Business Econometrics

ECO 601

Lecture Notes

As Delivered By

Dr Sayyid Salman Rizavi

On VU Television Network

Virtual University of Pakistan

ECO601 - BUSINESS ECONOMETRICS

Lecture 01

Overview of the Course

The course of Business Econometrics is designed for students of Business and Economics. It is an introductory level course but covers all useful topics. The course is not only suitable for students of Business, Commerce, Economics, and useful for Research students.

The presentation will be bilingual (English and Urdu) and is presented for a wide range of audience. It will include the uses software for estimations of the econometric models discussed. This includes the use of Microsoft Excel till the mid-term examination and later we plan to introduce stata (software for statistics and econometrics developed and supplied by Stata Corporation).

It will be supplemented with lecture notes, websites & learning modules of statistical software.

The course requires basic knowledge of statistics and probability. Understanding and use of calculus will be an added advantage. An average basic background of business and economics is also helpful.

Prescribed Text Books

- Wooldridge, J. M. (2007), Introductory Econometrics: A Modern Approach, 3rd Edition, Thomson-South Western
- Gujarati, D. N. (2003), Basic Econometrics, 4th ed. (McGraw-Hill: New York)
- Butt, A. Rauf, “Lest Square Estimation of Econometrics Models”, (National Book Foundation, Islamabad)

Supplementary Readings

- Green, William H. (2002), Econometric Analysis, 5th Edition, (New York University: New York).
- Salvatore, D. & Reagle, D. (2002), Statistics and Econometrics, 2nd Edition, Schaum’s outline series, (McGraw-Hill: New York).
- R.C. Hill, W.E. Griffiths and G.G. Judge (1993), Learning and Practicing Econometrics (Wiley: London). [More advanced.]

Additional Resources

http://www.wikihow.com/Run-Regression-Analysis-in-Microsoft-Excel

The above website provides a very good introduction to Use of a tool in Microsoft Excel to run regressions with some diagnostic test.

http://www.ats.ucla.edu/stat/stata/

The above website of University of California LA is a great collection of training material and modules to learn the statistical software that we intend to use. It provides video tutorials, lectures, training and learning material.

http://data.worldbank.org/data-catalog/world-development-indicators

The above website of The World Bank Group is a data archive for more than 180 countries. It provides macroeconomic and financial data on almost every aspect of the countries in the world for more than 60 years.

What is Econometrics or Business Econometrics?

Traditional Perception

- Econometrics is the branch of economics concerned with the use of mathematical methods (especially statistics) in describing economic systems.
- Econometrics is a set of quantitative techniques that are useful for making "economic decisions"
- Econometrics is a set of statistical tools that allows economists to test hypotheses using really world data. "Is the value of the US Dollar correlated to Oil Prices?", "Is Fiscal policy really effective?", "Does growth in developed countries stimulate growth in the developing countries?"
- The Economist's *Dictionary of Economics* defines Econometrics as "The setting up of mathematical models describing mathematical models describing economic relationships (such as that the quantity demanded of a good is dependent positively on income and negatively on price), testing the validity of such hypotheses and estimating the parameters in order to obtain a measure of the strengths of the influences of the different independent variables."

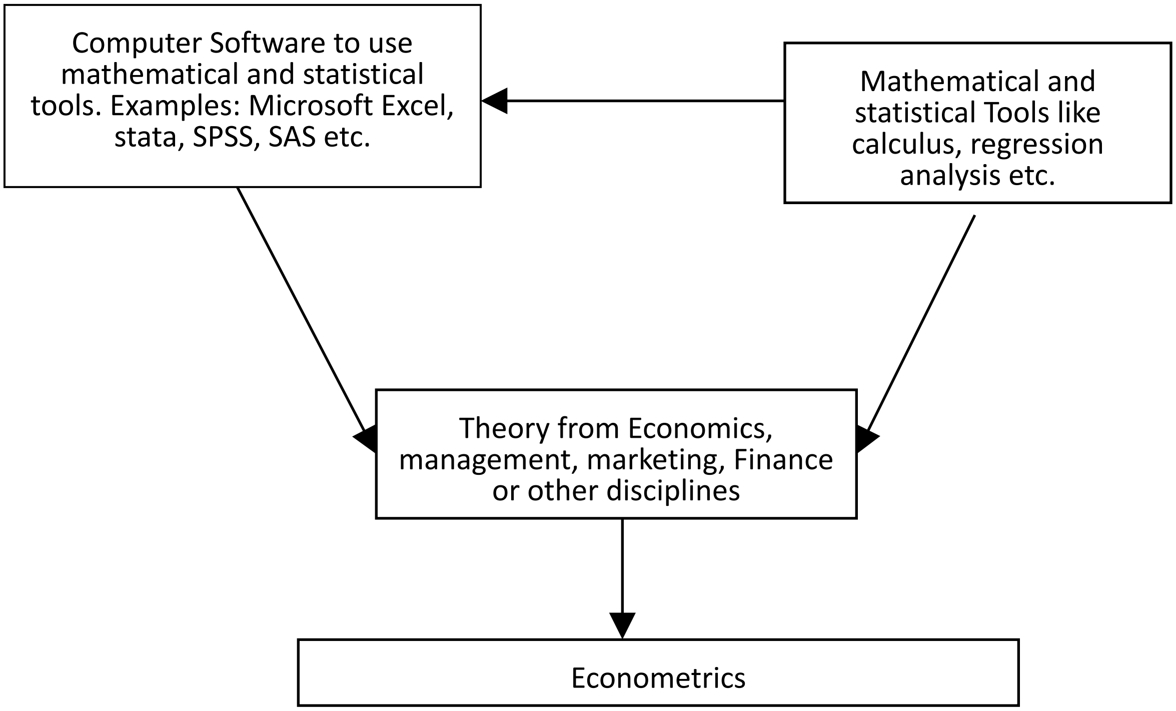
- Econometrics is the intersection of economics, mathematics, and statistics. Econometrics adds empirical content to economic theory allowing theories to be tested and used for forecasting and policy evaluation.
- Econometrics is the branch of economics concerned with the use of mathematical and statistical methods in describing, analyzing, estimating and forecasting economic relationships. Examples of Economic relationships or Business relations and interactions are:
  - Estimation of the market model (demand and supply)
  - Are oil prices and the value of US dollar correlated?
  - What are the determinants of growth?
  - How are liquidity and profitability related?

Modern View

- Econometrics is no more limited to testing, analyzing and estimating economic theory. Econometrics is used now in many subjects and disciplines like Finance, Marketing, Management, Sociology etc.
- Also, the advent of modern day computers and development of modern software has helped in estimation and analysis of more complex models. So computer programing is now an essential component of modern day econometrics.
- Econometrics is the application of mathematics, statistical methods, and, more recently, computer science, to economic data and is described as the branch of economics that aims to give empirical content to economic relations.
- It is no more limited to quantitative research but encompasses qualitative research. So we can finally arrive at a simple but modern and comprehensive definition as:

  Using the tools of mathematics, statistics and computer sciences, Econometrics analyses quantitative or qualitative phenomena (from Economics or other disciplines), based on evolution and development of theory, by recording observations based on sampling, related by appropriate methods of inference.

The following flow chart summarizes the above discussion


Computer Software to use mathematical and statistical tools. Examples: Microsoft Excel, stata, SPSS, SAS etc.
Mathematical and statistical Tools like calculus, regression analysis etc.
Theory from Economics, management, marketing, Finance or other disciplines
Econometrics

Why should you study Econometrics?

The following arguments can be presented to convince a student of business and economics to study Business Econometrics:

- Econometrics provides research tools for your subject.
- Econometrics provides empirical evidence for theoretical statements. Without empirical support the statements may have no value. The theories are tested based of different models and we can forecast the results and make predictions.
- Data never speaks for themselves; Econometrics makes Data speak
- From Idea to forecasting: First we may have an Idea that can be converted to a sound theory. To test the theory we need a functional form showing the relationship of the variables. After that we can go for specification in which we use mathematical equations to reflect the nature of relationship of the variables. The next step may be data collection. We then may use the data for estimation, testing, forecasting based on the model that we have specified.

The Methodology of Business Econometrics

The methodology of Business Econometrics may be described by the following steps:

- Creation of a statement of theory or hypothesis
- Collection of Data
- Model Specification
- Model Estimation
- Performing Diagnostic Tests
- Testing the Hypothesis
- Prediction or Forecasting

The creation of a statement of problem may be based on the existing theory of business and economics. We already know something about the interaction and relationship of variables. For example, we know that the quantity demanded may depend on price, income, prices of substitutes and complementary goods and some other variables. We collect data on these variables and specify our model based on demand theory. We can estimate the model with the help of some technique provided by Econometrics. The estimation may not be free form problems. Here some additional steps may be performed where we can check the validity of the model that we have specified by the use of various diagnostic tests to diagnose any possible problems in the estimation. For that, we test various hypothesis regarding the effectiveness and validity of the estimators. The ultimate result may be predicting or forecasting outcomes like economic and financial events of outcomes. If the technique and model applied is appropriate, the forecasts would be better.

Structure of Data

Cross-Sectional Data: Sample of entities at a given point in time

Time Series Data: Observations over time

Pooled Data / Pooled Cross Sections: Combined Cross Sections from different years

Panel / longitudinal Data: Time Series of each Cross Section, Same cross sectional units are followed over time

Example of Cross-Sectional Data

Monthly Income of a sample of individuals in 2014

Respondent	Income (Rupees)
Ali	75000
Faisal	42000
Iqbal	33000
Noreen	65000

Other Examples: GDP across countries, Annual Sales of different companies in 2014 etc.

Examples of Time Series Data

Monthly Income of a Person over time

Year	Average Monthly Income in Rupees
2010	35000
2011	42000
2012	47000
2013	51000
2014	55000

Other Examples: Pakistan's GDP from 1972 to 2012, Annual Sales of General Motors from 1985 to 2012 etc.

Time series data also need special attention. For example, many variables follow a time trend and we must take care of this while analyzing relationships of variables in time series data. Time series econometrics is evolving as a separate subject now.

Example of Pooled Data / Pooled Cross Sections

Monthly income of respondents from 2011 to 2013

Sample year	Respondent	Income (Rupees monthly average)
2011	Ali	75000
2011	Iqbal	42000
2012	Salma	74000
2012	Kumail	68000
2013	Sultan	80000
2013	Lubna	83000

Note that individual may change in different years

Examples of Panel or longitudinal Data

Exchange Rate of different countries over time

Source: Penn World Tables

Country	Year	Exchange Rate to US dollar
Indonesia	2008	9698.96
Indonesia	2009	10389.9
Indonesia	2010	9090.43
Pakistan	2008	70.40803
Pakistan	2009	81.71289
Pakistan	2010	85.19382
Sri Lanka	2008	108.3338
Sri Lanka	2009	114.9448
Sri Lanka	2010	113.0661

Note that Individual entities (countries) do not change over time

Some Sources of Data

You can just Google for the following and find economic and financial data

- World Development Report
- World Development Indicators
- International Financial Statistics
- Penn World Tables
- US time use Survey
- Panel Survey of Income Dynamics
- http://finance.gov.pk (Ministry of Finance, Pakistan)
- http://sbp.org.pk (State Bank of Pakistan)

File types that you may come across

For downloading and using data, e.g. on the websites like that of the World Bank Group, you may come across the following usual files containing data.

- Microsoft Excel (.xls or .xlsx)
- SPSS (.sav)
- Stata (.dat)
- .csv (Comma Separated values / character separated values)
- .xml (extensible markup language)

Lecture 02

The Summation Notation

The summation operator is heavily used in econometrics. This operator is used to show that we are summing up something e.g. an expression. The Greek letter ∑ (sigma) is used to indicate summation or addition. Usually ∑ is followed by an expression. Summation Notation is an effective and comprehensive way to describe a sum of terms. Let us take some examples to grasp the concept.

Let 'a', 'b' and 'k' denote constants.

Let 'X', 'Y' and 'i' symbolize variables.

X	Symbol
2	X_1
4	X_2
6	X_3
8	X_4
10	X_5
30	$\sum^{5} X$

In the example on the right, the sum of the column of the variable is given as

Sum of X = $X_1 + X_2 + X_3 + X_4 + X_5 =$

$$\sum_{i=1}^{5} X_i$$

Where i is a subscript and changes from 1 to 5

In general we write summation of X as

$$\sum_{i=1}^{n} X_i$$

Here n is a finite number.

Another Example: how summation notation makes life easy

Consider the expression containing different fractions like

$$\frac{2}{3} + \frac{3}{4} + \frac{4}{5} + \frac{5}{6} + \frac{6}{7}$$

Let $k = 2$, then the expression can be written as

$$\sum_{k=2}^{6} \frac{k}{k+1}$$

To see how, we need to let k=2 first which gives

$$\frac{2}{3}$$

If $k = 3$ the expression is

$$\frac{3}{4}$$

We will continue till $k = 6$ and sum up all terms which gives:

$$\frac{2}{3}+\frac{3}{4}+\frac{4}{5}+\frac{5}{6}+\frac{6}{7}$$

Now we need to specify the range of values of k which is 2 to 6. We also need to specify the we are summing up (not multiplying for instance) which we do by applying the letter ∑

The final expression is

$$\sum_{k=2}^{6}\frac{k}{k+1}$$

This gives

$$\frac{2}{3}+\frac{3}{4}+\frac{4}{5}+\frac{5}{6}+\frac{6}{7}$$

This is called expanding the summation expression

Practice Question 2.1: Try expanding the following expression and finding the value

$$\sum_{i=1}^{5}\frac{(i+1)^2}{i}$$

Practice Question 2.2: Try expanding the following expression and finding the value

$$\sum_{j=1}^{3}\frac{(2j+1)^2}{10j^2}$$

Practice Question 2.3: Try expanding the following expression and finding the value

$$\sum_{i=1}^{5} X^2$$

Where X assumes the values 5, 6, 7, 8 and 9

Practice Question 2.4: Try to write the following in summation notation

$$1 + 4 + 9 + 16 + 25 + 36 + 49 + 64 + 81 + 100$$

Practice Question 2.5: Try to write the following in summation notation

$$2 + \frac{3}{4} + \frac{4}{9} + \frac{5}{16} + \frac{6}{25} + \frac{7}{36}$$

Properties of the Summation Operator

Property 1

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} a_{\boldsymbol{i}} = na$$

Sum of 'a' = $a_1 + a_2 + a_3 + a_4 + a_5$

$$= \sum_{i=1}^{5} a_i$$

A	Symbol
2	a_1
2	a_2
2	a_3
2	a_4
2	a_5
10	$\sum^{5} a$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{5}} \boldsymbol{a_i} = 2 + 2 + 2 + 2 + 2 = 10$$

In fact it is five times 2 $= 5\, multiplyed\ by\ 2 = \ na \ = \ 10$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{5}} a_{\boldsymbol{i}} = 5a = na,\ where\ n = number\ of\ observations$$

Which can be generalized as

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} a_{\boldsymbol{i}} = na$$

IMPORTANT: We usually do not write subscript 'i' with a constant. This was just an example

Note that 'a' is a constant and all values of it are identical.

When ∑ is multiplied by a constant we can write 'n' instead of ∑

Property 2

$$\sum_{i=1}^{n} kX_i = k\sum_{i=1}^{n} X_i$$

Let k = 5

X	5X
1	5
2	10
3	15
4	20
5	25
Total: 15	**Total: 75**

In column 2,

$$5 + 10 + 15 + 20 + 25 = 75 = \sum_{i=1}^{5} 5X_i$$

This can also be computed as

$$5 \text{ x } 15 = 75 = 5\sum_{i=1}^{5} X_i$$

Hence A constant value can be factored out of the summation operator and we can write

$$\sum_{i=1}^{n} kX_i = k\sum_{i=1}^{n} X_i$$

Property 3

$$\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$$

X	Y	X + Y
1	5	6
2	12	14
3	18	21
4	22	26
5	27	32
Total: 15	**Total: 84**	**Total: 99**

$$\sum_{\boldsymbol{i=1}}^{\mathbf{5}} \boldsymbol{X_i^2} \neq (\sum_{i=1}^{5} X_i)^2$$

In column 3

$$\sum_{i=1}^{5}(X_i + Y_i) = 6 + 14 + 21 + 26 + 32 = 99$$

Which can also be computed as:

$$\sum_{i=1}^{5} X_i + \sum_{i=1}^{5} Y_i = 15 + 84 = 99$$

Extension: Combining property 2 & 3 we can also write

$$\sum_{i=1}^{n}(aX_i + bY_i) = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n}(aX_i + b) = \sum_{i=1}^{n}(aX_i) + \sum_{i=1}^{n} b = a\sum_{i=1}^{n} X_i + nb$$

What can NOT be done in the Summation Notation?

The summation algebra is not just identical to normal algebra. Some things that may seem obvious is normal algebra may not apply to summation algebra. Remember that the following expressions are NOT equal

$$\sum_{i=1}^{n}(X_i\,/Y_i) \neq \sum_{i=1}^{n} X_i \;\div \sum_{i=1}^{n} Y_i$$

Also

$$\sum_{i=1}^{n}(X_i\,Y_i) \neq \sum_{i=1}^{n} X_i\;.\sum_{i=1}^{n} Y_i$$

and

$$\sum_{i=1}^{n} X_i^2 \neq \;(\sum_{i=1}^{n} X_i)^2$$

Practice Question 2.6: *Construct a table to prove the first and second inequality* discussed above*.*

Application of Summation algebra

We can prove the following useful expression that may be used later.

Different forms of $\sum(\boldsymbol{X}-\overline{\boldsymbol{X}})(\boldsymbol{Y}-\overline{\boldsymbol{Y}})$

Subscripts (‘i') are omitted/ignored for simplicity

$$\sum(X-\bar{X})(Y-\bar{Y}) = \sum XY - \frac{\sum X \sum Y}{n} \quad = \sum XY - n\,\bar{X}\,\bar{Y}$$

$$\sum(X-\bar{X})(Y-\bar{Y}) = \sum[XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}]$$

$$= \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + n\bar{X}\bar{Y}$$

$$= \sum XY - \frac{\sum X}{n}\sum Y - \frac{\sum Y}{n}\sum X + n\frac{\sum X}{n}\frac{\sum Y}{n}$$

$$= \sum XY - \frac{\sum X \sum Y}{n} - \frac{\sum X \sum Y}{n} + \frac{\sum X \sum Y}{n}$$

$$= \sum XY - \frac{\sum X \sum Y}{n}$$

Also $\sum XY - \frac{\sum X \sum Y}{n} = \sum XY - n\frac{\sum X}{n}\frac{\sum Y}{n} = \sum XY - n\,\bar{X}\,\bar{Y}$

Different forms of $\sum(X-\bar{X})^2$

Subscripts (‘i') are omitted/ignored for simplicity

$$\sum(X-\bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$\sum(X-\bar{X})^2 = \sum[X^2 + \bar{X}^2 - 2X\bar{X}]$$

$$= \sum X^2 + n\bar{X}^2 - 2\bar{X}\sum X$$

$$= \sum X^2 + n\,\frac{(\sum X)^2}{n^2} - 2\frac{\sum X}{n}\sum X$$

$$= \sum X^2 + \frac{(\sum X)^2}{n} - 2\frac{(\sum X)^2}{n}$$

$$= \sum X^2 - \frac{(\sum X)^2}{n}$$

Double Summation

Double Summation or nested summation also can be used

Example:

$$\sum_{i=1}^{3}\sum_{j=1}^{2} X_{ij} = X_{11} + X_{12} + X_{21} + X_{22} + X_{31} + X_{32}$$

Example:

$$\sum_{i=1}^{3}\sum_{j=1}^{2} X_i Y_j = X_1Y_1 + X_1Y_2 + X_2Y_1 + X_2Y_2 + X_3Y_1 + X_3Y_2$$

Linear Functions

Most of you would be familiar to straight lines or linear functions. A variable may be a linear function of another if its plot produces a straight line. A linear function may be written as

$$Y = a + bX$$

a = intercept (the point where the line intersects the y-axis)

b = slope, rate of change, derivative

$$\text{As} \qquad Y = a + bX$$

$$\Delta Y = b\Delta X$$

$$b = \frac{\Delta Y}{\Delta X} = \text{marginal effect}$$

Function: Each domain value (X) represents a unique range value (Y)

Linear function: A function whose graph forms a straight line OR for which the rate of change ‘b’ is constant. Linear function can be with our without intercept. A straight line that is shown without intercept, when plotted, shows a line passing through the origin. Assuming linear relationship makes the models easy to solve.

Consider the following table

X	Y
1	7
2	9
3	11
4	13

5	15

As the linear equation is written as $Y = a + bX$, we need the values of a and b for this equation

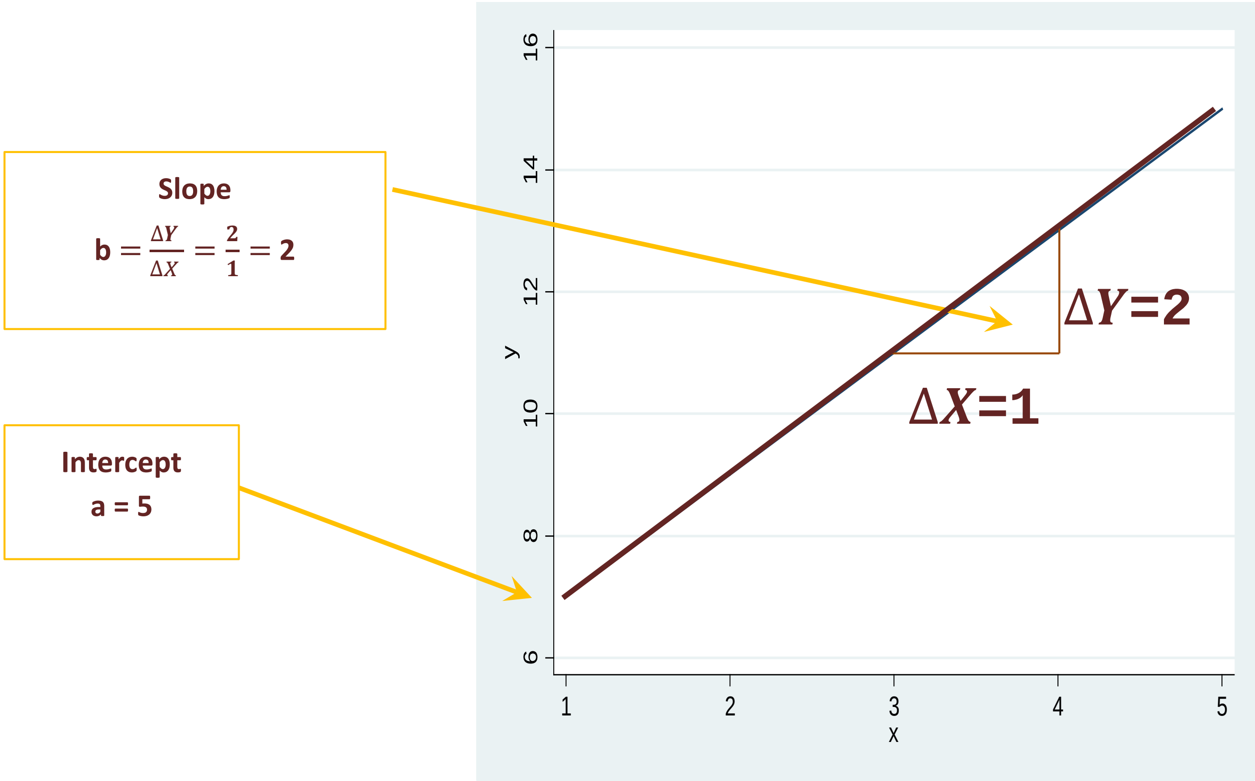
We can compute it from the first two rows as

$$b = \frac{\Delta Y}{\Delta X} = \frac{9-7}{2-1} = \frac{2}{1} = 2$$

Note that this ratio is the same for if we use row 2 and row 3 or any other two consecutive rows.

As $Y = a + bX$ we can get the value of a as $a = Y - bX$ and compute it from any row in the given table. Here $a = 7 - 2(1) = 5$ so the equation for the table above can be written as

$$Y = 5 + 2X$$



Slope

$$b = \frac{\Delta Y}{\Delta X} = \frac{2}{1} = 2$$

Intercept

a = 5

Simple examples of Linear Functions

Linear Demand Functions

The Demand Function: $Q_d = f(P,\ Y,\ P_s, P_c, A)$

Where $Q_d = Quantity\ Demanded$

$P = Price, Y = income,\ P_s = Price\ of\ Substitute, P_c = Price\ of\ complemantary\ good$

$A = Advertisement\ Expenditure$

Expression in terms of linear equation

$$Q_d = \mathrm{a} + \mathrm{b\ P} + \mathrm{c\ Y} + \mathrm{d}\ P_s + e\ P_c + f\ A$$

Simple Demand Function

$$Q_d = \mathrm{a} + \mathrm{b\ P}\ \ ,\ \text{Ceteris Paribus}$$

We estimate the parameters 'a' and 'b' from data. (Sometimes with the help of regression analysis)

What do we expect? The sign of 'b' is negative for 'normal' goods, sign of b is positive for 'Giffen' goods

Practice Question 2.7:

Assume $Q_d = 50 - 2\ \mathrm{P}\ \ ,\ \text{Ceteris Paribus}$

Activity: Assume valued of P (price) to be 1, 2, 3, 4 and 5

Compute Q_d and plot the 'Demand Curve'

NOTE: Here we have used a linear equation as a specification of a demand function, however Demand function may be non-linear in reality.

Simple examples of using Linear Equations

Example:

Some times we can 'linearize' equations

Simple linear regression: linear in variable functional form $Y = \beta_0 + \beta_1 X$

Marginal effect = β_1

Elasticity = ε = β_1 (X/Y)

Double log functional form

$$lnY = \beta_0 + \beta_1 lnX$$

Can be written as

$Y^* = \beta_0 + \beta_1 X^*$ where $Y^* = lnY, X^* = lnX$

Marginal effect: m = β2(Y/X)

Elasticity: ε = β_1

Example:

Linear-Log functional form

$$Y = \beta_0 + \beta_1 lnX$$

Can be written as

$Y = \beta_0 + \beta_1 X^*$where $X^* = lnX$

Marginal effect = $\frac{\beta_1}{X}$

Elasticity = ε = $\frac{\beta_1}{Y}$

Log-Linear functional form

$$lnY = \beta_0 + \beta_1 X$$

Can be written as

$Y^* = \beta_0 + \beta_1 X$ where $Y^* = lnY$

Marginal effect: m = $\beta_1 Y$

Elasticity: ε = $\beta_1 X$

Example:

Cobb-Douglas Production Function

$$Y = AL^{\alpha}K^{\beta}$$

Taking log on both sides,

$$\ln Y = \ln A + \alpha \ln L + \beta \ln K$$

Can be written as

$$Y^{*} = a + \alpha L^{*} + \beta K^{*}$$

where $L^{*} = \ln L$, $K^{*} = \ln K$ and $Y^{*} = \ln Y$

Which can be estimated as a linear equation

The equation is not linear but we can estimate it by transformation

Lecture 03

Quadratic Function

A quadratic function is a function of the form

$$f(x) = Y = aX^2 + bX + c \quad where\ a \neq 0$$

a, b and c are called coefficients

The graph forms a parabola. Each graph has either a maxima or minima

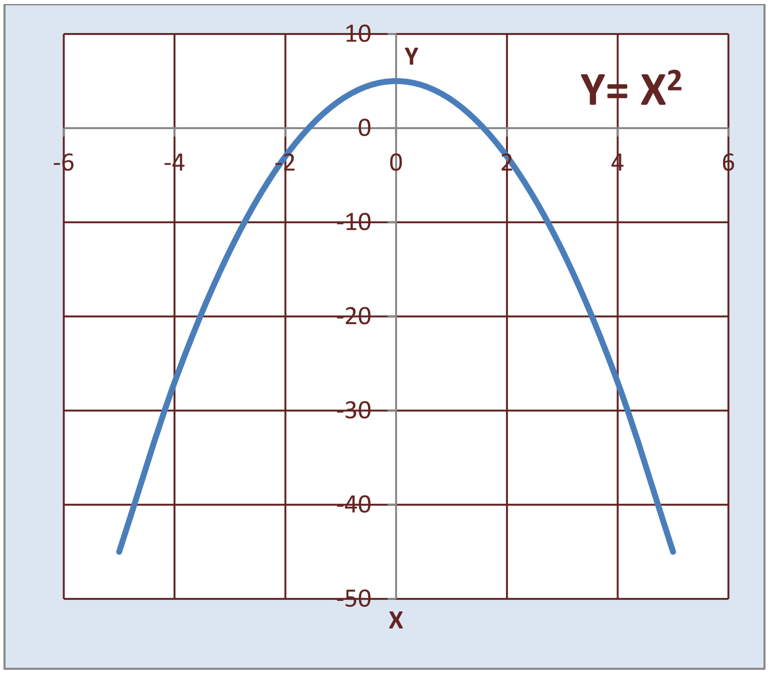
A line divides the graph in two parts creating symmetry

Examples:

- $Y = 2X^2 + 3X + 10$
- $Y = 3X^2 - 5X + 5$
- $Y = 10X^2 + 2X$
- $Y = 5X^2$

In the diagram:

- Axis of Symmetry: x = 0
- Here a = 1, b = 0, c = 0



Example:

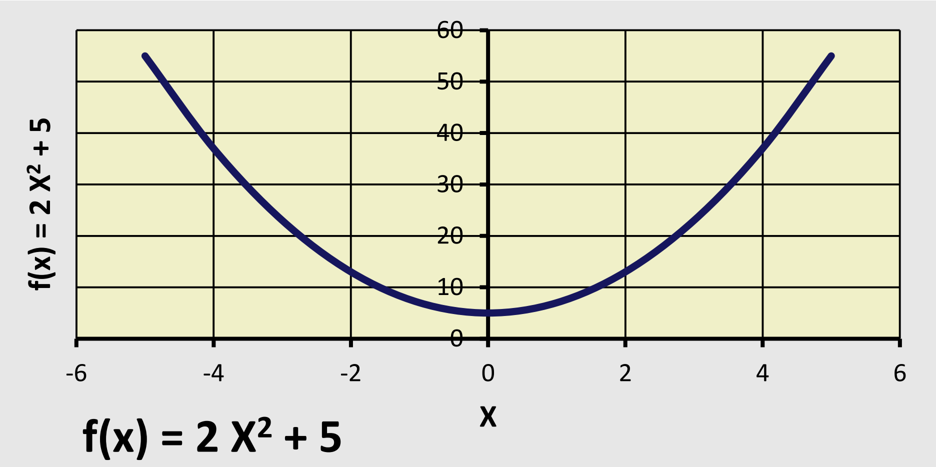
Form: $Y = aX^2 + bX + c$

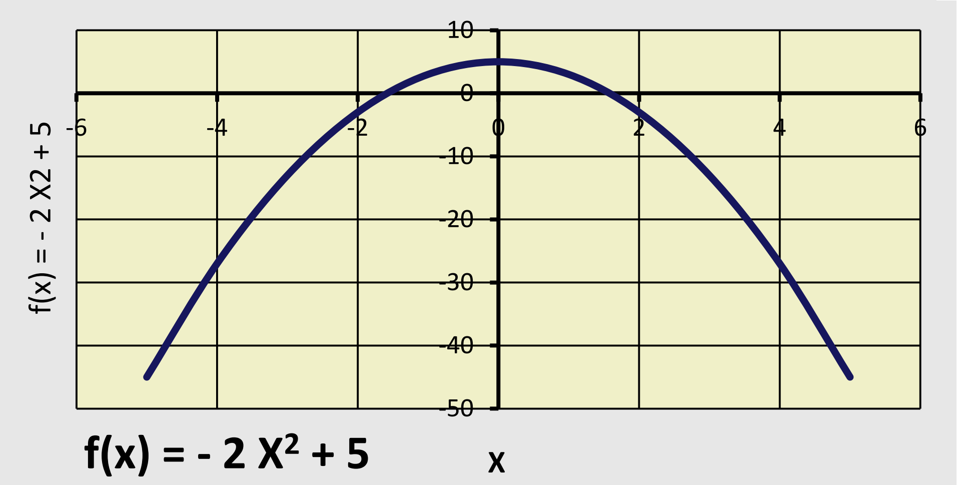
When a is positive, the graph concaves downward

When a is negative, the graph concaves upward (see the graph)

When c is positive, the graph moves up

When c is negative, the graph moves down.





Quadratic Function in econometrics

Let us consider some quadratic functions. The practical examples discussed here can be of inverted-U-shaped functions and U-shaped functions

Inverted U relationships

Liquidity and profitability

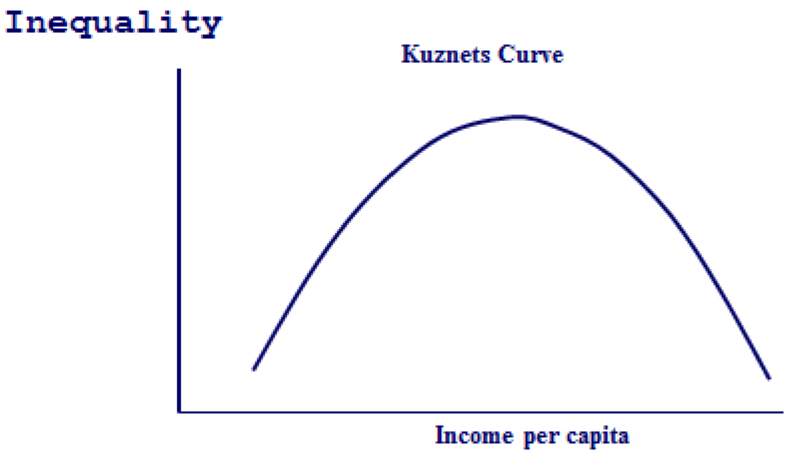
The profitability has many determinants including liquidity. For the liquidity of a firm, we use indicators like current ratio and quick ratio. Normally a range of 1 to 2 is fine for current ratio. This means that if the liquidity ratio is less than 1 then the firm has inadequate resources to meet her obligations. This may negatively affect profitability so, at this stage, an increase in liquidity may increase profits. However, if a current ratio of above 2 (excess liquidity) is observed, this means that the funds are not placed properly and are not contributing to profit. At this stage, and increase in liquidity may negatively affect profitability.

So, initially, increase in liquidity increases profit but later on an increase in liquidity may decrease profits. This can be dealt by showing the relationship as an inverted-U shape.

Competition and Innovation

Initially increase in competition is good and gives rise to innovation and modification in the products. But too much competition may decrease the possibility of innovation because too much competition gets the prices to a minimum level (break-even point in economics). With just a normal profit, the firms had no incentive to be innovative because they get the same price for the product.

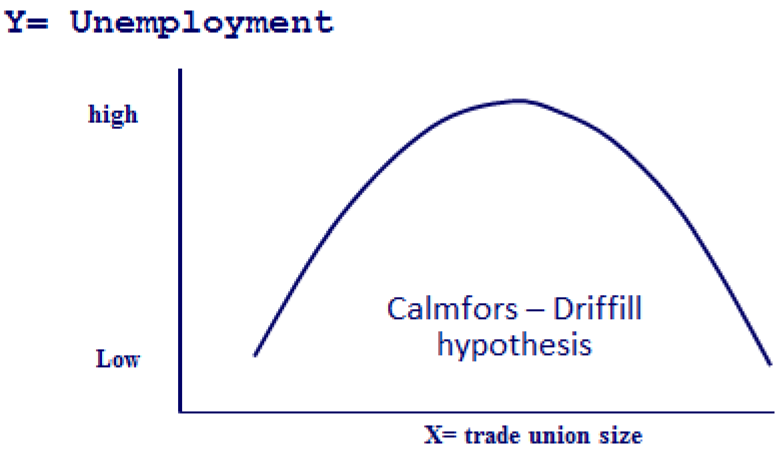
Kuznets Curve (income per capita & income inequality) Kuznets curve represents graphically the hypothesis of Simon Kuznets that with economic development, initially, economic inequality occurs naturally, and then decreases it after a certain average income is attained. This means that initially inequality increases with development but later, it decreases with further development.



Calmfors–Driffill Hypothesis

Inverted U relationships: Calmfors–Driffill hypothesis: Trade union size is a proxy for collective bargaining power. The following text is taken from Wikipedia.org

"The Calmfors–Driffill hypothesis is a macroeconomic theory in labor economics that states that there is a non-linear relationship between the degree of collective bargaining in an economy and the level of unemployment. Specifically, it states that the relationship is roughly that of an 'inverted U': as trade union size increases from nil, unemployment increases, and then falls as unions begin to exercise monopoly power. It was advanced by Lars Calmfors and John Driffill." (Source: Wikipedia.org)



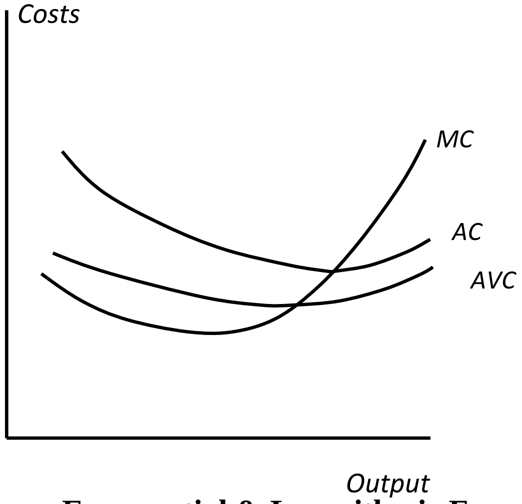
U shaped quadratic relationships

Economic Development and Fertility

As economic development takes place, fertility declines however with more economic development, countries may provide incentives for childbearing. When the cost of childbearing declines, fertility rates may start rising again. If the above is believed, it may be depicted by a quadratic form of equation.

Marginal Cost and Average Cost Curves

Both the marginal and average cost curves that are based on the Cost theory have a U-shape. This means that first marginal and average cost decline with increase in production but after a point they start rising when the production increases. The minimum point for both curves is different but the marginal cost curve intersects the average cost curve from the minimum average cost as seen in the following figure.



Exponential & Logarithmic Functions

Brief Description

- Exponential function are functions in which constant base 'a' is raised to a variable exponent x

$$Y = a^x \; where \; a > 0 \; and \; a \neq 1$$

- 'a' is the base and x is the exponent.
- The base can be any value including the value of e=2.7172828

- ‘e’ is the base of natural logarithm (Euler’s constant)

$$Y = a^x \text{ then } log_a Y = x \; is \; called \log \; to \; the \; base \; 'a'$$

And if

$Y = e^x$ then $log_e Y = \ln Y = x$ *(called natural logarithm)*

- *Some times the exponent can be an expression*

Exponential & Logarithmic Functions

Examples: Exponential Growth

At every instance, the rate of growth of the quantity is proportional to the quantity (population growth may be an example)

$$P(t) = 2e^{3t}$$

Continuous Compound Interest

$$C = Pe^{rt}$$

C = compounded balance after t years

P = Principal amount, t = number of years

r = rate of interest

Logarithmic equation

Equations of the type $ln\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} ln\boldsymbol{X}$ provide elasticity

Simple Derivative

The concept of differentiation

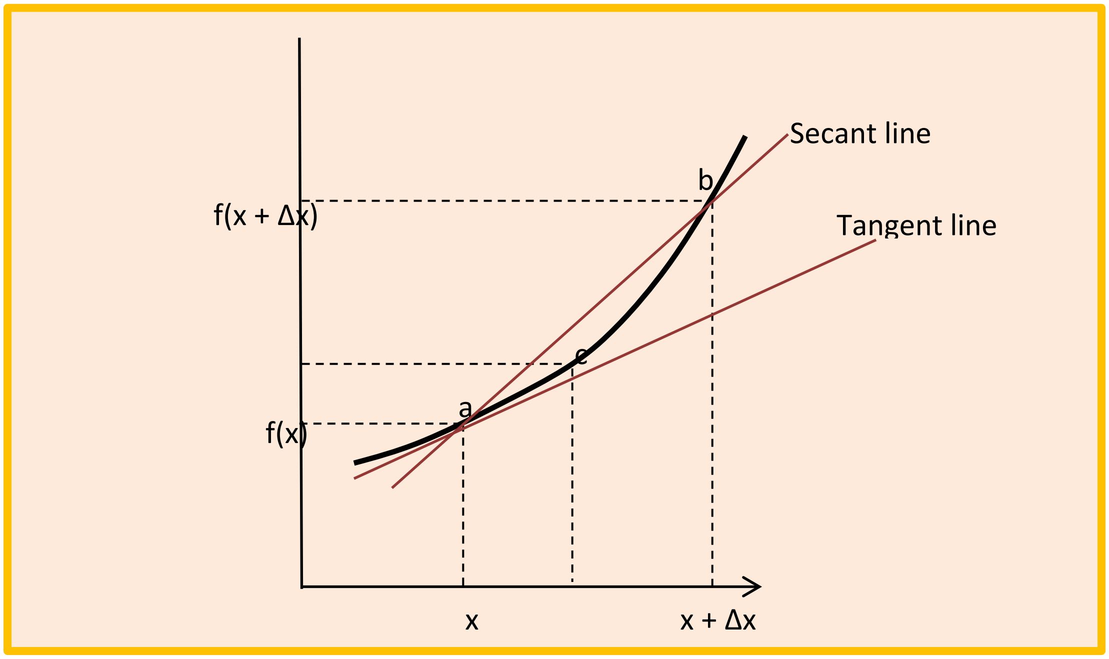
Consider $Y = f(x)$

The Rate of Change is defined as $= \frac{\Delta Y}{\Delta X}$

Derivative is the instantaneous rate of change of the dependent variable due to a very small change in the independent variable. The slope of the tangent line approximates the slope of the function at the point of tangency. The secant line approaches the tangent line by the definition of derivative (see the next slide)

For normal comprehension, derivative, slope of a function, marginal function (like MC as the derivative of TC) can be thought to be identical

$$\frac{dy}{dx} = \acute{y} = \acute{f}(x) = \lim_{\Delta x \to \infty} \frac{f(x+\Delta x) - f(x)}{f(x)}$$



Some Important things to note

Expression	Read as	Meaning
$\acute{f}(x)$	‘f prime x’	Derivative of ‘f’ with respect to x
$\frac{dy}{}$	‘dee why dee ecks’	Derivative of y with respect to x
$\acute{y}$	y prime	Derivative of y
$\frac{d}{dx}f(x)$	‘dee by dee ecks of f of x’	The derivative of the function of x
$'dx'$ *does not mean d multiplied by x* (*same for ‘dy’*)		
‘$\frac{dy}{dx}$’ does not mean $dy \div dx$		

Rules of differentiation

The Power Rule

$$If\ y = a\,x^n,\ \ \frac{dy}{dx} = anx^{n-1}$$

Example

$$y = 10x^3$$

$$\frac{dy}{dx} = \acute{y} = 10\,(3)x^{3-1} = 30\,x^2$$

Example

$$y = 5x^2$$

$$\frac{dy}{dx} = \acute{y} = 5\,(2)x^{2-1} = 10\,x$$

Example

$y = \frac{10}{x^2} = 10x^{-2}$ *(write as the format $a\,x^n$)*

$$\frac{dy}{dx} = \acute{y} = 10\,(-2)x^{-2-1} = -20\,x^{-3}$$

$$= \frac{20}{x^3}$$

The Constant Function Rule

$$If\ y = k\ where\ k\ is\ a\ constant,\ \ \frac{dy}{dx} = 0$$

Derivative is 'rate of change' and there is no change in a constant

This can be derived from the power rule!

The above can be written as $y = k = kx^0\ so\ \acute{y} = k\,(0)x^{0-1} = 0$

Example

$$y = 10,\ \acute{y} = 0$$

What is $y = x$?

$$y = x = 1.x^1$$

$$\frac{dy}{dx} = \acute{y} = (1)(1)x^{1-1} = 1x^0 = 1$$

Hence If $y = x \; then \; \frac{dy}{dx} = 1$

The Sum-Difference Rule

$$If \; y = f(x) \; \pm g(x), \;\; \frac{dy}{dx} = \acute{f}(x) \pm \acute{g}(x)$$

Example

$$y = 10x^3 + 5x^2$$

$$\frac{dy}{dx} = \acute{y} = 10\,(3)x^{3-1} + 5\,(2)x^{2-1}$$

$$= 30\,x^2 + 10\,x$$

Example

The above can be extended to more than two terms

$$y = 2x^3 - 3x^2 - 10x + 5$$

$$\frac{dy}{dx} = \acute{y} = \frac{d}{dx}(2x^3) - \frac{d}{dx}(3x^2) - \frac{d}{dx}(10x) + \frac{d}{dx}(5)$$

$$= 6x^2 - 6x - 10(1) + 0$$

$$= 6x^2 - 6x - 10$$

The Product Rule

$$If \; y = f(x)\,.\,g(x), \;\; \frac{dy}{dx} = g(x).\acute{f}(x) + f(x).\acute{g}(x)$$

The derivative of the product of two functions is equal to the second function times the derivative of the first plus the first function times the derivative of the second.

