistics, Vol. 71, No. 4 (Nov., 1989), pp. 687-692

[7] Tony Lancaster and Andrew Chesher (1983) An Econometric Analysis of Reservation Wages Econometrica, Vol. 51, No. 6 (Nov., 1983), pp. 1661-1676

[8] Sims Christopher A., Stephen M. Goldfeld, Jeffrey D. Sachs (1982) Policy Analysis with Econometric Models Brookings Papers on Economic Activity, Vol. 1982, No. 1 (1982), pp. 107-164

[9] Van Soest Arthur and Peter Kooreman (1987) A Micro-Econometric Analysis of Vacation Behaviour Journal of Applied Econometrics, Vol. 2, No. 3 (Jul., 1987), pp. 215-226

14 Tutorial Problems in Empirical Economics

14.0.10 Tutorial 1 (page 21)

Regress demand for a product (Y_i) on its own prices (X_i) as following

$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad i = 1 \ ...N$$

where e_i is a randomly distributed error term for observation i.

1. List the OLS assumptions on error terms e_i .
2. Derive the normal equations and the OLS estimators of $\widehat{\beta}_1$ and $\widehat{\beta}_2$.
3. A shopkeeper observed the data quantities and prices as given in Table 2 below. What are the OLS estimates of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ implied by these data? Is this a normal good?
4. What are the variances of e_i and Y_i?
5. What are R^2 and $\overline{R}^2$?
6. Determine the overall significance of this model by F-test at 5 percent level of significance. [Critical value of F for df(1,4) =7.71]
7. What are the variances and standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$?
8. Compute t-statistics and determine whether parameters $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are statistically significant at 5 percent level of significance [Critical value of t for five percent significance for 4 degrees of freedom is 2.776(i.e $t_{crit,0.05,4} = 2.777$)].
9. What is the prediction of Y when X is 0.5?
10. What is the elasticity of demand evaluated at the mean values of Y_i and X_i?
11. Reformulate the model to include price of a substitute product in the model. What will happen to this estimation if these two prices are exactly correlated?
12. How would you decide whether demand for this product varies by gender?

Table 16: Data on Quantities and Prices

Quantities (Y_i)	5	10	15	20	25	30
Prices (X_i)	10	8	6	4	2	1

Hints: $\left[\sum X_i = 31 \quad \sum X_i^2 = 221 \quad \sum Y_i^2 = 2275; \sum Y_i = 105 \quad \sum Y_i X_i = 380\right]$;
$(X'X)^{-1} = \begin{bmatrix} 0.605 & -0.085 \\ -0.085 & 0.0164 \end{bmatrix}$

Application

Test whether work-hours depend on weekly or annual pay among UK counties using data Unempl_pay-couties.csv.

14.1 Tutorial 2 (page 40)

A sport centre has a gym. A hypothetical data set on the monthly charges (X) and number of people using the gym (Y) are given in the following table with the values of cross products and square terms

Table 17: Monthy charges and number of customers

X_i	10	8	7	6	3	5	9	12	11	10
Y_i	60	75	90	100	150	120	125	100	80	65

1. Represent X and Y in a Scattered diagram.
2. Draw horizontal and vertical lines with the mean of X and Y in that diagram.
3. Draw a line by your hand that best represents all sample observations.
4. Write a classical linear regression model in which X causes Y.
5. Write the assumptions of the error terms.
6. Derive normal equations of the OLS estimator minimising sum of squared errors. Estimate parameters of the model using above information. Use the deviation technique in your estimation.
7. What is your prediction of Y when X is 13?
8. Calculate the sum of variation in Y.
9. Decompose this total variance into explained and residual components.
10. Find the coefficient of determination or the R-square of this model.
11. Find the variance and standard error of the slope parameter.
12. Calculate the t-statistics and determine its level of significance using the T-table.
13. Construct a 95 percent confidence interval for the slope parameter.

14. Find the variance and the standard error of the intercept parameter.

Application

Get the medal table in the Beijing summer Olympics from the following web (olympic.csv)

http://www.databaseolympics.com/

Determine whether the number of gold, silver or bronze medals won by a country are related to percapita GDP of the country. Get GDP percapita for these countries from the World Bank Development Indicators. Can you predict medal tables for London Olympics from this exercise?