Example

$$y = (10 - x)(5 + x)$$

$$Here \; f(x) = 10 - x, \; and \; g(x) = 5 + x$$

$$\frac{dy}{dx} = (5 + x)\frac{d}{dx}(10 - x) + (10 - x)\frac{d}{dx}(5 + x)$$

$$= (5 + x)(-1) + (10 - x)(1)$$

$$= -5 - x + 10 - x$$

$$\frac{dy}{dx} = 5 - 2x$$

Verification: $y = (10 - x)(5 + x) = 50 + 5x - x^2$

$$which\ gives\ \acute{y} = 5 - 2x\ (same\ as\ above)$$

The Quotient Rule

$$If\ y = \frac{f(x)}{g(x)},\ \ \frac{dy}{dx} = \frac{g(x).\acute{f}(x) - f(x).\acute{g}(x)}{(g(x))^2}$$

$(g(x))^2$ can be written as $g^2(x)$

Example

$$y = \frac{10 - x}{5 + x}$$

$$Here\ f(x) = 10 - x,\ and\ g(x) = 5 + x$$

$$\frac{dy}{dx} = \frac{(5 + x)\frac{d}{dx}(10 - x) - (10 - x)\frac{d}{dx}(5 + x)}{(5 + x)^2}$$

$$= \frac{(5 + x)(-1) - (10 - x)(1)}{(5 + x)^2}$$

$$= \frac{-5 - x - 10 - x}{(5 + x)^2} = \frac{-15 - 2x}{(5 + x)^2}$$

The Chain Rule: functions involving different variables

$let\ y = f\big(g(x)\big)\ where\ z = g(x)\ \ then\ \frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$

Remember: on the RHS dz does not cancel dz, (dy/dz) is ONE symbol

Example: $y = (5x^2 + 2x + 10)^3$ (we can make use of the chain rule)

$$let\ z = 5x^2 + 2x + 10\ then\ \frac{dz}{dx} = 5x + 2$$

$$y\ can\ be\ written\ as\quad y = z^3\ then\ \frac{dy}{dz} = 3z^2$$

Using the chain rule $\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$

$$\frac{dy}{dx} = (3z^2)(5x+2) = 3(5x^2+2x+10)^2\ (5x+2)$$

NOTE: we can directly apply power rule to such problems

$$\frac{dy}{dx} = 3(5x^2+2x+10)^{3-1}(derivative\ of\ the\ inner\ expression)$$

$$= 3(5x^2+2x+10)^2\ (5x+2)$$

Some Application of Simple Derivatives

Remember:

- Derivative is rate of change or slope or marginal function

Example: Finding Marginal functions

If total cost $C = \frac{1}{3}Q^3 - 2Q^2 + 120\,Q + 1000$

Then Marginal cost is the derivative of total cost

$$MC = \frac{d}{dx}\left(\frac{1}{3}Q^3 - 2Q^2 + 120\,Q + 1000\right)$$

$$MC = \frac{1}{3}(3)Q^{3-1} - 2(2)Q^{2-1} + 120(1) + 0$$

$$MC = Q^2 - 4Q + 120$$

Example: Applying the chain rule

If total cost $R = f(Q)\ \ and\ \ Q = g(L)$

$$\frac{dR}{dL} = \frac{dR}{dQ}.\frac{dQ}{dL}$$

$$= MR.\ MPP_L = MRP_L$$

Example: finding elasticity

$$Q_d = 100\ - 2P$$

(we will learn how to get the values of the above coefficients through regression)

$Price = P = 10,$ Using the above information, $Q_d =\ 100\ - 2(10) = 80$

Example: finding elasticity

$$Q_d = 100 \; - 2P \text{ gives } \frac{dQ}{dP} = -2$$

Price Elasticity of Demand = $E_p = \; \frac{dQ}{dp} . \frac{P}{Q}$

$= \text{(derivative of Q w.r.t. P) times (P/Q)}$

$$= (-2)\left(\frac{10}{115}\right) = -0.1739$$

Which means that for every one percent change in price, quantity demanded decreases by 0.1739 units

Higher Order Derivatives

What are higher order derivatives?

The derivative of a derivative is called second order derivative. The third order derivative is the derivative of the second order derivative. This may continue and are called Higher Order Derivative.

Meaning of the second order derivative: It show the rate of change of the rate of change.

Example: $\boldsymbol{y = 10x^3}$

$$y' = 10\,(3)x^{3-1} = 30\,x^2$$

$$y'' = \frac{d}{dx}(30\,x^2) = 60x$$

$$y''' = \frac{d}{dx}(60x) = 60$$

And so on

Lecture 04
Partial Derivatives

Multivariate Functions

Functions of more than one variable are called multivariate functions.

Examples:

- Quantity Demanded is a function of Price, Income, Prices of other goods and some other variables

$$Q_d = f\ (P,\ I,\ P_o, O)$$

- Profitability depends on liquidity, capital structure, government regulations, prices of raw material etc.

$$\pi = f\ (LQ,\ CS,\ GR,\ PR)$$

Partial Derivatives: Rate of change of the dependent variable with respect to change in one of the independent variables while the other independent variables are assumed to be constant (are held)

Partial Derivatives

Symbols

The mathematical symbol ∂ (partial or partial dee or del) is used to denote partial derivatives.

$$\frac{\partial z}{\partial x}$$

The above symbol is read as 'partial derivative of z with respect to x' (other variables are treated as constants)

Another symbol can also be used: $Z_x\ or\ Z_1$

For second order derivatives we can use the following symbols:

$$\frac{\partial^2 z}{\partial x^2}$$

Or Z_{xx} , Z_{xy} , Z_{11} , Z_{12} etc.

Partial Differentiation

Method to partially differentiate functions

- You have as many 'first order partial derivatives' as number of independent variables
- When we differentiate a variable with respect to any one independent variable, we treat all other variables as if they were constants.
- All the usual rules of differentiation are applicable.
- Higher Order Derivatives may be of two types
    - Direct Partial Derivative: differentiate twice w.r.t. the same variable (2nd order direct partial)
    - Cross Partial Derivatives: differentiate w.r.t. one variable and then w.r.t. another variable (2nd order cross partial)
    - Cross partial Derivatives are always equal (symmetry of second derivatives OR equality of mixed partial)

Partial Differentiation: Examples

Example:

$$Z = f(x,\ y) = 2x^2 + 3y^2 + 5xy + 20$$

- Three type of terms in the expression: That contain only x, That contain only y, That contain both x and y

$$\frac{\partial z}{\partial x} = \ \frac{\partial}{\partial x}(2x^2 + 3y^2 + 5xy + 20)$$

$$\frac{\partial z}{\partial x} = \ \frac{\partial}{\partial x}(2x^2) + \frac{\partial}{\partial x}(3y^2) + \frac{\partial}{\partial x}(5xy) + \frac{\partial}{\partial x}(20)$$

Now treat *y* as a constant while differentiating w.r.t. x

Remember: Derivative of a constant is zero, In case of coefficient do as in the power rule

$$\frac{\partial z}{\partial x} = 4x + 0 + 5y\ (1) + 0 = 4x + 5y$$

Here **y is treated as a constant so** $3y^2$ **is a constant and its derivative is zero**

And y is treated as a constant so $5y$ is a constant coefficient which is multiplied by the derivative of x i.e. by 1

$$Z = f(x,\ y) = 2x^2 + 3y^2 + 5xy + 20$$

Now let us differentiate w.r.t. y, treating x as a constant

$$\frac{\partial z}{\partial y} = \ \frac{\partial}{\partial y}(2x^2 + 3y^2 + 5xy + 20)$$

$$\frac{\partial z}{\partial y} = \ \frac{\partial}{\partial y}(2x^2) + \frac{\partial}{\partial y}(3y^2) + \frac{\partial}{\partial y}(5xy) + \frac{\partial}{\partial y}(20)$$

Now treat *x* as a constant while differentiating w.r.t. y

$$\frac{\partial z}{\partial y} = 0 + 6y + 5x\ (1) + 0 = 6y + 5x = 5x + 6y$$

Here

x is treated as a constant so $2x^2$ is a constant and its derivative is zero

x is treated as a constant so $5x$ is a constant coefficient which is multiplied by the derivative of y i.e. by 1

Example: $Z = f\big(x,\ y\big) = 2x^2y^2 + 5x^3y^4$

$$\frac{\partial z}{\partial x} = \ Z_x = \ \frac{\partial}{\partial x}(2x^2y^2 + 5x^3y^4)$$

$$\frac{\partial z}{\partial x} = Z_x = \frac{\partial}{\partial x}(2x^2y^2) + \frac{\partial}{\partial x}(5x^3y^4)$$

$$Z_x = 2y^2 . \frac{\partial}{\partial x}(x^2) + 5y^4 . \frac{\partial}{\partial x}(x^3)$$

Here $2y^2$ **is presently a constant so we factor it out and differentiate the variable part**

$$Z_x = 2y^2(2x) + 5y^4(3x^2)$$

$$Z_x = 4xy^2 + 15x^2y^4$$

$$Z = f(x,\ y) = 2x^2y^2 + 5x^3y^4$$

$$\frac{\partial z}{\partial y} = Z_y = \frac{\partial}{\partial y}(2x^2y^2 + 5x^3y^4)$$

$$\frac{\partial z}{\partial y} = Z_y = \frac{\partial}{\partial y}(2x^2y^2) + \frac{\partial}{\partial y}(5x^3y^4)$$

$$Z_y = 2x^2.\frac{\partial}{\partial y}(y^2) + 5x^3.\frac{\partial}{\partial x}(y^4)$$

$$Z_y = 2x^2(2y) + 5x^3(4y^3)$$

Here $2x^2$ **is presently a constant so we factor it out and differentiate the variable part**

$$Z_y = 4x^2y + 20x^3y^3$$

Example : Second Order Direct Partial Derivatives

Consider the previous example

$$Z = f(x,\ y) = 2x^2 + 3y^2 + 5xy + 20$$

$$\frac{\partial z}{\partial x} = Z_x = 4x + 5y$$

Differentiating again w.r.t. x

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial x}\right) = Z_{xx} = 4(1) + 0 = 4$$

Similarly

$$\frac{\partial z}{\partial y} = Z_y = 5x + 6y$$

Differentiating again w.r.t. y

$$\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial y}\right) = Z_{yy} = 0 + 6(1) = 6$$

Both are called ‘Second order DIRECT partial derivatives’

$$Z = f(x,\ y) = 2x^2 + 3y^2 + 5xy + 20$$

$$\frac{\partial z}{\partial x} = Z_x = 4x + 5y$$

After differentiating w.r.t. x first, we Differentiate w.r.t. y

$$\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right) = Z_{xy} = 0 + 5(1) = 5$$

Similarly

$$\frac{\partial z}{\partial y} = Z_y = 5x + 6y$$

Now Differentiating again w.r.t. x

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right) = Z_{yx} = 5(1) + 0 = 5$$

Both are called 'Second order Cross partial derivatives'

Note that $Z_{xy} = Z_{yx}$

Now let us find the second order direct partial derivatives

$$Z = f(x,\ y) = 2x^2y^2 + 5x^3y^4$$

$$Z_x = \ \frac{\partial}{\partial x}(2x^2y^2 + 5x^3y^4) = 4xy^2 + 15x^2y^4$$

Differentiating again w.r.t. x

$$Z_{xx} = \frac{\partial}{\partial x}(4xy^2 + 15x^2y^4)$$

$$Z_{xx} = 4y^2 + 30xy^4$$

Similarly

$$Z_y = \ \frac{\partial}{\partial y}(2x^2y^2 + 5x^3y^4) = 4x^2y + 20x^3y^3$$

Differentiating again w.r.t. y

$$Z_{yy} = \frac{\partial}{\partial y}(4x^2y + 20x^3y^3)$$

$$Z_{yy} = 4x^2 + 60x^3y^2$$

$$Z = f(x,\ y) = 2x^2y^2 + 5x^3y^4$$

$$Z_x = \ \frac{\partial}{\partial x}(2x^2y^2 + 5x^3y^4) = 4xy^2 + 15x^2y^4$$

Now Differentiating w.r.t. y

$$Z_{xy} = \frac{\partial}{\partial y}(4xy^2 + 15x^2y^4)$$

$$Z_{xy} = 8xy + 60x^2y^3$$

Similarly

$$Z_y = \frac{\partial}{\partial y}(2x^2y^2 + 5x^3y^4) = 4x^2y + 20x^3y^3$$

Now Differentiating w.r.t. x

$$Z_{yx} = \frac{\partial}{\partial x}(4xy^2 + 15x^2y^4)$$

$$Z_{yx} = 8xy + 60x^2y^3$$

$$Z_{xy} = Z_{yx}$$

An example with Chain Rule & Summation Algebra

Example : $Z = \sum(y - a - bx)^2$

This time let ‘a’ and ‘b’ act as the unknowns (you can think of them as variables)

Differentiating w.r.t. ‘a’

$$Z_a = 2\sum(y - a - bx)(0 - 1 - 0)$$

Here **Chain Rule is applied and we multiply the derivative of the inner expression. Here ‘a’ is the variable.**

$$Z_a = -2\sum(y - a - bx)$$

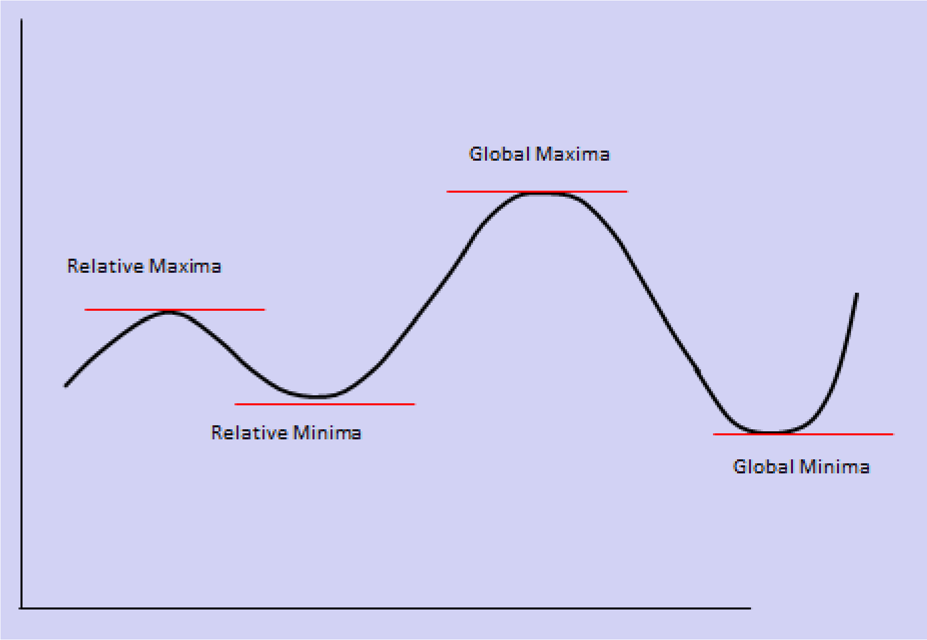
$$Z_a = -2(\sum y - na - b\sum x)$$

Similarly

$$Z_b = 2\sum(y - a - bx)(0 - 0 - x(1))$$

Here b is the variable and its derivative is ‘1’

Simple Optimization:Maxima and Minima



Finding Minima and Maxima

Note that the slope (derivative) of minima or maxima is zero so we can find the point by setting the first derivative equal to zero. This is called First Order Condition of Optimization

Also note that the derivative of the derivative (2nd order derivative) is positive in case of a minima and negative in case of a maxima.

This is called Second Order Condition for Optimization

So, to optimize a function of one variable, we can use two conditions

1. First Derivative = 0 (if $y = f(x),\ \acute{f}(x) = 0$
2. Second Derivative > 0 for minimization & is < 0 for maximization

   $(\acute{f}(x) > 0\ in\ case\ of\ minimization, \acute{f}(x) < 0$ in case of maximization.

Example:

$$If\ y = 40x - 2x^2$$

For maximization or minimization the first derivative should be set equal to zero

$$\frac{dy}{dx} = y' = 40\ - 4x = 0$$

$$40 = 4x$$

$$\bar{x} = \frac{40}{4} = 10$$

To know if it is a maxima or minima, we need to differentiate again

$$\frac{d^2y}{dx^2} = y'' = \ \frac{d}{dx}(40 - 4x) = -4\ < 0$$

As the second derivative is less than zero the function is maximized at $x = 10$,

$$the\ maximam\ value\ is\ Y_{max} = 40(10) - 2(10)^2 = 200$$

Example:

Consider the following profit function where Q is the output

$$\mathrm{\pi} = 100\ \mathrm{Q} - 120 - 2\mathrm{Q}^2$$

Frist Order Condition is

$$\mathrm{\pi}' = 100 - 4\mathrm{Q} = 0$$

$$4\mathrm{Q} = 100$$

$$\bar{Q} = 25$$

Second Order Condition is

$$\mathrm{\pi}'' = \ \frac{\mathrm{d}}{\mathrm{dx}}(100 - 4\mathrm{Q}) = \ - 4 < 0$$

Hence the profit function is maximized at Q = 4

$$\mathrm{\pi}_{\mathrm{MAX}} = 100\ (25) - 120 - 2\ (25)^2 = 2500 - 120 - 2(625)$$

$$\mathrm{\pi}_{\mathrm{MAX}} = 1130$$

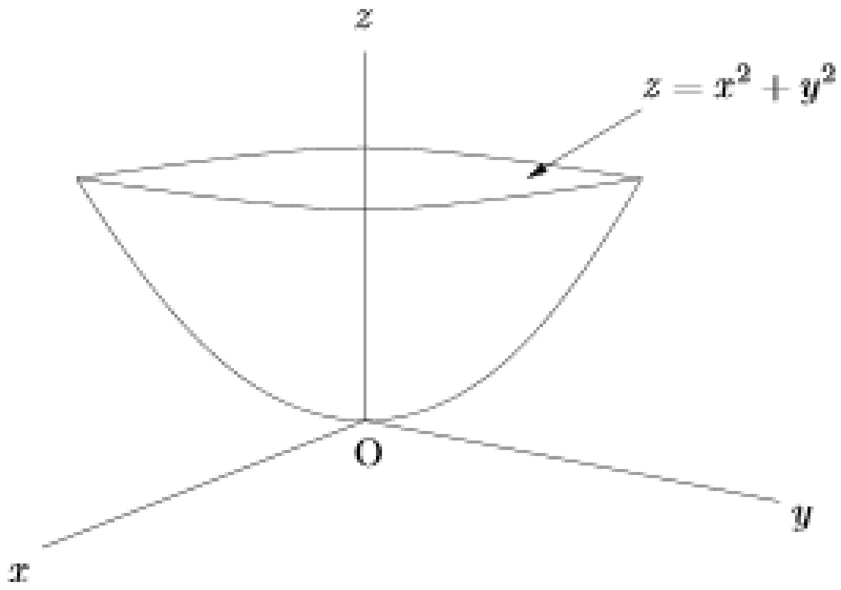
Lecture 05
Multivariate Optimization

Local Minimum

Consider the following diagram

- Point ‘O’ is a local minimum FROM ALL DIRECTIONS
- At point ‘O’, derivative of z w.r.t. x or w.r.t. y both are zero OR the slope of the tangents parallel to x-axis or the one parallel to y-axis at point ‘O’ are both zero i.e.

$$Z_x = 0\ AND\ Z_y = 0\ (First\ Order\ Condition)$$



Local Minimum: Second order condition

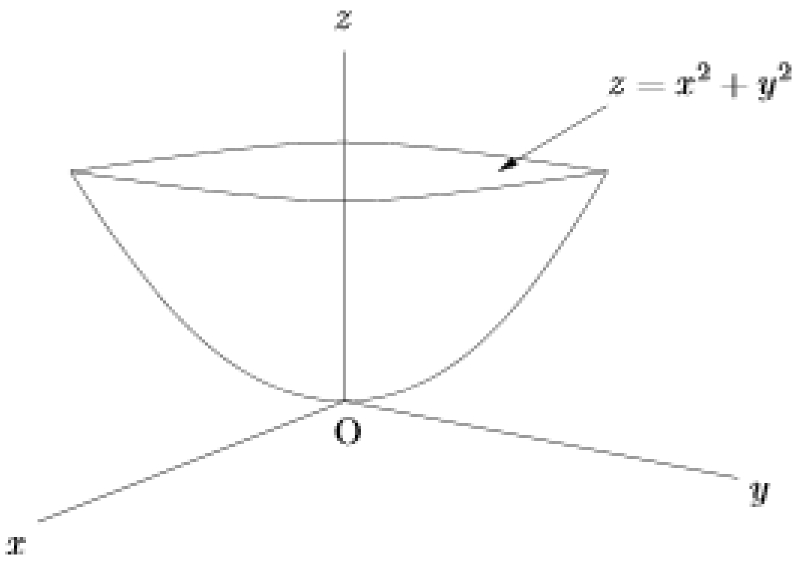
Now consider the point ‘O’ again

When we move the tangents parallel to x-axis or y-axis, there is a positive change in the derivative (derivative of the derivative is positive)

$$Z_{xx} > 0 \quad \& \quad Z_{yy} > 0$$

This gives us the Second Order Condition for minimization

(Both second order derivatives are positive)



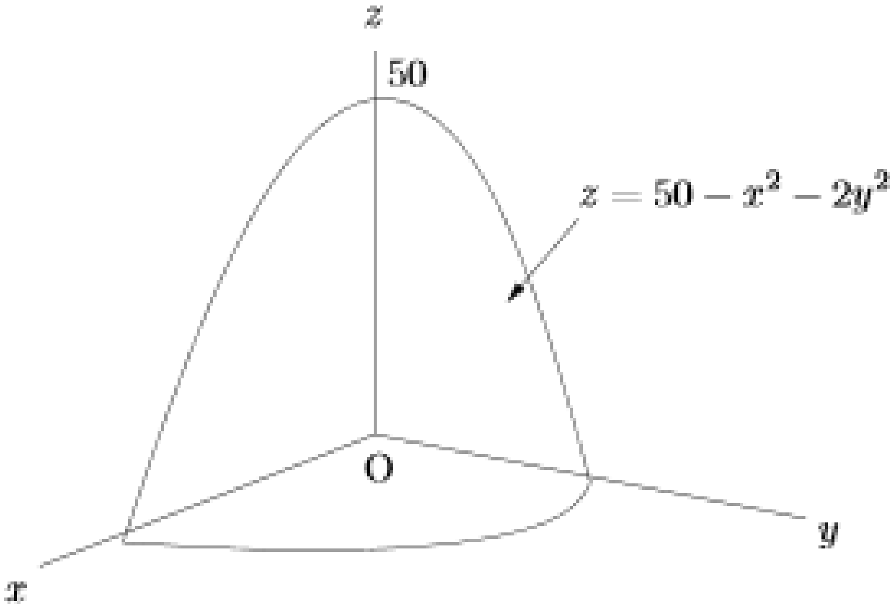
Local Maximum

Consider the following diagram

Point ‘a’ is a local maximum FROM ALL DIRECTIONS

At point ‘a’, derivative of z w.r.t. x or w.r.t. y both are zero OR the slope of the tangents parallel to x-axis or the one parallel to y-axis at point ‘a’ are both zero i.e.

$$Z_x = 0\ AND\ Z_y = 0\ (First\ Order\ Condition)$$



Local Maximum: Second order condition

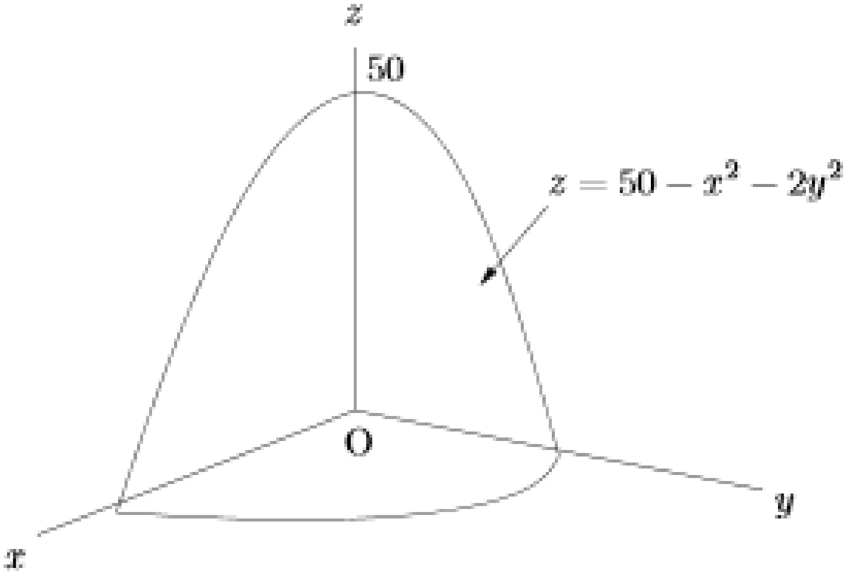
Now consider the point ‘a’ again

When we move the tangents parallel to x-axis or y-axis, there is a negative change in the derivative (derivative of the derivative is negative)

$$Z_{xx} < 0 \quad \& \quad Z_{yy} < 0$$

This gives us the Second Order Condition for maximization

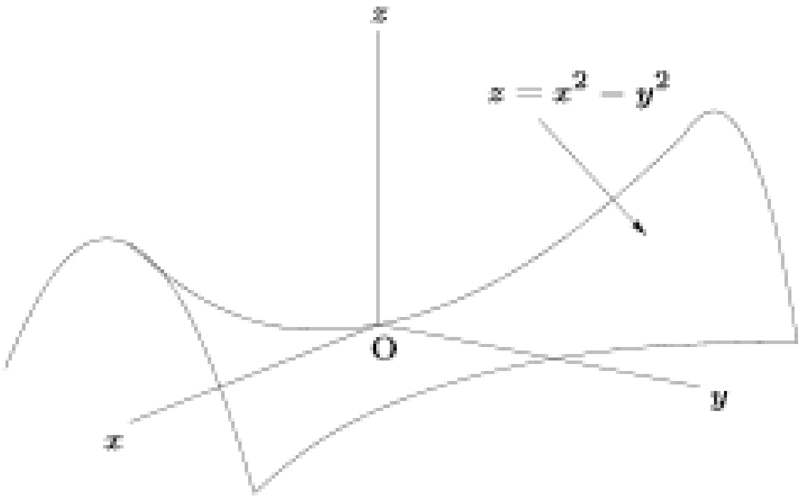
(both second order derivatives are negative)



Saddle Point: Second order derivatives have different signs

Consider the point ‘O’ in the following diagram

- A tangent at this point has a zero slope (first derivative is zero i.e. the first condition is met)
- If we shift the tangent in the direction of the x-axis, the slope of the tangent (derivative of the derivative) increases so this is a local minima form one direction (x-axis) $Z_{xx} > 0$
- But is we shift the tangent at point ‘O’ in the direction of the y-axis, its slope will decrease i.e. $Z_{yy} < 0$



A third condition: ruling out point of inflection

When evaluated at the critical point(s), the product of the second order partials must exceed the product of the cross partials. This condition rules out critical points that are neither points of maximum or minimum, but are points of inflection. A point of inflection is a where certain conditions of optima are met, but the function is not actually a maximum or minimum.

$$Z_{xx}.Z_{yy} > \boldsymbol{Z_{xy}^2}$$

$$\begin{vmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{vmatrix} > 0$$

We call the above a Hessian determinant or simply Hessian which shows that

$$Z_{xx}.Z_{yy} - \boldsymbol{Z_{xy}^2} > \mathbf{0}$$

Or

$$Z_{xx}.Z_{yy} > \boldsymbol{Z_{xy}^2}$$

Example: Maximization

Consider the following profit function where x & y are the levels of output

$$\pi = 80x - 2x^2 - xy - 3y^2 + 100y$$

$$\pi_x = 80 - 4x - y = 0$$

$$\pi_y = -x - 6y + 100 = 0$$

Solving the above two equations simultaneously gives the critical values

$$\bar{x} = 16.52 \quad and \quad \bar{y} = 13.91$$

Second order condition

$$\pi_{xx} = -4 < 0$$

$$\pi_{yy} = -6 < 0$$

Which confirms that profit is maximized from the principle direction at the critical points

Third condition

$$\pi_{xx}.\pi_{yy} = (-4)(-6) = 24\ AND \quad \boldsymbol{\pi_{xy}^2} = (-\mathbf{1})^{\mathbf{2}} = \mathbf{1}$$

$$Hence\ \pi_{xx}.\pi_{yy} > \boldsymbol{\pi_{xy}^2}$$

The profit function is maximized from all directions at the critical point. Maximum profit can be found by substituting the critical points in the profit function.

Example: Minimization

Consider the following marginal cost function where x and y are the level of output

$$MC = 5x^2 - 8x - 2xy - 6y + 4y^2 + 100y$$

$$MC_x = 10x - 8 - 2y = 0$$

$$MC_y = -2x - 6 + 8y = 0$$

Solving the above two equations simultaneously gives the critical values

$$\bar{x} = 1 \quad and \quad \bar{y} = 1$$

Second order condition

$$MC_{xx} = 10 > 0$$

$$MC_{yy} = 8 > 0$$

Which confirms that MC is minimized from the principle direction at the critical points

Third condition

$$MC_{xx}.MC_{yy} = (10)(8) = 80\ AND \quad \boldsymbol{MC_{xy}^2} = (-\mathbf{2})^{\mathbf{2}} = \mathbf{4}$$

$$Hence\ MC_{xx}.MC_{yy} > \boldsymbol{MC_{xy}^2}$$

The function is minimized from all directions at the critical point. Minimum MC can be found by substituting the critical points in the MC function.

Review of Probability

Probability: This is only a 'Review'

Random Experiment

Any process of observation or measurement that has more than one possible outcome and we are not certain about which outcome will materialize

Examples: Tossing a coin, throwing a pair of dice, drawing a card form deck of cards

Sample Space/Population

The set of all possible outcomes of an experiment

Example: when you toss a coin, S = {H, T}

Example: When you toss two coins, S = {HH, HT, TH, TT}

Sample Point

Each member of the sample space is a sample point

Event

Event is a particular collection of outcomes (a subset of the sample space)

Example: Event 'A' is occurrence of one head and one tail in the experiment of tossing two coins A = {HT, TH}

Mutually Exclusive Events: Occurrence of one event prevents the occurrence of the other event at the same time

Example: when we toss two coins, occurrence of two heads means that other three cannot occur at the same time

Example: when we toss a single coin, occurrence of a head means that the tail did not occur or can not occur at the same time

Equally Likely Events: if one event is as likely to occur as the other

Example: head and tail have the same possibility or chance of occurring

Collectively Exhaustive Events: if they exhaust all possible outcomes of an experiment

Example: Event is the sample space. A = Occurrence of a head or tail while tossing a single coin

Stochastic or Random Variable

A variable whose value is determined by the outcome of an experiment

Example: Let X = Number of heads in an experiment of tossing two coins, then X can have values of 0, 1, or 2 as the possibilities are no head, one head or two heads

A random variable can be discrete (can take only whole numbers and finite values) or continuous (can take any values between and interval either whole numbers or fractions e.g. height of an individual).

Classical Definition of Probability:

Probability of an event 'A' = $\mathrm{P(A)} = \frac{Number\ of\ favorable\ outcomes}{Number\ of\ total\ outcomes}$

Example: Total number of outcomes in tossing two coins is 4 {HT, HH, TH, TT}

Probability of getting exactly one head = $\frac{2}{4} = 0.5$

Probability Distribution

The possible values that a random variable can take with the number of occurrences (frequency) of those values.

Example: Probability Distribution of discrete random variable

Let X = Number of heads in an experiment of tossing two coins, the X can have values of 0, 1, or 2 as the possibilities are no head, one head or two heads as shown in table with the probabilities.

Probability Mass Function or simply Probability Function

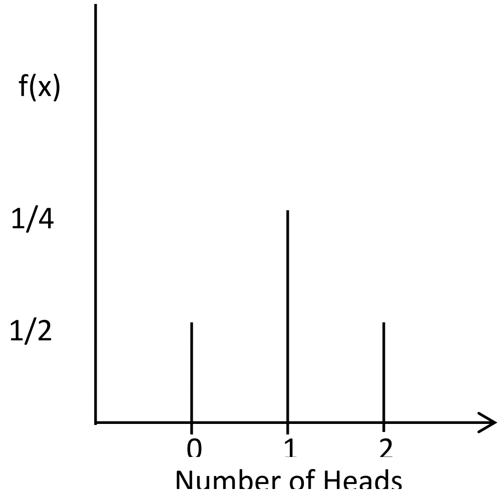
$$f(X = x_i) = P(X = x_i) \quad i = 1,2, \ldots.$$

$$= 0\ if\ X \neq x_i$$

$$0 \leq f(x_i) \leq 1$$

$$\sum_{x} f(x_i) = 1$$

X	$\boldsymbol{P[X = x_i]}$
0	¼=0.25
1	2/4=0.5
2	¼=0.25
Total	1



Probability Density Function (PDF)

Probability Distribution of a continuous random variable e.g. X = height of person measured in inches

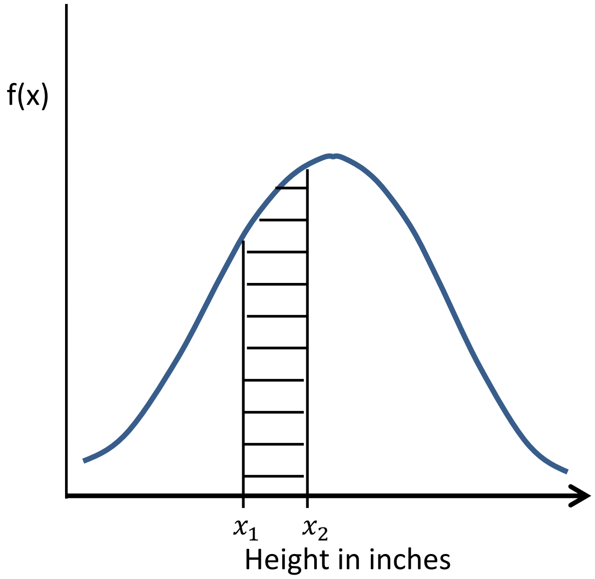
- X is a continuous random variable
- Probability of continuous random variable is always computed for a range not for a single value
- PDF is

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\, dx$$

This calculates the probability as the area under a curve between a range (x_1 to x_2)

Cumulative Distribution Function (CDF)

$$F(X) = P[X \leq x]$$



Important Probability Distributions

Some important probability distributions are Normal Distribution, t distribution, Chi square distribution and the F-distribution

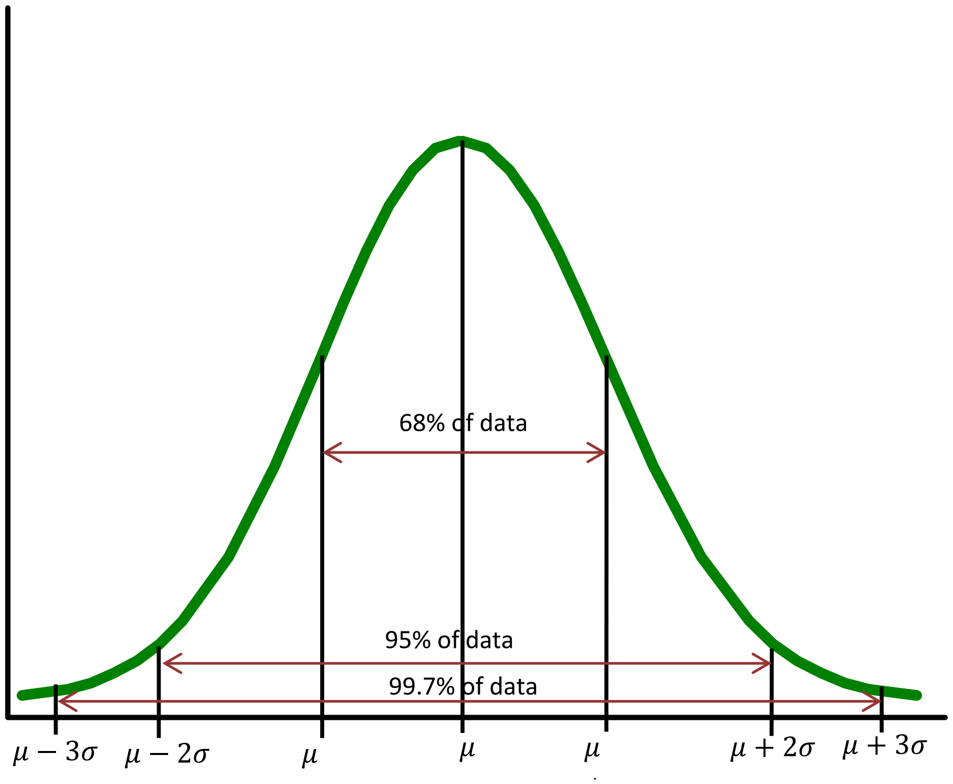
Normal Distribution

It is the most important probability distribution for a continuous random variable. It has a

Bell shaped curve (highest point at mean value) where

$$X \sim N(\mu_x, \boldsymbol{\sigma_x^2}), \; -\infty < X < \infty$$

Change in μ shifts the curve to right or left where change in σ increases of decreases the spread



of the curve. The function may be written as

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}$$

This gives a bell shaped curve with different centers and spreads depending on the values of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$

Mathematical Constants

π=3.14159

e=2.71828

$$p(Z) = \frac{1}{(1)\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{Z-0}{1})^2} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(Z)^2}$$

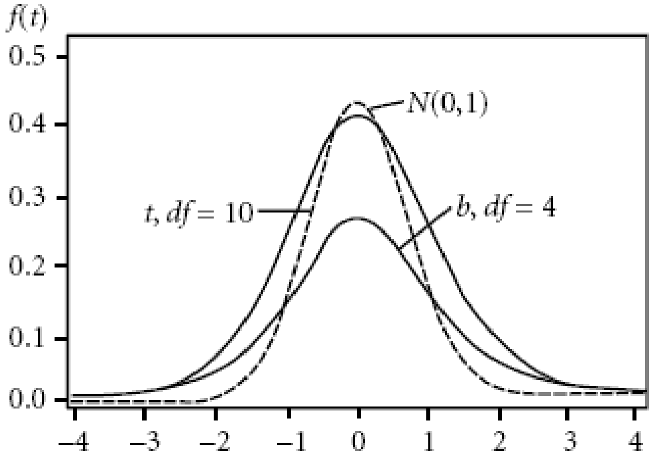
The probabilities or areas under standard normal curve are already calculated and available in the shape of tables (the Z-table)

Standard Normal Distribution

If $\mathrm{Z} = \frac{\mathrm{X}-\bar{\mathrm{X}}}{\sigma}$, then $\mu_{\mathrm{Z}} = 0$ and $\sigma_{\mathrm{Z}} = 1$. The distribution of Z is a 'Standard' normal distribution $\mathrm{Z} \sim \mathrm{N}(0,1)$

Student's t-distribution

t-distribution is a probability distribution of a continuous random variable when the sample size is small and the population variance is not known. Its curve is symmetric and bell shaped but flatter than normal distribution. The mean is zero but the variance is larger (heavier tails) than the variance of standard normal distribution (which is unity). It has only one parameter i.e. the degree of freedom. As the degree of freedom (or the number of observations) increases, the distribution approaches the normal distribution.



Chi-Square (χ^2) distribution

The Chi-square distribution has the following shape.

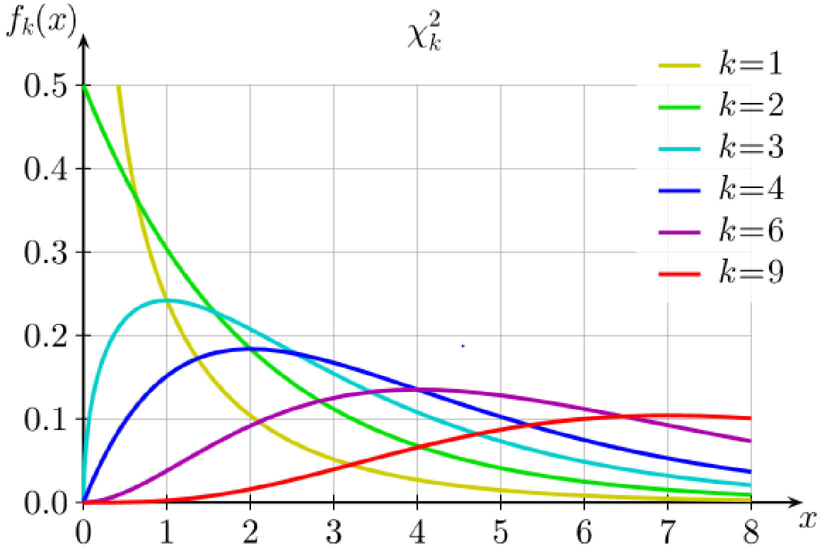
The square of a standard normal variable is distributed as a Chi-square probability distribution with one degree of freedom. Sampling distribution of samples means when the mean and variance is known as the Normal Distribution but when the variance is not known it is the t-distribution. If we need the sampling distribution of the sample variance we have the Chi-square distribution

$$Z^2 = \chi^2{}_{(1)}$$

You can say that Normal distribution and t-distributions are related to means but the Chi-square and F-distributions are related to variances.

Properties:

- Chi-square takes only positive values (zero to infinity)
- It is skewed (depending on the d.f.) unlike the normal distribution
- As the d.f. increases, the distribution approaches the normal distribution
- Its mean is k (=d.f.) and variance is 2k (variance is twice the mean)



F-distribution

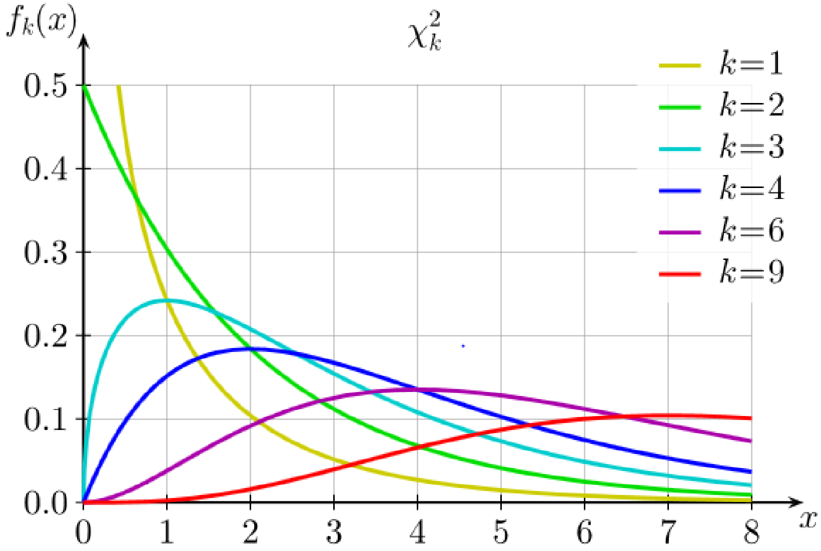
This is a variance ration distribution. (ratio of sample variances of two independent samples.). This is also equal to ratios of two Chi-squares. It has two parameters k_1 and k_2 (degrees of freedom in both samples i.e. numerator and denominator of F= $\frac{S_1^2}{S_2^2}$)

Uses

- Testing equality of variances
- Tests in Regression models like Goodness of fit test

Properties

- Skewed to the right between zero and infinity
- Approaches normal distribution as d.f. increases



Lecture 06
The Simple Regression Model

The Basic Concept

Regression is a statistical measure to determine the (strength of) relationship between a dependent variable (explained variable, response variable or regressand) and a list of independent variables (explanatory variables, predictor variables or regressors). It is a process of estimating the relationship among variables. It looks into the dependence of the regressand on the regressors). It is not the correlation but we want to know how and how much the dependent variable changes in response to changes in the dependent variable(s). We need, sometimes, to predict the values of the dependent variable with the help of the values of the regressors.

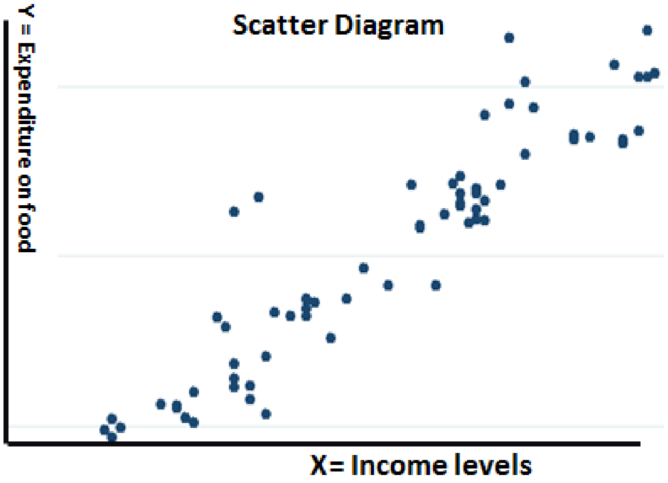
Consider the Demand Function. Demand theory suggests that the quantity demanded depend on various variables like price, income of the consumer, taste, prices of other variables etc. We want to know how the quantity demanded may change due to changes in some of the independent variables like price.

Usually we denote the dependent variable by Y and the regressors as X (or X_1, X_2 etc. in case of multiple regressors). In regression analysis, we try to explain the variable Y in terms of the variable X. Remember that the variable X may not be the only factor effecting Y. Also the relationship may not be exact e.g. for the same Y we may have different X values and for the same X value we may have different Y values. One row shows a pair of X and Y. We handle this by looking on the averages and try to know how the values of the variable Y change in response to changes in the variable X, *on the average*.

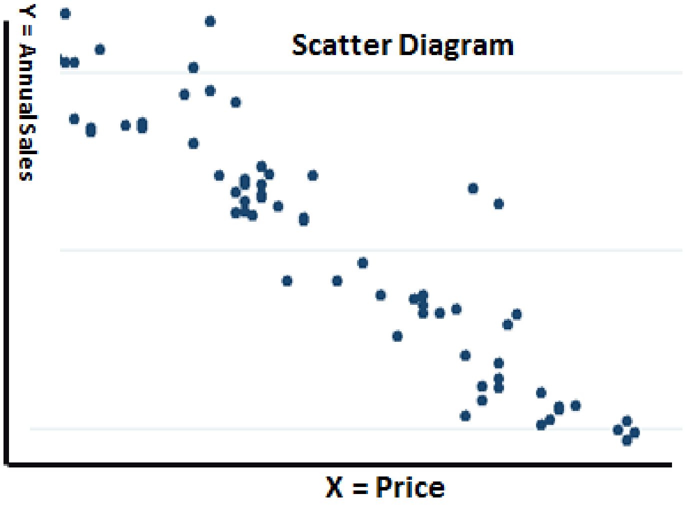
Before performing regression, we also need to have an idea about the nature of the functional relationship of the variable. The relationship may be linear, quadratic, exponential etc. There are many regression models and we select the model that closely approximates the relationship among the variables. We can have an idea about the type of relationship by looking into, what we call, a scatter diagram.

Scatter Diagram

Scatter diagram shows the pairs of actual observations. We usually plot the dependent variable against an explanatory variable to see if we can observe a pattern. If the pattern shows a linear relation, we use a linear regression model.



The above diagram shows that the expenditure on food is a direct (increasing) function of the income levels. The dots showing the plots of the pairs of observation resemble a linear shape (straight line). The points do not lie exactly on a straight line but are scattered around a hypothetical straight line. In the diagram below, the annual sales seem to be inversely related to the price of the commodity. This is because the dots of pairs of observation seem to be scattered around a (hypothetical) straight line that is negatively sloped.



Remember that straight lines are show by equations of the type $Y = a + b\,X$ where a is the y-intercept (the point where the straight line intersects the Y-axis) and b is the slope of the line (the change in the variable Y due to one unit change in the variable X or $\frac{\Delta Y}{\Delta X}$).

In simple regression, we try to estimate the best (explained later) values of a and b by applying appropriate techniques. One of the techniques is called Ordinary Least Square (OLS).

Simple Regression Line by OLS

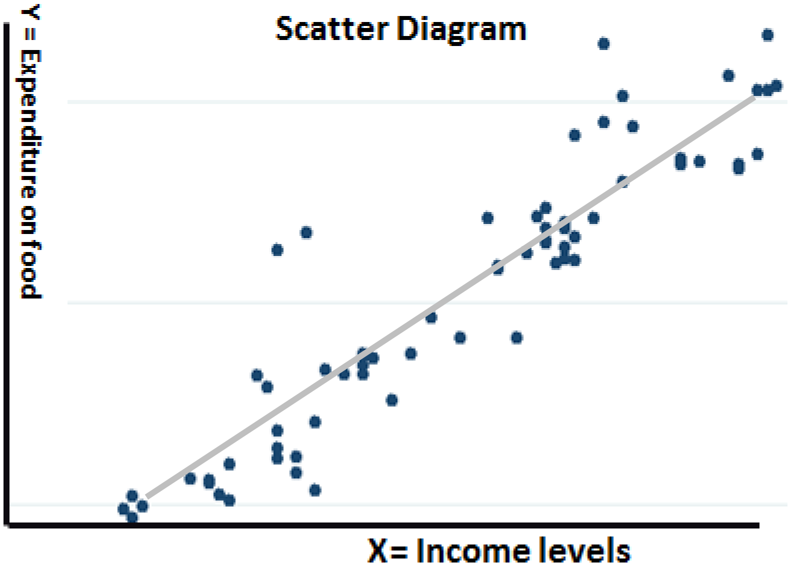
- The relationship seems to be 'linear' that can be captured with the equation of a straight line (Y = a + b X)
- We may need to predict Y if the value of X is given
- We capture the relation by writing a 'simple regression equation'

$Y = a + b\,X + e$ OR $Y = \beta_0 + \beta_1 X + e$

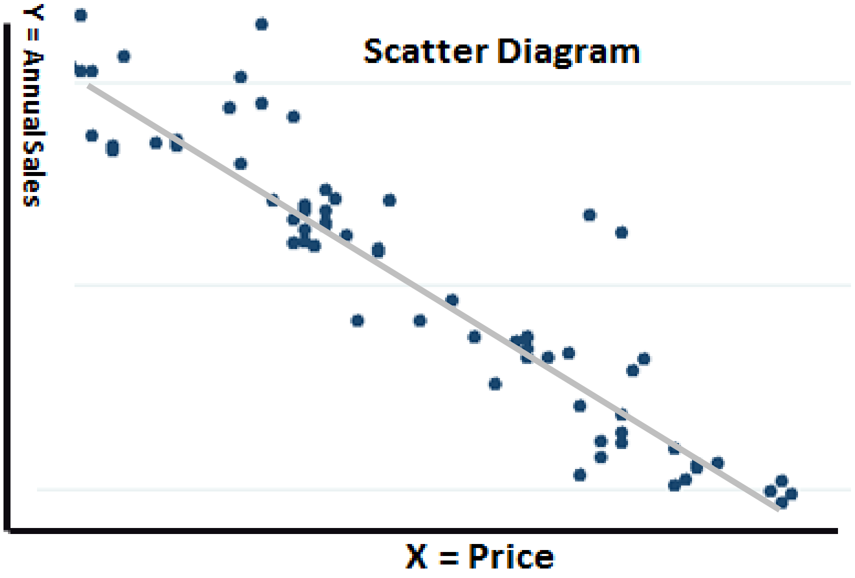
Residual: Note that we have added e which is called an error term or residual. We add this because the actual values do not exactly lie on a straight line but maybe scattered around it. To account for this difference, we capture it in the residual e. When we estimate the parameters 'a' and 'b', they do not provide exact estimates of the value of the dependent variable. The difference is called error term or residual

$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + e_i}$ **(with subscripts)**

Subscript: The subscript i shows that the variable may have multiple observations (as you learnt in summation algebra). $\boldsymbol{\beta_0}$ ***and*** $\boldsymbol{\beta_1}$ are written instead of a and b so that we follow the tradition of regression analvsis.



In both the diagrams above and below, we have imposed a straight line on the scatter diagram to show how the points are scattered around the straight line and if we move along the straight line, we approximate the relation of Y and X. A good technique applied on an appropriate situation may well approximate the relationship (with smaller values of)



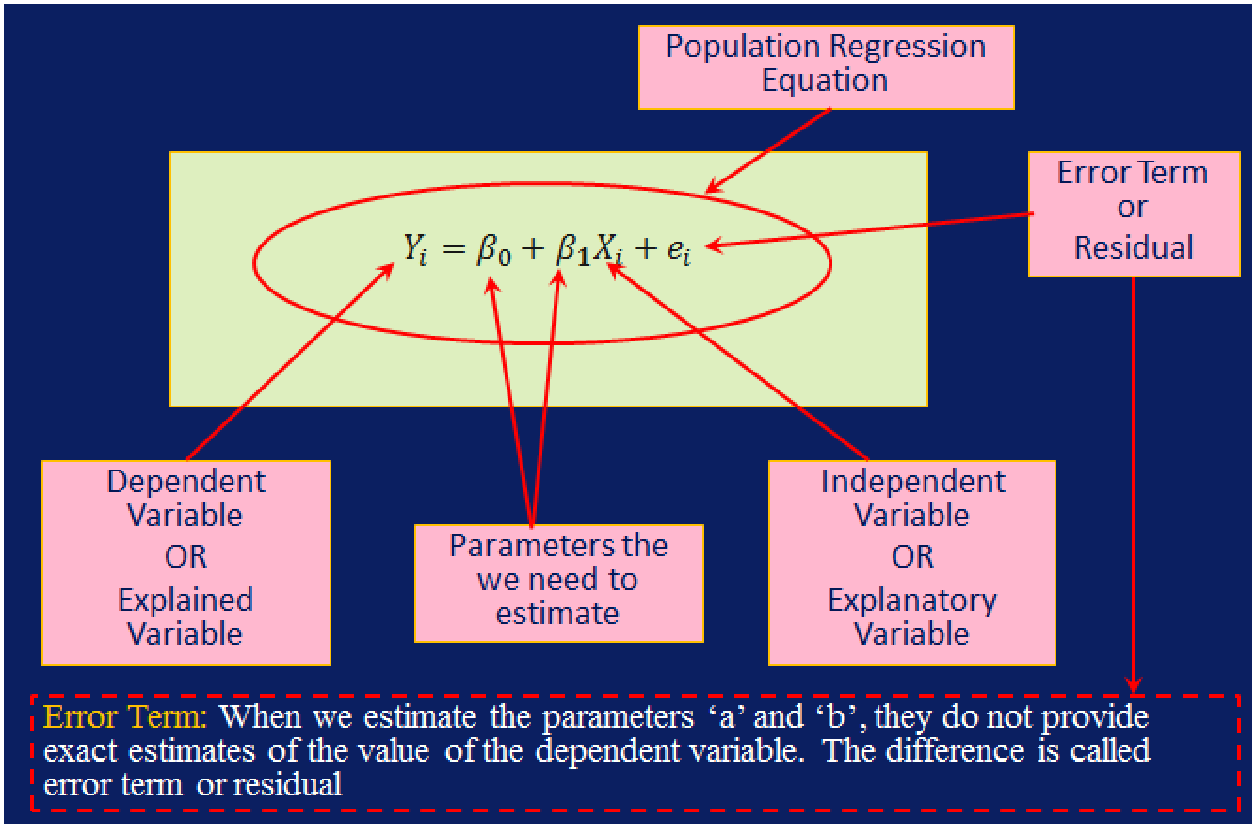
Regression Explained

Population Regression Equation is an assumed equation that may have possibly been estimated from a population. We will use samples to get the values of the parameters $\boldsymbol{\beta_0}$ $\boldsymbol{and}$ $\boldsymbol{\beta_1}$as all the population may not be available or observed.

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + e_i}$$

Here

Y_i = Dependent Variable or Explained Variable. $\boldsymbol{\beta_0}$ $\boldsymbol{and}$ $\boldsymbol{\beta_1}$ are Parameters the we need to estimate. X is the Independent Variable OR Explanatory Variable



Population Regression Equation

$Y_i = \beta_0 + \beta_1 X_i + e_i$

Error Term or Residual

Dependent Variable OR Explained Variable

Parameters the we need to estimate

Independent Variable OR Explanatory Variable

Error Term: When we estimate the parameters 'a' and 'b', they do not provide exact estimates of the value of the dependent variable. The difference is called error term or residual

Regression equation estimation

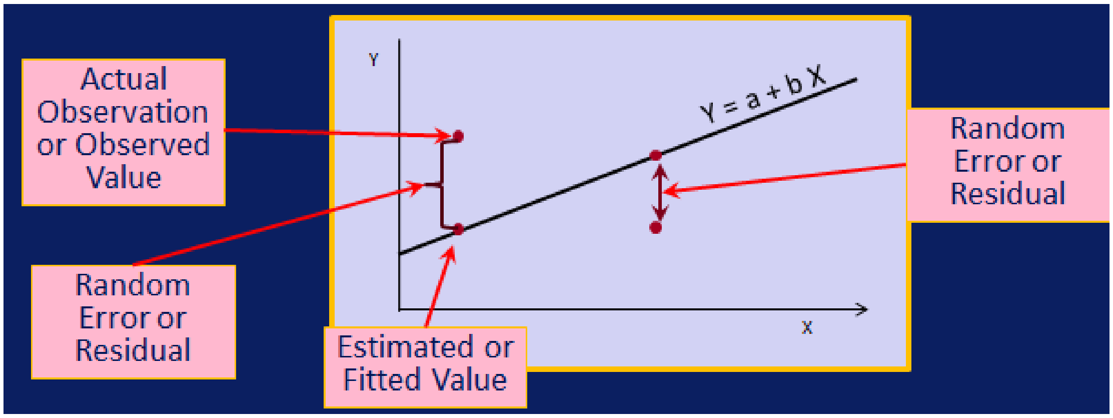
$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + e_i}$is the population regression equation

Let a and b be the estimated values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ respectively

We estimate a and b from a sample. The 'estimated value' of Y based on the estimated regression equation is

$\boldsymbol{\hat{Y} = a + b\,X}$ where $\boldsymbol{\hat{Y}}$ is the estimated value or 'Trend Value'

Then $\boldsymbol{e_i = Y - \hat{Y}}$


Actual Observation or Observed Value
Random Error or Residual
Estimated or Fitted Value
Random Error or Residual
Y
X
Y = a + b X

Regression equation estimation

It is good to have low values of errors (residuals). Negative and Positive Errors cancel each other and we want to 'Magnify' larger errors so we focus on the 'Square of Errors' and try to minimize their sum. In least square estimation, we minimize the 'Sum of Squared Residuals' or 'Sum of Square of Errors' .

We try to estimate the parameters a and b for which we have the minimum possible 'Sum of Square of Residuals'.

Other values of 'a' and 'b' may provide larger SSR.

- *Finding the values of 'a' and 'b' in the regression equation is a minimization problem*

$$Min \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2}$$

NOTE: We will ignore the subscript 'i' for convenience

Remember that

- $e = Y - \hat{Y}$
- $\hat{Y} = a + bX$

Also Remember that

- *For 'Optimization' we take the first derivative and set it equal to zero*

Important: Here although X and Y are variables but for this minimization problem only we will consider 'a' and 'b' to be the unknowns as we are trying to estimate the values of 'a' and 'b'

The above minimization becomes

$$Min\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \hat{Y})^{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \boldsymbol{a} - \boldsymbol{bX})^{\boldsymbol{2}}$$

Where 'a' and 'b' are the unknowns we focus on.

Let Z denote our expression so that we need to minimize

$$Min\, Z = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \boldsymbol{a} - \boldsymbol{bX})^{\boldsymbol{2}}$$

Ignoring subscripts and partially differentiating Z w.r.t. 'a' and setting equal to zero

$$Z_a = \sum (Y - a - bX)^{2-1}. \frac{\partial}{\partial a}(Y - a - bX) = 0$$

Chain rule is used for differentiation

$$\sum (Y - a - bX)(-1) = 0$$

$$\sum (Y - a - bX) = 0$$

$$\sum Y - na - b\sum X = 0$$

Summation Algebra is used when we multiply the Summation symbol

$$\sum Y = na + b\sum X$$

This is called the **first normal equation**

Ignoring subscripts and partially differentiating Z w.r.t. 'b' and setting equal to zero

$$Z_b = \sum (Y - a - bX)^{2-1}. \frac{\partial}{\partial a}(Y - a - bX) = 0$$

Chain rule is used for differentiation

$$\sum (Y - a - bX)(-X) = 0$$

This time 'b' is the unknown and X is the coefficient of 'b'.

The derivative of 'b' is 1 and X will be retained as its coefficient

$$\sum (Y - a - bX)(X) = 0$$

$$\sum(X)(Y-a-bX)=0$$

$$\sum XY\ -a\sum X\ -b\sum X2\ =0$$

Summation Algebra is used when we multiply the Summation symbol

$$\sum XY\ =a\sum X\ +b\ \sum X^2$$

This is called the **second normal equation**

In summary, to estimate a linear regression line $\boldsymbol{Y_i=\ \beta_0+\ \beta_1X_i+e_i}$

Where 'a' is a sample estimate of $\boldsymbol{\beta_0}$ and 'b' is a sample estimate of $\boldsymbol{\beta_1}$,

We minimized the Sum of Squared Residuals

$$Min\,Z=\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{Y-a-bX})^{\boldsymbol{2}}$$

As a result we got two normal equations

$$\sum Y\ =na+b\sum X$$

$$\sum XY\ =a\sum X\ +b\ \sum X^2$$

To find the values of 'a' and 'b' we need some observations of X and Y. We can solve the normal equations and find the values of 'a' and 'b'. Solving these equations gives the values of parameters a and b.

Finding the values of parameters directly

Instead of solving two normal equations, we can derive expressions to directly find the values of a and b.

$$\sum Y\ =na+b\sum X\qquad\qquad \dots\dots\dots\dots\dots(1)$$

$$\sum XY\ =a\sum X\ +b\ \sum X^2\qquad\quad \dots\dots\dots(2)$$

Dividing equation (1) by n,

$$\frac{\sum Y}{n}=a+b\ \frac{\sum X}{n}\ which\ gives\ \ a=\frac{\sum Y}{n}-b\ \frac{\sum X}{n}$$

We can also write this as

$$a = \bar{Y} - b\bar{X}$$

Substituting this value of ‘a’ in equation (2)

$$\sum XY = \left(\frac{\sum Y}{n} - b\,\frac{\sum X}{n}\right)\sum X + b\sum X^2$$

$$\sum XY = \frac{\sum Y \sum X}{n} - b\,\frac{\sum X \sum X}{n} + b\sum X^2$$

$$\sum XY - \frac{\sum Y \sum X}{n} = b(\sum X^2 - \frac{\sum X \sum X}{n})$$

$$b = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$b = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

& $a = \bar{Y} - b\bar{X}$

Example:

Consider the following example where X = Income in thousand rupees and Y = expenditure on food items (thousand rupees)

Observation #	**X**	**Y**	**XY**	**X^2**
1	25	20	500	625
2	30	24	720	900
3	35	32	1120	1225
4	40	33	1320	1600
5	45	36	1620	2025
Totals	**175**	**145**	**5280**	**6375**
	$\sum X$	$\sum Y$	$\sum XY$	$\sum X^2$

The normal equations are

$$\sum Y = na + b\sum X \quad \& \quad \sum XY = a\sum X + b\sum X^2$$

Substituting values in the normal equations gives us:

$$145 = 5a + b\,(175) \quad \&$$

$$5280 = \mathrm{a}(175) + \mathrm{b}(6375)$$

OR

$$145 = 5a + 175\,b$$

$$5280 = 175\ \mathrm{a} + 6375\ \mathrm{b}$$

Solving them simultaneously gives us:

$\mathrm{a} =$ 0.3 and $\mathrm{b} =$ 0.82

We can write the regression line as

$$Y = 0.3 + 0.82\,X$$

Interpretation

The value of a is the Y-intercept and the value of b is the slope of the line (rate of change of Y w.r.t. X or derivative of Y w.r.t. X)

Alternative Method

The values of a and b can also be found by substituting in any one of the expressions that we derived.

$$b = \frac{n\sum XY\ - \sum X\ \sum Y}{n\sum X^2 - \left(\sum X\right)^{\ 2}}$$

$$b = \frac{5(5280) - (175)(145)}{5(6375) - (175)^2}$$

$$b = 0.82$$

$$a = \ \bar{Y} - b\bar{X}$$

$$a = \left(\frac{145}{5}\right) - 0.82\ \left(\frac{175}{5}\right)$$

$a = 0.3$

‘b’ is called the slope coefficient.

b = 0.82 means that a one unit change in X (income level) brings 0.82 unit changes in Y (expenditure of food), on the average.

OR Change if 1000 rupees (one unit is in thousands) increase in income may increase the expenditure on food items by 820 rupees.

Trend Values and Errors:

We can substitute the values of X in the estimated regression equation and find Trend Values

Observation \#	X	Y	$Tend\ Value$ $\hat{Y}$	Residual or Error $e = Y - \hat{Y}$	Square of Residuals e^2
1	25	20	20.8	-0.8	0.64
2	30	24	24.9	-0.9	0.81
3	35	32	29	3.0	9
4	40	33	33.1	-0.1	0.01
5	45	36	37.2	-1.2	1.44
Totals	**175**	**145**	**145**	**Zero**	**11.9**
	$\sum X$	$\sum Y$	$\sum \hat{Y} = \sum Y$	$\sum e$	$\sum e^2$

The First trend value is computed as Y = 0.3 + 0.82 (25) = 20.8 and so on. If you change the values of a and b and compute new squares of errors, the new value would be larger than the value here (Least Square of errors)

The Error Term

We assume that error are normally distributed with zero mean and constant variance

$$e \sim N(0, \sigma^2)$$

As you must have noticed while estimating regression parameters,

$$\sum_{i=1}^{n} \boldsymbol{e_i = 0}$$

Also, you can verify easily that

$$\sum_{i=1}^{n} \boldsymbol{e_i X_i = 0}$$

And as we got the regression equation by minimization process,

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2} \ is\ minimum$$

Nature of the Error Term

- Error Term may represent the influence the variables NOT included in the model. (Missing Variables)
- Even if we are able to include all variables or determinants of the dependent variable, there will remain randomness in the error as human behavior is not rational and predictable to the extent of 100%
- e may represent 'Measurement Error'; When data is collected we may round some values or observe values in ranges or some variables are not accurately measured

Assumptions of OLS estimators

Gauss-Markov assumptions

1. Linear in Parameters
2. Random Sampling of n observations
3. Sample variation in explanatory variable (X_i). are not all the same value
4. Zero conditional mean: The error e has an expected value of 0, given any values of the explanatory variable
5. Homoskedasticity: The error has the same variance given any value (in subsets) of the explanatory variable.

BLUE: Best Linear Unbiased Estimators

Under the Gauss-Markov Assumptions the OLS estimators are Best, Linear and Unbiased in the Model $Y = \beta_0 + \beta_1 X + e$ where a and b are sample estimates of $\beta_0 and\ \beta_1$ respectively.

Linear: The model is linear in parameters. However variables can have powers not equal to one.

- $Y = a + b X$ is linear but $Y = a + b^2 X$ is not
- $Y = a + b X + c X^2$ is fine

- Y = a + ln(bX) is not OLS
- Y = a + b ln(X) is OLS

In fact 'linear' means that we can express the slope coefficient as a linear function of Y

Unbiased: A parameter is unbiased if the average value of the estimator in repeated samples is equal to the true population parameter.

In our case $E\left(b_j\right) = \beta_{1j}$

Best / Efficient: A parameter is best if its variance is less than any other estimator of the parameter

$$Var\ (b) \leq Var\ \left(\tilde{b}\right)$$

$$where\ \tilde{b} \text{ is any other unbiased estimatro of } \beta_1$$

We will learn later how to compute Var (b)

Exercise

1. Prove that the above expression for 'b' can also be written as

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$

2. Estimate a linear regression Y on X from the following data. Find the trend values and compute the errors

   X = 1, 5, 6, 9 , 9, 10, 9, 11, 10, 12

   Y= 45, 42, 41, 37, 36, 31, 33, 36, 29, 27

3. For the data and results of question 2, See if the following is true

$$\boldsymbol{\sum_{i=1}^{n} e_i = 0\ and\ \sum_{i=1}^{n} e_i X_i = 0}$$

Lecture 07
Estimation and Testing in Regression Analysis

Example:

Consider the following example where X = Income in thousand rupees and Y = expenditure on food items (thousand rupees)

Sr. #	X	Y	XY	X^2	$\boldsymbol{Tend\ Value}$ $\hat{Y}$	Error $e = Y - \hat{Y}$	Square of Residuals e^2
1	25	20	500	625	20.8	-0.8	0.64
2	30	24	720	900	24.9	-0.9	0.81
3	35	32	1120	1225	29	3.0	9
4	40	33	1320	1600	33.1	-0.1	0.01
5	45	36	1620	2025	37.2	-1.2	1.44
Totals	**175**	**145**	**5280**	**6375**	**145**	**Zero**	**11.9**
	$\sum X$	$\sum Y$	$\sum XY$	$\sum X^2$	$\sum \hat{Y} = \sum Y$	$\sum e = 0$	$\sum e^2$

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{5(5280) - (175)(145)}{5(6375) - (175)^2} = 0.82$$

$$a = \bar{Y} - b\bar{X} = \left(\frac{145}{5}\right) - 0.82\left(\frac{175}{5}\right) = 0.3$$

Interpretation of Regression Coefficients

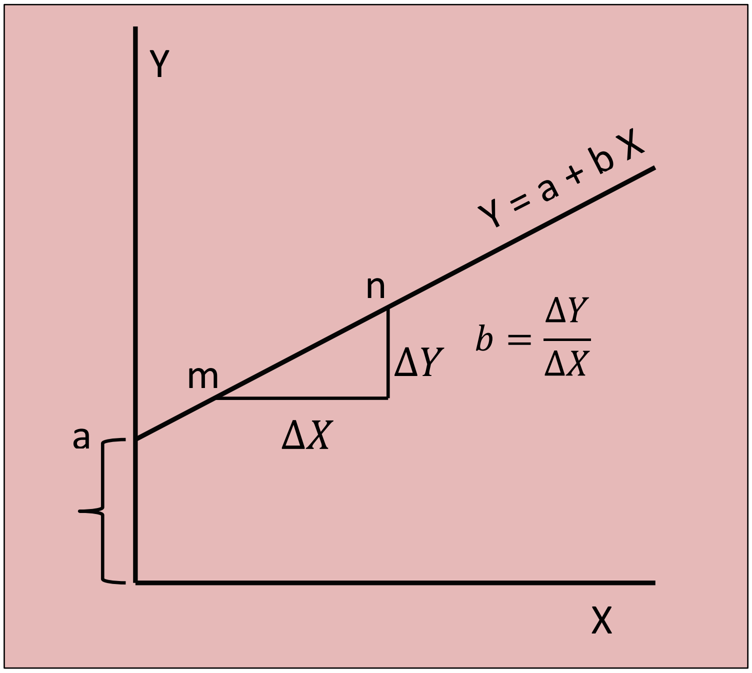
The Intercept 'a' (usually also denoted by β_0)

This is the y-intercept of the straight line and indicates the value of Y when X is zero; Usually the base or the initial value

Examples: (Assuming linear relationships)

- Autonomous Consumption in the Keynesian Consumption function

- Autonomous Investment in the Investment function



The Slope Coefficient 'b' (Usually denoted also by $\boldsymbol{\beta_1}$)

In linear regression lines, it shows the average unit change in the dependent variable due to one unit change in the independent variable. In a linear equation, It can also be called the derivative of the dependent variable w.r.t. the independent variable.

The slope coefficient b is an unbiased estimate of the population regression coefficient β_1

Standard Error of Estimate/Standard Error of Regression

The standard error of the estimate is a measure of the accuracy of predictions.

It is the standard deviation of errors and defined as

$$\boldsymbol{\hat{\sigma}_e = \sqrt{\frac{\sum e_i^2}{N-k}}}$$

Where N = Number of observations

k = number of restrictions imposed which is equal to number of parameters

$$N - k = degree\ of\ freedom$$

Standard Error of ‘b’ (for regression with one independent variable)

$$se(b) = \frac{\boldsymbol{\hat{\sigma}_e}}{\sqrt{\sum(X-\bar{X})^2}}$$

Or

$$se(b) = \frac{\boldsymbol{\hat{\sigma}_e}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}}$$

In the current example

$$\boldsymbol{\hat{\sigma}_e = \sqrt{\frac{\sum e_i^2}{N-k}}}$$

$$\boldsymbol{\hat{\sigma}_e = \sqrt{\frac{11.9}{5-2}} = \sqrt{\frac{11.9}{3}}}$$

$$\boldsymbol{\hat{\sigma}_e = 1.9916}$$

Also

$$se(b) = \frac{\boldsymbol{\hat{\sigma}_e}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}}$$

$$= \frac{1.9916}{\sqrt{6375 - \frac{(175)^2}{5}}} = \frac{1.9916}{15.81139}$$

$$= 0.12596$$

Testing for the significance of slope coefficient: Individual variable significance test

The Procedure is as follows:

$$\text{H0:} \quad \text{b} \ = \ 0$$

$$H1\text{:} \quad \text{b} \ \neq 0$$

$$\alpha = 0.05 \ \ (or\ 0.01)$$

Level of significance: Probability of type I error (rejecting a true hypothesis

$$Test\ statistic$$

$$t = \frac{b}{se(b)}$$

If you do not fully understand all this then you need to revise the topic 'Test of Hypothesis' in your basic statistics course.

$$Region\ of\ Rejection$$

$$|t| > t_{\frac{\alpha}{2},n-k}$$

As this is a two tailed test so we search for the value if t in the table of t-distribution corresponding to $\alpha/2$ and n-k

In this example

$$t = \frac{b}{se(b)} = \frac{0.82}{0.12596} = 6.51$$

Looking into the table of t-distribution

$$t_{\frac{\alpha}{2},n-k} = t_{0.025,3} = 3.182$$

As $|t| > t_{\frac{\alpha}{2},n-k}$

$$6.51 > 3.182$$

So we reject H_0 and conclude that 'b' is significant and the variable X has a significant impact on the variable Y

Reading the value of t-distribution

You can download statistical tables from the internet

e.g. http://wps.aw.com/wps/media/objects/15/15512/stat_tables.pdf

To read the value of $t_{0.025,3}$ look in the column below $t_{0.025}$ (for 2 tailed test 0.5 is divided in two parts so we look at 0.025 or 2.5%) and 3 degrees of freedom. You get the value 3.182. You may find slightly different tables so read the instruction provided with the table.

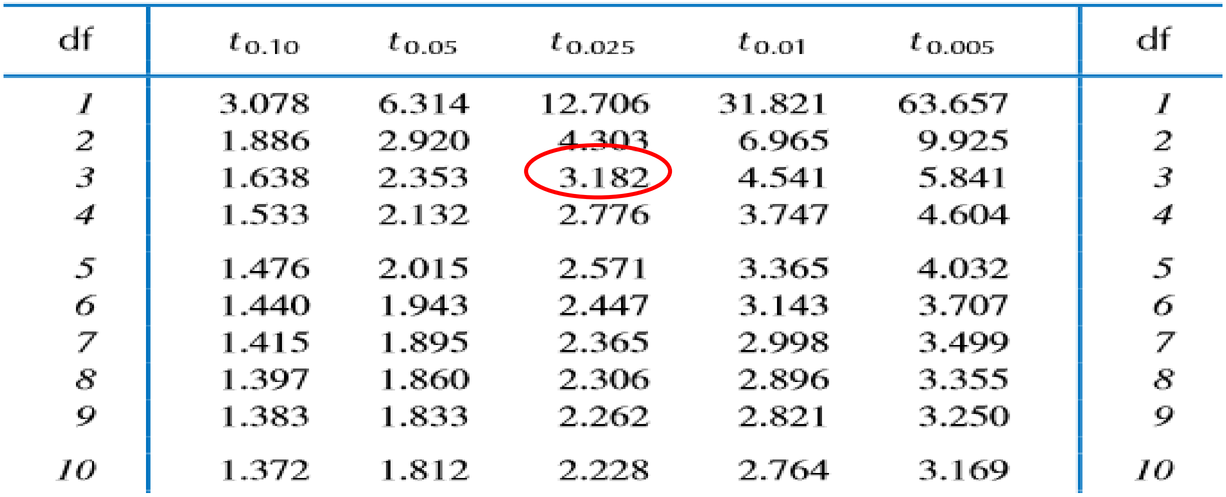
Reading the value of $t_{0.025,3}$ in Microsoft Excel (2007 or later)

Type in any cell and press ENTER

= T.INV.2T(0.05,3)

The cell will display the value 3.182

This formula has two parameters. The first one is the level of significance and the second is the d.f. (N – K)



df	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$	df
1	3.078	6.314	12.706	31.821	63.657	*1*
2	1.886	2.920	4.303	6.965	9.925	*2*
3	1.638	2.353	3.182	4.541	5.841	*3*
4	1.533	2.132	2.776	3.747	4.604	*4*
5	1.476	2.015	2.571	3.365	4.032	*5*
6	1.440	1.943	2.447	3.143	3.707	*6*
7	1.415	1.895	2.365	2.998	3.499	*7*
8	1.397	1.860	2.306	2.896	3.355	*8*
9	1.383	1.833	2.262	2.821	3.250	*9*
10	1.372	1.812	2.228	2.764	3.169	*10*

The Coefficient of Determination: Explanatory power of the model

The proportion of variation in Y that is explained by X

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}$$

Where

Total Variation

=Explained Variation +Unexplained variation

So the above can be written as

$$R^2 = \frac{Total\ Variation\ - Unexplained\ Variation}{Total\ Variation}$$

$$= 1 - \frac{Unexplained\ Variation}{Total\ Variation} = 1 - \frac{\sum \boldsymbol{e^2}}{\sum(Y-\bar{Y})^2}$$

Which also can be transformed as

$$R^2 = 1 - \frac{\boldsymbol{N\sum e^2}}{N\sum Y^2 - (\sum Y)^2}$$

In our current example

$$R^2 = 1 - \frac{\boldsymbol{N\sum e^2}}{N\sum Y^2 - (\sum Y)^2}$$

$$R^2 = 1 - \frac{5(11.9)}{5(4385) - (145)^2}$$

$$R^2 = 0.934$$

Interpretation

The explanatory power of the model is 93.4%

Or

With the simple regression model, The variations in X can explain 93.4% of variation in Y

Note: The coefficient of determination is equal to the square of the correlation coefficient only in case of the simple regression line with one independent variable.

The Goodness of Fit Test: Using the F-distribution

The Procedure is as follows:

$$H0: \quad The\ Fit\ is\ not\ good$$

$$H0: \quad The\ fit\ is\ good$$

$$\alpha = 0.05 \ \ (or\ 0.01)$$

$$Test\ statistic$$

$$F = \frac{R^2}{(1-R^2)}\frac{N-k}{k-1}$$

Where

$$R^2 = 1 - \frac{\boldsymbol{\sum e^2}}{\sum(Y-\bar{Y})^2} = 1 - \frac{\boldsymbol{N\sum e^2}}{N\sum Y^2 - (\sum Y)^2}$$

Note: the Coefficient of Determination R^2 is equal to the square of the correlation coefficient only in case of simple regression line (one independent variable)

$$Region\ of\ Rejection$$

$$F > F_{\alpha,\, k-1, N-k}$$

In our current example

$$F = \frac{R^2}{(1-R^2)}\frac{N-k}{k-1}$$

$$= \frac{0.934}{1-0.934}\left(\frac{3}{1}\right) = 42.4$$

Looking into the table of F-distribution

$$F_{\alpha,\, k-1, N-k} = F_{0.05,\, 1,3} = 10.13$$

As

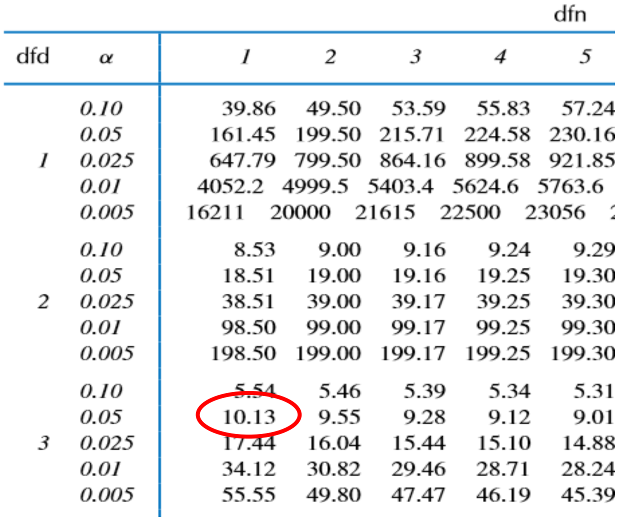
$$F > F_{\alpha,\, k-1, N-k}$$

$$42.4 > 10.13$$

So we reject H_0 and conclude that The Fit is Good .

Reading the value of F-distribution

See table VI from http://wps.aw.com/wps/media/objects/15/15512/stat_tables.pdf



						dfn
dfd	α	1	2	3	4	5
	0.10	39.86	49.50	53.59	55.83	57.24
	0.05	161.45	199.50	215.71	224.58	230.16
1	0.025	647.79	799.50	864.16	899.58	921.85
	0.01	4052.2	4999.5	5403.4	5624.6	5763.6
	0.005	16211	20000	21615	22500	23056
	0.10	8.53	9.00	9.16	9.24	9.29
	0.05	18.51	19.00	19.16	19.25	19.30
2	0.025	38.51	39.00	39.17	39.25	39.30
	0.01	98.50	99.00	99.17	99.25	99.30
	0.005	198.50	199.00	199.17	199.25	199.30
	0.10	5.54	5.46	5.39	5.34	5.31
	0.05	10.13	9.55	9.28	9.12	9.01
3	0.025	17.44	16.04	15.44	15.10	14.88
	0.01	34.12	30.82	29.46	28.71	28.24
	0.005	55.55	49.80	47.47	46.19	45.39

To read the value of $F_{0.05,\, 1,3}$look in the column below 1 (k-1) and 3 (N-k) and look for the value for 0.05 (α). The value is 10.13.

Reading the value of$F_{0.05,\, 1,3}$in Microsoft Excel (2007 or later)

Type in any cell and press ENTER

= F.INV.RT(0.05,1,3)   The cell will display the value 10.13

This formula has three parameters. The level of significance, k-1 & N-k

Using basic Microsoft Excel Formulas: intercept and slope

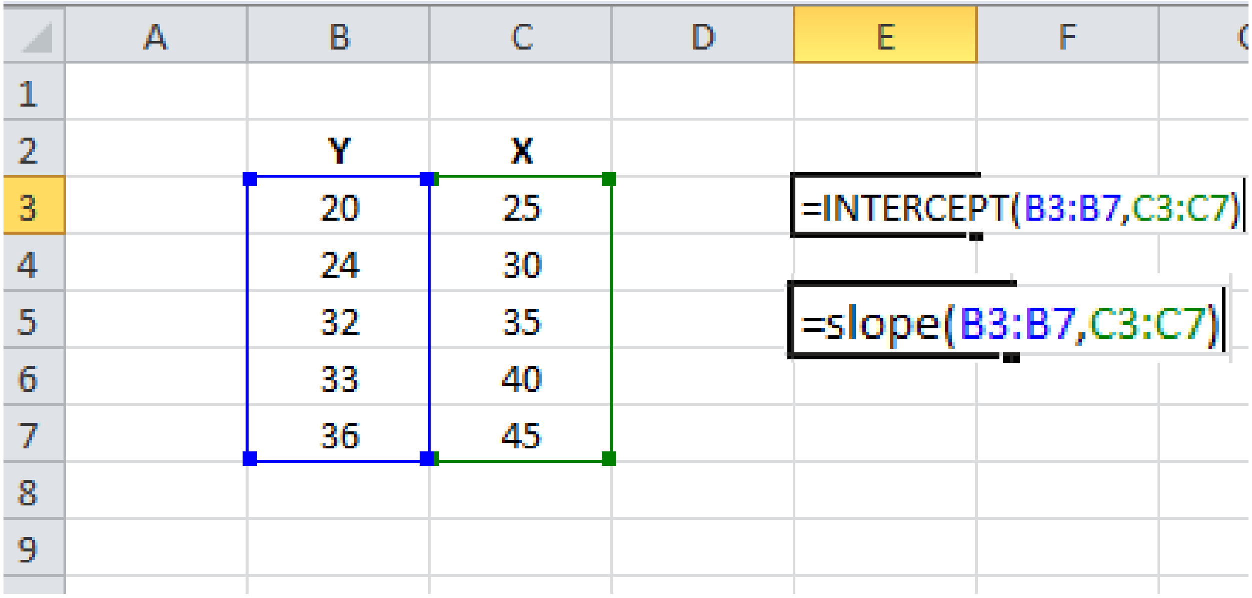
We will use basic formulas here. An additional tool called DATA ANALYSIS tool pack will be discussed later. Intercept and slope: To Estimate Y = a + b X + e, We can find the values of the parameter 'a' (intercept) and parameter 'b' (slope) in Microsoft Excel.

Value of 'a'

= intercept (Cell range of Values of Y, Cell range of Values of X)

Value of 'b'

= slope (Cell range of Values of Y, Cell range of Values of X)



	A	B	C	D	E	F
1						
2		Y	X			
3		20	25		=INTERCEPT(B3:B7,C3:C7)	
4		24	30			
5		32	35		=slope(B3:B7,C3:C7)	
6		33	40			
7		36	45			
8						
9						

The cells are showing the formulas. When you will press ENTER after writing these formulas, the values will be displayed as 0.3 and 0.82

Using basic Microsoft Excel Formulas: the LINEST formula

Remember that We will use basic formulas here. An additional tool called DATA ANALYSIS tool pack will be discussed later. **Syntax: LINEST(known_y's, [known_x's], [const], [stats])**

Steps:

- This is an array command so we select the cells D3 to E7 (2 columns and five rows to display results)
- When the cells are selected then type = LINEST(B3:B7,C3:C7,TRUE,TRUE)
- Press and hold Ctrl + SHIFT and press ENTER (Ctrl + SHIFT + ENTER)

The results are displayed in the cells that you had selected like this

value of 'b'	value of 'a'
standard Error of 'b'	standard error of 'a'
R-squared	standard error of estimate
F-statistic	N-k
Regression Sum of Squares	Residual SS = Sum of square of errors

This example is for simple regression line. For details see Microsoft Help on LINEST

Understanding the LINEST formula

LINEST(known_y's, [known_x's], [const], [stats])

The range of Y values B3:B7

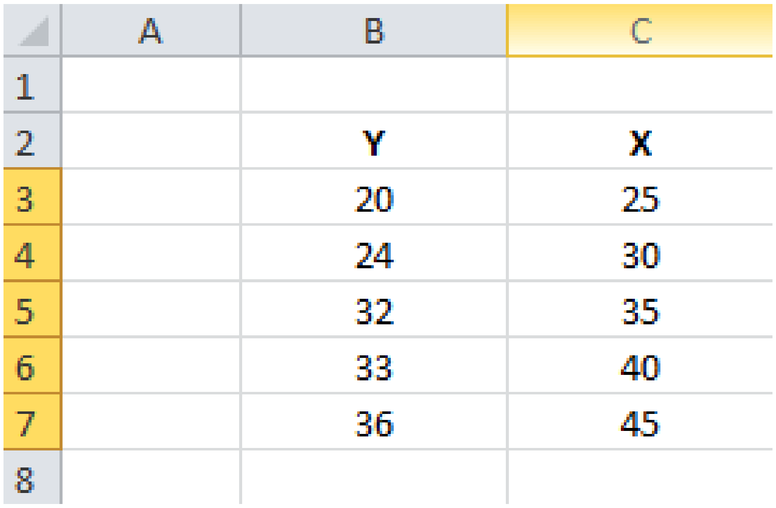
The range of X values C3:C7

If TRUE then the constant intercept ‘a’ is calculated normally. If FALSE then ‘a’ is set equal to zero

If TRUE then additional statistics are displayed in an array (D3 to E7)

STEP 1

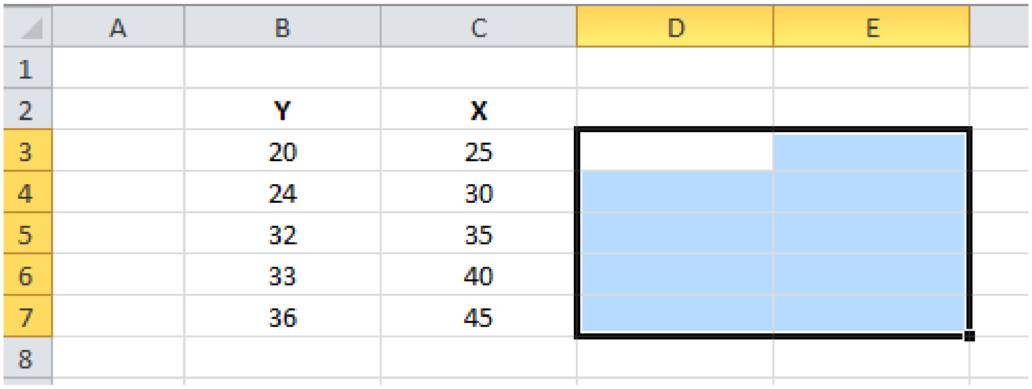
First Enter your data in cells B3 to C7 (you can add title Y and X in cells B2 and C2 respectively



	A	B	C
1			
2		**Y**	**X**
3		20	25
4		24	30
5		32	35
6		33	40
7		36	45
8			

STEP 2

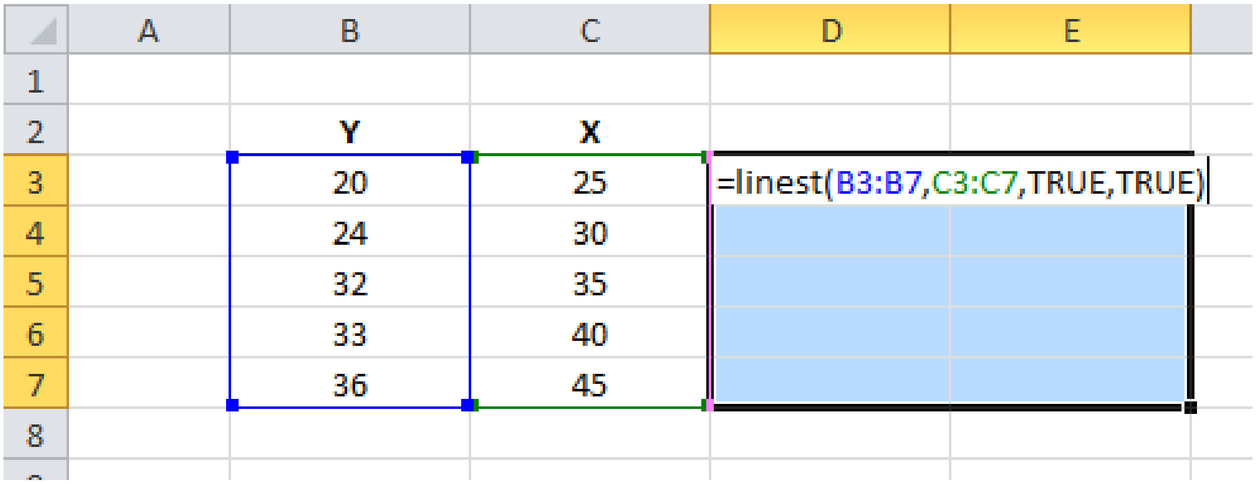
Now, whatever the number N, select any two columns and five rows to display the result in an array (10 different statistics will be displayed)



	A	B	C	D	E
1					
2		**Y**	**X**		
3		20	25		
4		24	30		
5		32	35		
6		33	40		
7		36	45		
8					

STEP 3

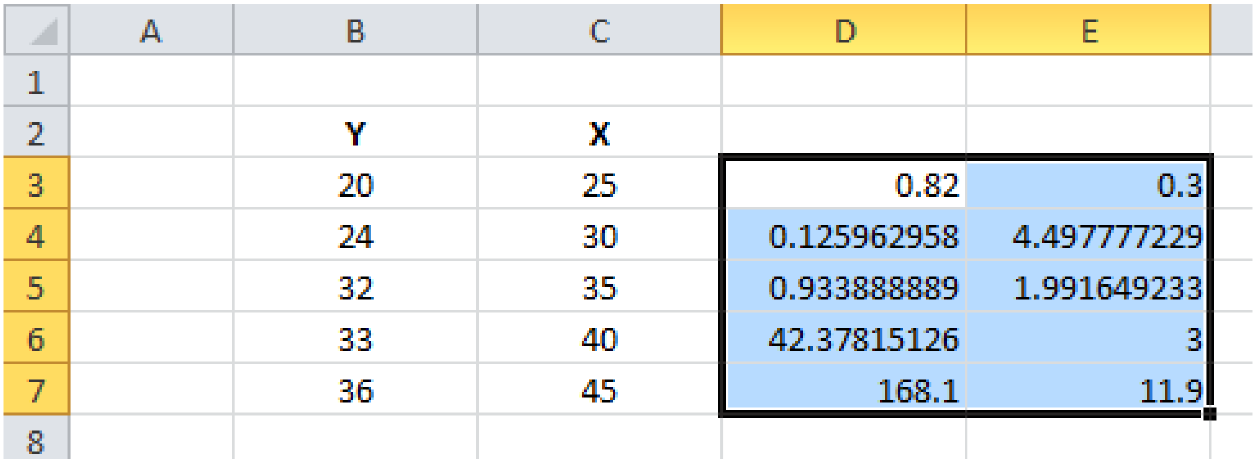
While the cells remain selected, type the following then hold the Ctrl and SHIFT keys and press ENTER. We do this so that the result is displayed in the selected array.