14.2 Tutorial 3 (page 52)

Q1. Suppose you have the following data set on number of tickets sold in a football match (Y), price of tickets (X_1) and income of the customers (X_2). and Y are measured in 10 thousand pounds. You want to find out the exact relation between tickets sold and prices and income of people watching football games.

Table 18: Price, Income and Sales

$X_{1,i}$	11	7	6	5	3	2	1
$X_{2,i}$	2	2	4	5	6	5	4
Y_i	1	2	3	4	5	6	7

1. Write a simple regression model to explain the number of tickets sold in terms of the price of the ticket. Explain briefly underlying assumptions and expected signs of the parameters in this model.

2. Estimate the slope and intercept parameters. Calculate cross products and squared terms needed for estimation from the above data table.

3. Use your estimates in (b) find the explained squared sum $\sum \widehat{y}_i^2$, sum of squared errors $\sum \widehat{e}_i^2$ and the $R^2$and $\overline{R}^2$.

4. Estimate the variance of the error term and the slope coefficient. Explain its importance.

5. Test whether the slope term is significant at 5% confidence level.

6. Build 95 percent confidence interval for estimate of slope and intercept terms.

7. Discuss how reducing type I error may cause increase in type II errors.

8. Calculate the elasticity of demand for football around the mean of Y and X_1.

9. Write a multiple regression model to explain the number of tickets sold in terms of the price of the ticket and the income of individuals going to the football game. What additional assumption(s) do you need while introducing an additional variable.

10. Estimate the parameters of that multiple regression model.

11. What is your prediction of the number of tickets sold if X_1= 5 and X_2=4?

12. Introduce dummy variables in your multiple regression model to show differences in demand for football ticket based on gender differences (1 for male and 0 for females), four seasons (autumn, winter, spring and summer) and interaction between gender and income.

Q2. Suppose that a leading supermarket in the city centre requests to estimate a demand function for beef. Your are considering estimating a model where demand for beef depends on price of beef , price of pork, price of chicken and consumer income as following:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i \quad i = 1 \ldots N \tag{567}$$

where is Y_i demand for beef, $X_{1,i}$ is price of beef, $X_{2,i}$ is price of pork, $X_{3,i}$ is price of chicken,$X_{4,i}$ is income of consumer and ε_i is a normally and $\varepsilon_i \sim N(0, \sigma^2)$ identically distributed random variable.

1. Using your knowledge of microeconomics, write down the expected signs of β_0,β_1,β_2,β_3,and β_4 in this model and explain why?

2. Write major assumptions of the ordinary least square approach to this model.

3. Suppose you have a data set on these variables over last 35 years and you want to estimate parametersβ_0,β_1,β_2,β_3,and β_4. Derive normal equations that you will use get OLS estimators of these parameters?

4. Compute the variances of parameters β_1,β_2,β_3,and β_4.

5. Compute variance-covariance matrix for the random term.

6. Construct a confidence interval on β_1,β_2,β_3,and β_4 and predicted Y_i.

7. How would your result be affected if you find that $X_{1,i} = 0.6X_{2,i}$?

8. How would you modify your model to correct a problem in reported in (g)?

14.3 Tutorial 4 (page 61)

Suppose that you are interested in estimating the demand for beer in Yorkshire pubs and consider the following multiple regression model:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_{1,i}) + \beta_2 \ln(X_{2,i}) + \beta_3 \ln(X_{3,i}) + \beta_4 \ln(X_{4,i}) + \varepsilon_i \quad i = 1 \ldots N \tag{568}$$

where Y_i is the demand for beer, $X_{1,i}$ is the price of beer, $X_{2,i}$ is the price of other liquor products, $X_{3,i}$ is the price of food and other services, $X_{4,i}$ is consumer income. Coefficients β_0,β_1,β_2,β_3,and β_4 are the set of unknown elasticity coefficients you would like to estimate. Again assume that errors ε_i are independently normally distributed, $\varepsilon_i \sim N(0, \sigma^2)$.