	A	B	C	D	E
1					
2		**Y**	**X**		
3		20	25	=linest(B3:B7,C3:C7,TRUE,TRUE)	
4		24	30		
5		32	35		
6		33	40		
7		36	45		
8					

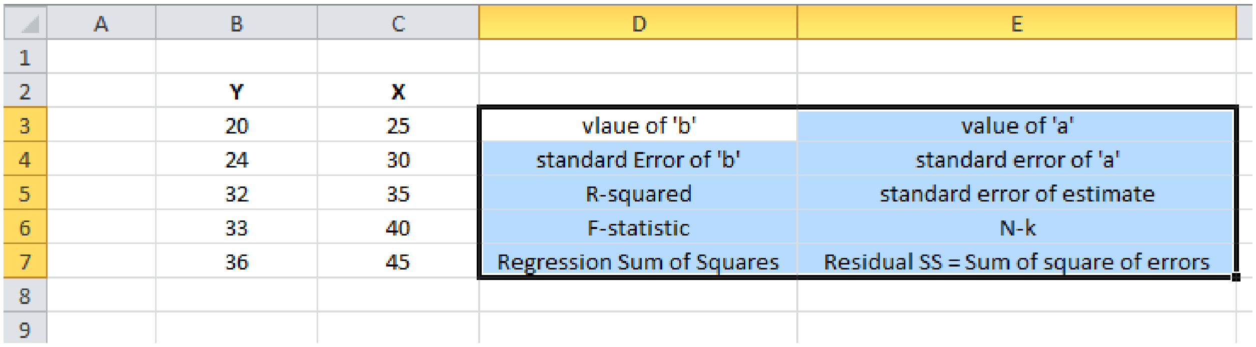
Results

Hold Ctrl and SHIFT keys and press ENTER, the results will be as follows



	A	B	C	D	E
1					
2		**Y**	**X**		
3		20	25	0.82	0.3
4		24	30	0.125962958	4.497777229
5		32	35	0.933888889	1.991649233
6		33	40	42.37815126	3
7		36	45	168.1	11.9
8					

The results may be read in the following order



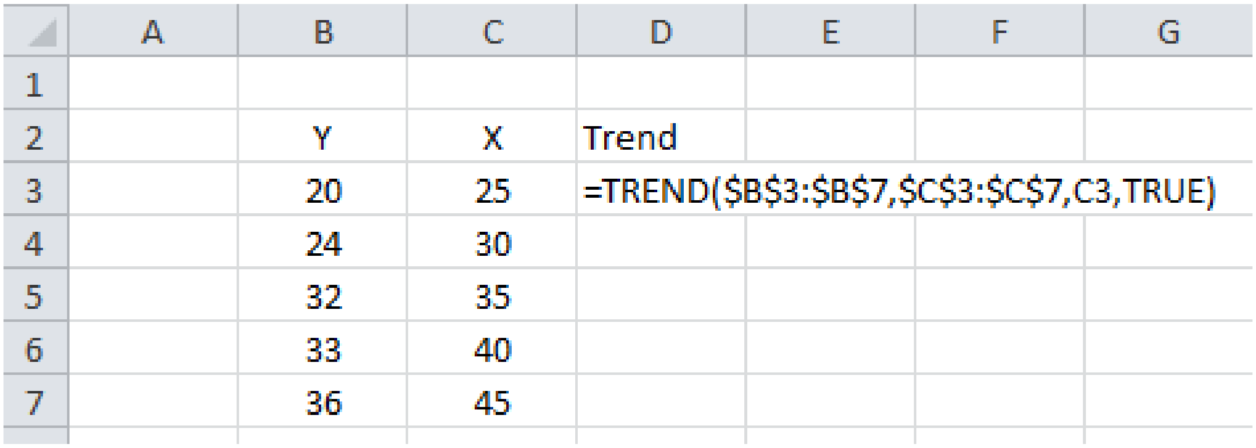
	A	B	C	D	E
1					
2		**Y**	**X**		
3		20	25	vlaue of 'b'	value of 'a'
4		24	30	standard Error of 'b'	standard error of 'a'
5		32	35	R-squared	standard error of estimate
6		33	40	F-statistic	N-k
7		36	45	Regression Sum of Squares	Residual SS = Sum of square of errors
8					
9					

For example a = 0.3, b = 0.82, F = 42.37815126 etc.

Note: The value of t could be calculated

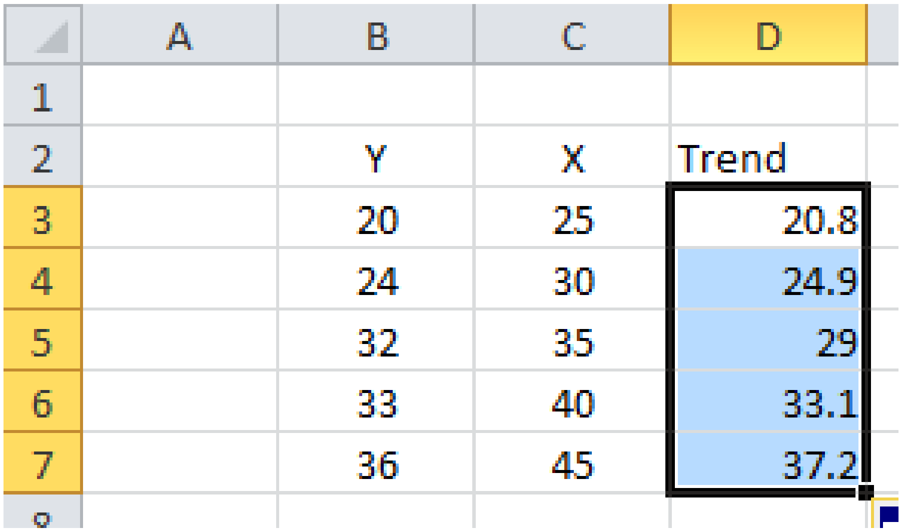
Understanding the TREND formula

You may compute all the trend values without calculating the regression coefficients by using this formula



	A	B	C	D	E	F	G
1							
2		Y	X	Trend			
3		20	25	=TREND($B$3:$B$7,$C$3:$C$7,C3,TRUE)			
4		24	30				
5		32	35				
6		33	40				
7		36	45				

Note the $ sings; they are inserted around the letter so that the reference does not change when you extend the formula to other cells by dragging



	A	B	C	D
1				
2		Y	X	Trend
3		20	25	20.8
4		24	30	24.9
5		32	35	29
6		33	40	33.1
7		36	45	37.2

Drag the formula to get other trend values

Lecture 08

Examples Using Microsoft Excel

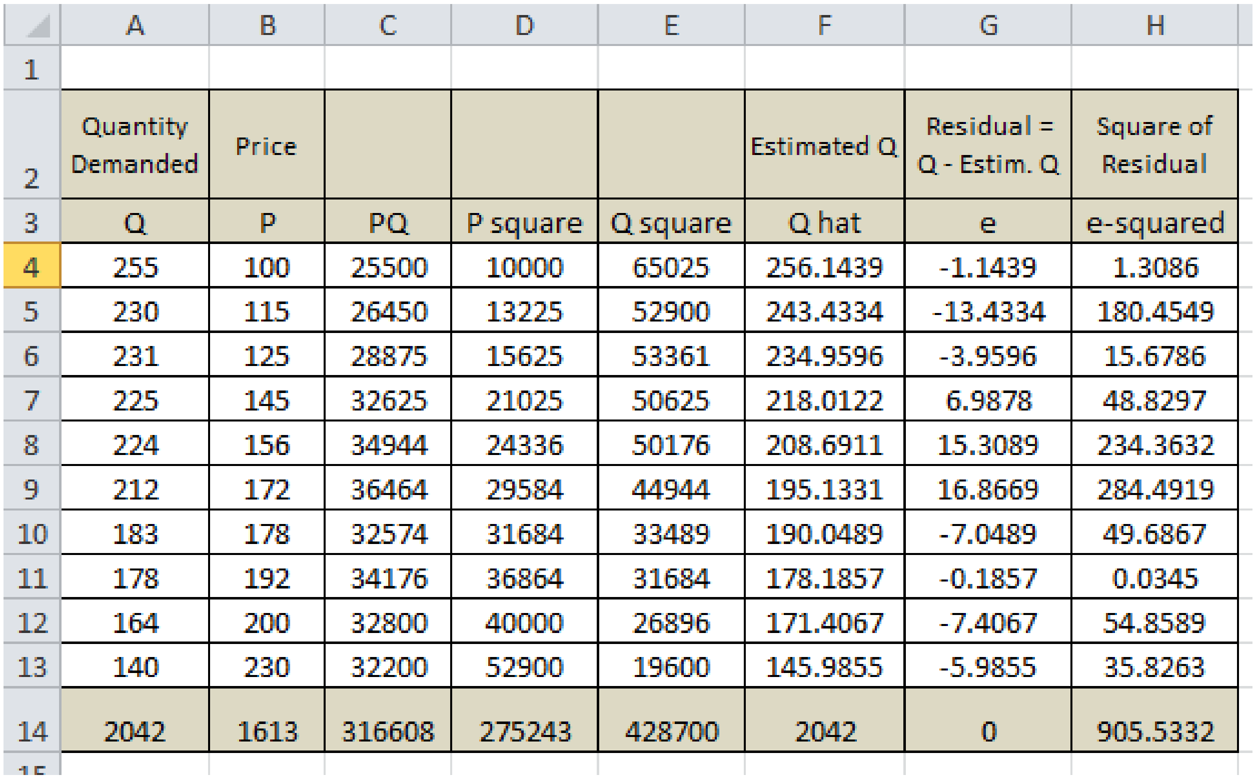
This is only to demonstrate the procedure of OLS. In Future you would be using the Data Analysis Tool to perform regressions. Prerequisite: The student must be familiar with Microsoft Excel.

Model Exam Question

Consider the following data. You may open Microsoft Excel and enter the Data and create the required sums

- Calculate the values of parameters ‘a’ and ‘b’ for the regression Y=a + b X + e using the formulas you have learnt.
- Calculate the standard error of estimate and standard error of b.
- Calculate t-values and draw your conclusions regarding the individual variable significance
- Calculate the Coefficient of determination
- Calculate the F-statistic and draw your conclusion regarding the goodness of fit
- Use the LINEST Excel formula and verify all your results

NOTE: Interpret your results at each step, write formulas using equation editor (you can write formulas / Sums in Microsoft Word and paste in Excel sheet if you feel easy like that)



	A	B	C	D	E	F	G	H
1								
2	Quantity Demanded	Price				Estimated Q	Residual = Q - Estim. Q	Square of Residual
3	Q	P	PQ	P square	Q square	Q hat	e	e-squared
4	255	100	25500	10000	65025	256.1439	-1.1439	1.3086
5	230	115	26450	13225	52900	243.4334	-13.4334	180.4549
6	231	125	28875	15625	53361	234.9596	-3.9596	15.6786
7	225	145	32625	21025	50625	218.0122	6.9878	48.8297
8	224	156	34944	24336	50176	208.6911	15.3089	234.3632
9	212	172	36464	29584	44944	195.1331	16.8669	284.4919
10	183	178	32574	31684	33489	190.0489	-7.0489	49.6867
11	178	192	34176	36864	31684	178.1857	-0.1857	0.0345
12	164	200	32800	40000	26896	171.4067	-7.4067	54.8589
13	140	230	32200	52900	19600	145.9855	-5.9855	35.8263
14	2042	1613	316608	275243	428700	2042	0	905.5332

= A4*B4

= A4*A4 OR =A4^2

= TREND($A$4:$A$13,$B$4:$B$13,B4)

= A4 – F4

= Sum(A4:A13)

Consider P as X and Q as Y so we do not need to change the usual formula

Calculate the following in Excel by entering formulas

$$b = \frac{n\sum XY \ - \sum X \ \sum Y}{n\sum X^2 - (\sum X)^{\ 2}}$$

$$= (10 * (C14) - (B14) * (A14))/(10 * (D14) - (B14)\text{^}2)$$

$$= -0.84737$$

$$a = \ \bar{Y} - b\bar{X}$$

$$= (A14/10) \ - \ K6 * (B14/10)$$

$$= 340.8812$$

Verification

Instead of calculating by formula, you can use the SLOPE and INTERCEPT Excel Formula as well

=SLOPE(A4:A13,B4:B13)

And for the intercept

=INTERCEPT(A4:A13,B4:B13)

The regression Equation may now be written as

$$Y = 340.8812 \; - 0.84737\, X$$

Interpretation of 'b' (the slope coefficient): *For every on unit change in X, there may be, on the average, 0.84737 unit change in Y in the inverse direction.*

Computations in Microsoft Excel

Standard Error of Estimate

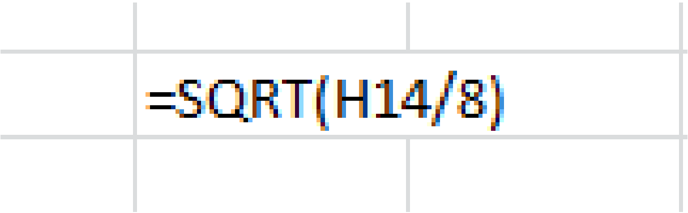
$$\boldsymbol{\hat{\sigma}_e} = \sqrt{\frac{\sum \boldsymbol{e_i^2}}{\boldsymbol{N - k}}} = \mathbf{10.63916}$$

Also

$$se(b) = \frac{\boldsymbol{\hat{\sigma}_e}}{\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}}$$

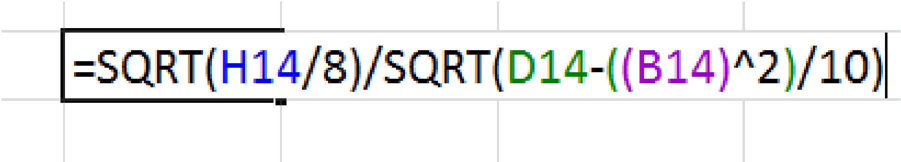
$$= 0.08668$$

Cell H14 contains the value of $\sum e^2$



=SQRT(H14/8)

Cell D14 contains the value of $\sum X^2$



```
=SQRT(H14/8)/SQRT(D14-((B14)^2)/10)
```

Cell B14 contains the value of $\sum X$

Individual Variable Significance Test

$$H0: \quad \text{b} = 0$$

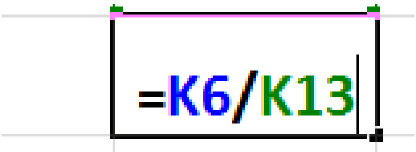
$$H1: \quad \text{b} \neq 0$$

$$\alpha = 0.05 \;\; (or\ 0.01)$$

$$Test\ statistic$$

$$t = \frac{b}{se(b)} = -9.77614$$

If the value of ‘b’ was calculated in Cell K6 and If Standard Error of b was calculated in Cell K14

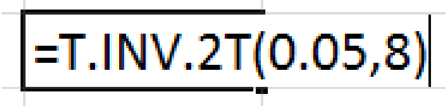


```
=K6/K13
```

$$Region\ of\ Rejection$$

$$|t| > t_{\frac{\alpha}{2},n-k}$$

$$t_{\frac{\alpha}{2},n-k} = 2.306$$



```
=T.INV.2T(0.05,8)
```

As the absolute value of ‘t-statistic’ (9.77614) is greater than the value in the table so we reject $H1$. and conclude that ‘b’ is significant.

The Goodness of Fit Test: Using the F-distribution

The Procedure is as follows:

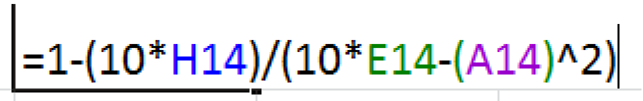
$$H0: \quad The\ Fit\ is\ not\ good$$

$$H0: \quad The\ fit\ is\ good$$

$$\alpha = 0.05$$

$Test\ statistic$ (F)

$$R^2 = 1 - \frac{\boldsymbol{N\sum e^2}}{N\sum Y^2 - (\sum Y)^2} = 0.9228$$



```
=1-(10*H14)/(10*E14-(A14)^2)
```

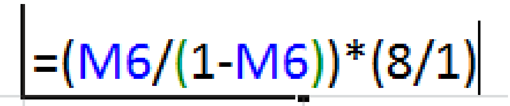
Cell H14 contains the value of $\sum e^2$

Cell E14 contains the value of $\sum Y^2$

A14 contains the value of $\sum Y$

Which gives F

$$F = \frac{R^2}{(1-R^2)}\frac{N-k}{k-1} = 95.573$$



```
=(M6/(1-M6))*(8/1)
```

$k\ -1$=*1*

We calculated the value of R^2in Cell M6

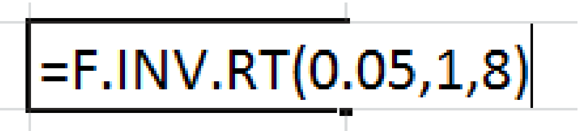
$$N - k = 10 - 2 = 8$$

$$Region\ of\ Rejection$$

$$F > F_{\alpha,\,k-1,N-k}$$

or

$$F > 5.3177$$



```
=F.INV.RT(0.05,1,8)
```

As 95.573 > 5.3177, We reject H$_0$ and conclude that the FIT IS GOOD

Verifying the results with the LINEST formula

Remember that We will use basic formulas here. An additional tool called DATA ANALYSIS tool pack will be discussed later.

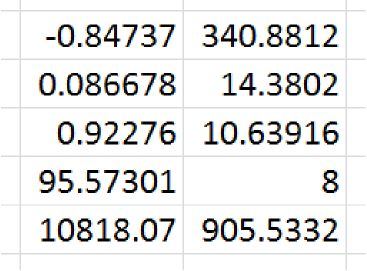
Syntax: LINEST(known_y's, [known_x's], [const], [stats])

The results are displayed in the cells that you Select like this

value of 'b'	value of 'a'
standard Error of 'b'	standard error of 'a'
R-squared	standard error of estimate
F-statistic	N-k
Regression Sum of Squares	Residual SS =Sum of square of errors

Select Cells D18:D22

- Start Typing =LINEST(A4:A13,B4:B13,TRUE,TRUE)
- hold Ctrl + SHIFT and press ENTER (Ctrl + SHIFT + ENTER)



-0.84737	340.8812
0.086678	14.3802
0.92276	10.63916
95.57301	8
10818.07	905.5332

The results are displayed and all the value are identical to our calculations. The value of t could be calculated by dividing the value of ‘b’ by the lower cell that contains the value of the standard error of b

Change of Unit of Measurement

Sometimes we change the unit of measurement for larger or smaller values. This makes the calculation easy and also makes the meaning of the coefficients useful. (Multiplying or dividing by a factor can be an example)

Examples:

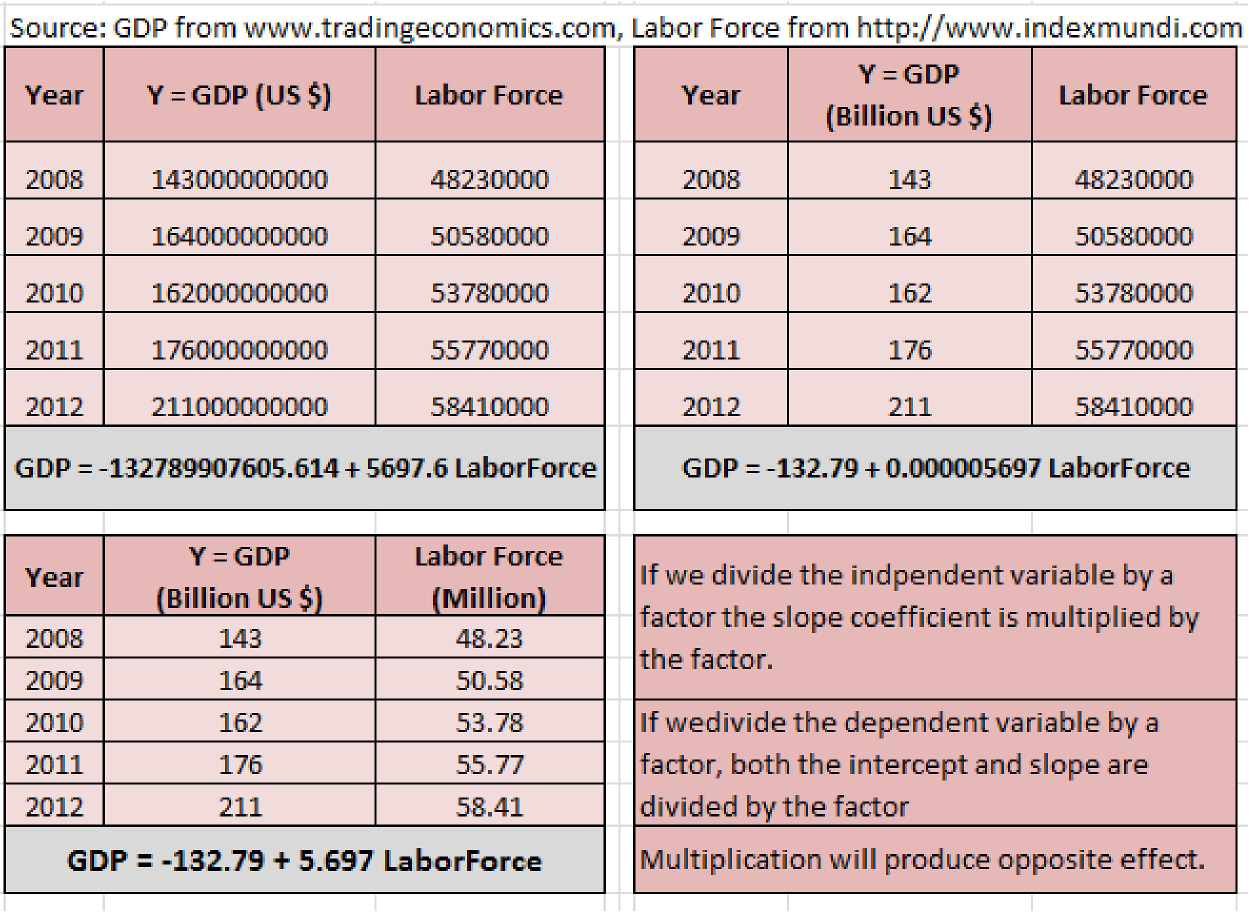
- Population can be recorded or displayed in Millions

- Prices may be displayed in thousand rupees if required

In fact we are dividing the values of Population by 1000000 (one million) and the values of Prices by 1000

We can call this a linear transformation. A linear transformation preserves linear relationships between variables. Therefore, the correlation between x and y would not change.

Change of Unit of Measurement



Source: GDP from www.tradingeconomics.com, Labor Force from http://www.indexmundi.com

Year	Y = GDP (US $)	Labor Force
2008	143000000000	48230000
2009	164000000000	50580000
2010	162000000000	53780000
2011	176000000000	55770000
2012	211000000000	58410000

GDP = -132789907605.614 + 5697.6 LaborForce

Year	Y = GDP (Billion US $)	Labor Force
2008	143	48230000
2009	164	50580000
2010	162	53780000
2011	176	55770000
2012	211	58410000

GDP = -132.79 + 0.000005697 LaborForce

Year	Y = GDP (Billion US $)	Labor Force (Million)
2008	143	48.23
2009	164	50.58
2010	162	53.78
2011	176	55.77
2012	211	58.41

GDP = -132.79 + 5.697 LaborForce

If we divide the indpendent variable by a factor the slope coefficient is multiplied by the factor.

If wedivide the dependent variable by a factor, both the intercept and slope are divided by the factor

Multiplication will produce opposite effect.

Transformation of models and Use of OLS

Some times we need to estimate models that are not linear but we can transform them and use OLS to estimate them.

Non Linear Transformation:

Nonlinear transformation changes (increases or decreases) linear relationships between variables. Correlation between variables changes.

Examples:

Taking logs of natural logs

Taking the square or square root or reciprocals of a variable

Cobb Douglas Production Function

$$Q = A\,L^{\alpha}K^{\beta}$$

Reducing to one variable (to make example of simple regression)

$$Q = AK^{\beta}$$

Taking log on both sides

$$\ln Q = \ln A + \beta \ln K$$

Let $Y = \ln Q,\ \alpha = \ln A\ and\ X = \ln K,$ then the above can be written as

$$Y = \alpha + \beta X$$

That can be estimated by OLS

Transformation of models and Use of OLS

Non Linear Transformation: Nonlinear transformation changes (increases or decreases) linear relationships between variables. Correlations between variables change.

Method	Transformation	Regression equation
Standard linear regression	Not required	y = b_0 + b_1x
Exponential model or log-linear functional form	Dependent variable = log(y)	log(y) = b_0 + b_1x
Logarithmic model or Linear-Log functional form	Independent variable = log(x)	y= b_0 + $b_1$log(x)
Double log functional form	Dependent variable = log(y), Independent variable = log(x)	log(y)= b_0 + $b_1$log(x)
Cobb Douglas Production Function (like double log form)	$Y = A\,L^{\alpha}K^{\beta}$	$\ln Y = \ln A + \alpha \ln L + \beta \ln K$

Interpretation of different functional forms using OLS

Model	Interpretation	Marginal Effect	Elasticity
Linear in variable $Y = a + bX$	One unit change in X will cause, on the average, ‘b’ units of change in Y	b	$b\frac{X}{Y}$
Double log form (log-log) $ln\, Y = a + b\; lnX$	One percent change in X will cause, on the average, ‘b’ % change in Y	$b\frac{Y}{X}$	b
Level-Log $Y = a + b\; ln\, X$	One percent change in X is expected to change Y by $\frac{100}{b}$ units	$\frac{b}{X}$	$\frac{b}{Y}$
Log-Level form $lnY = a + b\, X$	When X changes by one unit, Y will change by approximately (b*100)%	bY	bX
For interpretation, we assume that Gauss Markov assumption hold and parameters are significant			
Marginal effect of X is defined as the partial derivative of Y w.r.t. X			
marginal effect and elasticity ($\frac{dy}{dx}\frac{X}{Y}$) may be computed at mean values of X and Y			

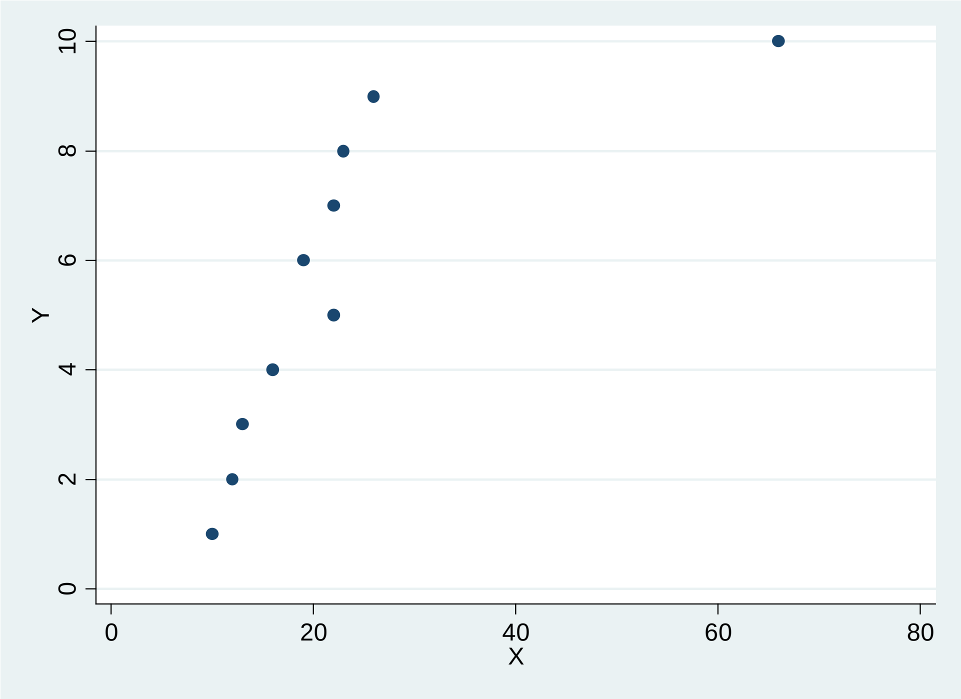
Outliers in Regression

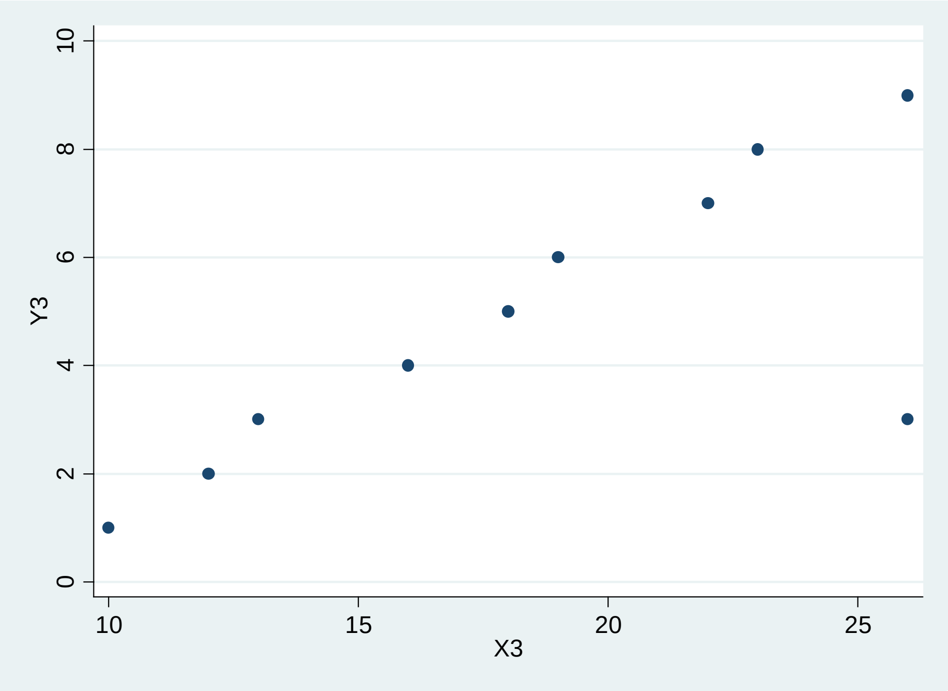
Outliers are the points that diverge a lot from the data in general and may affect the slope of the regression equation or the predictive power of the model.

Types

1. Extreme X values
2. Extreme Y values
3. Extreme X and Y
4. Distant point with normal X or Y value

If the removal of outliers changes the slope or changes the coefficient of determination a lot, this may be called an influential point



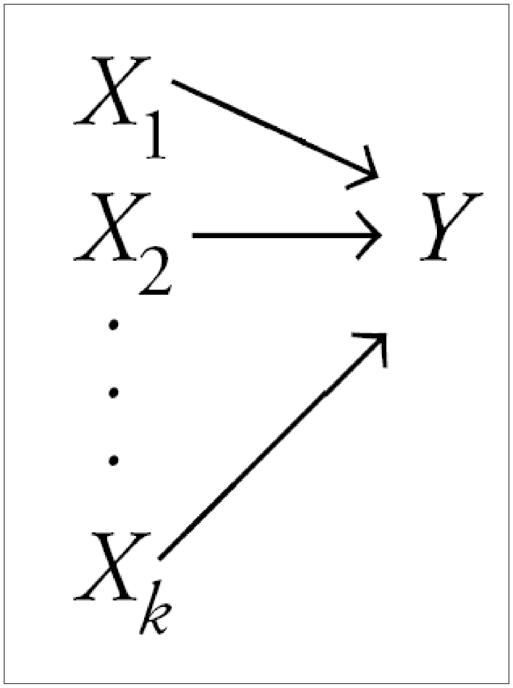


In both the diagrams the dots that appear far from the cluster of dots are outliers.

Lecture 09
Multiple Regressions

General Idea

Simple Regression considers the relation between single independent variable and the dependent variable. The Multiple Regression considers relation between one dependent variable and two or more independent (explanatory) variables



We intend to look into the impact of one independent variable on the dependent variable while other independent variables remain constant (are held).

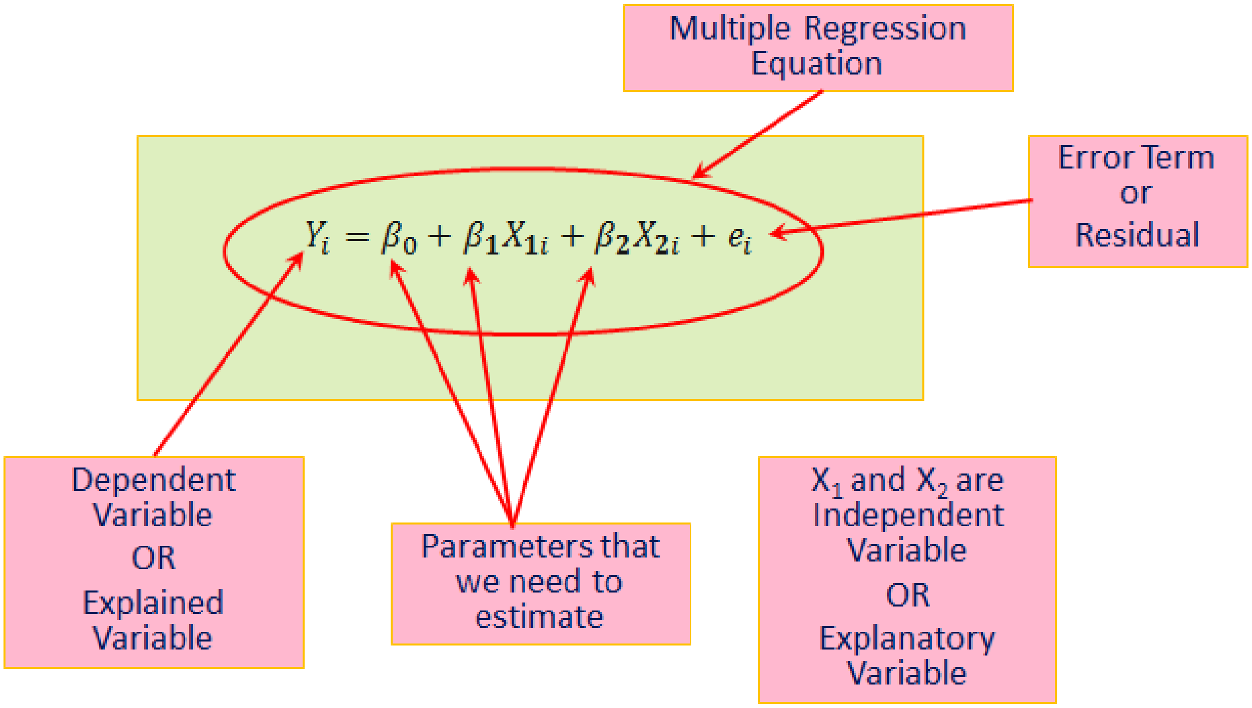
Multiple Regression: Examples

- Earnings may depend on both the educational level and the experience
- Quantity demanded may depend on price, income, prices of substitutes and other variables.
- GDP may be related to Labor Force, Capital Stock, Human Resources, and Openness etc.

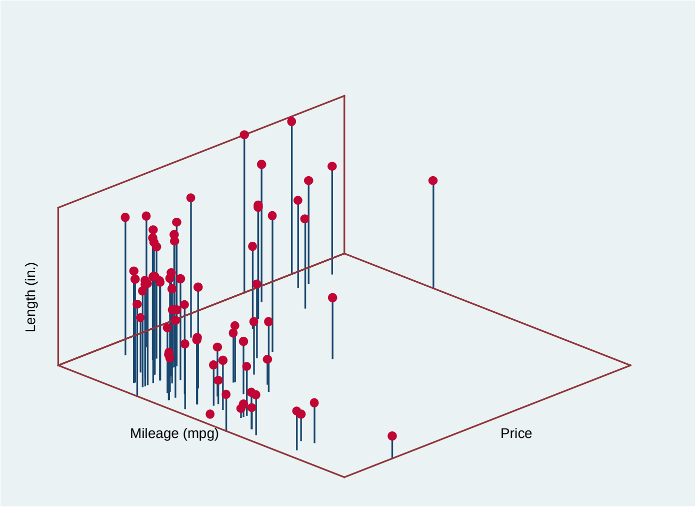
Multiple Linear Regression: 2 independent variables

Multiple Regression Equation

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i}$$


Multiple Regression Equation
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$
Error Term or Residual
Dependent Variable OR Explained Variable
Parameters that we need to estimate
X_1 and X_2 are Independent Variable OR Explanatory Variable

- Simple Regression fits a regression line in 2-dimensional space
- The Multiple Regression with two independent variables fits a line in 3-dimensional space



Scatter Diagram in Multiple Regression

In Multiple Regressions we can:

- Use several variables at the same time to explain the variation in a continuous dependent variable.
- Isolate the unique effect of one variable on the continuous dependent variable while taking into consideration that other variables are affecting it too. (remember the concept of partial derivatives)
- Write a mathematical equation that tells us the overall effects of several variables together and the unique effects of each on a continuous dependent variable. (Multiple Regression line)

Estimating a Multiple Regression Line

Least Square Estimation

Consider a regression equation

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

We need to find the values of β_0, β_1 and β_2 using OLS

Remember that we estimate the parameters by minimizing the sum of squared residuals so we face a minimization problem

$$Min \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2}$$

NOTE: We will ignore the subscript 'i' for convenience

Remember that $e = Y - \hat{Y}$ *and* $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ *(subscript 'i' ignored)*

Also remember that for 'Optimization' we take the first derivative and set it equal to zero

Important: Here although X_1, X_2 *and Y are variables but for this minimization problem only we will consider* β_0, β_1 *and* β_2 *to be the unknowns as we are trying to estimate their values. The minimization problem is*

$$Min \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \hat{Y})^{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^{\boldsymbol{2}}$$

Where 'β_0', 'β_1' and 'β_2' are the unknowns we focus on.

Differentiating w.r.t. the unknown parameters and setting equal to zero, we get THREE normal equations. (If you have difficulty then refer to the lecture with minimization problem of the simple regression line)

$$\sum Y = n\beta_0 + \beta_1 \sum X_1 + \beta_2 \sum X_2$$

$$\sum X_1 Y = \beta_0 \sum X_1 + \beta_1 \sum \boldsymbol{X_1^2} + \beta_2 \sum X_1 X_2$$

$$\sum X_2 Y = \beta_0 \sum X_2 + \beta_1 \sum X_1 X_2 + \beta_2 \sum \boldsymbol{X_2^2}$$

Solving the above ***three normal equations*** *will provide the values of the parameters.*

The Deviation form in Regression

We can derive expression to estimate the parameters of the multiple regression equation using the three normal equations.

For making calculations and working easy, we sometimes use what we call deviation form where deviations are taken from the arithmetic mean.

Here we will use SMALL letters to indicate a variable in deviation form.

$$\boldsymbol{x_{1i} = X_{1i} - \bar{X}}$$

$$\boldsymbol{x_{2i} = X_{2i} - \bar{X}}$$

$$\boldsymbol{y_i = Y_i - \bar{Y}}$$

Then the Regression equation (ignoring subscript 'i') can be written as

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2}$$

And (Applying Summations and dividing first normal equation by N

$$\boldsymbol{\bar{Y} = \beta_0 + \beta_1 \overline{X_1} + \beta_2 \overline{X_2}}$$

First minus second equation gives (note that $\beta_0$is cancelled out)

$$\boldsymbol{Y - \bar{Y} = \beta_1 (X_1 - \overline{X_1}) + \beta_2 (X_2 - \overline{X_2})}$$

That can be written in deviation form as

$$\boldsymbol{y = \beta_1 x_1 + \beta_2 x_2}$$

The above is a regression equation in deviation form.

Now, as we have a regression equation in deviation form

$$\boldsymbol{y = \beta_1 x_1 + \beta_2 x_2}$$

where

$$\boldsymbol{x_{1i} = X_{1i} - \bar{X}}$$

$$\boldsymbol{x_{2i} = X_{2i} - \bar{X}}$$

$$\boldsymbol{y_i = Y_i - \bar{Y}}$$

We can minimize the sum of squared residual in deviation form. The equation has two unknowns and, as a result, we have two normal equations (in deviation form)

$$\sum x_1 y \;= \beta_1 \sum \boldsymbol{x_1^2} \;+ \beta_2 \sum x_1 x_2$$

$$\sum x_2 y \;= \beta_1 \sum x_1 x_2 \;+ \beta_2 \sum \boldsymbol{x_2^2}$$

Normal equations are derived by minimizing sum of squared residual

One option is to substitute the summation values and solve the equations to get β_1 and β_2 but we can find expression for β_1 and β_2

Deriving Expression for $\boldsymbol{\beta}$ $\mathbf{coefficients}$

$$\sum x_1 y \;= \beta_1 \sum \boldsymbol{x_1^2} \;+ \beta_2 \sum x_1 x_2$$

$$\sum x_2 y \;= \beta_1 \sum x_1 x_2 \;+ \beta_2 \sum \boldsymbol{x_2^2}$$

From the first equation, we get

$$\beta_2 = \frac{\sum x_1 y \;- \beta_1 \sum \boldsymbol{x_1^2}}{\sum x_1 x_2}$$

Substituting this in the second equation and solving for β_1 gives

$$\beta_1 = \frac{\sum x_1 y \; \sum \boldsymbol{x_2^2} - \sum x_2 y \sum x_1 x_2}{\sum \boldsymbol{x_1^2} \sum \boldsymbol{x_2^2} - (\sum x_1 x_2)^2}$$

Similarly

$$\beta_2 = \frac{\sum x_2 y \sum \boldsymbol{x_1^2} - \sum x_1 y \sum x_1 x_2}{\sum \boldsymbol{x_1^2} \sum \boldsymbol{x_2^2} - (\sum x_1 x_2)^{\boldsymbol{2}}}$$

Also, using the first normal equation of the NORMAL form

$$\boldsymbol{\beta_0 = \overline{Y} - \beta_1 \overline{X_1} - \beta_2 \overline{X_2}}$$

Notice the symmetry. If we just replace $x_1 with$ x_2

$\mathrm{and x_2 with\ x_1}$ in the expression for $\beta_1$we get the expression for β_2 and vice versa.

Manual procedure to solve for Multiple Regression

The least time consuming manual procedure to estimate a regression line

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

Is to use the expression (in deviation form)

$$\beta_1 = \frac{\sum x_1 y \sum \boldsymbol{x_2^2} - \sum x_2 y \sum x_1 x_2}{\sum \boldsymbol{x_1^2} \sum \boldsymbol{x_2^2} - (\sum x_1 x_2)^{\boldsymbol{2}}}$$

$$\beta_2 = \frac{\sum x_2 y \sum \boldsymbol{x_1^2} - \sum x_1 y \sum x_1 x_2}{\sum \boldsymbol{x_1^2} \sum \boldsymbol{x_2^2} - (\sum x_1 x_2)^{\boldsymbol{2}}}$$

and

$$\boldsymbol{\beta_0 = \overline{Y} - \beta_1 \overline{X_1} - \beta_2 \overline{X_2}}$$

Where

$$\boldsymbol{x_{1i} = X_{1i} - \overline{X}}$$

$$\boldsymbol{x_{2i} = X_{2i} - \overline{X}}$$

$$\boldsymbol{y_i = Y_i - \overline{Y}}$$

You can either generate columns having the values in deviation form (like $X_{1i} - \overline{X}$) or you can calculate the summations in deviation form, using expressions shown in the next slide

Conversion from Normal to Deviation form

Notice that we need several summations in deviation form to estimate the regression

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

where

$$\boldsymbol{x_{1i} = X_{1i} - \bar{X}}$$

$$\boldsymbol{x_{2i} = X_{2i} - \bar{X}}$$

$$\boldsymbol{y_i = Y_i - \bar{Y}}$$

We need

$\sum x_1 y$, $\sum x_2 y$, $\sum x_1 x_2$, $\sum \boldsymbol{x_1^2}$, $\sum \boldsymbol{x_2^2}$

We also would be needing $\sum y^2$ in future although not required yet. Remember that we are using deviation form so

$$\sum x_1 y \ = \sum (X_1 - \overline{X_1})\,(Y - \bar{Y})$$

$$\sum \boldsymbol{x_1^2} = \sum (\boldsymbol{X_1} - \overline{X_1})^2$$

etc.