1. (a) Estimate the unknown parameters of this model using data in Beer1.csv.
   (b) How would you determine the overall significance of this model? Write down your test criterion. Compare that test statistic with another test statistic that you would use to test whether a particular coefficient, such as β_3, is statistically significant or not.
   (c) How would you establish whether a particular variable is helping to explain the variation in beer consumption?
   (d) Further suppose that you have some non-sample information on the relation between the price and income coefficients as following:
      i. sum of the elasticities equals zero: $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$.
      ii. two cross elasticities are equal: $\beta_3 = \beta_4$ or $\beta_3 - \beta_4 = 0$
      iii. income elasticity is equal to unity: $\beta_5 = 1$
   (e) How do you test whether these restrictions are valid or not ?
   (f) In addition to the variables listed in the above model you suspect that gender and level of education of individuals are important determinants of beer consumption. Explain how you could incorporate these variables in this model.
   (g) The income of an individual also depends upon his/her age. Income in turn determines the consumption of beer. Thus age interacts with income. How would you introduce this age-income interaction effect in the above model?

Instructions for testing linear restrictions in PcGive for cross section data like this:

a. regress Y on $X_{1,i}$ $X_{2,i}$, $X_{3,i}$ and $X_{4,i}$.

b. click on test/linear restriction, put the restrictions in the matrix box. one line for each restriction. For instance if $\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$. to be tested then type 1 1 1 1 1 0 , then click ok , it will test validity of that restriction. If there are two restrictions

β_0+β_1+β_2+β_3+ $\beta_4 = 0$ and β_3- $\beta_4 = 0$ then
1 1 1 1 1 0
0 0 0 1 -1 0
put this input in the matrix box, then click OK. This will test for both restrictions.

14.4 Tutorial 5 (page 67-68)

Q1. Data on income (y), performance indicator (x_1) and quality of workers (x_2) in a certain reputable company is given as following.

Table 19: Data on income, performace and quality of work

y	3	5	7	6	9	6	7
x_1	1	2	3	4	5	6	7
x_2	5	10	15	20	25	30	35

Fit a regression model $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$ for this company. If any problem suggest remedial measures.

Q2. Some international macroeconomists argue that the devaluation has expansionary effect on output through its positive impact on exports and negative impacts on imports. Others think that devaluation has contractionary impact on output. As an econometrician you would like to test which one of these two claims bear close relation to the empirical facts. Based on literature review you come up with the following model

$$g_{y,t} = \beta_0 + \beta_1 time + \beta_2 \left(\frac{G}{Y}\right) + \beta_3 \left[\Delta \ln M - \Delta lm M^*\right] + \beta_4 TOT + \beta_5 RE_t + \varepsilon_t \tag{569}$$

Where $g_{y,t}$ is the growth rate of real output, *time* is time trend, $\frac{G}{Y}$ is the ratio of government expenditure to GNP, M is the money supply, M^* is expected money supply , TOT is the term of trade as provided by the ratio of indices of price of exports to the prices of imports, RE is the real exchange rate. Terms β_0,β_1,β_2,β_3,and β_4 are unknown coefficients to be estimated. As before is the error term, it has a zero mean and constant variance, $\varepsilon_i \sim N(0, \sigma^2)$.

Relevant data are provided in juk.xlsx file (update this data if you can). Estimate the above parameters and answer following questions studying the regression results.

1. (a) Explain significance of coefficients β_0,β_1,β_2,β_3,$\beta_4$and β_5 in the above model and state whether the estimates are consistent with the economic theory. Is a devaluation, an increase in RE, contractionary or expansionary from the results of this model?

(b) Explain how you can test three of the following restrictions (1) separately and (2) jointly in this model.

i. Restriction 1: $\beta_5 = 0$

ii. Restriction 2: $\beta_2 = 0$ and $\beta_4 = 0$

iii. Restriction 3: $\beta_3 + \beta_4 = 0$

iv. Discuss your test statistic for (i) to (iv).

(c) If the data series used in this model is non-stationary, mention how does it affect the estimates of the parameters? What would you do to correct it?

Instructions for testing linear restrictions in PcGive for cross section data like this:

a. regress Y on $X_{1,i}$ $X_{2,i}$, $X_{3,i}$ and $X_{4,i}$.

b. click on test/linear restriction, put the restrictions in the matrix box. one line for each restriction. For instance if β_0+β_1+β_2+β_3+ $\beta_4 = 0$. to be tested then type 1 1 1 1 1 0 , then click ok , it will test validity of that restriction. If there are two restrictions

β_0+β_1+β_2+β_3+ $\beta_4 = 0$ and β_3- $\beta_4 = 0$ then

1 1 1 1 1 0

0 0 0 1 -1 0

put this input in the matrix box, then click OK. This will test for both restrictions.