We can use the following to convert summations from normal to deviation form

$$\sum x_1 y = \sum X_1 Y - \frac{\sum X_1 \sum Y}{n}$$

$$\sum x_2 y = \sum X_2 Y - \frac{\sum X_2 \sum Y}{n}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{\sum X_1 \sum X_2}{n}$$

$$\sum \boldsymbol{x_1^2} = \sum \boldsymbol{X_1^2} - \frac{(\sum X_1)^2}{\boldsymbol{n}}$$

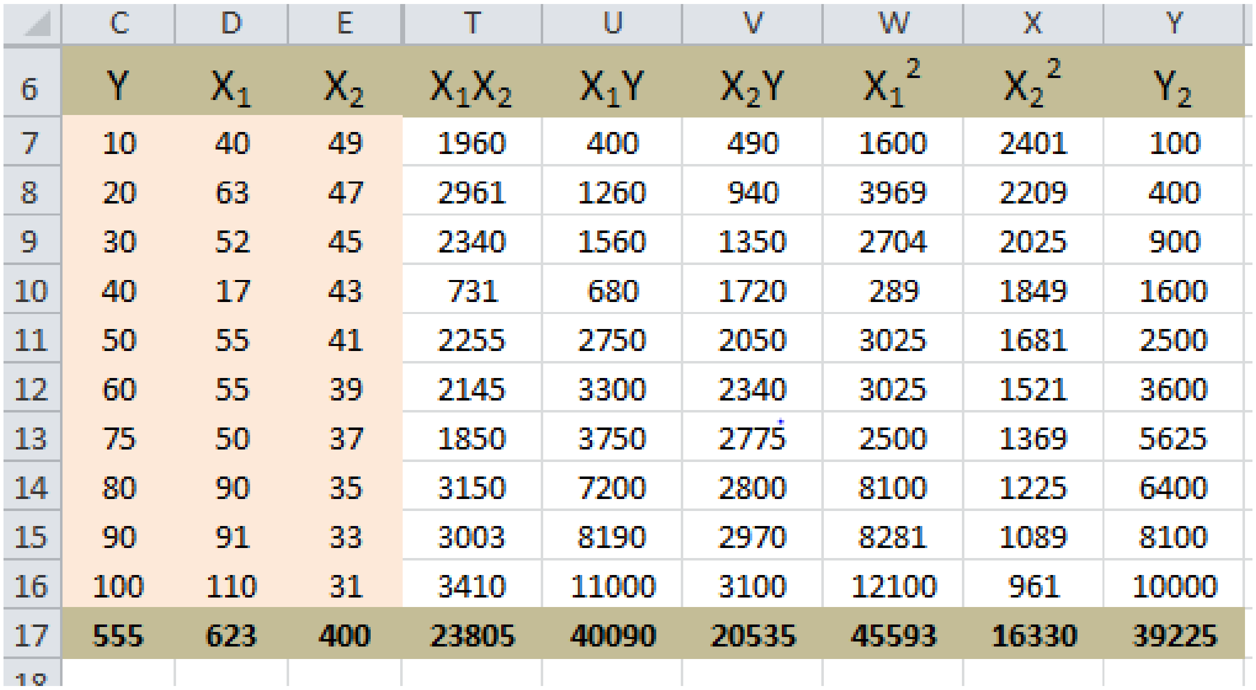
$$\sum \boldsymbol{x_2^2} = \sum \boldsymbol{X_2^2} - \frac{(\sum X_2)^2}{\boldsymbol{n}}$$

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

As An exercise, you can try to prove these expression using summation algebra

Example

Although we can use data analysis tool in Microsoft Excel but for better understanding, we present an example in Microsoft Excel that first manually estimates the regression line and then verifies it with the LINEST formula



	C	D	E	T	U	V	W	X	Y
6	Y	X_1	X_2	X_1X_2	X_1Y	X_2Y	$X_1{}^2$	$X_2{}^2$	Y_2
7	10	40	49	1960	400	490	1600	2401	100
8	20	63	47	2961	1260	940	3969	2209	400
9	30	52	45	2340	1560	1350	2704	2025	900
10	40	17	43	731	680	1720	289	1849	1600
11	50	55	41	2255	2750	2050	3025	1681	2500
12	60	55	39	2145	3300	2340	3025	1521	3600
13	75	50	37	1850	3750	2775	2500	1369	5625
14	80	90	35	3150	7200	2800	8100	1225	6400
15	90	91	33	3003	8190	2970	8281	1089	8100
16	100	110	31	3410	11000	3100	12100	961	10000
17	**555**	**623**	**400**	**23805**	**40090**	**20535**	**45593**	**16330**	**39225**

Example

From the example, calculating using Microsoft Excel, we have

$$\sum x_1 y = \sum X_1 Y - \frac{\sum X_1 \sum Y}{n} = 5513.5, \sum x_2 y = \sum X_2 Y - \frac{\sum X_2 \sum Y}{n} = -1665$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{\sum X_1 \sum X_2}{n} = -1115, \ \sum \boldsymbol{x_1^2} = \sum \boldsymbol{X_1^2} - \frac{(\sum X_1)^2}{\boldsymbol{n}} = 6780.1$$

$$\sum \boldsymbol{x_2^2} = \sum \boldsymbol{X_2^2} - \frac{(\sum X_2)^2}{\boldsymbol{n}} = 330, \ \sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} = 8422.5$$

Using these values, we compute

$$\beta_1 = \frac{\sum x_1 y \sum \boldsymbol{x_2^2} - \sum x_2 y \sum x_1 x_2}{\sum \boldsymbol{x_1^2} \sum \boldsymbol{x_2^2} - (\sum x_1 x_2)^{\boldsymbol{2}}} = -0.0372$$

$$\beta_2 = \frac{\sum x_2 y \sum \boldsymbol{x_1^2} - \sum x_1 y \sum x_1 x_2}{\sum \boldsymbol{x_1^2} \sum \boldsymbol{x_2^2} - (\sum x_1 x_2)^{\boldsymbol{2}}} = -\boldsymbol{5.1713}$$

and $$\boldsymbol{\beta_0 = \overline{Y} - \beta_1\overline{X_1} - \beta_2\overline{X_2} = 264.67}$$

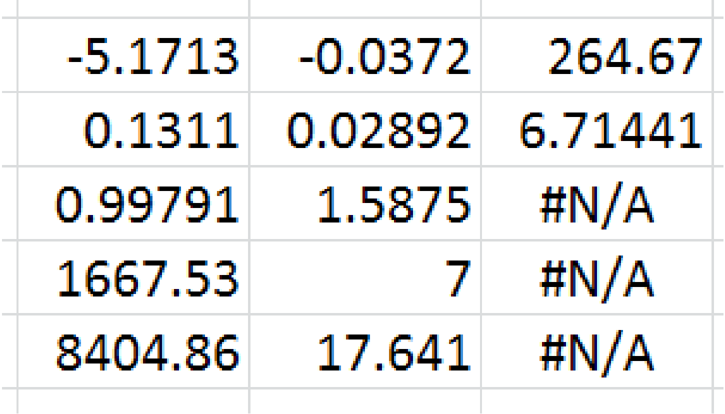
$$Y = 264.67 - 0.0372X_1 - 5.1713X_2$$

Using LINEST in Multiple Regression

Now we use the LINEST formula for multiple regression. It is used like as you did in simple regression but with additional columns. You need three columns (as you have three parameters) and five rows. So we Select Cells C19:E23

- Start Typing =LINEST(C7:C16,D7:E16,TRUE,TRUE)
- hold Ctrl + SHIFT and press ENTER (Ctrl + SHIFT + ENTER)

The results are displayed in the cells that you select like this

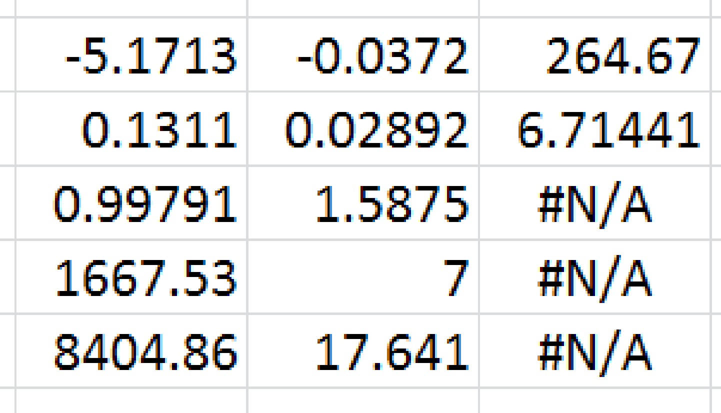


-5.1713	-0.0372	264.67
0.1311	0.02892	6.71441
0.99791	1.5875	#N/A
1667.53	7	#N/A
8404.86	17.641	#N/A

Using LINEST in Multiple Regression

Values will be displayed like this

value of β_2	value of β_1	value of β_0
standard Error of β_2	standard error of β_1	standard error of β_0
R-squared	standard error of estimate	
F-statistic	N-k	
Regression Sum of Squares	Residual SS =Sum of square of errors	



-5.1713	-0.0372	264.67
0.1311	0.02892	6.71441
0.99791	1.5875	#N/A
1667.53	7	#N/A
8404.86	17.641	#N/A

Values of t-statistic can be computed as a ratio of parameter and its standard error

Simple Vs Multiple Regression

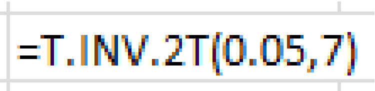
As the number of independent variable increases the degree of freedom decreases. In our examples the d.f. was 8. In the current example with 10 observations and two independent variables, the degree of freedom N-k = 7

In simple regression the coefficient of determination had a value identical to the square of the correlation coefficient. In multiple regressions, the coefficient of determination has a value different from the square of the correlation coefficient.

We performed a test of significance for the parameter 'b' in simple regression. Here we need to perform two tests of significance as we have two slope coefficients. The formula for standard error changes in case of multiple regressions.

Diagnostic Tests for Regression Analysis

- We performed a test of significance for the parameter 'b' in simple regression.
- Here we need to perform two tests of significance as we have two slope coefficients
- The formula for standard error changes in case of multiple regression



=T.INV.2T(0.05,7)

$$t_{\frac{\alpha}{2},n-k} = t_{0.025,7} = 2.36$$

We need to perform two tests for individual variable significance. One for $\beta_1$and the second for β_2

Test for β_1

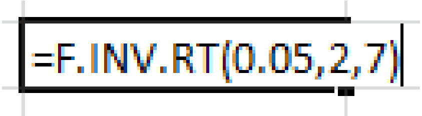
$$t = \frac{\beta_1}{se(\beta_1)} = \frac{-0.0372}{0.02892} = -1.287$$

As $|t| > t_{\frac{\alpha}{2},n-k}$ is not satisfied so β_1 is not significant. The variable X_1does not seem to have a significant impact on Y.

Test for β_2

$$t = \frac{\beta_2}{se(\beta_2)} = \frac{-5.1713}{0.1311} = -39.45$$

As $|t| > t_{\frac{\alpha}{2},n-k}$ is satisfied so β_2 issignificant. The variable X_2 seems to have a significant impact on Y.



=F.INV.RT(0.05,2,7)

$$F_{\alpha,k-1,n-k} = F_{0.05,2,7} = 4.74$$

$$R^2 = 1 - \frac{\boldsymbol{N\sum e^2}}{N\sum Y^2 - (\sum Y)^2} = 0.9979$$

$$F = \frac{R^2}{(1-R^2)}\frac{N-k}{k-1} = 1667.53$$

$$F_{\alpha,\, k-1,N-k} = F_{0.05,\, 1,3} = 4.74$$

As $F > F_{\alpha,\, k-1,N-k}$ is satisfied, we conclude that the fit is good.

Lecture 10

Linear Regression Estimation with K-independent variables

We define a linear regression equation with K independent variables as

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\;\;.\; + \;\beta_K X_{Ki} + e_i}$$

Where $1 < K < M$ and M is a finite number. The variables have their usual meaning.

If we minimize the sum of squared residuals, we get K+1 normal equations to be solved simultaneously. The software may use Matrix algebra to solve these equations. There is nothing new in such regression. The usual process may be followed. We can use the LINEST formula or Data Analysis Add-in in Microsoft Excel. Other software also may be used.

However you may notice that some things change by adding independent variables

- The degree of freedom (N-k) decreases.
- You may observe some problems like multicollinearity
- The formula for standard errors of the coefficients change.
- You may need larger number of observations.
- The coefficient of determination (R^2) increases by adding more independent variables.
- The value of the regression coefficients may change by adding a new variable.
- Software have their limitations.

Analysis ToolPak: DATA ANALYSIS Add-In in Microsoft Excel

To perform statistical and econometric analysis, we can save lot of time and effort by using software. The Analysis ToolPak in Microsoft Excel is an example of that. This tool uses the LINEST function to perform regressions. It produces and output containing basic regression statistics in tabular form. This tool has a limit of maximum 16 independent variables. This tool comes with Microsoft Excel but you need to install or add it from within Excel

Analysis ToolPak: Installing or Adding

Typically it is already installed. You just need to add it for which you will not need a CD. If it is not installed, you may need an Office CD. This is required only once.

Excel 2003 and earlier: select in Excel the Tools Menu and the menu item Add-ins.

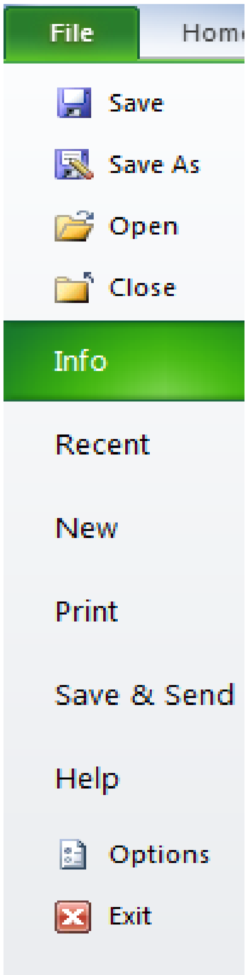
Excel 2007: Office Button , Excel Options, Add-ins, Manage Excel Add-ins in the selection box, then click GO

Excel 2010: Green File, Options, Manage Excel Add-ins in the selection box, then click GO

We will demonstrate in Microsoft Excel 2010

1. Click on FILE and you will see a list

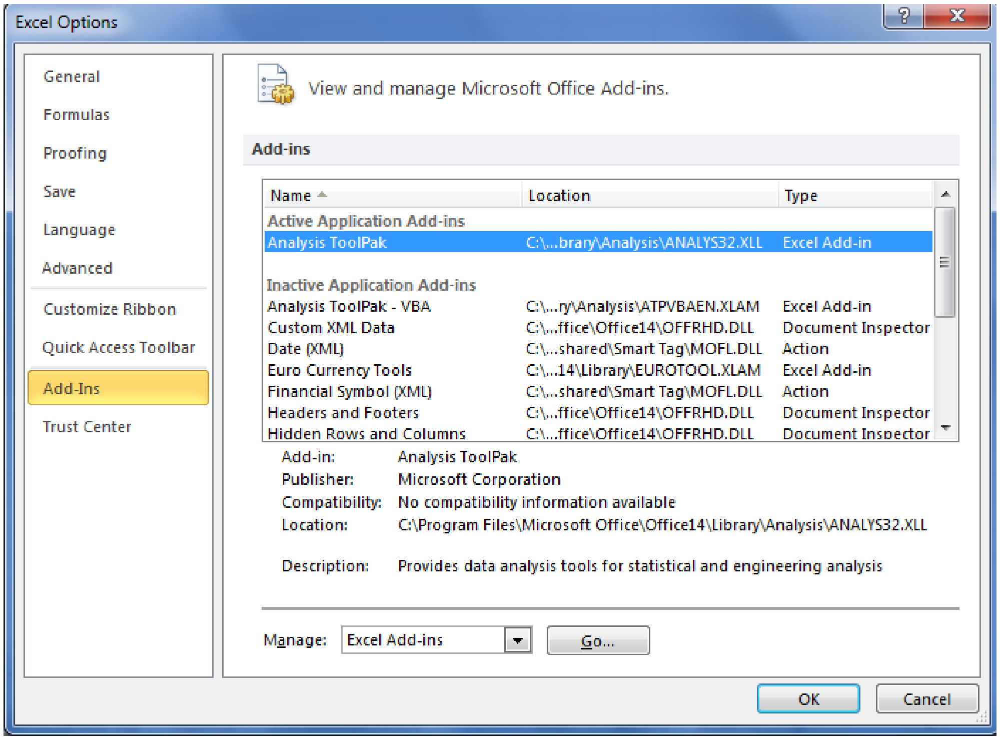
2. Now click on Options (the second last item)



3. In the Excel options window, click on Add-In

4. Now Click on Analysis ToolPak (not with the VBA)

5. Click GO



View and manage Microsoft Office Add-ins.

Add-ins

Name	Location	Type
Active Application Add-ins		
Analysis ToolPak	C:\...brary\Analysis\ANALYS32.XLL	Excel Add-in
Inactive Application Add-ins		
Analysis ToolPak - VBA	C:\...ry\Analysis\ATPVBAEN.XLAM	Excel Add-in
Custom XML Data	C:\...ffice\Office14\OFFRHD.DLL	Document Inspector
Date (XML)	C:\...shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...14\Library\EUROTOOL.XLAM	Excel Add-in
Financial Symbol (XML)	C:\...shared\Smart Tag\MOFL.DLL	Action
Headers and Footers	C:\...ffice\Office14\OFFRHD.DLL	Document Inspector
Hidden Rows and Columns	C:\...ffice\Office14\OFFRHD.DLL	Document Inspector

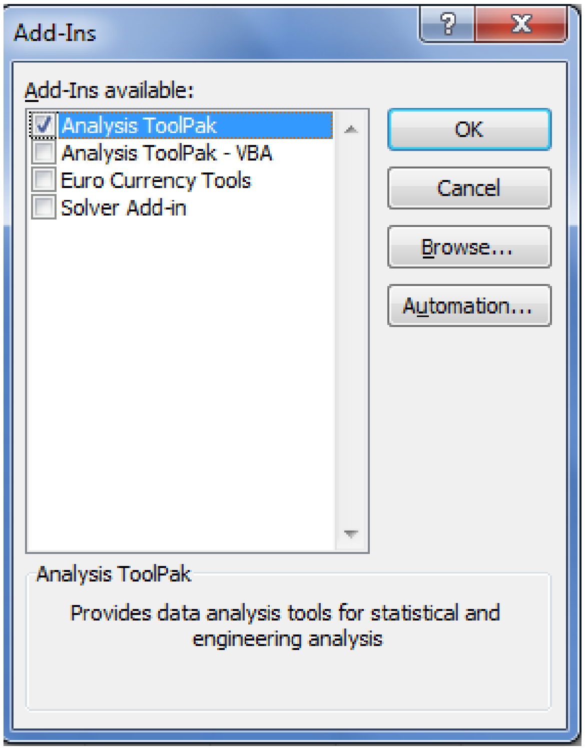
Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information available
Location: C:\Program Files\Microsoft Office\Office14\Library\Analysis\ANALYS32.XLL

Description: Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins

6. Check the box for Analysis ToolPak

7. Click OK



Add-Ins available:

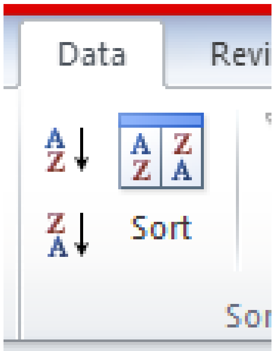
☑ Analysis ToolPak
□ Analysis ToolPak - VBA
□ Euro Currency Tools
□ Solver Add-in

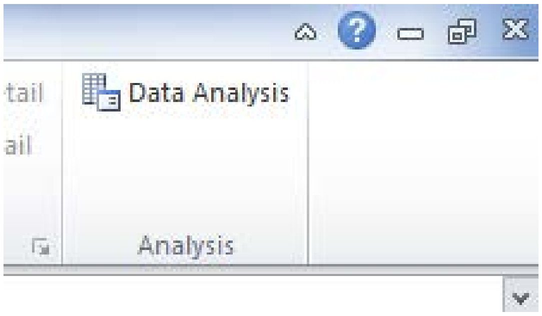
Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis

8. You may be prompted that currently this is not installed, do you want to install it so here click YES

9. Installation will start. After few seconds you can see if it is installed by clicking on the DATA ribbon and seeing if you find Data Analysis

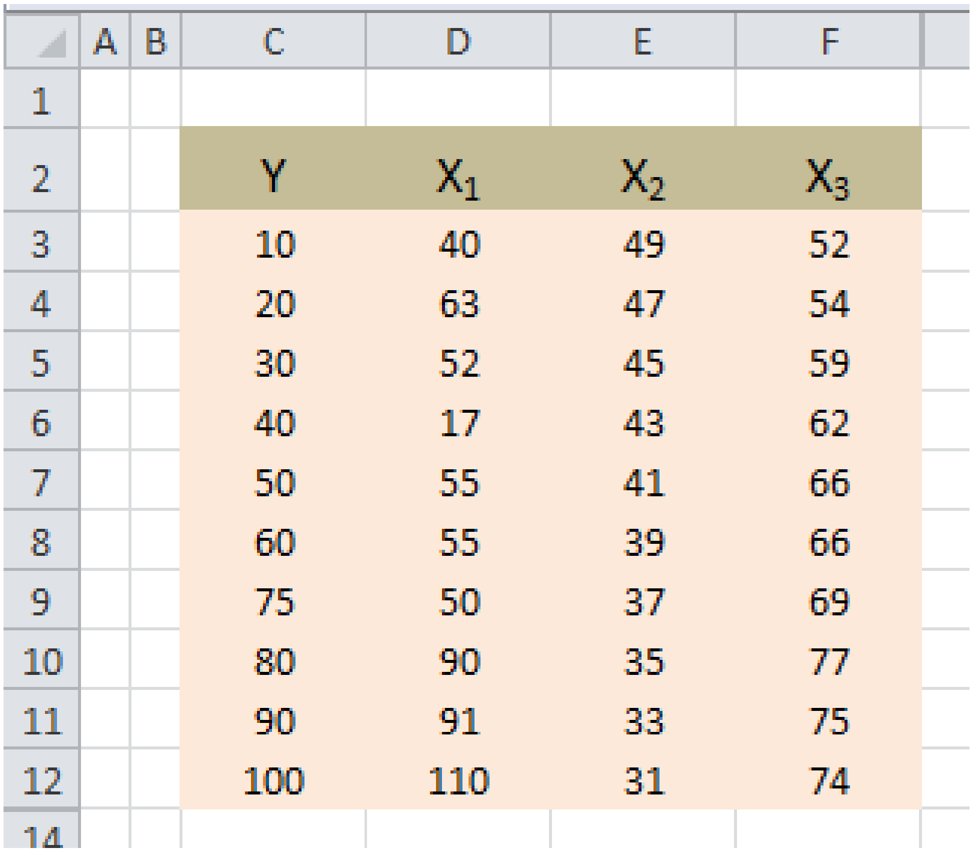




Regression with The Analysis ToolPak

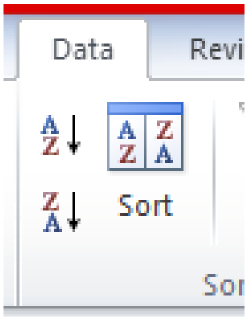
There are 19 options available in the Add-in but presently we will use and perform Regression.

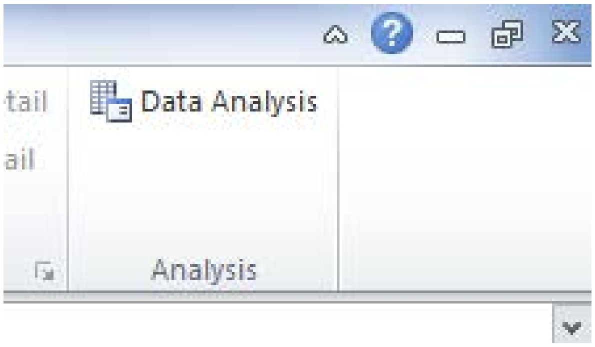
First Enter the following data (10 observations each) in your Excel Sheet. Use the same cells as we have done



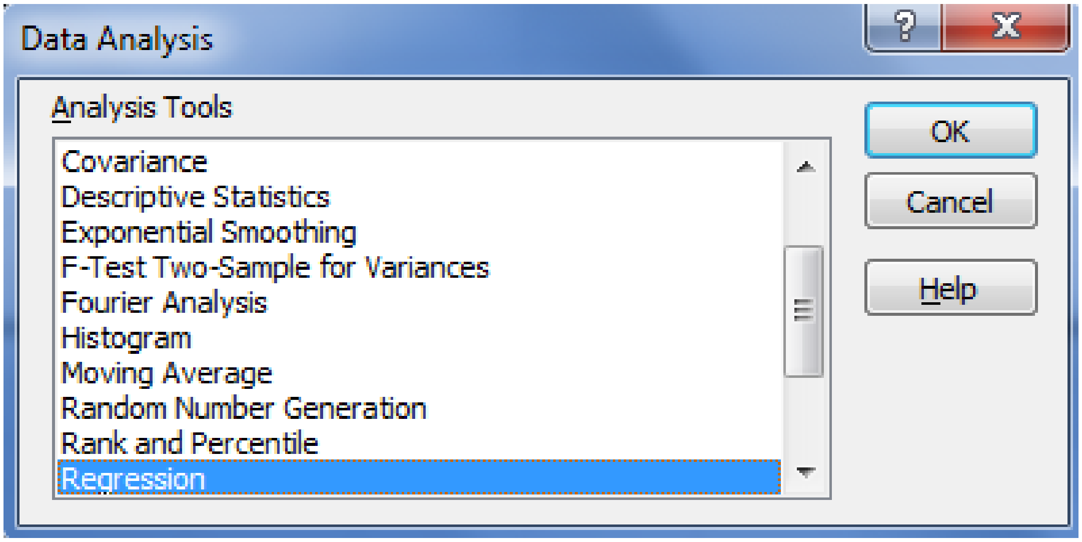
	A	B	C	D	E	F
1						
2			Y	X_1	X_2	X_3
3			10	40	49	52
4			20	63	47	54
5			30	52	45	59
6			40	17	43	62
7			50	55	41	66
8			60	55	39	66
9			75	50	37	69
10			80	90	35	77
11			90	91	33	75
12			100	110	31	74
14						

In the Data Ribbon, Click on Data Analysis (right most item)

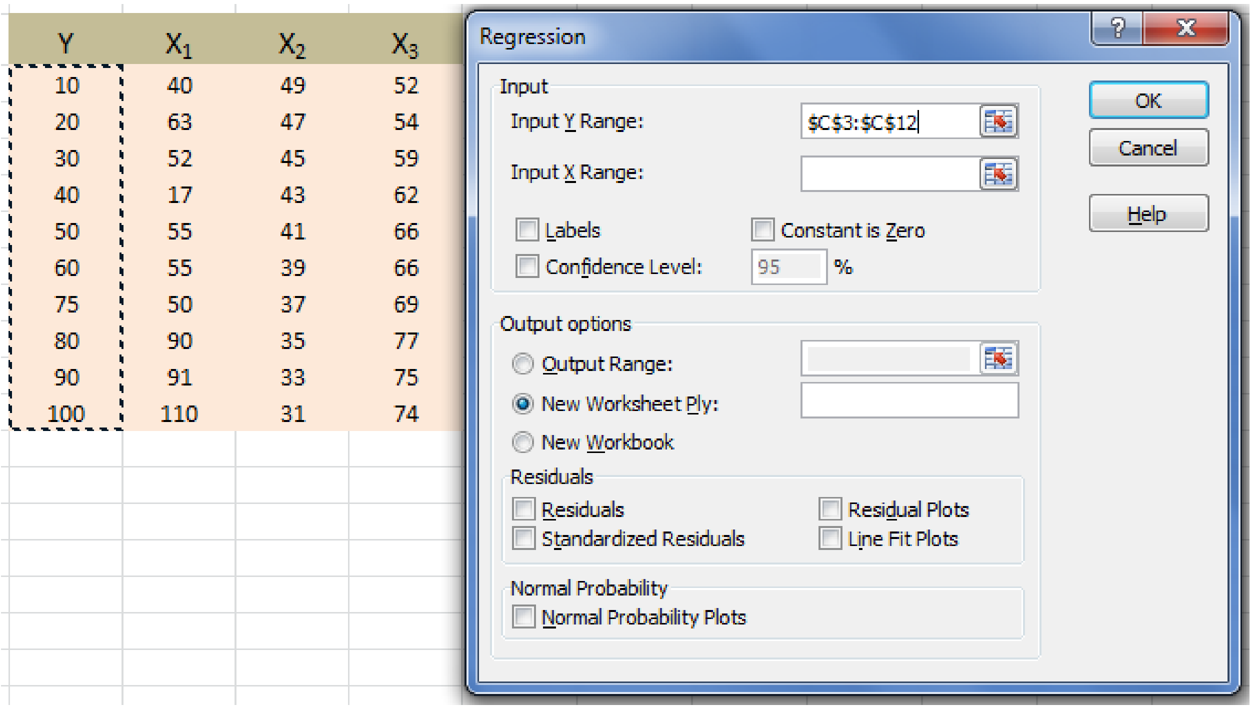




In the window that appears, scroll down to Regression and click OK



In the window that appears now, click in the Input Y Range and select by dragging your mouse on the Y values



Y	X_1	X_2	X_3
10	40	49	52
20	63	47	54
30	52	45	59
40	17	43	62
50	55	41	66
60	55	39	66
75	50	37	69
80	90	35	77
90	91	33	75
100	110	31	74

Regression

Input

Input Y Range: $C$3:$C$12

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

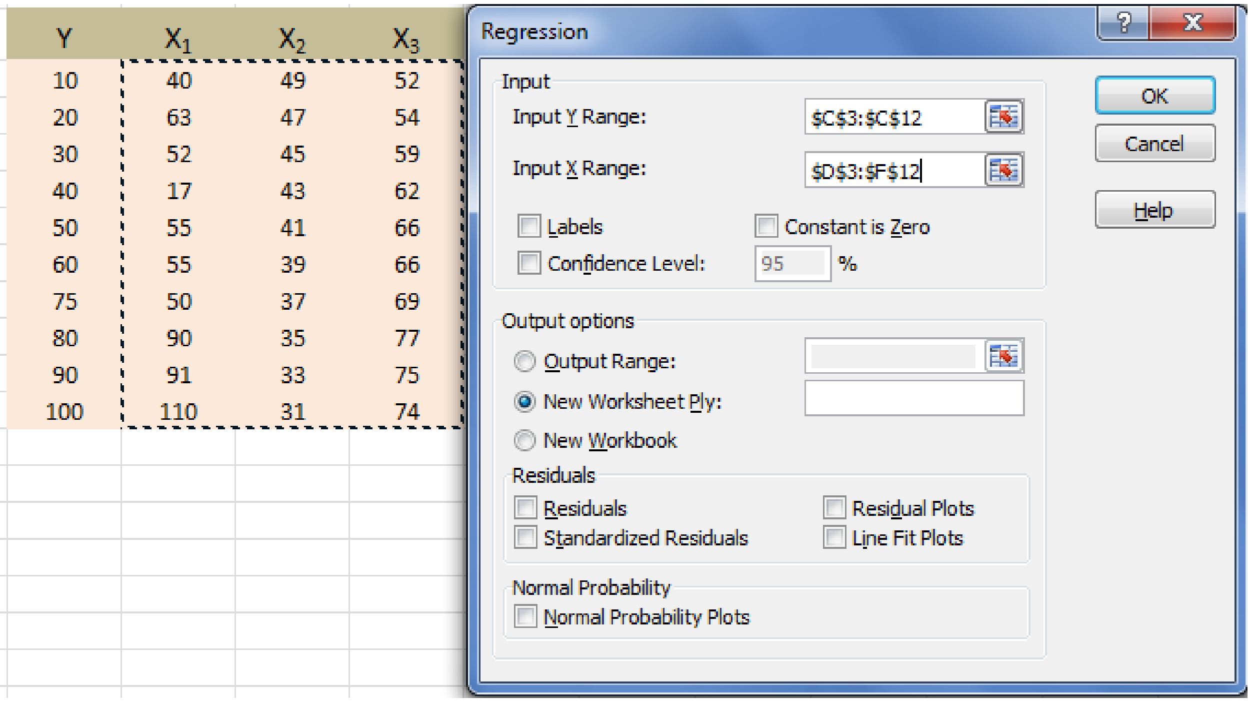
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Now click in the Input X Range and select by dragging your mouse on the X values (all the columns of X values)

Check (if not already selected) on New Worksheet. Then click OK



Y	X_1	X_2	X_3
10	40	49	52
20	63	47	54
30	52	45	59
40	17	43	62
50	55	41	66
60	55	39	66
75	50	37	69
80	90	35	77
90	91	33	75
100	110	31	74

Regression

Input

Input Y Range: $C$3:$C$12

Input X Range: $D$3:$F$12

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

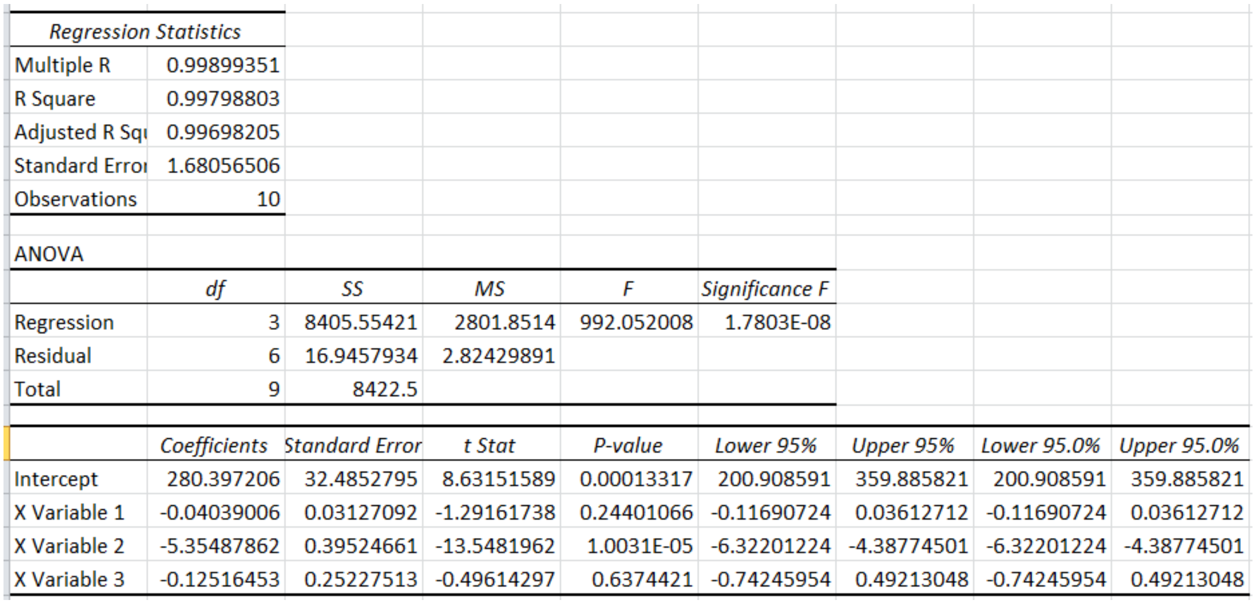
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

The results will be displayed in a new worksheet. Just resize the columns to see the results in a better way.

You have got all your usual regression statistics in a tabular form



Regression Statistics	
Multiple R	0.99899351
R Square	0.99798803
Adjusted R Squ	0.99698205
Standard Error	1.68056506
Observations	10

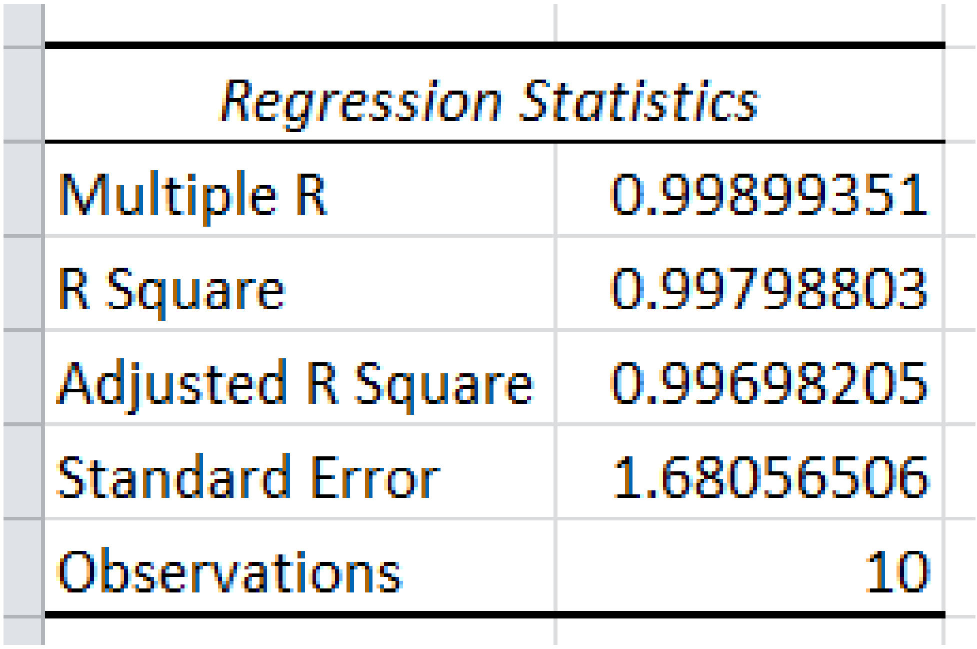
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	8405.55421	2801.8514	992.052008	1.7803E-08
Residual	6	16.9457934	2.82429891		
Total	9	8422.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	280.397206	32.4852795	8.63151589	0.00013317	200.908591	359.885821	200.908591	359.885821
X Variable 1	-0.04039006	0.03127092	-1.29161738	0.24401066	-0.11690724	0.03612712	-0.11690724	0.03612712
X Variable 2	-5.35487862	0.39524661	-13.5481962	1.0031E-05	-6.32201224	-4.38774501	-6.32201224	-4.38774501
X Variable 3	-0.12516453	0.25227513	-0.49614297	0.6374421	-0.74245954	0.49213048	-0.74245954	0.49213048

Understanding the Regression Results

In the third part of the table given above, we have the values of the coefficients, the standard errors, the t-statistic, the significance of t-statistic etc. Look at the P-values. These values show the level of significance at which, based on the t-statistic, we can reject the H_0 that the coefficient is not significant. Only the value corresponding to the variable X_2 is below 1% (it is 0.00001003). This shows that, at 1% level of significance, the coefficient of X_2 is significant. Other coefficients are not significant. However, the intercept is also significant as the p-value is below 1% .



Regression Statistics	
Multiple R	0.99899351
R Square	0.99798803
Adjusted R Square	0.99698205
Standard Error	1.68056506
Observations	10

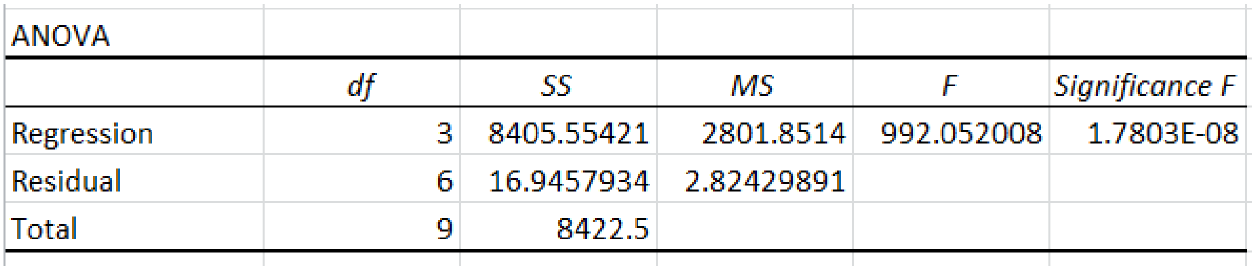
Multiple correlations is computable form partial correlation coefficients and is the Square Root of $\boldsymbol{R^2}$. Coefficient of Determination is Explanatory power of the model. Here it is 0.99 or 99%.

Adjusted Coefficient of Determination (Adjusted to degrees of freedom)

$$Adjusted\ R^2$$

$$= R^2 - (1 - R^2) * \frac{k-1}{N-k}$$

As the number of independent variables increases the coefficient of determination is overstated so it may be sometimes better to use the adjusted R^2 for goodness of fit test



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	8405.55421	2801.8514	992.052008	1.7803E-08
Residual	6	16.9457934	2.82429891		
Total	9	8422.5			

The above table shows the sums of square. If we divide the SS (Sum of square) by the df (degree of freedom), we get the MS (Means Sum of Square). The Value of F-statistic that is required in the goodness of fit test is also given. The value under the significance of F which is 0.0000000178 (approximately zero) shows that the F-Statistic is significance below 1% so we can conclude at 1% level of significance that the model is a good fit.

P-value equals the Pr{|t| > t-Stat} where t is a t-distributed random variable with n-k degrees of freedom and t-Stat is the computed value of the t-statistic given in the previous column.

You can say that it is the level of significance at which the calculated t-statistic becomes larger than the table value of t-statistic

LIMITATIONS of working with Data Analysis Toolpak

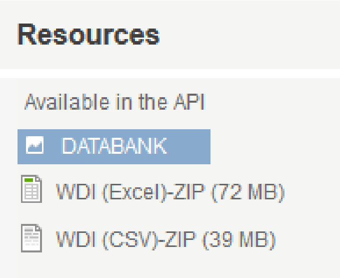
- Excel restricts the number of independent variables to 16.
- Excel requires that all the independent variables variables be in adjoining columns.
- Excel standard errors and t-statistics and p-values are based on the assumption that the error is independent with constant variance.
- Excel does not provide alternative models like the ones with robust standard errors So more powerful software such as STATA, EVIEWS, SPSS may be needed

Practicing Regression Estimation

Let us practice regression on real life data.

Go to http://data.worldbank.org/data-catalog/world-development-indicators

THE WORLD BANK
Working for a World Free of Poverty



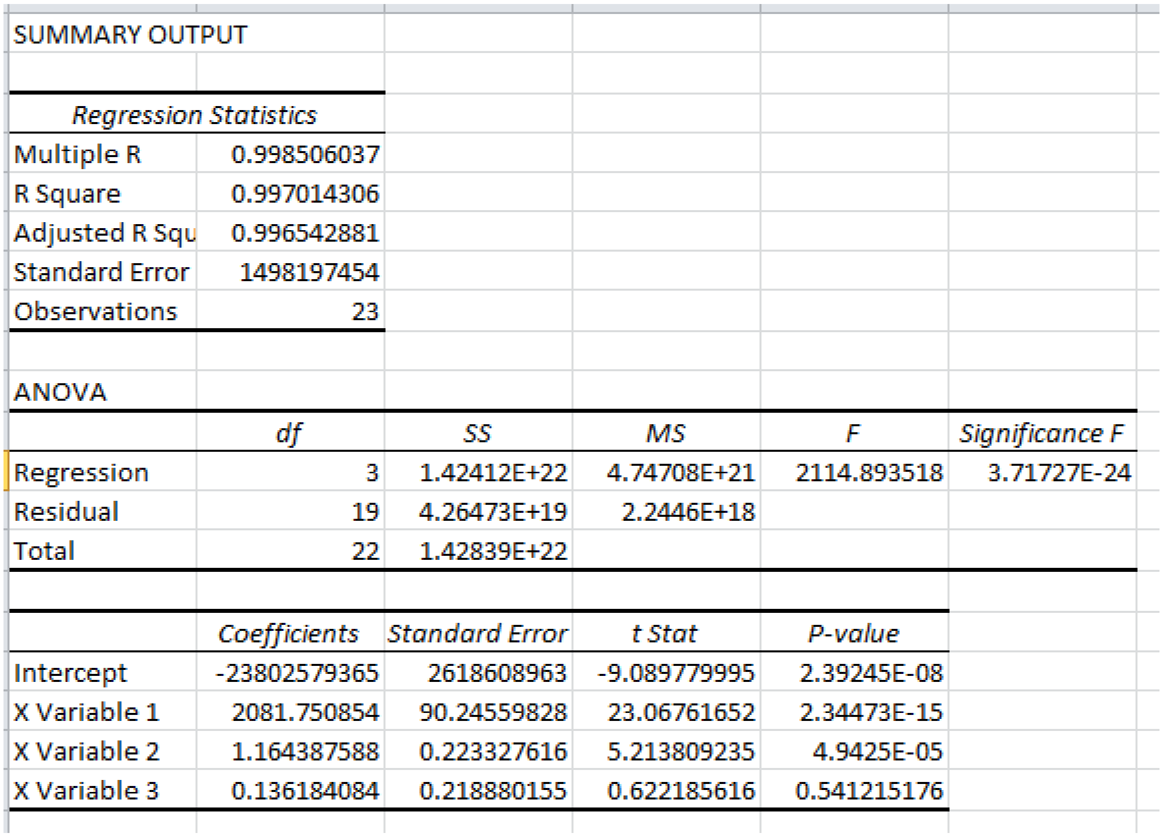
On the right, under RESOURCES, click on DATA BANK. When the page loads, select PAKISTAN in COUNTRY. Click on SERIES and when it loads, select the following variables:

- GDP (constant 2005 US$)
- Labor force, total
- Gross fixed capital formation (constant 2005 US$)
- Exports of goods and services (constant 2005 US$)

You can use the filter given to search for variables. Now click on TIME and select the years from 1991 to 2012. Now on top right of the page, click DOWNLOAD. You will be asked about the data format; Select EXCEL and click DOWLNOAD.

Practicing Regression Estimation

Save the file on your hard disk. Open the file and you will have the data you downloaded. First you need to reshape the data for use in Data Analysis ToolPak. Select all data and copy. Now press Ctrl+V to paste but use paste special and check on TRANSPOSE. Now you can use the DATA ANALYSIS ToolPak. REMEMBER: All independent variables should be in adjacent columns. Let us run a regression where we think that GDP has a linear relationship with Labor Force, Stock of Capital (proxy is GFCF), and Exports. The result is displayed below (for discussion view the video lecture)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.998506037
R Square	0.997014306
Adjusted R Squ	0.996542881
Standard Error	1498197454
Observations	23

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1.42412E+22	4.74708E+21	2114.893518	3.71727E-24
Residual	19	4.26473E+19	2.2446E+18		
Total	22	1.42839E+22			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-23802579365	2618608963	-9.089779995	2.39245E-08
X Variable 1	2081.750854	90.24559828	23.06761652	2.34473E-15
X Variable 2	1.164387588	0.223327616	5.213809235	4.9425E-05
X Variable 3	0.136184084	0.218880155	0.622185616	0.541215176

Possible Problems

- The relationship may not be linear. (Wrong specification)
- We may have missed some important variables
- Some of the indicators may not be appropriate
- There may be outliers (e.g. abnormal years)
- The errors may not be normally distributed
- There may be problems like multicollinearity, heterskedasticity or autocorrelation (to be discussed later)

Lecture 11

Transformation for Regression

Linear in the 'Linear Regression' means that the model is 'linear in parameters' and not necessarily linear in variables. Linearity is a poor approximation of truth. This is because the relationships may be non-linear e.g. quadratic, cubic, logarithmic etc. The scatter plot may indicate the type of relation.