14.5 Tutorial 6 (page 115)

Q1. What are elasticities of Y_i with respect to X_i in the following regression models

(a) $Y_i = \beta_1 + \beta_2 X_i + e_i$

(b) $ln\,(Y_i) = \beta_1 + \beta_2 X_i + e_i$.

(c) $ln\,(Y_i) = \beta_1 + \beta_2 \ln(X_i) + e_i$.

(d) $Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + e_i$.

(e) $Y_i = \beta_1 + \beta_2 X_i + \beta_2 X_i^2 + e_i$

II. Output (Y_i) of a firm depends non-linearly on physical capital (K_i), labour (L_i) and energy (E_i)as given by

$$Y_i = \beta_0 K_i^{\beta_1} L_i^{\beta_2} E_i^{\beta_3} e_i \tag{570}$$

(a) How can OLS be applied to estimate $\beta_0, \beta_1, \beta_2$ and β_3?.

(b) How can one test constant return to scale assumption ($\beta_1 + \beta_2 + \beta_3 = 0$) using this regression?

(c) What are the elasticities of output with respect to physical capital (K_i), labour (L_i) and energy (E_i) in this model? .

Q2. Consider a multiple regression model of certain product in which quantity supply (Y_i) depends on its own price ($X_{1,i}$) and prices of two inputs ($X_{2,i}, X_{3,i}$) as

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + e_i \quad i = 1 \ ...N \tag{571}$$

(a) Write normal equations to derive the OLS estimators of $\beta_0, \beta_1, \beta_2$ and β_3.

(b) Derive estimators of $\beta_0, \beta_1, \beta_2$ and β_3. (hint: use deviation form and the Cramer's rule).

(c) Write expression of the variance of the error term.

(d) What is the covariance matrix of $\beta_0, \beta_1, \beta_2$ and β_3? Write expression of variances of β_1, β_2 and β_3.

(e) Determine the test statistics for significance of each of $\beta_0, \beta_1, \beta_2$ and β_3 and to test the general hypothesis that $\beta_0 = 0, \beta_1 = 0, \beta_2 = 0$ and $\beta_3 = 0$.

(f) What would be OLS estimators derived above if $X_{2,i} = 5X_{3,i}$? What can be done to improve the validity of model like this?

(g) Assume that instead of estimating $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + e_i$ a researcher estimated $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + e_i$. How could you determine the statistical significance of the estimated model?

Take a simple linear regression model of the following form.

$$Y_i = \beta_1 + \beta_2 X_i + e_i \quad i = 1 \ ...N \tag{572}$$

Where the variance of the error term differs for different observations of X_i.

Q3 (a) Discuss how the graphical method be used to detect the heteroskedasticity.

(b) Prove that parameters $\beta_1$and β_2 are still unbiased.

(c) Analyse consequences of heteroskedasticity on the BLUE properties of the OLS estimators.

(d) Discuss how the Goldfeld and Quandt and Glesjer tests can be used to determine existence of the heteroskedasticity problem.

(e) Illustrate procedure for the White test of heteroskedasticity.

(f) Illustrate any two remedial measures of removing the heteroskedasticity when the variance is known and when it is unknown.

(g) From a sample of 6772 observations on pay work-hours and taxes contained in PAYHRTX.csv determine whether heteroskedasticity exists or not on the basis of cross section estimates from the the PcGive. Feel free to use Shazam if you know and prefer it.

(h) From a sample of 201 counties of Great Britain contained in Unempl_pay_counties.csv regress work-hours on annual pay and determine whether heteroskedasticity is present in this estimation using the White test.

(i) Suggest remedial measures to remove heteroskedasticity in models like above.

(j) Explains concepts of ARCH and GARCH models briefly.

14.6 Tutorial 7 (page 89)

Suppose that you are estimating a log linear consumption function of the following form:

$$\ln(C_t) = \beta_0 + \beta_1 \ln(Y_t) + \beta_2 \ln(P_t) + \varepsilon_t \quad t = 1 \ldots T \tag{573}$$

where C, Y and P are consumption, income and prices and ε_t is the random error

term. Use information in conyp.xls (update this to quarterly series from 1960 using OECD or Eurostat New Cronos) to estimate unknown parametersβ_0,$\beta_1$and β_2 and answer following questions using these results.

Q1. (a) What are the estimates of $\beta_1$and β_2? Do these estimates have signs as you expected and why?