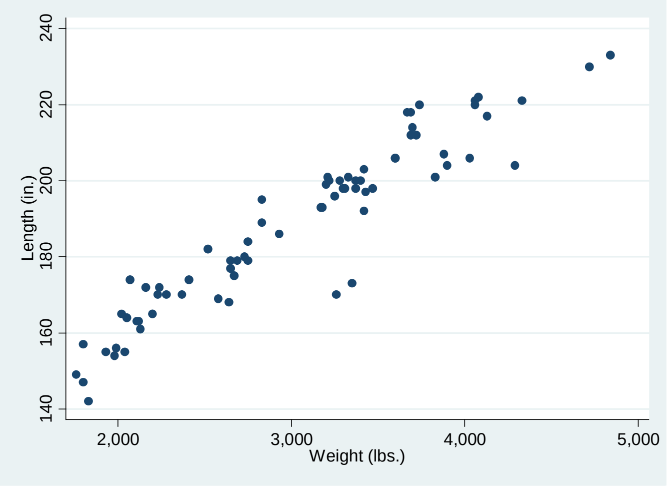
Examples:

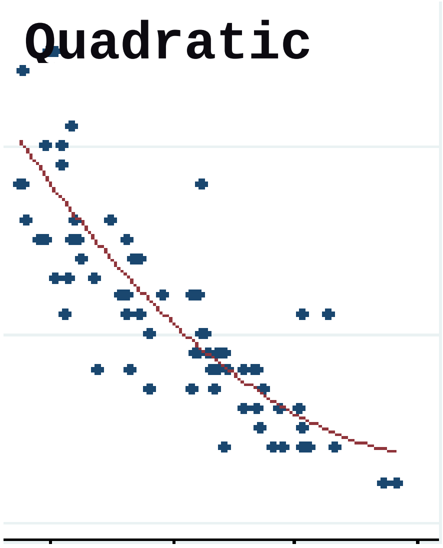
- Marginal Cost Curves: U shaped curve; a quadratic relationship
- Total Cost Curves: The curve is upward sloping, changes its nature at a point of inflection. This may be shown by a cubic equation
- Firms face diminishing returns: This means that the total product curve is not linear. There are three stages of production (study law of variable proportions for detail). In each stage the relationship may be different
- Diminishing marginal utility
- Elasticity changes with price (Demand Curve is not linear)

Other Functional Forms

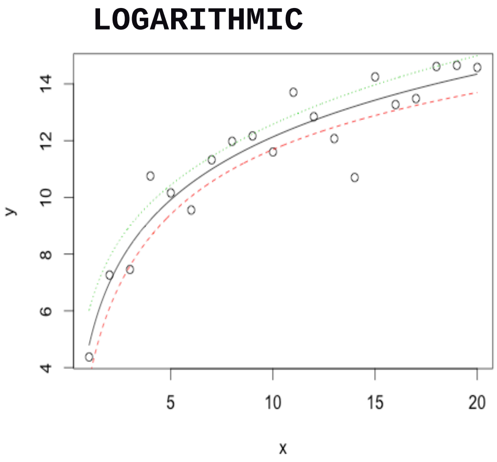
- The scatter plot may indicate the type of relation.

Linear

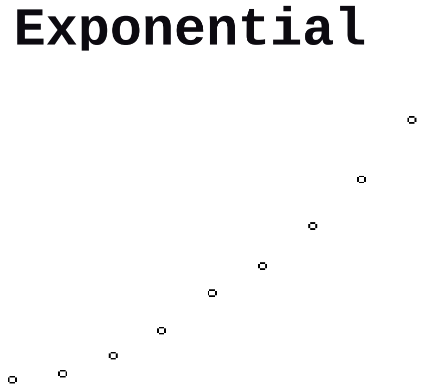




Quadratic



LOGARITHMIC



Exponential

Exponential and Logarithmic Function

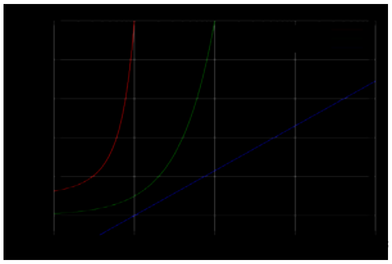
- The exponential functional form is $y = b^x$ where $b > 0$ It can be estimated as $\ln y = x \ln b$
- $y^* = b^* x \;\; (b^* = \ln b\,, y^* = \ln y)$
- This is a regression equation without an intercept(another case: $y = ab^x \; would \; be \; with \; intercept)$
- When estimated, we need to compute the $inverse \; log$ to get the original parameter b

- Logarithmic functions are related to exponential functions
- From the above case, by the definition of logarithm to the base a logarithmic function can be written as $x = log_a y$
- Uses
    - When looking at growth or decay like models of economic growth
    - investment that increases by a constant percentage each time period
    - sales of a company that increase at a constant percentage each year
    - models of the spread of an epidemic

Logs of variables

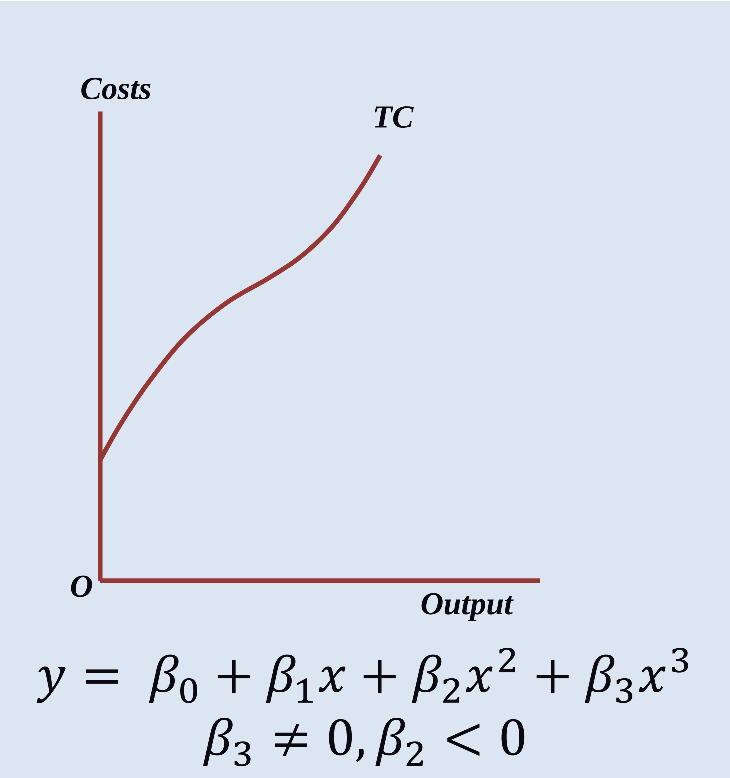
Linear-log, log-linear and log-log forms

- Linear log form: $y = a + b \ln x + e$
- Log-linear or semi-log model: $\ln y = a + bx + e$
- For log-log form please refer to the Cobb-Douglas production function discussed earlier
- The graph of y against x is curved
- The graph of y against $\ln x$ is a straight line
- Example:
    - Short run production: Y experiences diminishing marginal returns with respect to increases in X
    - Bends a concave curve to a straight line

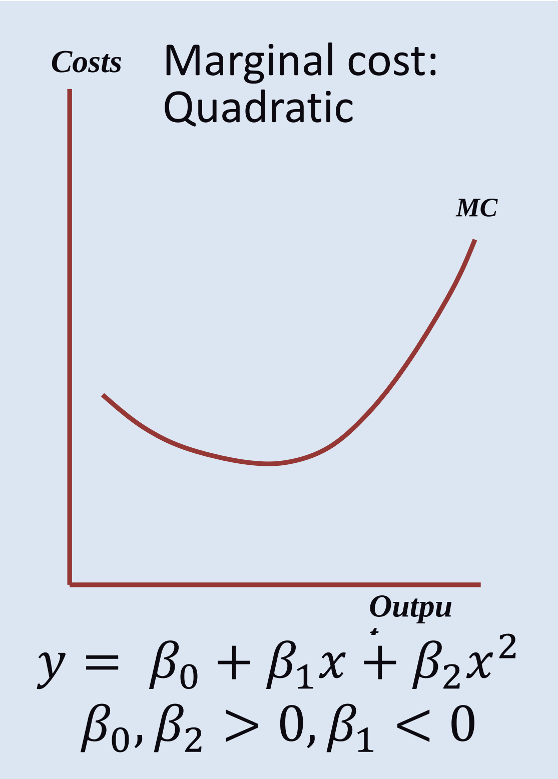


Polynomial functions of higher order

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \,.\,.\,.\,.\,.\,.\,.\,.\,.\,. \beta_k x^k + e$$



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$
$$\beta_3 \neq 0, \beta_2 < 0$$


Marginal cost: Quadratic

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$\beta_0, \beta_2 > 0, \beta_1 < 0$$

The model is linear in parameters and not in variables. Explanation of the squared term with some examples

Dynamic Models: Models with lagged independent Variable

Example 1.

Some times we see that a variable has an impact that continues in time e.g. the multiplier effect (EXPLAIN). Suppose that a change in investment took place in the period $t-2$. It had an impact in time $t-2,\ in\ t-1\ and\ t$

$$\Delta Q = \ \beta_0 + \beta_1 \Delta I_t + \beta_2 \Delta I_{t-1} + \beta_3 \Delta I_{t-2} + e$$

Example 2. Consider the Supply of an agricultural crop depends on the price of the previous time period. The farmers sow keeping in mind the price that they observed last year

$$Q_s = a + bP_{t-1} + e$$

Example 3. A variable may depend on the previous value of itself

$$y_t = a + b\, y_{t-1} + e$$

What is Interaction term?

Example 1:

Suppose that the impact of one independent variable on the dependent variable depends on another independent variable. For example suppose that marks in econometrics (M) depend on your IQ level (I) and on your hard work (H =time spend on study). What if the impact of hard work on marks of econometrics depends on the IQ level. (More intelligent student may need less hard work to achieve the same result).

This can be captured by an interaction term $(I * H)$

$$M = a + b\, I + c\, H + d\, (I * H) + e$$

$(I * H)$ is a new variable computed as the product of $I\ and\ H$.

The derivative of M w.r.t. H depends on $I.\,(c + dI)$

Example 2:

Consider the impact of hypertension (B = blood pressure) and diabetes (S= average level of blood sugar) on the heart of individuals (H = some index of health of the heart). Both are '*Risk Factors*'. The impact of having both risk factors may be greater than the total of the average impact of individual risk factors. This can be captured by an interaction term ($B * S$)

$$H = a + b\,B + c\,S + d\,(B * S) + e$$

$(B * S)$ is a new variable computed as the product of $B\ and\ S$.

The coefficient of the interaction term captures the additional impact of having BOTH risk factors

Lecture 12

Regression on standardized variables

The units in which the regressand and the regressor are expressed effect the interpretation of the coefficients. (diff functional forms). We can avoid this if we standardize our variables. A variable is standardized if we subtract its mean from it and divide by the standard deviation

$$\boldsymbol{X_i^* = \frac{X_i - \bar{X}}{S_X}} \quad , \quad \boldsymbol{Y_i^* = \frac{Y_i - \bar{Y}}{S_Y}}$$

Then the variables will have zero means and unit variances

$$\boldsymbol{\overline{X_i^*} = \overline{Y_i^*} = 0}$$

$$Var(\boldsymbol{X_i^*}) = Var(\boldsymbol{Y_i^*}) = \boldsymbol{1}$$

It does not matter in what unit the variables are expressed

Running the regression $\boldsymbol{Y_i^* = a + b\, X_i^* + e}$ will give ‘a’ and ‘b’ as standardized coefficients often called BETA coefficients. The intercept term will always be zero as $\boldsymbol{\overline{X_i^*} = \overline{Y_i^*} = 0}$

$$(As\ a = \boldsymbol{\overline{Y_i^*} - b\overline{X_i^*}})$$

Interpretation: if the (standardized) regressor increases by one standard deviation, on average, the standardized) regressand increases by β standard deviation units.

Advantages

Standard coefficients' advocates note that the coefficients ignore the independent variable's scale of units, which makes comparisons easy.

Disadvantages

Such standardization can be misleading. The meaning of a standard deviation may vary markedly between non-normal distributions (e.g. skewed)

Theory behind regression

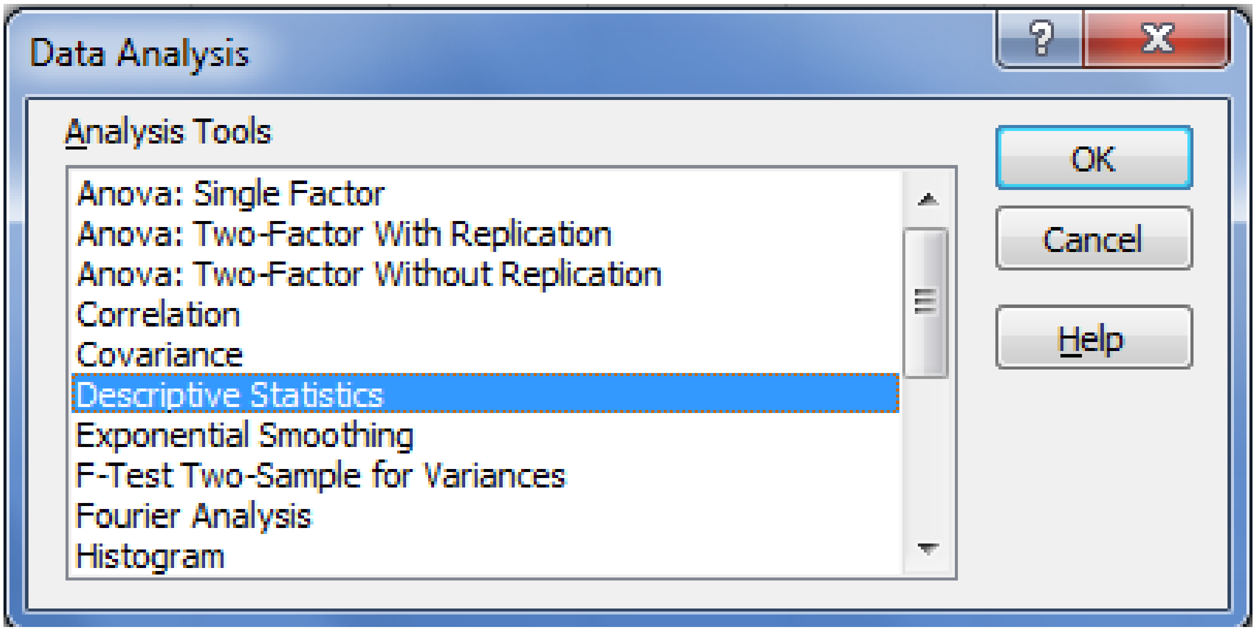
Regression makes sense if there is a sound theory behind. Make Sure to include all necessary predictor variables *(depends on problem statement, theory and previous knowledge)*

Some variables may measure the same things. Either keep one of them or combine them.

Also, Consider the possible interactions

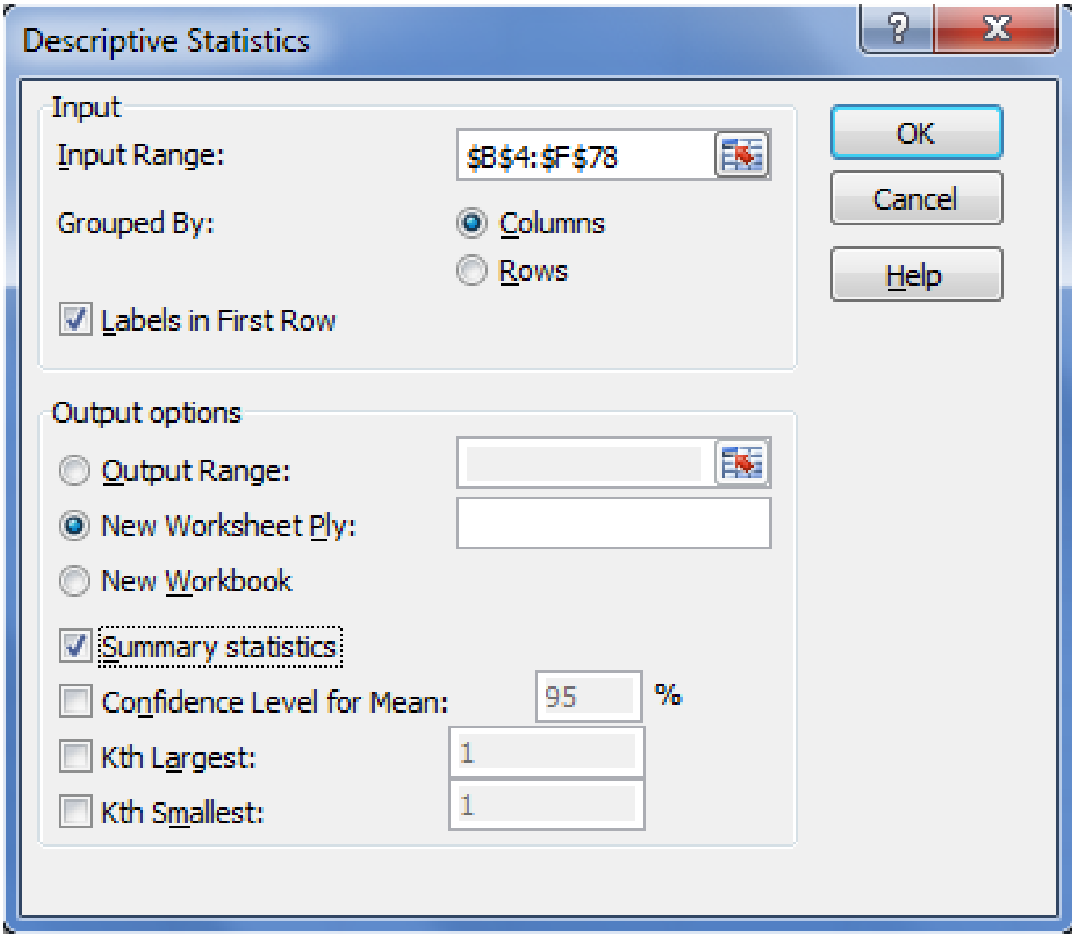
Summary Statistics

After you have entered your data, click Data Analysis in the Data Ribbon in Microsoft Excel, select Descriptive Statistics and click OK



Summary Statistics before regression

Click in the Input Range. Select your data including variable names by dragging your mouse. Check the option Labels in First Row. Check the option Summary Statistics and click OK



Descriptive Statistics

Input

Input Range: $B$4:$F$78

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

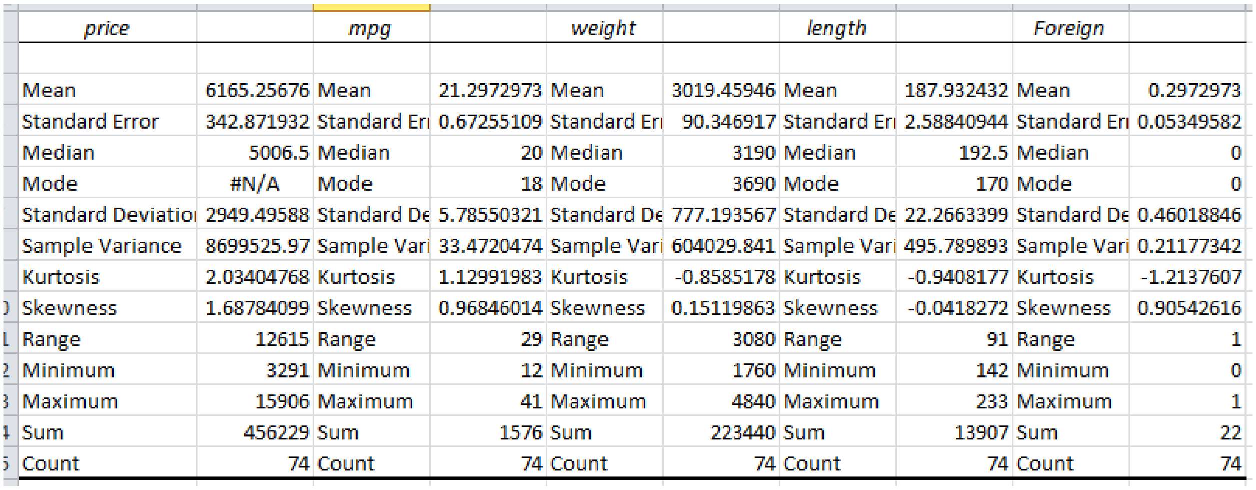
☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

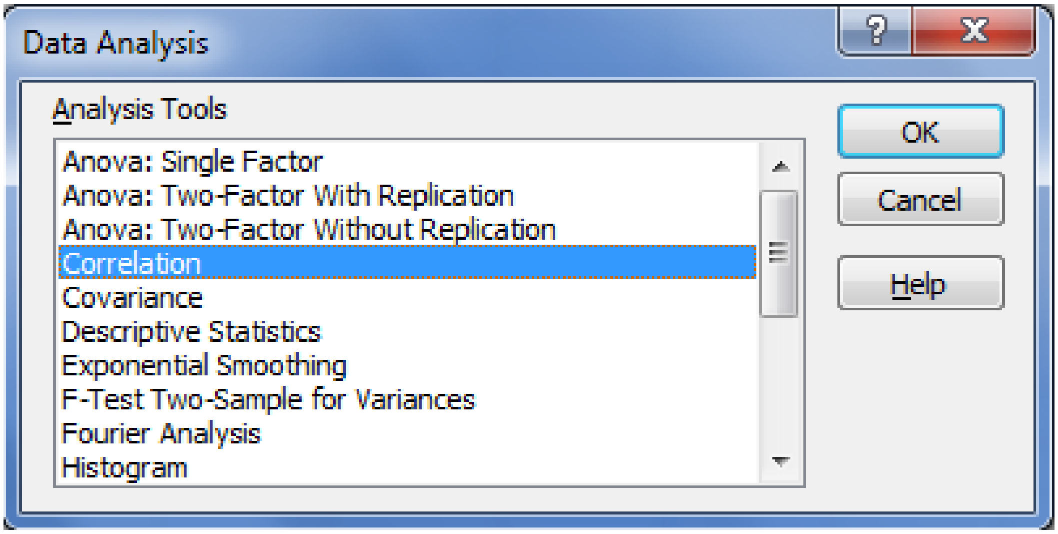
You will get the summary statistics in a new worksheet.



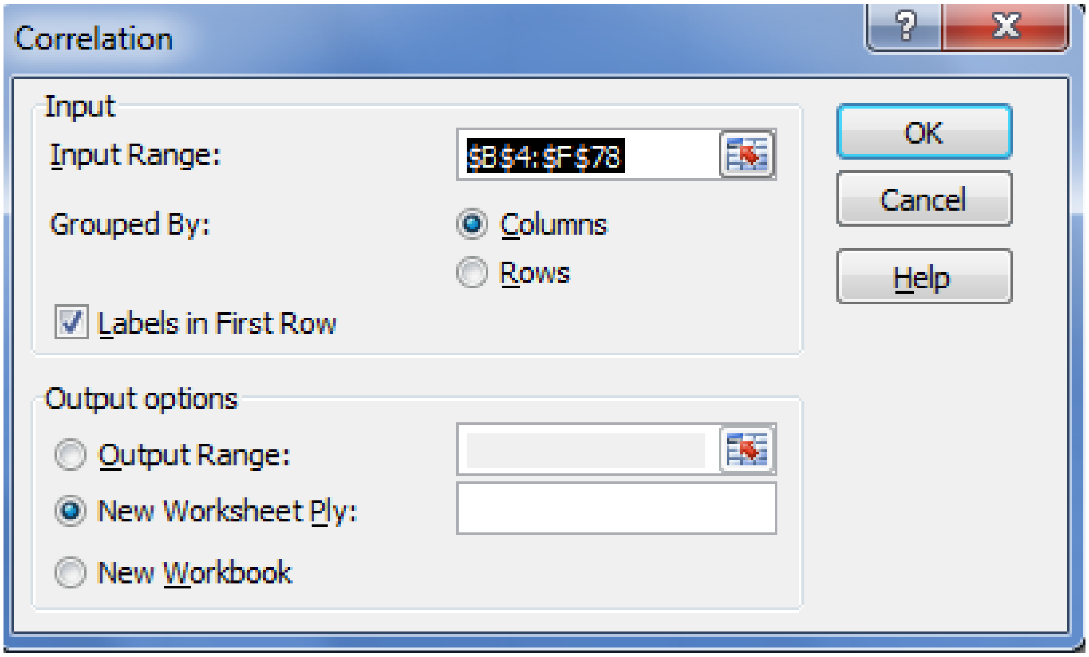
price		mpg		weight		length		Foreign	
Mean	6165.25676	Mean	21.2972973	Mean	3019.45946	Mean	187.932432	Mean	0.2972973
Standard Error	342.871932	Standard Er	0.67255109	Standard Er	90.346917	Standard Er	2.58840944	Standard Er	0.05349582
Median	5006.5	Median	20	Median	3190	Median	192.5	Median	0
Mode	#N/A	Mode	18	Mode	3690	Mode	170	Mode	0
Standard Deviatio	2949.49588	Standard De	5.78550321	Standard De	777.193567	Standard De	22.2663399	Standard De	0.46018846
Sample Variance	8699525.97	Sample Vari	33.4720474	Sample Vari	604029.841	Sample Vari	495.789893	Sample Vari	0.21177342
Kurtosis	2.03404768	Kurtosis	1.12991983	Kurtosis	-0.8585178	Kurtosis	-0.9408177	Kurtosis	-1.2137607
Skewness	1.68784099	Skewness	0.96846014	Skewness	0.15119863	Skewness	-0.0418272	Skewness	0.90542616
Range	12615	Range	29	Range	3080	Range	91	Range	1
Minimum	3291	Minimum	12	Minimum	1760	Minimum	142	Minimum	0
Maximum	15906	Maximum	41	Maximum	4840	Maximum	233	Maximum	1
Sum	456229	Sum	1576	Sum	223440	Sum	13907	Sum	22
Count	74	Count	74	Count	74	Count	74	Count	74

Correlation Matrix

After you have entered your data. Click Data Analysis in the Data Ribbon in Microsoft Excel, select Correlation and click OK



Click in the Input Range. Select your data including variable names by dragging your mouse. Check the option Labels in First Row and click OK.



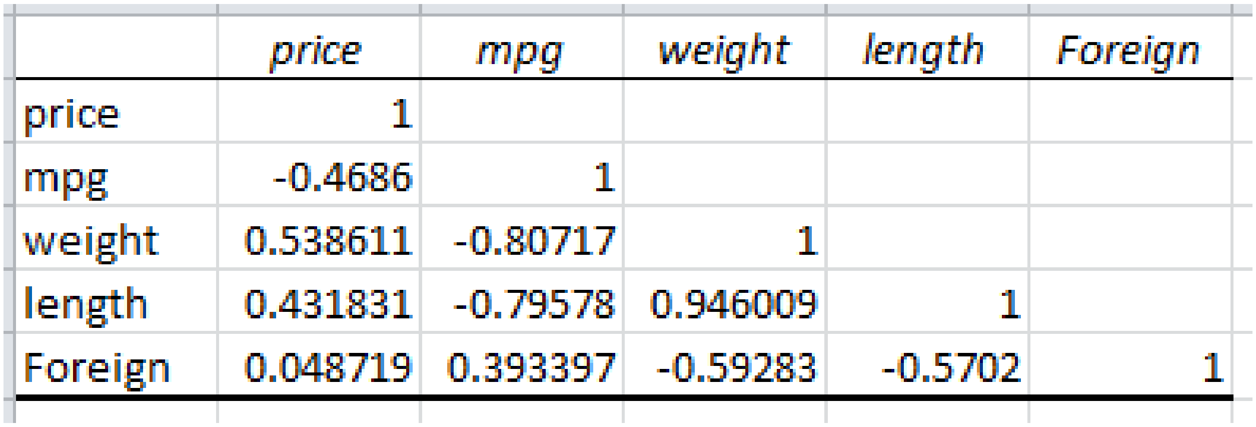
Input Range: $B$4:$F$78

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

● New Worksheet Ply:

The resulting matrix of correlations may help you to understand the relation between variables

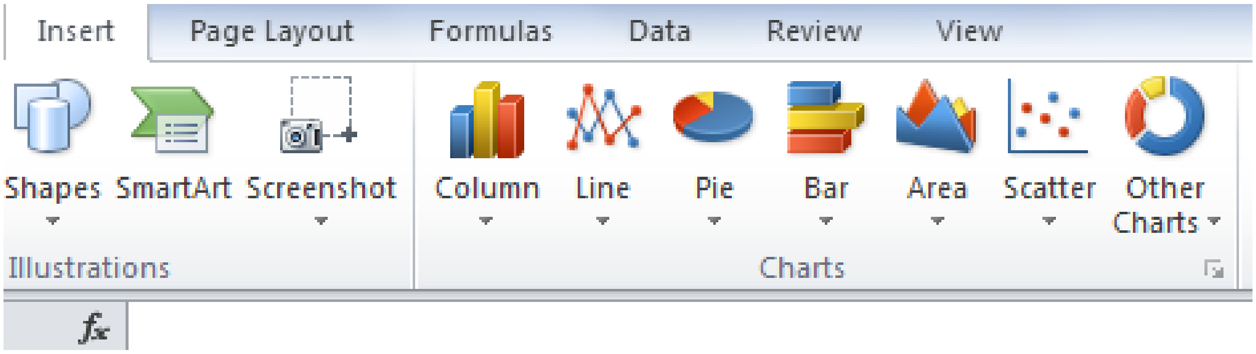


	price	*mpg*	*weight*	*length*	*Foreign*
price	1				
mpg	-0.4686	1			
weight	0.538611	-0.80717	1		
length	0.431831	-0.79578	0.946009	1	
Foreign	0.048719	0.393397	-0.59283	-0.5702	1

Graphs

Some Graphs before regression include Scatter Diagrams, Scatter Diagrams with line plots, Histograms, Pie Charts etc.

In the Insert Ribbon, click on the chart that your require, provide the required information and get your chart



How to present Regression Results in research papers

Every paper uses slightly different strategy. Standard information to report in a regression table includes:

- Dependent variable
- Explanatory variables
- Number of observations, sample period, data labels etc.
- Estimates of intercept and other coefficients
- Standard errors of estimate
- Significance of variables/coefficients
- R-squared and other required statistics

It is common to present more than one regression results.

You can add or drop variables and perform regressions.

Keeping or Dropping Variables

Regression results by adding or including variables vary markedly between non-normal distributions (e.g. skewed). A good strategy would be:

- Keep significant predictors
- Keep insignificant predictors with the expected sign
- Drop insignificant predictors with unexpected sign
- For significant predictors with unexpected sign, keep after review and including or excluding other variables

Examples of presented results

Dependent Variable	**Model 1** **Car Price**	**Model 2** **Car Price**
Millage	855.25 *	852.55 *
Weight	1256.9 **	1246.1 **
Foreign (=1 for imported cars)	950.5 *	855.25 *
Price of Oil	-	-132.25
_constant	46.5*	41.5 *
Number of observations	2500	2500
R-Squared	0.65	0.69

*, ** significance at 1% and 5% respectively

Regression related function in Microsoft Excel: A Brief review

1. The LINEST (50 independent Variables)
2. Data Analysis ToolPak (16 independent variables) RSQ
3. Basic Formulas that are available include: INTERCEPT, SLOPE, TREND

Steps before Regression

- Theory
- Graphs

- Summary statistics
- Model specification
- Regression
- Estimation
- Change of scale if required
- Different functional forms
- Interaction
- Post Regression: Change of scale if required, Different functional forms, Interaction

Lecture 13

Qualitative Independent Variable / Dummy Variables

Remember that we are yet NOT considering Qualitative DEPENDENT variable

Dummy variables have different names

- Categorical Variable
- Qualitative Independent variables (mostly called dummy variables)
- Indicator variable
- Binary variable / dichotomous variable (most often cases)
- Polytomous dummy variables

Such variable may divide the data into mutually exclusive Categories

Dummy Variables

Dummy variables are the variables that help encoding the qualitative variables. They may be binary (with values zero or one)

Examples:

- Gender: We can create a variable with value 0 for female and 1 for male
- Yes/No for existence of a fact (win, admit, disease, marital status. specific years etc.)
- Control group / experiment group / treatment group
- Race
- Educational Categories

It does not matter for what status you assign zero and for what one. For example, We can assign zero for females as well as males in the Gender variable

What happens if we ignore Gender?

Scenario 1: Education and Gender are not correlated

We will have correct Slope estimate but larger errors

Scenario 2: Education and Gender correlated

We will have Biased Slope Estimates and larger errors

Possible Solutions

1. Run separate regression for male and female

What if we need to test the gender difference?

2. Solution

Use dummy variables to capture the influence of gender

Example of binary dummy variable:

- Consider the hourly wages (W) as a function of years of education (E), years of experience (X) and gender (G). Set G=1 if gender="FEMALE" else G=0
- Gender is a qualitative variable (binary) encoded as 0 = male, 1 = female

$$G = \begin{cases} 1 & for\ Women \\ 0 & for\ Men \end{cases}$$

- Performing a linear regression

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + e$$

The data dummy.xlsx provided with lecture notes gives the result as

$$W = 77.76 + 8.29\ E + 28.52\ X - 26.6\ G$$

Interpret other variable

- -26.6 means that female respondent's wage, on the average, is 26.6 rupees less than males

Suppose that E=14, X=5

For a female $W = 77.76 + 8.29\ (14) + 28.52\ (5) - 26.6\ (1) = 309.82$

For a male $W = 77.76 + 8.29\ (14) + 28.52\ (5) - 26.6\ (0) = 336.42$

336.42 − 309.82 = 26.6 is the average difference in wages of male and female with identical characteristics

Exam Model Question

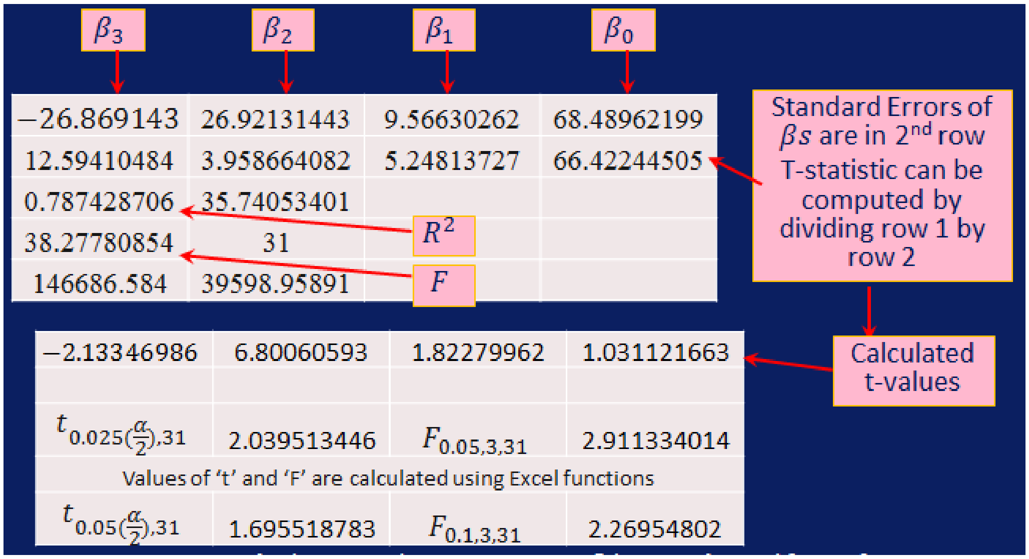
The file in your lesson notes (dummy.xlsx) provides the following information after applying the LINEST function for the regression

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + e$$

−26.869143	26.92131443	9.56630262	68.48962199
12.59410484	3.958664082	5.24813727	66.42244505
0.787428706	35.74053401		
38.27780854	31		
146686.584	39598.95891		

Interpret the results and apply the individual variable significance test and goodness of fit test

For the regression, remember how LINEST provides the results. You need to calculate t-statistic and the table values of t-statistic and the F-statistic. For this, use Microsoft Excel. The results are as follows:



β_3	β_2	β_1	β_0
−26.869143	26.92131443	9.56630262	68.48962199
12.59410484	3.958664082	5.24813727	66.42244505
0.787428706	35.74053401		
38.27780854	31		
146686.584	39598.95891		

Standard Errors of βs are in 2nd row T-statistic can be computed by dividing row 1 by row 2

Calculated t-values

−2.13346986	6.80060593	1.82279962	1.031121663
$t_{0.025(\frac{\alpha}{2}),31}$	2.039513446	$F_{0.05,3,31}$	2.911334014
Values of 't' and 'F' are calculated using Excel functions			
$t_{0.05(\frac{\alpha}{2}),31}$	1.695518783	$F_{0.1,3,31}$	2.26954802

Now you can interpret your results using the above calculations (the coefficient of the dummy variable is significant). Female earn, on average 26.87 less than the 'reference' i.e. Males

Polytomous qualitative variables: Qualitative variables with more than two categories

Race: White, Black, Asian

- We need 3 (binary) dummy variables
- $D_1 = \begin{cases} 1 & if\ Black \\ 0 & in\ NOT\ Black \end{cases}$
- $D_2 = \begin{cases} 1 & if\ Asian \\ 0 & in\ NOT\ Asian \end{cases}$
- $D_3 = \begin{cases} 1 & if\ white \\ 0 & in\ NOT\ white \end{cases}$

D1 (Black=1)	D2 (Asian=1)	D3 (White=1)	Race
0	0	1	White
0	1	0	Asian
0	1	0	Asian
0	1	0	Asian
0	0	1	White
1	0	0	Black
0	1	0	Asian
↓	↓	↓	↓
↓	↓	↓	↓

We will include 2 dummy variables in the regression i.e. one less than the total categories. We will include 2 dummy variables in the regression i.e. one less than the total categories. If we include on two dummies in regression the third one is called reference variable.

Race	D_1	D_2
White	0	0
Asian	0	1
Black	1	0

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4 D_1 + \beta_5 D_2 + e$$

For White

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4(0) + \beta_5(0) + e$$

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + e$$

For Asian

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4(0) + \beta_5 D_2 + e$$

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_5 D_2 + e$$

For Black

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4 D_1 + \beta_5(0) + e$$

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4 D_1 + e$$

Interpretation

Example: dummy.xlsx provided with your lecture notes is used to perform such regression

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4 D_1 + \beta_5 D_2 + e$$

Reference variable is ‘White’

The results are

Dependent Variable: hourly wages			
Observations	35	R-squared	0.85
		F	33.83
	Coefficient	S. Error	t-values
constant	74.84	60.11	1.25
Education	11.53**	4.67	2.47
Experience	26.11*	3.51	7.44
Female	-31.49*	11.10	-2.84
Black (ref. White)	-45.17*	13.15	-3.44
Asian (ref. White)	-35.43**	13.96	-2.54
*, ** significant at 1% & 5% respectively			

Interaction with Dummy Variables

Interaction:

- Use the interaction when the effect of one independent variable depends on the value of the other independent variable.

Example:

- Effect of race may be greater in case of black race
- Effect of education on hourly wages may be different for males and female respondents

Interaction & Correlation:

- Correlation: If independent variables are related to each other
- Interaction: If the effect of one independent variable depends on some other independent variable

- Variables may interact weather they are correlated or no

Example:

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4 (G * E) + e$$

The new regressor is a function of G and E but not a linear function so we should not fear of perfect collinearity

For Men (G = 0)

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 (0) + \beta_4 (0 * E) + e$$

Or

$$W = \beta_0 + \beta_1 E + \beta_2 X + e$$

For Women (G = 1)

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 (1) + \beta_4 (1 * E) + e$$

Or

$$W = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) E + \beta_2 X + e$$

In this case we need to take interaction as a product with all dummy variables

Example:

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4 D_1 + \beta_5 D_2 + e$$

With interaction of race and gender becomes

$$W = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 G + \beta_4 D_1 + \beta_5 D_2 + \beta_6 (G * D_1) + \beta_7 (G * D_2) + e$$

For Men (G = 0)

White Men: $W = \beta_0 + \beta_1 E + \beta_2 X + e$ (as both D_1& D_2 are zero)

Black Men: $W = (\beta_0 + \beta_4) + \beta_1 E + \beta_2 X + e$ (as $D_1 = 1$)

Asian Men: $W = (\beta_0 + \beta_5) + \beta_1 E + \beta_2 X + e$ (as $D_2 = 1$)

For Women (G = 1)

White Women: $W = (\beta_0 + \beta_3) + \beta_1 E + \beta_2 X + e$

Black Women: $W = (\beta_0 + \beta_3 + \beta_4 + \beta_6) + \beta_1 E + \beta_2 X + e$ (as $\mathrm{G} = D_1 = 1$)

Asian Women: $W = (\beta_0 + \beta_3 + \beta_5 + \beta_7) + \beta_1 E + \beta_2 X + e$ (as $\mathrm{G} = D_2 = 1$)

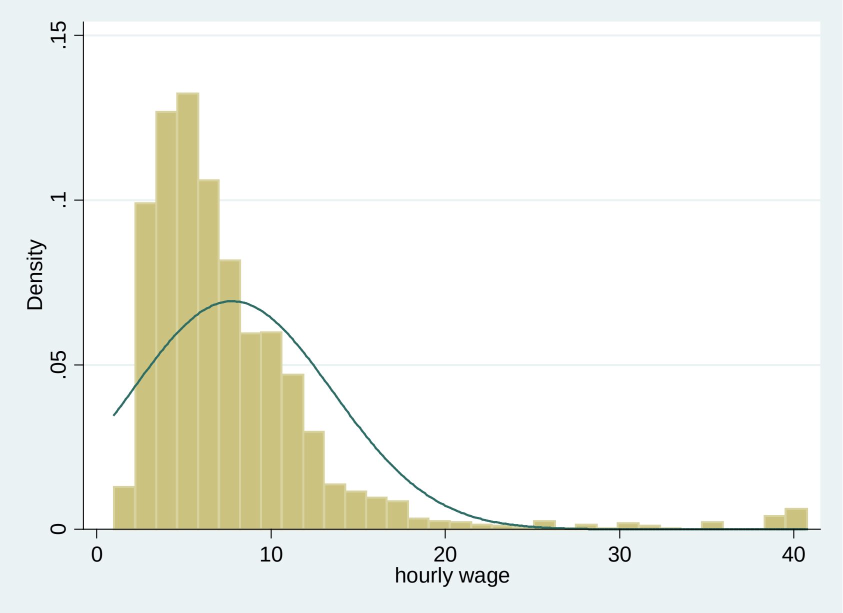
Lecture 14

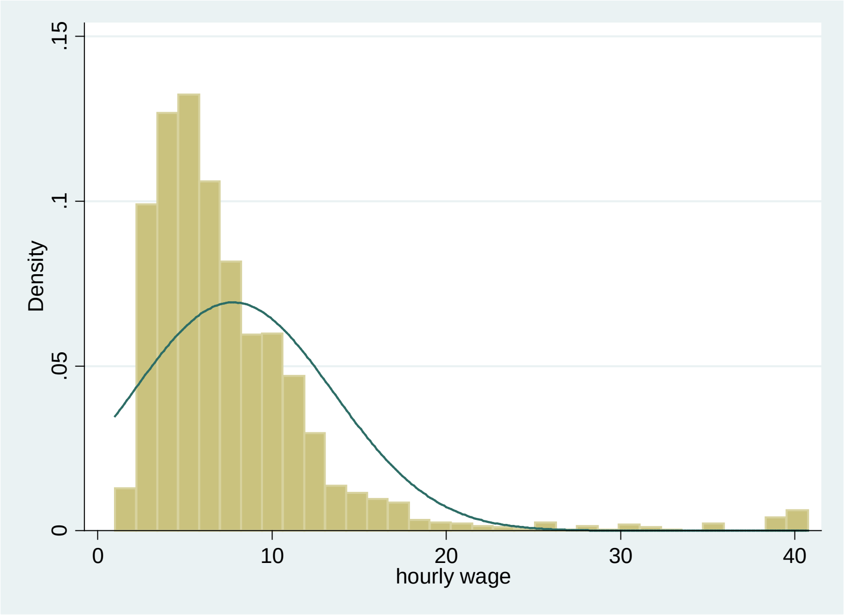
Transforming Variables in Regression

Using logs in regression analysis

Logarithms could be used because of:

- Positively Skewed Distribution of Variable (take natural log)
- When residuals are skewed
- When change in either or both (LHS & RHS) are related in percentage terms (when theory indicates)
- To linearize a relationship





Logs in regression

Consider relation of wages with tenure and a dummy ‘race’ using nlsw88.dta (provided with stata)

Dependent	**Wage**	**ln(Wage)**	**ln(Wage)**
R-squared	0.039	0.1043	0.1244
F-Statistic	45.56	129.66	154.69
variable	**coefficients**		
tenure	0.1898*	0.0317*	
ln(tenure)			0.1624*
race	-1.07	-0.144*	-0.1366*
*** Significant at 1%**			

Treatment of Missing Data

Why data may be missing?