(b) Does the Durbin-Watson Statistic show evidence of autocorrelation in the model? If so how does it affect the properties of the OLS estimators of $\beta_1$and β_2?

(c) What is the 95 and 90 percent of confidence interval estimate of $\beta_1$and β_2?

(d) How well does this model can explain variation in consumption? How do you decide overall fit of this model? What statistics do you use to decide at least there is one significant variable in the model?

Q2. Consider a simple linear regression model.

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad t = 1 \ldots T \tag{574}$$

Now assume that errors are correlated to each other over time with AR(1) process as:

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t \tag{575}$$

where v_t is identically and normally distributed error term with zero mean and constant variance, $v_t \sim N(0, \sigma^2)$.

1. Illustrate how the graphical method can be applied to detect autocorrelation in a simple regression model like above?
2. What are consequences of autocorrelation in a regression model? Show how the existence of such autocorrelation among the error terms affects the BLUE properties of the OLS estimators.
3. Define and derive the Durbin-Watson test statistics. Show how it can test for existence or non existence of autocorrelation in a given estimation?
4. How the autocorrelation can be removed if the ρ is known?
5. What is a spurious regression? Why does it arise and how does it affect the usefulness of estimation from an OLS regression? What can be done to correct it?
6. Estimate aggregate supply function for UK using data in UKsupply.csv and determine whether autocorrelation exists in it using the Durbin-Watson statistics. Use remedial measures as necessary.

Application:
Read data on growth rate of per capita GDP, exchange rate and inflation rates from the www.imf.org for year 1980 to 2003 for China, India, South Africa, UK, USA and Brazil as contained in PERCAP6.csv. Test whether inflation and the exchange rate are the significant variables in explaining the growth rate of per capita output (in PPP) in these economies. Determine whether heteroskedasticity and autocorrelation exist in this regression using PcGive. Feel free to use Shazam if you know and prefer it. Suggest a remedy for autocorrelation in a model like this.

14.7 Tutorial 8 (page 89)

Stationarity, Unit Root and Cointegarion

1. Study the monthly data on unemployment rate and inflation since 1972:1 to 2004:8 as given in "unmnth.xls" file. Use GiveWin PcGive to

- Draw diagrams to represent the rates of unemployment among males and females and the RPI over this period.
- Ascertain whether unit root exists in the overall unemployment rate, URT and RPI at 5% and 1% level of significance in level, in log and in the first difference of these series.
- Detrend the data with Hodrik-Prescott filter and conduct stochastic volatility tests.

2. Regress unemployment rate on inflation rate in levels and in the first differences. Test whether these series are cointegrated using the Engle-Granger procedure. (hint: stationarity of residuals).

3. The time series and represent the underlying data generating processes (DGP) of consumption $\{C_t\}$ and income $\{Y_t\}$. Answer the following questions regarding the properties these series.

   (a) What is meant by saying that$\{C_t\}$ and $\{Y_t\}$ are stationary series? Why is it important that the series are stationary for a robust regression analysis?

   (b) How do you determine whether $\{C_t\}$ and $\{Y_t\}$ are stationary series, or not?

   (c) Analyse the properties of these series when they follow a random walk, or have a unit root.

   (d) What is the meaning of the order of integration in this respect? Discuss any three different methods of checking for stationarity.

   (e) What is the meaning of cointegration between the series and ? How would you decide whether these series$\{C_t\}$ and $\{Y_t\}$ are co-integrated, or not?

   (f) If the original series $\{C_t\}$ and $\{Y_t\}$ are not co-integrated, what transformation can be applied to achieve co-integration? How do you decide the order of co-integration?

   (g) Use time series of consumption and income contained in Quarterly_cons.xls. Determine the order of integration for both consumption and income. Is there an evidence of cointegration between consumption and income in levels or in the first differences?

14.8 Tutorial 9 (page 94-95)

Q1. Suppose that you have a simple model of consumption and income as following

Consumption function:

$$C_t = \beta_0 + \beta_1 Y_t + u_t \tag{576}$$

National income identity:

$$Y_t = C_t + I_t \tag{577}$$

1. Use rank and order conditions to find whether the consumption function is identified in this model.
2. Write a reduced form for this system. Show how you could retrieve the structural coefficientsβ_0 and β_1 if you applied OLS to this reduced form.
3. Show that application of OLS to (1) generates a biased estimate of β_1.
4. What other method would you recommend to get an unbiased and best estimator for this model? Write steps to be followed until you get the structural coefficients β_0 and β_1.
5. Write a short note on how this model could be used to make a historical simulation of consumption and income series.
6. Estimate a simple macromodel using data in macro08_uk.csv

Empirical Procedure in PcGive

- construct data set in macroeocnomic variables (Y, C, I , G, T , X, M, MS, i, inflation, wage rate, exchange rate etc)
- save data in *.csv format; e.g. macro.csv
- Start GiveWin and PcGive and open data file
- choose multiple equation dynamic modelling
- determine endogenous and exogenous variables and run simultaneous equation using 3SLS or FIML
- Study coefficients
- Change policy variables and construct few scenarios

Q2. Consider a market model for a particular product.
Demand: $Q_t^d = \alpha_0 + \alpha_1 P_t + \alpha_2 I_t + u_{1,t}$ (1)
Supply: $Q_t^s = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + u_{2,t}$ (2)
Here Q_t^d is quantity demanded and Q_t^s is quantity supplied, P_t is the price of commodity,P_{t-1} is price lagged by one period, I_t is income of an individual, $u_{1,t}$ and $u_{2,t}$ are independently and identically distributed (iid) error terms with a zero mean and a constant variance.Q_t and P_t are endogenous variables and $P_{t-1}$and I_t are exogenous variables α_0,α_1,α_2,and β_0,β_1,$\beta_2$are six parameters defining the system.

1. How can simultaneity bias occur if one tries to apply OLS to each of the above equations.

2. Use rank and order conditions to judge whether each of these two equations are over-, under- or exactly identified.

3. Write down the reduced form for this system.

4. How would you estimate the coefficients of the reduced form equations? Write down the estimator.

5. If equations are identified explain how you may retrieve the structural parameters α_0,α_1,α_2,and β_0,β_1,β_2, and from the coefficients of the reduced form equations.

14.9 Tutorial 10 (page 101)

Q1. Consider a panel data regression model aimed to measure the impacts of FDI on economic growth as following:

$$y_{i,t} = \alpha_i + \beta_1 y_{i,t-1} + \beta_2 F_{i,t} + \beta_3 T_{i,t} + \beta_4 I_{i,t-1} + e_{i,t} \qquad e_{i,t} \sim IID\left(0, \sigma_e^2\right) \tag{578}$$

where $y_{i,t}$ is the growth rate $F_{i,t}$ FDI ratio to GDP, $T_{i,t}$ is the ratio of tax revenue, $I_{i,t-1}$ is the ratio of investment. Use data in panel_growth_inflow_outflow.csv to estimate this model using panel package in PcGive. Interprete your results.

Q2. Consider the cross-regional variation of expenditure on food in the UK. For simplicity, it is assumed that food expenditure depends only on wage and salary income in each region.

1. Formulate a model relating expenditure on food (F) and income (Y) that takes account of region specific effects. Note that the equations for each region are independent but that there is contemporaneous correlation among the error terms across the regions. State the major assumptions of the model.

2. Represent the model in terms of a system of stacked regressions that takes account of both individual and system specific effects. What is the structure of the covariance matrix of the error terms in this system?

3. Show how the SURE or GLS estimator system can be applied to estimate the structural parameters of this model. Write out their covariance structure in the matrix form.

4. This model has been estimated using a pooled time series and cross section data set (with the sample size of T=14 and N=13) available from the web site of the Office of the National Statistics (food_exp_UK_regional_panel.csv: hhttp://www.statistics.gov.uk). The estimated coefficients, by region, are given in the following table. Analyse and interpret these results.

Q3. Construct a panel data on growth rate of per capita income, investment ratio, population growth, export, imports, exchange rate, inflation rate for any five country of your choice. Suggest a panel growth model to be estimated. (datafile Panel_growth_Exchangerate(1).csv).

Action:

Construct data on growth rate of per capita GDP, exchange rate and inflation rates from the www.imf.org for year 1980 to 2009 for China, India, South Africa, UK, USA and Brazil from the World Economic Outlook Database. Test whether inflation and the exchange rate are the significant variables in explaining the growth rate of per capita output (in PPP) in these economies using random or fixed effect models. (See percap6.csv)

Q4. Study house price and related variables for economic regions in UK contained in HousePrice_regional.csv file. Do SURE estimation and interprete the results. [In in PcGive use 3SLS routine in simultaneous equation model].