- Attrition due to natural process
    - Death (panels)
    - Dropouts
    - Migration
    - New entries (like new countries)
- Data not available legally (in some countries)
- No Response / Refusal of respondents
- Conditional questions in surveys
    - If student then - - -
    - If employed then - - -
    - If living in couple then - - -
- Data Collection Issues
- Encoding and Recoding

Probability of Missing Data

- Some groups are more likely to have missing values
    - Businessmen (particularly service) as compared to salaried individuals
    - Rural areas as compared to Urban (in developing countries)
    - Less education as compared to educated
    - Developing vs developed countries (documentation)
- Some Variables may have more missing values

  - Income (high income groups)
  - Area-specific variable (like use of AC in rural areas)
  - Variables related to Certain taboos (e.g. drinking in Pakistan, prostitution, drugs)

Missing Data Mechanism: probability distributions

1. MCAR: Missing Completely at Random

Probability that Y values missing is neither dependent on Y nor on X

$$Pr(Y\ is\ missing\ |X,Y) = Pr(Y\ is\ missing)$$

2. MAR: Missing at Random

Probability that Y values missing does not depend on Y but depends on X

$$Pr(Y\ is\ missing\ |X,Y) = Pr(Y\ is\ missing|X)$$

Example: the probability of missing income depends on occupation, but within each occupation, the probability is not dependent on income

3. Ignorable MAR

Missing data mechanism is said to be ignorable if data is Missing At Random but the parameters governing the missing data are distinct from them ones being estimated

- It is, in fact , MAR

4. MNAR: Missing Not at Random

Probability that Y values missing depends on a variable that is missing

- Heckman Regression for sample selection
- Estimation of NMAR missing data
  - Data contains no information
  - Results are sensitive to choice of model

Dealing with missing data: Deletion Methods

1. List wise deletion (complete case analysis)

Only analyze row of data where there are no missing values

i.e. delete or do not include the row in which there are missing values (software like stata may automatically do that)

Example: all variable of the response with even one missing is deleted all together

Advantages: Simple

Disadvantages: Reduces statistical power

2. Pairwise deletion (available case analysis; analysis to analysis)

We do not only delete the row in which there is missing value but we delete on analysis to analysis base

Example: diff observations for correlation, regression

Advantages: uses all information possible with each case

Disadvantages: Sample is different each time

Imputation

1. **Single Imputation**

a) Substitute Mean / Median / Mode

Method:

- Just substitute constant in place of missing values

Problems:

- *Run complete case*
- *Weaker covariance & Reduced variability*

b) Dummy variable control

Method:

- Generate indicator of rows with missing value
- Use single imputation to fill in
- Perform regression with dummy variable control

Problems:

- *Estimates are biased and this method is not theoretically driven*

c) Regression Imputation

Method:

- Use regression to estimate value to substitute

Problems:

- *Overestimates model fit*
- *Weaker variance*

Model Based Imputation

1. Maximum Likelihood

Method:

- Identify set of parameters that produces the maximum log likelihood.

ML: value most likely to have resulted in the observed data

- *Uses full estimation; unbiased estimates*
- *Standard Errors biased downwards*

2. Multiple Imputation

Method:

- Data filled-in by applying specific regressions

- Experiment is repeated n number of times generating separate data sets
- Perform regression on each dataset
- Pool the datasets to performed regression

Advantages

- *Good variability*

Disadvantages

- *Errors possible when specifying regression models*

Lecture 15

Multicollinearity

Perfect Multicollinearity: Definition

The above can be defined as Perfect or exact linear relationship between a pair or more of the explanatory variables. K-variable linear regression ($X_0 = 1$ for all observations to allow the intercept term)

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots\ldots \beta_k X_k + e_i$$

where we are going to ignore subscripts for convenience. Exact or perfect linear relationship exists if $\boldsymbol{\lambda_0 X_0 + \lambda_1 X_1 + \lambda_2 X_2 + \cdot\ \cdot\ \cdot + \lambda_k X_k = 0}$ Where $\boldsymbol{\lambda i}$ cannot be zero at the same time.

Near or Imperfect Multicollinearity: Definition

In practice, we rarely observe perfect Multicollinearity but a degree of Multicollinearity

For $Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots\ldots \beta_k X_k + e_i$ where we are going to ignore subscripts for convenience.

Near or imperfect linear relationship exists if $\lambda_0 X_0 + \lambda_1 X_1 + \lambda_2 X_2 + \cdot\ \cdot\ \cdot + \lambda_k X_k + v_i = 0$

Understanding Perfect Multicollinearity

Exact or perfect linear relationship exists if $\lambda_0 X_0 + \lambda_1 X_1 + \lambda_2 X_2 + \cdot\ \cdot\ \cdot + \lambda_k X_k = 0$

Where $\boldsymbol{\lambda i}$ can not all be zero at the same time. If we solve for any variable, it will be in a perfect linear combination of other independent variables e.g. for convenience, let all $\boldsymbol{\lambda i}$ be zero except $\boldsymbol{\lambda 1}$ and $\boldsymbol{\lambda 2}$

Then $\boldsymbol{X1 = \frac{\lambda 2}{\lambda 1} X2}$ **(**e.g. $\boldsymbol{X1 = 5X2}$)

Then both variables can be expressed in linear combinations of each other. Let us see what happens in this case

$$\boldsymbol{X1 = \frac{\lambda 2}{\lambda 1} X2} \text{ (e.g. } \boldsymbol{X1 = 5X2}\text{) Or } \boldsymbol{X2 = \frac{1}{5} X1}$$

Then $\quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$ becomes

$$Y = \beta_0 + \beta_1(5X_2) + \beta_2 X_2 + e_i$$

$$Y = \beta_0 + (5\beta_1 + \beta_2)X_2 + e_i$$

$$Y = \beta_0 + \beta_m X_2 + e_i$$

i.e. a simple regression where Y depends on X_2. Hence we cannot estimate all the parameters. Also Software may drop one of the variables. In the example the variable X_1 (X_2 = $5X_1$) has been dropped and a zero coefficient is shown. In fact a simple regression line Y on X_2 has been estimated.

Y	**X_1**	**X_2**
10	10	50
20	11	55
30	16	80
40	15	75
50	19	95
60	21	105
70	22	110
80	24	120
90	27	135
100	31	155
Function used =LINEST(F8:F17,G8:H17,TRUE,TRUE)		
0.882638215	0	-31.4985451
0.051409833	0	5.301697428
0.973576698	5.220060351	
294.7630701	8	
8032.007759	217.9922405	
t-statistic calculated by us are below		
17.16866536		-5.94121893

Let us now slightly change the values of X. The new data is:

Y	**X_1**	**X_2**
10	10	52
20	11	54
30	16	77
40	15	78
50	19	93
60	21	104
70	22	109
80	24	119
90	27	136
100	31	153
=LINEST(J8:J17,K8:L17,TRUE,TRUE)		
1.49	-2.93	-33.31
0.86	4.24	4.85
0.98	4.67	
185.94	7.00	
8097.58	152.42	
t-statistic calculated by us are below		
1.74	-0.69	-6.86

Note the following:

- We have slightly changed the values of X_2
- This time all the parameters are estimated
- Standard errors are large (less precision in estimating parameters)
- t-statistic are low
- R-square is high but the t-statistic are low
- Coefficients of parameters do not seem to be significant but he F-Test show a good fit.

Sources of Multicollinearity

1. Narrow Subspace or The data collection method: sampling over a limited range or subgroups of population (correlation may exist only in the subgroup
2. Natural Constraints on the model or population :
    - e.g. If we regress GDP on exports and imports, high imports normally means high exports as well in most of the countries
    - in a sample survey both income and status of house are included as explanatory variables.
    - Model specification: for example, adding polynomial terms especially when X has a small range
3. Over-determined model: Large number of explanatory variables with very low degree of freedom
4. Common Time Series Trend: the explanatory variables may share a common trend in time.

Theoretical consequences of Multicollinearity

1. OLS estimators remain unbiased: no violation of BLUE property
2. Interpretation of coefficients is not independent: simple interpretation does not seem to be valid
3. Multicollinearity is a data deficiency problem: sometimes it is difficult to increase the sample size (cost, time)
4. Multicollinearity is a sample phenomenon: especially with non-experimental data (occurring naturally like GDP and its determinants.
5. We need larger samples: larger than without MC

Practical consequences of Multicollinearity

- Large variances and standard errors of coefficients: when correlation between pairs of explanatory variables is high; or there is fall in precision of estimators
- Wider confidence intervals: as a consequence of larger standard errors
- Insignificant t-ratios: as a consequence of larger standard errors (H_0 accepted)
- High R^2 but low t: seem to be contradictory results.
- Estimators and standard errors are very sensitive to changes in data: unstable
- Wrong signs of coefficients: May not be according to economic and finance theory.
- Difficulty in assessing the individual contribution of regressors to the explained variation: due to correlated regressors

Detecting Multicollinearity

Remember that

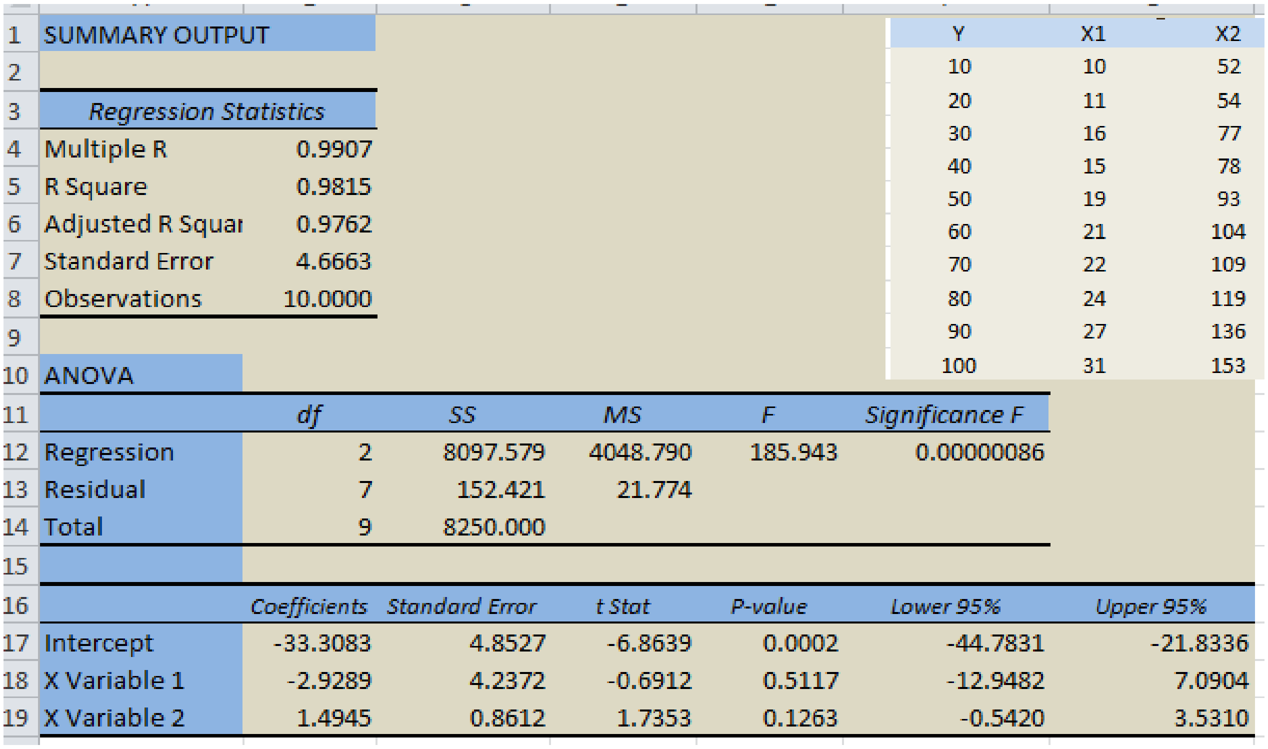
1. Multicollinearity is question of DEGREE not of Presence alone
2. Multicollinearity is a sample phenomenon
3. It is a data deficiency problem

There are three ways.

1. Look at the symptoms
2. Look at the correlation matrix
3. Calculate VIF or Tolerance

1. Look at the symptoms / indicators

- High R^2 but low t-statistic
- Large standard errors
- Individual variables are not significant but model seems to be a good fit



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9907
R Square	0.9815
Adjusted R Squar	0.9762
Standard Error	4.6663
Observations	10.0000

Y	X1	X2
10	10	52
20	11	54
30	16	77
40	15	78
50	19	93
60	21	104
70	22	109
80	24	119
90	27	136
100	31	153

ANOVA

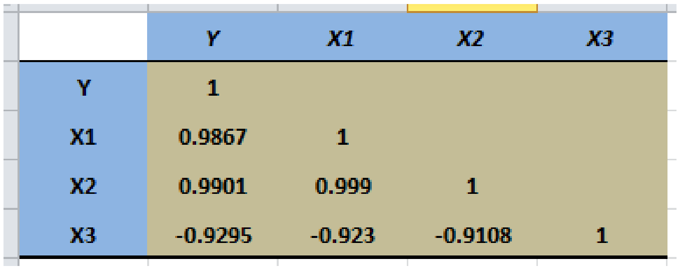
	df	SS	MS	F	Significance F
Regression	2	8097.579	4048.790	185.943	0.00000086
Residual	7	152.421	21.774		
Total	9	8250.000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-33.3083	4.8527	-6.8639	0.0002	-44.7831	-21.8336
X Variable 1	-2.9289	4.2372	-0.6912	0.5117	-12.9482	7.0904
X Variable 2	1.4945	0.8612	1.7353	0.1263	-0.5420	3.5310

2. Look at the Correlation Matrix

High pairwise correlations between the explanatory variables (e.g. greater than 0.8)

Y	X1	X2	X3
10	10	52	26
20	11	54	23
30	16	77	23
40	15	78	25
50	19	93	22
60	21	104	21
70	22	109	19
80	24	119	18
90	27	136	19
100	31	153	17



	Y	*X1*	*X2*	*X3*
Y	1			
X1	0.9867	1		
X2	0.9901	0.999	1	
X3	-0.9295	-0.923	-0.9108	1

3. Estimate and look at VIF or Tolerance

Method

- Auxiliary Regressions: Regress each explanatory variable on other explanatory variables and find the coefficient of determination
- You can test the significance of R^2 (F or goodness of fit)
- Estimate Tolerance or VIF

$$Tolerance = 1 - R^2$$

- $Variance\ Inflation\ Factor$

$$VIF = \frac{1}{1 - R^2}$$

- Decide according to the values of Tolerance / VIF

Decision about the degree of Multicollinearity

	Tolerance	VIF
Problematic Multicollinearity	≤ 0.1	≥ 10
Mild Multicollinearity	$0.1 < T < 0.2$	$5 < VIF < 10$
Nearly No Multicollinearity	≥ 0.2	≤ 5

Important Points to note

- VIF or Tolerance for the variable is calculated by the R^2 of the auxiliary regression of the variable regressed on all other explanatory variables
- One single variable is not responsible so we may estimate auxiliary regressions one less than the explanatory variable

Lecture 16

Multicollinearity: Remedial Measures

1. Do nothing

If Multicollinearity is mild or If the purpose is only forecasting. If data deficiency is the problem, we have no choice over data. It is better to try to increase the data by extending the sample if possible. Multicolinearity is not a problem if theory permits us to estimate the missing coefficient e.g. in Cobb-Douglas production function, if we assume constant returns to scale the either of alpha and beta can be estimated if one is estimated by regression.

2. Drop one of the variables

Drop the one that is less significant or drop the one with larger VIF. But this may lead to wrong model specification or may go against theoretical considerations (e.g. dropping price of substitute in demand function). An example of dropping variables is of import and export in GDP equation

3. Transform the variable

Combine the variables (we just add exports and imports to get a new variable labeled as openness). Another option is to convert the variables (import = f(GNP, CPI) we can divide by CPI (real imports = f (real GNP)); but error term may become heteroskedastic. Another way is to use first difference form (loss of one observation)

$$\boldsymbol{Yt - Yt-1 = \beta 2(X2t - X2, t-1) + \beta 3(X3t - X3, t-1) + vt}$$

(may not be appropriate in cross sectional data; has no sense)

Other options include:

4. Get additional data and increase the sample size or 5. Combine cross section and time series (pool) or

6. Use of panel data or 7. Use ridge regression, factor analysis etc.

Multicollinearity

Examples from Business and Economics

Consider the data given in **MC.xlsx**. It contains data on quantity demanded, prices, and monthly income in thousands and prices of two different substitutes. First example has 50 observations in total. But first let us consider a small example

The data is as follows:

Y	X_1	X_2	X_3
10	**10**	**52**	**26**
20	**11**	**54**	**23**
30	**16**	**77**	**23**
40	**15**	**78**	**25**
50	**19**	**93**	**22**
60	**21**	**104**	**21**
70	**22**	**109**	**19**
80	**24**	**119**	**18**
90	**27**	**136**	**19**
100	**31**	**153**	**17**

Auxiliary regressions are produce using Microsoft Excel:

<table>
<tr><th colspan="2">Auxiliary Regressions</th><th>Important</th></tr>
<tr><td colspan="2">Regression: X1 on X2 & X3</td><td rowspan="15"><b>• First two variables show MC</b>
<b>• Look at F</b>
<b>• Look at T/VIF</b></td></tr>
<tr><td colspan="2">$X1 = 4.977 + 0.188\,X2 - 0.176\,X3$</td></tr>
<tr><td>R^2 = 0.998</td><td>F=1848.5</td></tr>
<tr><td>Tolerance =</td><td>0.002</td></tr>
<tr><td>VIF =</td><td>500</td></tr>
<tr><td colspan="2">Regression: X_2 on X_1 & X_3</td></tr>
<tr><td colspan="2">$X2 = -22.11 + 5.24\,X1 + 0.794\,X3$</td></tr>
<tr><td>R^2 = 0.9978</td><td>F= 1608.6</td></tr>
<tr><td>Tolerance =</td><td>0.0022</td></tr>
<tr><td>VIF =</td><td>454.5455</td></tr>
<tr><td colspan="2">Regression: X_3 on X_1 & X_2</td></tr>
<tr><td colspan="2">$X3 = 28.97 - 2.03\,X1 + 0.3296\,X2$</td></tr>
<tr><td>R^2 = 0.89</td><td>F= 28.43</td></tr>
<tr><td>Tolerance =</td><td>0.11</td></tr>
<tr><td>VIF =</td><td>9.090909</td></tr>
</table>

Multicollinearity: Example-1 from Business and Economics

Let us use the file MC.xlsx and Fit different Regressions. First Look at the correlation matrix.

Correlation Matrix

	Q	*price*	*income (000)*	*price of sub.-1*	*price of sub.-2*
Q	1				
Price	-0.94144	1			
income (000)	-0.04567	0.309407	1		
price of sub.-1	0.071323	0.234771	0.793596	1	
price of sub.-2	0.017076	0.274298	0.79072	0.960501	1

Below are different models that we ran. The first one contains only 10 observations. The second increases the sample size and the third is when we drop a variable. Let us see what happens.

Model 1 (N=10)

R Square=0.9106   F=12.73**

	Coefficients	*VIF*	*Comment*
Intercept	977.7725 ***		
Price	-0.2560***	2.5	No MC
Income	2.33**	5.12	No MC
Price of Substitute-1	-0.2384	33.04	MC, wrong sign, insignificant
Price of Substitute-2	0.6310	22.75	MC, insignificant

Model 2 (N=50)

R Square=0.978 F=508.41*

	Coefficients	*VIF*	*Comment*
Intercept	993.6106 *		
Price	-0.499*	1.13	No MC
Income	0.498***	2.88	No MC
Price of Substitute-1	0.1924*	13.74	MC
Price of Substitute-2	-0.0912	13.6	MC, wrong sign

In the second model note that the VIF has decreases but still is greater than 10. Now let us drop one variable.

Model 3 (N=50)

R Square=0.978 F=687.98*

	Coefficients	*VIF*	*Comment*
Intercept	985.77*		
Price	-0.5*	1.13	No MC
Income	.4768***	2.88	No MC
Price of Substitute-1	0.1663*	13.6	No MC

Multicollinearity Example-2 from Business and Economics

- Data on GDP, Population, GFCF, Exports, Imports (2000-2012 Pakistan)
- Downloaded from WDI

Dependent Variable	**GDP**		
R-square	0.9927	F	271.4
Expl Variables		significance	VIF
Population	1367.5	*	4.61
GFCF	0.196		7.45
Exports	-0.247		10.28
Imports	0.879		14.36
Openness			
constant	large value		

Result: No significance, high VIF, wrong signs

After dropping imports the situation becomes:

Dependent Variable	**GDP**		
R-square	0.9897	F	289.17
Expl Variables		significance	VIF
Population	1307.9	*	4.53
GFCF	1.118	**	4.15
Exports	0.254		2.82
Imports			
Openness			
constant	large value	*	

New Results show: better significance, low VIF, signs correct now

Another option is to transform the variables. Let us generate a new variable, volume of trade as a proxy for openness and define it as the sum of imports and exports.

Dependent Variable	GDP		
R-square	0.7947	F	19.36
Expl Variables		significance	VIF
GFCF	1.06		5.36
Openness	1.809	**	5.36
constant	large value		

Results: openness is significance, low VIF indicates no Multicollinearity, the signs of the coefficients are also correct now however we may need other variables.

Lecture 17
Heteroskedasticity

Heteroskedasticity: Background

Consider the multiple regression line

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ \ldots\ldots\ldots\ldots \beta_k X_k + e_i$$

If we estimate the coefficients by OLS, we assume that

$$Var(e_j) = \sigma^2 \ for\ all\ j$$

or

$$Var(e_j | X_1, X_1, . \ . \ . \ . \ . X_1) = \sigma^2$$

This is called Homoskedasticity (Absence of Heteroskedasticity). The alternative spelling is use are Homoscedasticity.

Heteroskedasticity: Meaning

$$Var(e_j | X_1, X_1, . \ . \ . \ . \ . X_1) = \sigma^2$$

This may means that the variance of errors in subgroups of the sample may not remain the same

So if

$$Var(e_j | X_1, X_2, . \ . \ . \ . \ . X_K) \neq \sigma^2$$

$$or$$

$$Var(e_j | X_1, X_2, . \ . \ . \ . \ . X_K) = \boldsymbol{\sigma_i^2}$$

The subscript 'i' shows the variance to be variable and this is called Heteroskedasticity.

Hetero means 'different' and *Skedasis* means 'dispersion', Spread or variance. A *homoskedastic* error is one that has constant variance. Equivalently, this means that the dispersion of the observed values of Y around the regression line is the same across all observations. Hence a *heteroskedastic* error is one that has a non-constant variance.

Heteroskedasticity: Causes and Examples

I- Variance of Errors may increase as the value of explanatory variable increases

Examples 1:

Expenditure on vacation = f (family income)

$$Low\ income\ \rightarrow low\ expenditure\ on\ vacation$$

$$High\ income\ \rightarrow possibly\ greater\ expenditure\ on\ vacation$$

With high income the variability as well as errors increases. High income is a necessary but not sufficient condition for greater expenditure on education. HSK is likely in such situation.

Examples 2:

Profitability = f (annual sales, liquidity, CCC)

$$small\ firms \rightarrow smaller\ annual\ sales \rightarrow less\ variation$$

$$Large\ Firms \rightarrow Larger\ sales \rightarrow greater\ variation$$

Being a large firm is a necessary but not sufficient condition for greater annual sales. HSK is likely in such situation.

II- Subpopulation may have different effect

Examples 1:

Expenditure on vacation = f (family income)

The Effect of income on expenditure on vacation may be different in different Localities or may be different for different races so we need to use dummy variables to capture this difference.

Examples 2:

Profitability = f (annual sales, liquidity, CCC)

The Effect of sales on profitability may be different in different types of organizations due to different cost conditions or may be different for different areas due to different tax regulations so here also we need to use dummy variables

III- Measurement Errors

Example:

Some respondents may provide more accurate information for example high income groups may report income less than actual

IV- Wrongly specified model

Examples:

Instead of using Log of Y, you may be using Y. Instead of using both X and square of X, you may be using only X

V- Missing Variables

Examples:

Important Variables may be missing or instead of using both X and square of X, you may be using only X

VI- HSK is more likely in Cross-sectional data

Examples:

Savings = f (income, wealth)

People with higher income are likely to save a greater percentage of income

VII- Different Quality of Data

Examples:

In cross country data, different quality of data may be reported e.g. some developing countries may provide lower or higher values than actual.

VIII- Time Dependence of variables

Examples:

Seasonal component in variables: In some countries the variation in electricity prices increases in summer

Heteroskedasticity: Summary of Causes

I- Variance of Errors may increase with explanatory variable

II- Subpopulation may have different effect

III- Measurement Errors

IV- Wrongly specified model

V- Missing Variables

VI- HSK is more likely in Cross-sectional data

VII- Different Quality of Data

VIII- Time Dependence of variables

Consequences of Heteroskedasticity

1. OLS estimators remain unbiased and consistent
2. However the distribution of estimators is effected increasing the variance of the distributions

   (Inefficient estimators: minimum variance property violated)
3. Estimates of Variance are biases (formula require changes)
4. Standard Errors / Confidence intervals not correct
   i. wrong conclusions on t and F
   ii. significance tests too high or too low
   iii. OLS gives more weight to larger errors (min Sum E square) SO IF HSK then overemphasized extreme values
   iv. Extreme values, in fact, contain less information so should be given less importance

Detection of Heteroskedasticity

The following can be used to detect Heteroskedasticity

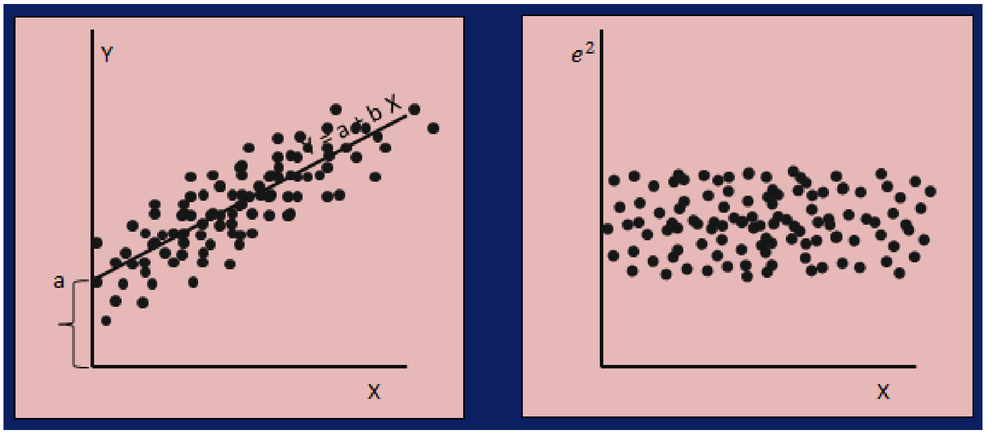
1. Visual Inspection (Graphs)
2. Formal Tests (Various Tests are available)

Detection of Heteroskedasticity: Visual Inspection

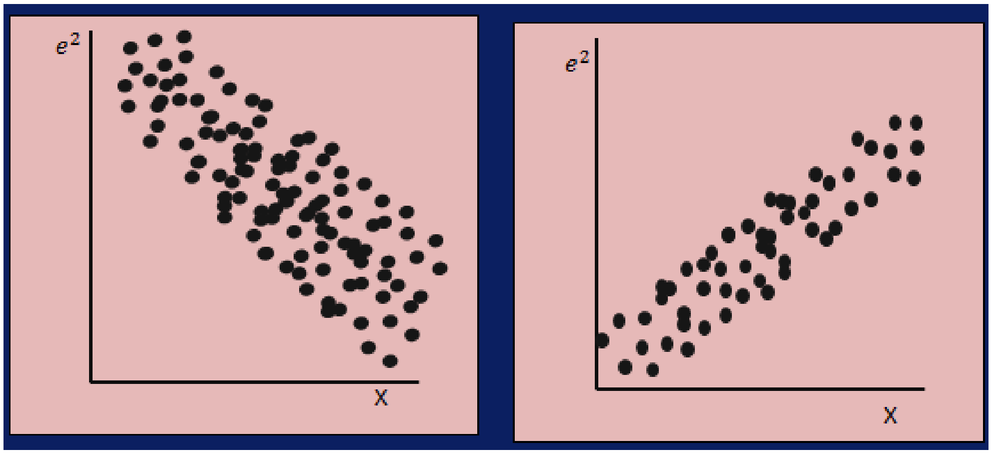
Normally graphs are used to detect Heteroskedasticity

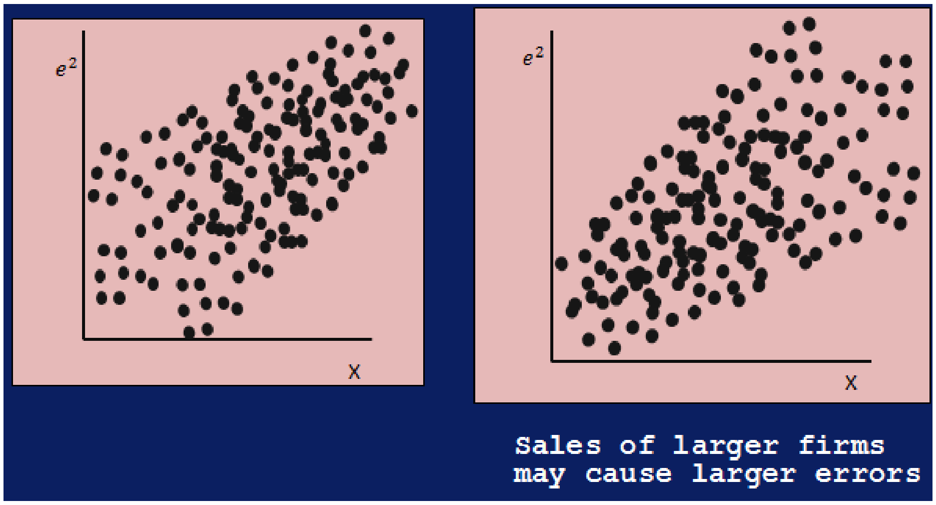
- Without HSK the errors do not vary with the independent variable
- With HSK errors may systematically be correlated with an independent variable
- With HSK we may get variety of patterns
- We can plot Y and X and look at the channel of errors
- We can use either of error or the square of error on the Y-axis and plot it against each independent variable

Homoskedasticity (No HSK): note that dots are evenly spread around the regression line. In the right panel X and square of error do not show any pattern.

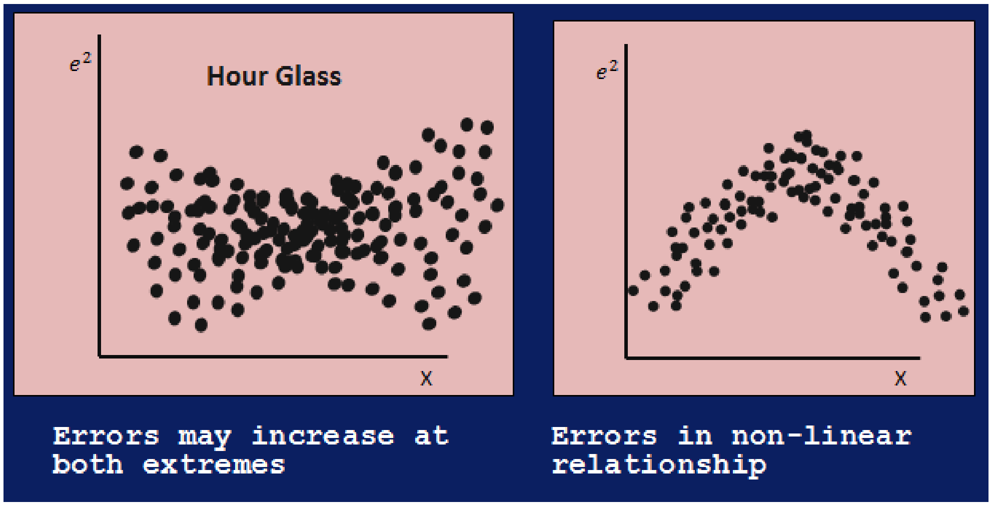


Heteroskedasticity: the following diagrams show the presence of Heteroskedasticity due to patterns or trend that is clearly visible.





Sales of larger firms
may cause larger errors



Errors may increase at
both extremes

Errors in non-linear
relationship

Lecture 18
Detection of Heteroskedasticity: Formal Tests

The Goldfeld-Quandt Test: Stephen Goldfeld, Richard Quandt

The following steps are performed for this test.

Step 1.

Arrange the data from small to large values of the explanatory variable X_j (the one we suspect responsible for HSK).

Step 2.

Omit the middle C observations (C = roughly 20% observations)

Step 3.

Run two separate regressions, one for small values of X_j and one for large values of X_j, omitting C middle observations and record the residual sum of squares RSS for each regression: RSS_1 for large values of X_j and RSS_2 for small values of X_j

Step 4.

Calculate the ratio

$$F = \frac{RSS1}{RSS2}$$

Degree of Freedom $= \frac{N-C}{2} - K$ both in the numerator and the denominator, where N = total number of observations, C is the number of omitted observations, and K is the Number of explanatory variables + one.

Step 5.

Test if $F > F_{\alpha, \frac{N-C}{2}-K, \frac{N-C}{2}-K}$ to reject homoskedasticity ($H0$)

Example: GoldFeld-Quandt Test

$H_0: Homoskedasticity$

$H_1: Heteroskedasticity$

$\alpha = 0.05 \; or \; 0.01$

$Test \; Statistic$

$$F = \frac{RSShigh}{RSSlow}$$

If you think that var(e) is increasing function of X

$Region \; of \; Rejection$

$$F > F_{\alpha,\frac{N-C}{2}-K,\frac{N-C}{2}-K}$$

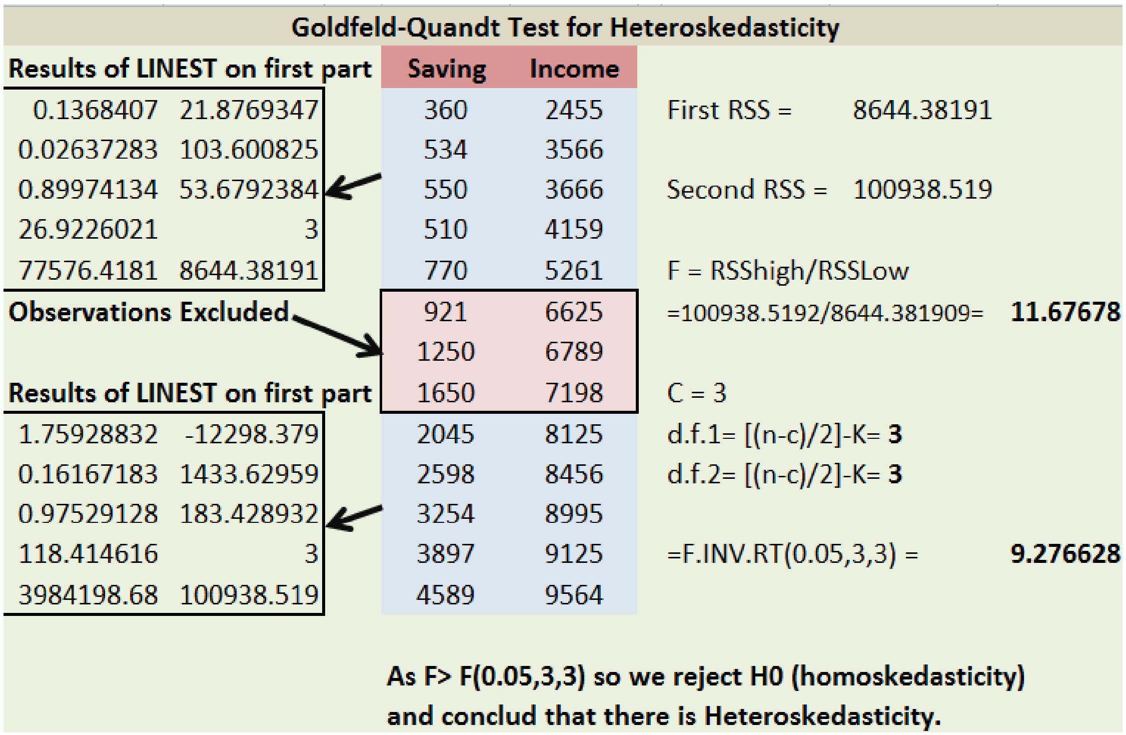
$RSS = Residual \; Sum \; of \; Squares$

$N = Number \; of \; observations$

$C = central \; observations \; excluded$

$K = Number \; of \; parameters \; estiamted = Number \; of \; explanatory \; variables \; plus \; one$

The Goldfeld-Quandt Test: Example



Goldfeld-Quandt Test for Heteroskedasticity					
Results of LINEST on first part		**Saving**	**Income**		
0.1368407	21.8769347	360	2455	First RSS =	8644.38191
0.02637283	103.600825	534	3566		
0.89974134	53.6792384	550	3666	Second RSS =	100938.519
26.9226021	3	510	4159		
77576.4181	8644.38191	770	5261	F = RSShigh/RSSLow	
Observations Excluded		921	6625	=100938.5192/8644.381909=	**11.67678**
		1250	6789		
Results of LINEST on first part		1650	7198	C = 3	
1.75928832	-12298.379	2045	8125	d.f.1= [(n-c)/2]-K= **3**	
0.16167183	1433.62959	2598	8456	d.f.2= [(n-c)/2]-K= **3**	
0.97529128	183.428932	3254	8995		
118.414616	3	3897	9125	=F.INV.RT(0.05,3,3) =	**9.276628**
3984198.68	100938.519	4589	9564		

As F> F(0.05,3,3) so we reject H0 (homoskedasticity) and conclud that there is Heteroskedasticity.

The Goldfeld-Quandt Test: Drawbacks

It cannot handle situations where several variables jointly cause Heteroskedasticity. There is no fix rule to know how many middle observations should be excluded. It is also possible that the difference in variance of errors may be observed in subsamples with different number of observations. The middle C observations are lost. It accounts only for linear relationship of independent variable and the variance of errors.

The Park LM Test

The following steps are performed in this test.

Step 1. Run the required regression e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots\dots.\beta_K X_K + e$, obtain residuals e, compute $\ln e^2$

Step 2. Run the auxiliary regression

$$lne^2 = \alpha_0 + \alpha_1 lnX_1 + \alpha_2 lnX_2 + . \ . \ . \ . \ . \ . \ . \ . \alpha_K lnX_K + u$$

Step 3. Compute $LM = N.R^2$ ($N \ and \ R^2$are from the auxiliary regression)

Step 4. If $LM > \boldsymbol{\chi^2_{K-1}}$ then reject Null hypothesis and conclude that there is significant evidence of Heteroskedasticity

Harvey-Godfrey Test

The following steps are performed in this test.

Step 1. Run the required regression e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots\dots.\beta_K X_K + e$, obtain residuals e, compute $\ln e^2$

Step 2. Run the auxiliary regression [assume $\sigma^2 = \exp(\alpha_0 + \alpha_K X_K)$]

$$lne^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + . \ . \ . \ . \ . \ . \ . \ . \alpha_K X_K + u$$

Step 3. Compute $LM = N.R^2$ ($N \ and \ R^2$are from the auxiliary regression)

Step 4. If $LM > \boldsymbol{\chi^2_{K-1}}$ then reject Null hypothesis and conclude that there is significant evidence of Heteroskedasticity

Glesjer Test

The following steps are performed in this test.

Step 1. Run the required regression e.g. $Y = \beta_0 + \beta_1 X_1 + e$, obtain residuals e, compute $\ln e^2$

Step 2. Run the auxiliary regressions [assume $\sigma^2 = \exp(\alpha_0 + \alpha_K X_K)$]

$$|e_i| = \alpha_0 + \alpha_1 X_1 + u_i$$

$$|e_i| = \alpha_0 + \alpha_1 \sqrt{X_1} + u_i$$

$$|e_i| = \alpha_0 + \alpha_1 \frac{1}{X_1} + u_i$$

Step 3. Compute $LM = N.R^2$ ($N\ and\ R^2$are from the auxiliary regression)

Step 4. If $LM > \boldsymbol{\chi^2_{K-1}}$ then reject Null hypothesis and conclude that there is significant evidence of Heteroskedasticity

Breusch - Pagan Test

The following steps are performed in this test.

Step 1. Run the required regression e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots\dots.\beta_K X_K + e$, obtain residuals e, compute $\ln e^2$

Step 2. Run the auxiliary regression [assume $\sigma^2 = \exp(\alpha_0 + \alpha_K X_K)$]

$$e^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + .\ .\ .\ .\ .\ .\ .\ .\ \alpha_K X_K + u$$

Step 3. Compute $LM = N.R^2$ ($N\ and\ R^2$are from the auxiliary regression)

Step 4. If $LM > \boldsymbol{\chi^2_{K-1}}$ then reject Null hypothesis and conclude that there is significant evidence of Heteroskedasticity

OR

test F-statistic for the above regression in step 2. (Goodness of fit)

Problems with Breusch-Pagan and others

Specification of model for variance dependence is needed e.g. Breusch Pagan assume linear relation. If the errors are not normally distributed, then these tests may not be valid. Breusch Pagan Test has been shown to be sensitive to any violation of the normality assumption. Three other popular LM tests: the Glejser test; the Harvey-Godfrey test, and the Park test, are also sensitive to such violations

The White Test: Most popular

The following steps are performed in this test.

Step 1. Run the required regression e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$, obtain residuals e, compute e□2

Step 2. Run the auxiliary regression

$$e^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_2 \boldsymbol{X_1^2} + \alpha_2 \boldsymbol{X_2^2} + \alpha_2 X_1 X_2 + u$$

(Include all square and product terms)

OR

Run the regression

$$e^2 = \gamma_0 + \gamma_1 \hat{Y} + \gamma_2 \hat{Y} 2 + v$$

Step 3.

Compute $LM = N.R^2$ ($N\ and\ R^2$are from the auxiliary regression)

OR

Compute the F-Statistic

Step 4.

If $LM > \boldsymbol{\chi^2_{K-1}}$ then reject Null hypothesis and conclude that there is significant evidence of Heteroskedasticity

OR

test F-statistic for the above regression in step 2. (Goodness of fit)

Lecture 19

Examples of detecting Heteroskedasticity

General Procedure for all the tests

$H_0: Homoskedasticity$

$H_1: Heteroskedasticity$

$\alpha = 0.05\ or\ 0.01$

$Test\ Statistic$

$LM = N.\ R^2$

$OR\ \ F = \frac{R^2}{1-R^2}\frac{N-K}{K-1}$

$OR\ \ (for\ Goldfeld\ Quandt\ Test)\ F = \frac{RSShigh}{RSSlow}$

$Region\ of\ Rejection$

$LM > \chi^2_K\ \ \ OR\ \ \ F > F_{\alpha,K-1,\,N-K}$

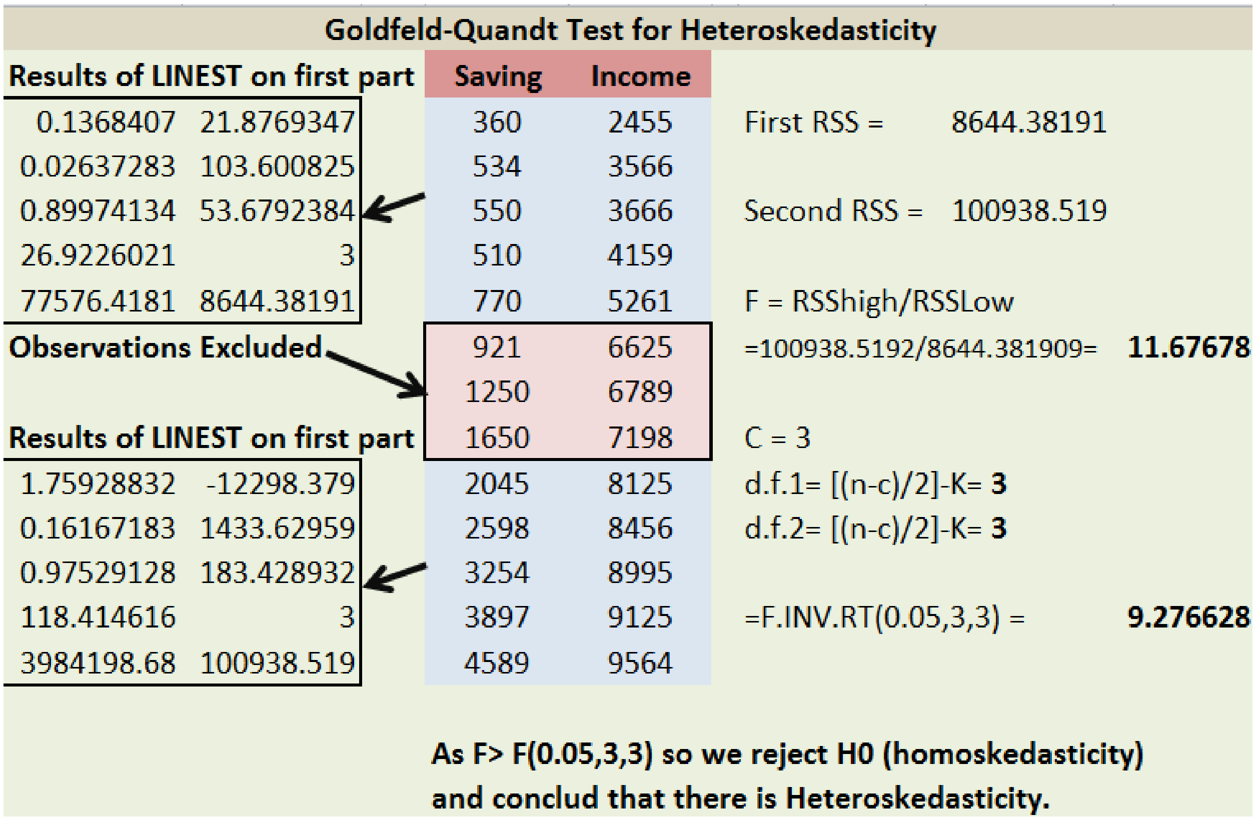
$OR\ \ (for\ Goldfeld\ Quandt\ Test)$

$F > F_{\alpha,\frac{N-C}{2}-K,\,\frac{N-C}{2}-K}$

Examples

For the below examples use the file **HSK.xlsx** for data

The Goldfeld - Quandt Test: Example Revision



Goldfeld-Quandt Test for Heteroskedasticity					
Results of LINEST on first part		**Saving**	**Income**		
0.1368407	21.8769347	360	2455	First RSS =	8644.38191
0.02637283	103.600825	534	3566		
0.89974134	53.6792384	550	3666	Second RSS =	100938.519
26.9226021	3	510	4159		
77576.4181	8644.38191	770	5261	F = RSShigh/RSSLow	
Observations Excluded		921	6625	=100938.5192/8644.381909=	**11.67678**
		1250	6789		
Results of LINEST on first part		1650	7198	C = 3	
1.75928832	-12298.379	2045	8125	d.f.1= [(n-c)/2]-K= **3**	
0.16167183	1433.62959	2598	8456	d.f.2= [(n-c)/2]-K= **3**	
0.97529128	183.428932	3254	8995		
118.414616	3	3897	9125	=F.INV.RT(0.05,3,3) =	**9.276628**
3984198.68	100938.519	4589	9564		

As F> F(0.05,3,3) so we reject H0 (homoskedasticity)
and conclud that there is Heteroskedasticity.

Example: The Park LM Test (ln (square of errors) and the ln(income)

First we run the regression savings on income, find ln (square of errors) and the ln(income), while Square of errors are generated from the regression line Saving on Income.

Saving	Income	ln (income)	Square of Errors	Ln e-square
360	2455	7.80588204	497370.7375	13.1171
534	3566	8.179199798	86591.3042	11.3690
550	3666	8.206856428	66363.1361	11.1029
510	4159	8.33302994	1761.6205	7.4740
770	5261	8.568076402	131198.9807	11.7845
921	6625	8.798605651	863800.5063	13.6691
1250	6789	8.823058934	471640.0397	13.0640
1650	7198	8.881558489	252118.1070	12.4377
2045	8125	9.002701007	354277.5972	12.7778
2598	8456	9.042631528	46870.3204	10.7551
3254	8995	9.104424146	24243.0246	10.0959
3897	9125	9.118773178	533268.0951	13.1868

4589	9564	9.165761329	1418726.1336	14.1653

Now we run The Park LM Test: auxiliary regression: $lne^2 = \alpha_0 + \alpha_1 \ln(income) + u$

Park Test for Heteroskedasticity (LM test and F-Test)
Run a regression ln(square of Errors) on ln(income)
R_2 = 0.0634
We can use both LM and the F-statistic for framing the conclusion.
LM = N. R^2=13(0.0634) = **0.8243**
ChSQ(0.05,1) = CHISQ.INV.RT(0.05,1)= 3.841458821
As LM > Chi-SQ is not met, we cannot reject H_0 and conclude that errors are not Heteroskedastic (There is no Heteroskedasticity)
F = $[R^2/(1-R^2)]/[11/1]$ = 0.7446
$F_{0.05,1,11}$ = F.INV.RT(0.05,1,11)= 4.844335675
As F > $F_{0.05,1,11}$ is not met so we cannot reject H_0 and conclude that errors are not Heteroskedastic

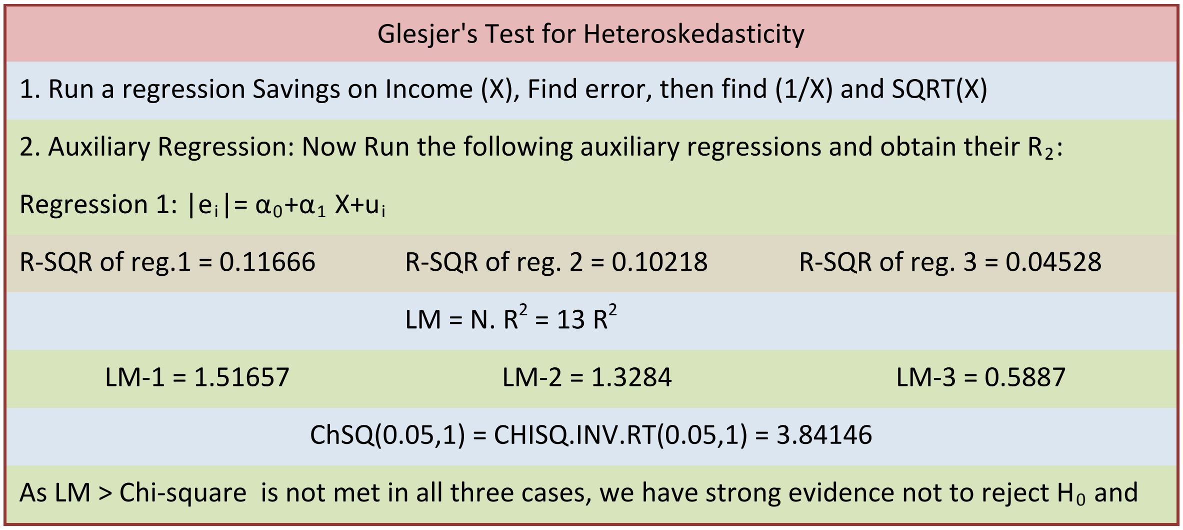
In the above table both LM and F-test are applied and the result is that the errors are not Heteroskedastic.

Example: Harvey-Godfrey Test for Heteroskedasticity

For the following results, see the file **HSK.xlsx** for detailed data

Harvey-Godfrey Test for Heteroskedasticity	
1. Run a regression Savings on Income, Find error, then ln (square of error)	
2. Auxiliary Regression: Now Run a regression ln(square of Errors) on Income and obtain R^2	
R^2 obtained from the auxiliary Regression using LINEST() = 0.087076159	
LM = N. R^2=	1.13199
ChSQ(0.05,1) = CHISQ.INV.RT(0.05,1) = 3.84146	
As LM > Chi-square value, we reject H_0 and conclude that errors are Heteroskedastic	
F = $[R^2/(1-R^2)]/[11/1]$ =	1.04919786607434
$F_{0.05,1,11}$	= F.INV.RT(0.05,1,11) = 4.84434
As F < $F_{0.05,1,11}$ so we cannot reject H_0 and conclude that errors are not Heteroskedastic	

Example: Glesjer's Test for Heteroskedasticity



Glesjer's Test for Heteroskedasticity

1. Run a regression Savings on Income (X), Find error, then find (1/X) and SQRT(X)

2. Auxiliary Regression: Now Run the following auxiliary regressions and obtain their R_2:

Regression 1: $|e_i| = \alpha_0 + \alpha_1 X + u_i$

R-SQR of reg.1 = 0.11666 R-SQR of reg. 2 = 0.10218 R-SQR of reg. 3 = 0.04528

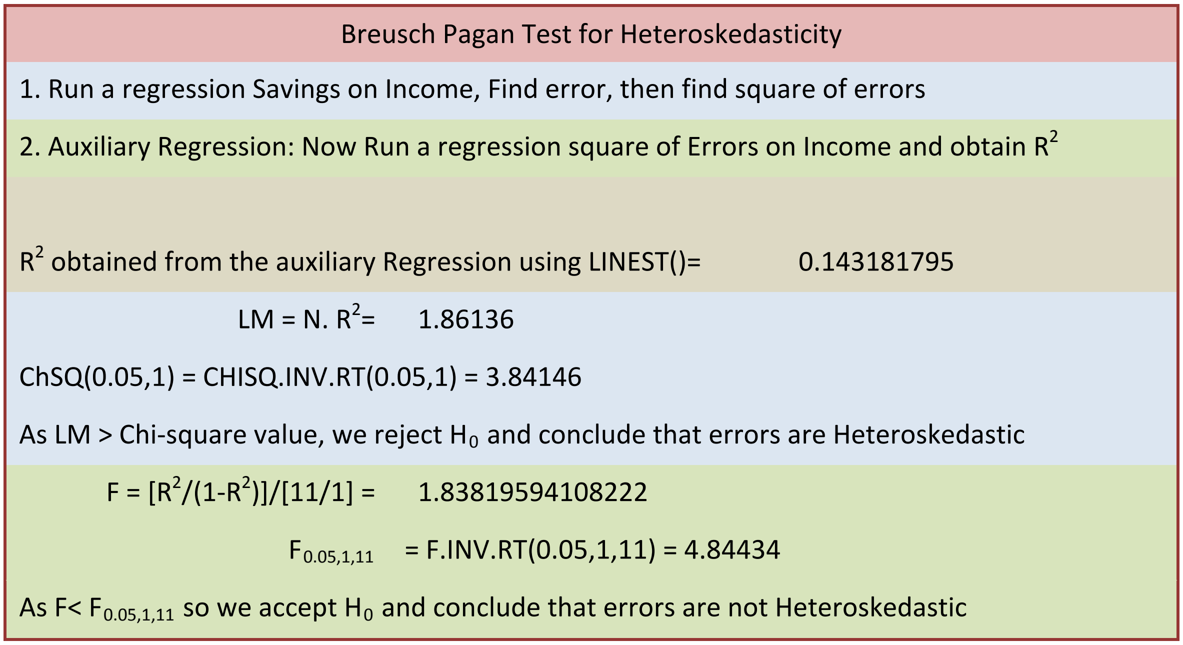
$LM = N.\ R^2 = 13\ R^2$

LM-1 = 1.51657 LM-2 = 1.3284 LM-3 = 0.5887

ChSQ(0.05,1) = CHISQ.INV.RT(0.05,1) = 3.84146

As LM > Chi-square is not met in all three cases, we have strong evidence not to reject H_0 and

Example: Breusch Pagan Test for Heteroskedasticity



Breusch Pagan Test for Heteroskedasticity

1. Run a regression Savings on Income, Find error, then find square of errors

2. Auxiliary Regression: Now Run a regression square of Errors on Income and obtain R^2

R^2 obtained from the auxiliary Regression using LINEST()= 0.143181795

$LM = N.\ R^2 =$ 1.86136

ChSQ(0.05,1) = CHISQ.INV.RT(0.05,1) = 3.84146

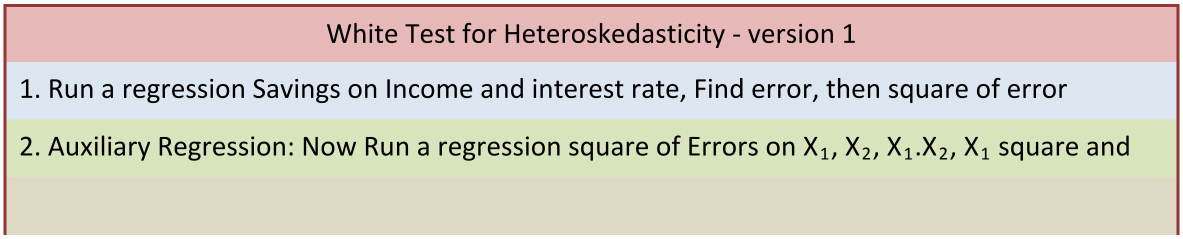
As LM > Chi-square value, we reject H_0 and conclude that errors are Heteroskedastic

$F = [R^2/(1-R^2)]/[11/1] =$ 1.83819594108222

$F_{0.05,1,11}$ = F.INV.RT(0.05,1,11) = 4.84434

As F< $F_{0.05,1,11}$ so we accept H_0 and conclude that errors are not Heteroskedastic

Example: White Test for Heteroskedasticity - version 1



White Test for Heteroskedasticity - version 1

1. Run a regression Savings on Income and interest rate, Find error, then square of error

2. Auxiliary Regression: Now Run a regression square of Errors on X_1, X_2, $X_1.X_2$, X_1 square and

R^2 obtained from the auxiliary Regression using LINEST()= 0.5021

LM = N. R^2= 6.5273

ChSQ(0.05,5) = CHISQ.INV.RT(0.05,5) = 11.0705

As LM > Chi-square value, we can not reject H_0 . We conclude that errors are not

F = $[R^2/(1-R^2)]/[7/5]$ = 0.470603200107117

$F_{0.05,4,7}$ = F.INV.RT(0.05,4,7) = 4.12031

As F < $F_{0.05,1,11}$ so we cannot reject H_0 . There is no HSK

Handling Heteroskedasticity

Method 1: Change the model specification / transform variables

The Relation of variables may not be linear; change the model. Some important variable may be missing; find & use them. If there are subgroup differences, use dummy variables. If possible use panel data techniques

Method 2: Use Huber / White Standard Errors

This type of standard error is also called Heteroskedasticity consistent standard errors. With Heteroskedasticity our standard errors are incorrect so t would be incorrect (Another formula is required for SE). White's idea: Use simple OLS and correct the SE. For a simple regression line, the *white estimator of error variance* is

$$White\ Variance\ (\beta_1) = \frac{\sum(X-\bar{X})^2\sigma_i^2}{\left(\sum(X-\bar{X})^2\right)^2}$$

Estimate $\sigma_i^2 = \hat{e}_i^2$ (using the squared residual for each observation as the estimate of its variance)

Method 3: GLS / WLS

Two cases; One where the variance of errors is known, Second where variance of errors is not known

GLS / WLS (variance of errors known)

Consider the model

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + e_i}$$

- Suppose we know σ_i^2 i.e. $Variance\ (\beta_1)$

- Divide all by this error Standard Deviation,

$$\frac{\boldsymbol{Y_i}}{\sigma_i} = \frac{\boldsymbol{\beta_0}}{\sigma_i} + \frac{\boldsymbol{\beta_1}}{\sigma_i}\boldsymbol{X_i} + \frac{\boldsymbol{e_i}}{\sigma_i}$$

- New model variance

$$Var\left(\frac{\boldsymbol{e_i}}{\sigma_i}\right) = \frac{1}{\sigma_i^2}\, var(\boldsymbol{e_i}) = \frac{1}{\sigma_i^2}\sigma_i^2 = 1$$

GLS / WLS (Variance of errors NOT known)

Make assumptions about the variance. Always transform the model dividing by the Standard Deviation of errors

What does WLS do?

OLS minimizes the sum of squared errors. OLS gives equal weight (importance) to all observations. WLS: observations with larger error variance will get less weight. WLS minimizes a weighted sum of square of errors

e.g. minimizes $\frac{\sum e^2}{var(e_i)}$

Transformation Examples		
Assumption	Comments	Transformation
$Var(e_i) = \sigma^2 X_i$	Variance increases linearly with X	$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_0}{\sqrt{X_i}} + \frac{\boldsymbol{\beta_1}}{\sqrt{X_i}}X_i + \frac{\boldsymbol{e_i}}{\sqrt{X_i}}$
$Var(e_i) = \sigma^2 X_i^2$	Variance is related to square of X	$\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \frac{\boldsymbol{\beta_1}}{X_i}X_i + \frac{\boldsymbol{e_i}}{X_i}$

Example: Handling Heteroskedasticity

Consider the model $\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + e_i}$, from the file **HSK-Removal.xlsx**

observations	20		
K	2	$F_{0.05,\ K-1,\ N-K}$	4.4138734
N-K	18	ChiSQ$_{0.05,\ K-1}$	3.8414588
K-1	1		

Breusch Pagan test			
From Reg e-squared on X			
R^2 =	0.2283669	LM=N.R^2 =	4.5673372
		F =	5.3271473
Both F and LM indicate the presence of Heteroskedasticity			
White Test			
From Reg e-squared on X and X^2			
R^2 =	0.3495372	LM=N.R^2 =	6.9907444
		F =	4.5676194
Both F and LM indicate the presence of Heteroskedasticity in non-linear fashion.			

Now let us try to estimate without heteroskedastic errors

Consider the model $Y = \beta_0 + \beta_1 X_i + e_i$

Assume the relation $Var(e_i) = \sigma^2 X_i^2$

Transform the model to estimate $\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \frac{\beta_1}{X_i} X_i + \frac{e_i}{X_i}$

Or $(\frac{Y_i}{X_i}) = \frac{\beta_0}{X_i} + \beta_1 + \frac{e_i}{X_i}$

Or $\frac{Y_i}{X_i} = \beta_1 + \beta_0 \frac{1}{X_i} + \frac{e_i}{X_i}$

Note the intercept is $\boldsymbol{\beta_1}$ and the slope here is $\boldsymbol{\beta_0}$

This can be estimated by OLS

For $\frac{Y_i}{X_i} = \beta_1 + \beta_0 \frac{1}{X_i} + \frac{e_i}{X_i}$

We get $\frac{Y_i}{X_i} = 11.51 + 25.61 \frac{1}{X_i}$

The intercept 11.51 is in fact the slope of the original equation and the slope in this regression line is the intercept of the original regression line. But they may now be Heteroskedasticity Free (this can be checked by applying the White test to this regression)

Test for Heteroskedasticity for the new transformed regression $\frac{Y_i}{X_i} = 11.51 + 25.61\frac{1}{X_i}$ (**data in HSK-Removal.xlsx**). Now testing for Heteroskedasticity for the new model gives the following results that are free from the problem.

observations	20
K	3
N-K	17
K-1	2
$F_{0.05,\ K-1,\ N-K}$	3.5915306
$ChiSQ_{0.05,\ K-1}$	5.9914645

White Test		
From Reg e-squared on1/X and $(1/X^2)$		
	R^2 =	0.0389161
	LM=N.R^2 =	0.7783225
	F =	0.3441812
Conclusion: Heteroskedasticity does not		

Lecture 20

Autocorrelation

What is Autocorrelation / Serial Correlation?

Gauss Markov Assumptions: BLUE include

- Errors are normally distributed with zero mean and constant variance
- Errors are independent
- Independence: one error term is not correlated to any other error term
- Violation of above is called serial correlation

But with Serial Correlation

- Error terms are correlated with one another
- Error term of different time periods (usually adjacent) or different cross sectional observations are correlated
- if we know something about the error term of one observation, we know something about the error term of another observation
- Errors associated in one time period carry over to future time periods (example of lagged models)
- Serial correlation is usually associated with time series data so we will use a subscript t instead of i

Error terms may be correlated more with nearby observations as compared to distant observations

$$\boldsymbol{\rho(e_t, e_{t-i}) > \rho(e_t, e_{t-j})\ where\ i < j}$$

Where $\boldsymbol{\rho}$ (rho) is the autocorrelation coefficient

In case serial correlation exists (in time or space),

$$\boldsymbol{\rho(e_t, e_{t-i}) \neq 0\ where}\ t \neq 0\ \ in\ time\ series\ data$$

$$\boldsymbol{\rho(e_i, e_j) \neq 0\ \ where\ i \neq j\ \ in\ cross\ sectional\ data}$$

$$-\mathbf{1} \leq \boldsymbol{\rho} \leq \mathbf{1}$$

First Order Serial Correlation

Let us consider one specific type i.e. first order (linear) serial correlation. In this case Error in one time period are correlated with the previous time period

In case of first order serial correlation

$$\boldsymbol{\rho}(\boldsymbol{e_t}, \boldsymbol{e_{t-1}}) \neq \mathbf{0}$$

$$where \quad -\mathbf{1} \leq \boldsymbol{\rho} \leq \mathbf{1}$$

In dynamic models $\boldsymbol{e_t} = \boldsymbol{\rho} \boldsymbol{e_{t-1}} + \boldsymbol{u_t}$

Where $\mathrm{u_t}$ is called white noise and is independently and identically distributed with zero mean and constant variance

Possible Causes of Autocorrelation

1. Missing Variables

Error term may include all the variables not included in the regression equation. Change in any of the unobserved variable in one time period may impact the errors in different time periods. Errors may follow the patterns/ trends of the unobserved variables.

***Examples**:* Y depends on X_{t1} and X_{t2} and X_{t2} is not included OR Sales depend on seasonal changes and it is not included

2. Inertia or sluggishness

Speed of Change in variables depends on time. Business Cycles (GDP, Prices etc. follow) are an example. Speed of change in price may depend on how far it is from the equilibrium price. This is a common phenomenon in time series data

3. Incorrect Functional Form

Autocorrelation may exist if Linear models are specified when non-linear are required. Linear-in-variable or simple models are used when log form is needed. This is called model specification error

4. Cobweb phenomenon / reaction with lag

When the dependent variable acts with a lag, we may observe Autocorrelation.

Examples

$$\boldsymbol{Q_{st} = a + bP_{t-1} + e_t}$$

$$C_t = ay_{t-1} + by_{t-2} + e_t$$

Overproduction in one year may lead to underproduction in the next

5. Lagged Relationship

One reason may be that dependent variable may depend on its previous value

Examples

- Stock Prices
- Consumption $\boldsymbol{C_t = a + bC_{t-1} + e_t}$

6. Ratchet effect

Tendency of people to be influenced by the previous (high, low or best) level of a variable may cause the ratchet effect to exist. Another example is of consumption. Consumption changes quickly upward (when income goes up) but does not come down easily if income declines (over or underestimation)

7. Data Manipulation

- Data manipulated in the following ways may cause the problem:
- Averaging or smoothing
- Converting quarterly data to annual
- Converting monthly data to quarterly
- Finding mid-points when faced with ranges
- Sometimes we need to do the above because of measurement error in, e.g., monthly data but the byproduct is autocorrelation

8. Systematic Measurement Error

Measurement error in one time period may be carried forward. Measurement error in inventories, Measurement error in stock of capital, Measurement error in asset value all can cause the problem of autocorrelation. This is why Time series is a noisy proxy.

9. Variable Behavior

Electricity consumption in different hours of the day; similar temperature patterns may cause patterns in errors.

Spatial Autocorrelation

This is a special cause of autocorrelation in cross sectional data.

"Everything is related to everything else but near things are more related than distant things"

First law of Geography by WALDO TOBLER

Spatial autocorrelation is a correlation of a variable with itself through space. It is due to systematic pattern of spatial distribution of a variable. Sometimes the nearby areas are more alike: positive Sp. AC. Values in sample do not remain independent. Occurrence of one event in an area makes it more likely in other areas. Areas with higher concentration of events will have more impact on results (spatially clustered observations)

Example:

- In Karachi, due to high crime rate, more Police and Rangers are deployed
- Due to Karachi operation the crime rate decreases in Karachi
- Crime rate in nearby cities will increase although they did not decrease Police etc.

Consequences of Autocorrelation

- Estimators remain unbiased and consistent
- Estimators are no more efficient
- Standard Error of Estimate / variance of error is likely to be underestimated resulting in Overestimated R-square

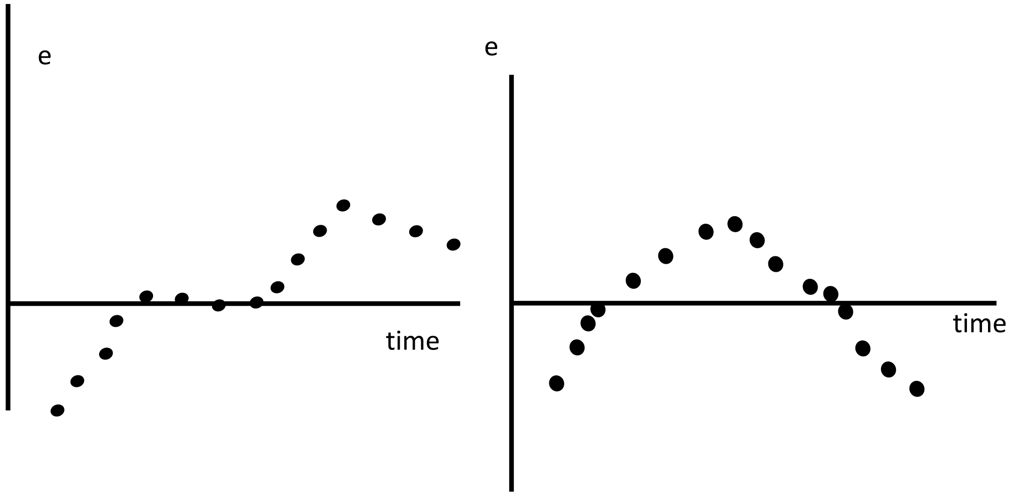
- Variances of estimators are biased (see next slide)
- Forecasts are unbiased but inefficient (with larger variances)
- In case of positive AC, standard errors (of coefficients) are too small resulting in overestimation of t-statistic
- In case of negative AC, standard errors (of coefficients) are too large resulting in underestimation of t-statistic

Lecture 21
Detection of Autocorrelation

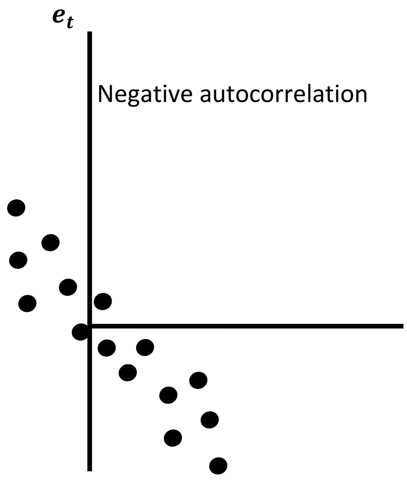
Detection (Testing for serial correlation)

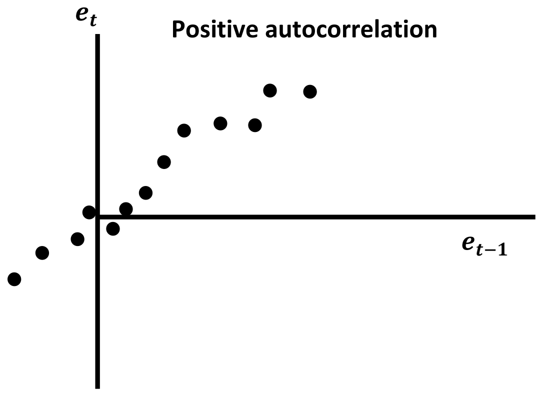
- Graphic method
- AR(1) test with strictly exogenous regressors
- AR(1) test without strictly exogenous regressors
- Durbin Watson *d* Test
- Durbin *h* Test
- LM Test

Looking at the graph, we may see patterns or trends in residuals w.r.t. time or w.r.t. previous values of the errors.



The above errors are plotted against time. You can see that a pattern is visible. This shows that the errors are not random. The graphs below show the errors of residuals plotted agains the past values of errors. In this case also we can observe some patterns and trends which is an indication of serial correlation.





Detecting Autocorrelation by formal tests

AR(1) test with strictly exogenous regressors

This tests first order AC with exogenous regressors

Procedure

- Run the regression Y on $X1,\ X2,\ \ldots\ XK$ and obtain residual e_t
- Run the regression $\hat{e}t$on $\hat{e}t - 1\, for\ N\ =\ 2\ to\ n$

$$e_t = \rho e_{t-1} + u_t$$

- Apply t-test (individual variable significance test) to test the hypothesis H$_0$: $\rho = 0$

The data and some estimated columns are given below:

Time Period	Y	X	e_t	e_{t-1}
1	41	3.1	-6.69226	.
2	51	3	3.57925	-6.69226
3	55	3.3	6.76473	3.579245
4	58	4	7.86418	6.764725
5	56	4.3	5.04966	7.864179

6	51	4.7	-1.03637	5.049659
7	46	5.4	-7.93691	-1.03637
8	46	4.9	-6.57938	-7.93691
9	48	6	-7.56595	-6.57938
10	57	7.5	-2.63855	-7.56595
11	58	6.9	-0.00951	-2.63855
12	57	7.7	-3.18157	-0.00951
13	62	7.9	1.27542	-3.18157
14	64	8.1	2.73241	1.27542
15	71	8.6	8.37487	2.732407

1. Example of AR(1) test with strictly exogenous regressors

The test is given below with all the required steps performed using Microsoft Excel.

AR(1) test with strictly exogenous regressors

Step 1: Run a Regression Y on X and obtain residuals

Step 2: Run a Regression $e_t = a + \rho\, e_{t-1} + u_t$

Results of using '=linest(E52:E65,F52:F65,true,true)'

0.6494361	0.8665148
0.2328061	1.2152534
0.393384	4.5171056
7.7818704	12
158.78317	244.85092

Compute t-value = beta/se(beta) = 0.6494/0.2328= 2.789600395

Testing H_0: ρ = 0

t = 2.7896004 , t_c=T.INV.2T(0.05,13) = 2.160368656

As the calculated value of *t* is larger than the tabulated value so we reject H_0 and conclude that the coefficient ρ is significant. This means that there is a problem of Autocorrelation.

2. Example of AR(1) test without strictly exogenous regressors

This tests first order AC without exogenous regressors (where we suspect that explanatory variables may be correlated to residuals)

Procedure

- Run the regression Y on $X1,\ X2,\ \ldots\ XK$ and obtain residual e_t
- Run the regression $\hat{e}t$on $X1,\ X2,\ \ldots\ XK$ and $\hat{e}t - 1\, for\, N\ =\ 2\ to\, n$
- $e_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \rho e_{t-1} + u_t$
- Apply t-test (individual variable significance test) to test the hypothesis H$_0$: $\rho = 0$

AR(1) test without strictly exogenous regressors		
Step 1: Run a Regression Y on X and obtain residuals		
Step 2: Run a Regression e$_t$ = a + b X + ρ e$_{t-1}$ + u$_t$		
You will need to rearrange the columns to put et-1 and X together		
The regression results are		
-0.05608296	**0.645754025**	1.193999801
0.700460158	**0.247399143**	4.282505484
0.393737269	4.716588768	
3.571974441	11	
158.9257843	244.7083057	
Compute t-value		
beta/se(beta) = 0.64575/0.247399 = 2.610170824		
Testing H$_0$: ρ=0		
t= 2.610170824		
Critical value of t=T.INV.2T(0.05,11) = 2.20098516		
As the calculated value of t is larger than the tabulated value so we reject H$_0$ and conclude that the coefficient ρ is significant. This means that there is a problem of Autocorrelation.		

3. Example of Durbin Watson d Test

This tests first order AC with exogenous regressors. It is not good when regressors are not exogenous. It is not good when the model contains lagged dependent variable

Assumptions:

- Regression Model has an intercept
- Errors are generated by first order autoregressive scheme

$$e_t = \rho e_{t-1} + u_t$$

- Regression Model does not have lagged dependent variable as regressor
- There are no missing observations

Procedure:

$H_0\text{: } \rho = 0$

$H_A\text{: } \rho \neq 0$

$\alpha = 0.05\ (5\%)$

$$Test\ Statistic$$

$$\boldsymbol{d} = \frac{\sum_{t=2}^{n}(\boldsymbol{e_t} - \boldsymbol{e_{t-1}})^2}{\sum_{t=1}^{n}\boldsymbol{e_t^2}}$$

This has an inconclusive region, rejection and acceptance region

It has some limit or range of values:

$$d = \frac{\sum_{t=2}^{n}(\boldsymbol{e_t} - \boldsymbol{e_{t-1}})^2}{\sum_{t=1}^{n}\boldsymbol{e_t^2}}$$

$$d = \frac{\sum e_t^2 + \sum e_{t-1}^2 - 2\sum e_t \sum e_{t-1}}{\sum e_t^2}$$

For large samples

$$\sum e_t^2 = \sum e_{t-1}^2$$

So $d = 2(1 - \frac{\sum e_t e_{t-1}}{\sum e_t^2})$

$$d = 2(1 - \frac{\sum e_t \, e_{t-1}}{\sum e_t^2})$$

But $\rho = \frac{\sum e_t \sum e_{t-1}}{\sum e_t^2}$ So

$$d = 2(1 - \rho)$$

As we know that $\boldsymbol{-1 \leq \rho \leq 1},$ So d will range from zero to four

$$0 \leq d \leq 4$$

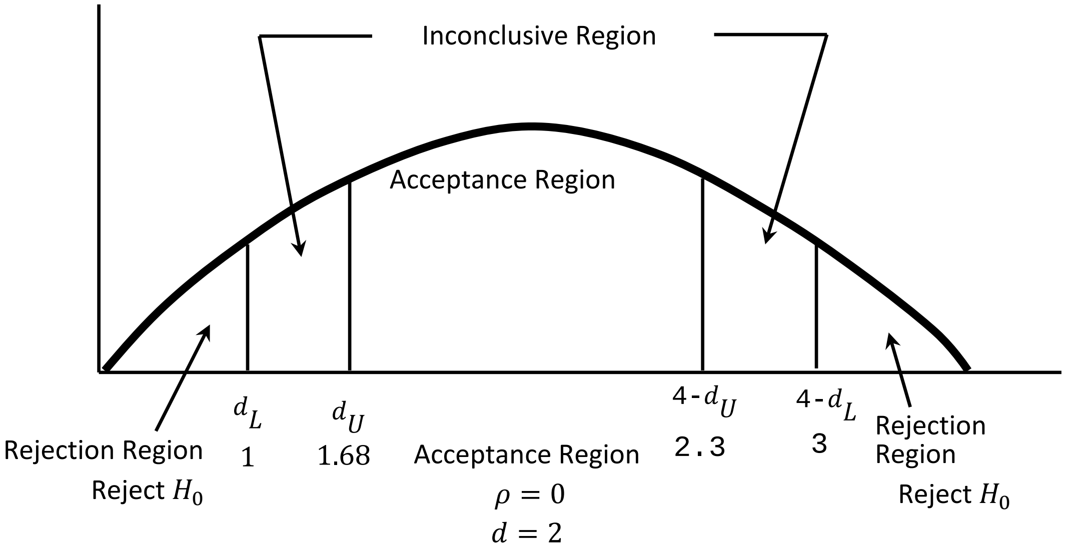
$$if\boldsymbol{\rho} = 0, \ \mathrm{d} = 2(1 - 0) = 2$$

$$if\boldsymbol{\rho} = -1 \text{ (negative)}, \ \mathrm{d} = 2\big(1 - (-1)\big) = 4$$

$$if\boldsymbol{\rho} = 1 \text{ (positive)}, \ \mathrm{d} = 2\big(1 - (1)\big) = 0$$

The table of d at 0.05 level of significance and $K - 1$ degrees of freedom will provide two values of d. A lower value dL and an upper value dU.

Example: dL=1, dU =1.68 (at 5% and $k - 1 = 3$)



Based on the previous diagram, we may frame our conclusions as follows

Accept H_0 if $d_u < d < 4 - d_u$ (no autocorrelation)

Reject H_0

if $d < d_L$ (positive autocorrelation)

if $d > 4 - d_L$ (negative autocorrelation)

Test is inconclusive if

if $d_L < d < d_U$

if $4 - d_U < d < 4 - d_L$

General Rules about Durbin Watson Statistic

- d lies between 0 and 4. As d is closer to 2, the chances of AC decrease
- If d < 2, This may indicate positive AC, IF d > 2, This may indicate negative AC

Use the file ***AC.xlsx*** for data, run regression Y on X, obtain residuals

e_t is in cells E51:E65, e_{t-1} is in cells F52:F65		**Formula used**
$\sum e_t^2 =$	451.6195	=SUMSQ(E51:E65)
$\sum (e_t - e_{t-1})^2 =$	307.333	=SUMXMY2(E52:E65,F52:F65)
$d = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2}$	d = 0.680513	=SUMXMY2(E52:E65,F52:F65)/SUMSQ(E51:E65)
	d_L = 1.08	Value read from table (n=15, K=2)
	d_U = 1.36	Value read from table (n=15, K=2)
		K= number of parameters estimated
As $d < d_L$, we reject H_0 and conclude that there exists autocorrelation (positive)		

Durbin *h* Test

Procedure

Run the regression

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \gamma Y_{t-1} + u_t$$

Compute Durbin h, where

$$h = \hat{\rho}\sqrt{\frac{N}{1 - N(Var(\gamma))}}$$

Where $\hat{\rho} = 1 - \frac{d}{2}$ (provided the sample is large)

Note: h is normally distributed with unit variance so conclusions may be formed by looking at the normal distribution table.

Problem: Cannot be computed if $N\big(Var(\gamma)\big) > 1$

Simple LM Test: Testing for AR(1) model

Procedure

To test first order serial correlation

- Run the regression

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + e_t$$

- Obtain the residuals $\hat{e}t$
- Regress $\hat{e}t$ on all explanatory variables and $\hat{e}t - 1$
- Now compute $LM = (n-1)\,R2$ (because we loose one observation so n-1
- Test using Chi-square (0.05, 1)

Similar to AR(1) in case of NOT Exogenous regressors but uses LM instead of F.

Lecture 22
Treating Autocorrelation

Some consequences of ignoring autocorrelation are

- Coefficients unbiased and consistent but not efficient (not BLUE) even in large samples. (like if there is heteroskedasticity)
- Standard error estimates are inappropriate which leads to wrong inferences.
- t-statistic overestimated
- Regression coefficients appearing significant are, in fact, not significant.
- R^2 inflated and residual variance underestimated for positively correlated residuals (if X grows over time).
- Forecasts are unbiased but with large variances

Remedial Measures for the problem of Autocorrelation

We may have two situations

- When ρ is known. . . . GLS
- When ρ is not known. . . . .estimate rho. . . .GLS

Generalized Differencing / Generalized Least Square

Consider a two variable model

$$Y_t = \beta_0 + \beta_1 X_t + e_t$$

Assume that the error term follows AR(1) scheme

$$e_t = \rho e_{t-1} + v_t$$

Where

$$-1 < \rho < 1$$

Where v_t satisfies OLS assumptions and ρ is either known or estimated

For estimation without autocorrelation, transform the model

Procedure

write the regression with one period lag

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

Multiplying by ρ

$$\rho Y_{t-1} = \rho\,\beta_0 + \rho\,\beta_1 X_{t-1} + \rho\, e_{t-1}$$

Subtracting this equation from the original regression (without lags)

$$Y_t - \rho Y_{t-1} = \beta_0 - \rho\,\beta_0 + \beta_1 X_t - \rho\,\beta_1 X_{t-1} + e_t - \rho\, e_{t-1}$$

Which gives us

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + e_t - \rho\, e_{t-1}$$

Or

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

Now we can estimate the AC free model

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

Where

$$Y_t^* = Y_t - \rho Y_{t-1}$$

$$\beta_0^* = \beta_0(1-\rho)$$

$$X_t^* = X_t - \rho X_{t-1}$$

Generalized Differencing: Special Cases

Special Case: $\boldsymbol{\rho = +1}$:

write the regression with one period lag

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

Multiplying by ρ

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

Subtracting this equation from the original regression (without lags)

$$Y_t - Y_{t-1} = \beta_0 - \beta_0 + \beta_1 X_t - \beta_1 X_{t-1} + e_t - e_{t-1}$$

Which gives us

$$\Delta Y_t = \beta_1 \Delta X_t + \Delta e_t$$

Where

$$\Delta Y_t = Y_t - Y_{t-1},\ \Delta X_t = X_t - X_{t-1}$$

And we estimate an equation without intercept

Special Case: $\boldsymbol{\rho = -1}$:

write the regression with one period lag

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

Multiplying by ρ

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

Subtracting this equation from the original regression (without lags)

$$Y_t + Y_{t-1} = \beta_0 + \beta_0 + \beta_1 X_t + \beta_1 X_{t-1} + e_t + e_{t-1}$$

Which gives us

$$Y_t^* = 2\beta_0 + \beta_1 X_t^* + v_t$$

Where

$$Y_t^* = Y_t + Y_{t-1}, X_t^* = X_t - X_{t-1}$$

Prais-Winsten Transformation

We lose one observation due to differencing. Usually the first observation is lost. In small samples we may estimate the first observation as

$$Y_1^* = \sqrt{1-\rho^2}\,Y_1$$

$$X_1^* = \sqrt{1-\rho^2}\,X_1$$

However this does not need to be done in large samples. We can estimate the AC free model

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

Suppose we know that

$$\rho = 0.6$$

$$Y_t^* = Y_t - \rho Y_{t-1}$$

$$X_t^* = X_t - \rho X_{t-1}$$

Where $\beta_0^* = \beta_0(1-\rho)$

The following example is a small sample (Prais-Winsten transformation may be required)

Time Period	**Y**	**X**	**e$_t$**	**Y***	**x***
1	41	3.1	-6.69226		
2	51	3	3.57925	26.40	1.14
3	55	3.3	6.76473	24.40	1.50
4	58	4	7.86418	25.00	2.02
5	56	4.3	5.04966	21.20	1.90
6	51	4.7	-1.03637	17.40	2.12
7	46	5.4	-7.93691	15.40	2.58
8	46	4.9	-6.57938	18.40	1.66
9	48	6	-7.56595	20.40	3.06
10	57	7.5	-2.63855	28.20	3.90
11	58	6.9	-0.00951	23.80	2.40
12	57	7.7	-3.18157	22.20	3.56
13	62	7.9	1.27542	27.80	3.28
14	64	8.1	2.73241	26.80	3.36
15	71	8.6	8.37487	32.60	3.74

We knew that $\rho = 0.6$

We transformed the model accordingly and estimated

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

Using Microsoft Excel functions of intercept and slope we find the regression line to be

$$Y_t^* = 17.995 + 2.1554\, X_t^*$$

The slope 2.1554 is the value of β_1**. As** $\beta_0^* = \beta_0(1-\rho)$, We can estimate β_0 as $\beta_0 = \frac{\beta_0^*}{1-\rho} = \frac{17.995}{1-0.6} = 44.9877$

REMEMBER that the above procedure is based on an assumed value of ρ to explain the procedure which may not be true.

Removing Autocorrelation when ρ is not know

When is not known, we need to estimate it first (before treating autocorrelation). We will discuss three methods here

- Durbin Watson d
- Cochrane-Orcutt Method
- Hildreth-Lu Procedure

Using Durbin-Watson d

A relationship already established between ρ and d is

$$d = 2(1-\rho)$$

$$\text{So } \rho \cong 1 - \frac{d}{2}$$

$$\text{As} \qquad 0 \leq d \leq 4$$

$$-1 \leq \rho \leq 1$$

Above relation is approximate. For small samples Theil-Nagar suggest

$$\rho = \frac{N^2\left(1-\frac{d}{2}\right) + K^2}{N^2 + K^2}$$

Now perform GLS.

EXAMPLE (use ACRemoval.xlsx)

As this is a small sample, we use Theil-Nagar estimation

$$\rho = \frac{N^2\left(1-\frac{d}{2}\right) + K^2}{N^2 + K^2}$$

N = 14, K = 2, Durbin-Watson d is estimated as 0.6805

$$\rho = \frac{14^2\left(1 - \frac{0.6805}{2}\right) + 2^2}{14^2 + 2^2} = 0.667$$

The value of ρ can be used to apply Generalized difference model as described in the procedure for known-rho

Cochrane-Orcutt Method

This method uses Residuals to estimate ρ. It is an iterative process and is applicable with First-Order Autocorrelation only

STEPS

Consider $Y_t = \beta_1 + \beta_2 X_t + e_t$

- Round 1: apply OLS, find residuals, compute $\rho = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$
- Round 2: Perform GLS

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + e_t - \rho\, e_{t-1}$$

Now obtain second order residual and ρ again

- Round 3 (if required): Perform GLS again with the new ρ and continue in the same way

We continue until the estimated ρ from two successive rounds is almost equal. We also may perform Durbin Watson test and stop iteration if AC is not detected.

Cochrane-Orcutt Method: EXAMPLE

We will just demonstrates some rounds for your help

Cochraine-Orcutt method			Round 1			
Time Period	**Y**	**X**	**e_t**	**et-1**	**$e_t.e_{t-1}$**	**e_t^2**
1	41	3.1	-6.69226			44.78636
2	51	3	3.57925	-6.69226	-23.95325	12.81100
3	55	3.3	6.76473	3.57925	24.21261	45.76151
4	58	4	7.86418	6.76473	53.19901	61.84531
5	56	4.3	5.04966	7.86418	39.71142	25.49906
6	51	4.7	-1.03637	5.04966	-5.23330	1.07406
7	46	5.4	-7.93691	-1.03637	8.22556	62.99460
8	46	4.9	-6.57938	-7.93691	52.21998	43.28825
9	48	6	-7.56595	-6.57938	49.77929	57.24366
10	57	7.5	-2.63855	-7.56595	19.96317	6.96196
11	58	6.9	-0.00951	-2.63855	0.02510	0.00009
12	57	7.7	-3.18157	-0.00951	0.03027	10.12237
13	62	7.9	1.27542	-3.18157	-4.05783	1.62670
14	64	8.1	2.73241	1.27542	3.48497	7.46605
15	71	8.6	8.37487	2.73241	22.88356	70.13851
					240.49056	406.83312
					Sum	Sum

Consider $Y_t = \beta_1 + \beta_2 X_t + e_t$

We have just demonstrates some rounds for your help

Run OLS Y on X and obtain e	
OLS:Y on X	
Intercept	39.27556
Slope	2.715066
First Round Rho = 0.591128	
round 2: transformed model	
OLS:Y on X	
Intercept	18.33661
Slope	2.166349
Second Round Rho = 0.520366	
round 3:transformed model	
Intercept	19.55304
Slope	2.794671

We will transform the model again and continue till the rho from two consecutive iterations is almost the same. This may be a lengthy process.

Cochraine Orcutt Transformed variables & Round 2					
Y*	**X***	**et**	**et-1**	**et.et-1**	**et2**
26.76374	1.167502	5.89791			
24.85246	1.526615	3.208663	5.89791	18.9244	10.29552
25.48794	2.049277	2.711883	3.208663	8.701517	7.354307
21.71456	1.935487	-0.81499	2.711883	-2.21017	0.664215
17.89682	2.158148	-5.1151	-0.81499	4.168775	26.16425
15.85246	2.621697	-8.16367	-5.1151	41.75797	66.64546
18.8081	1.707907	-3.22844	-8.16367	26.35589	10.42281
20.8081	3.103471	-4.25172	-3.22844	13.7264	18.0771
28.62584	3.95323	1.725152	-4.25172	-7.33486	2.97615
24.30569	2.466538	0.625692	1.725152	1.079415	0.391491
22.71456	3.621215	-3.46687	0.625692	-2.16919	12.01918
28.30569	3.348312	2.715461	-3.46687	-9.41415	7.37373
27.35005	3.430087	1.582668	2.715461	4.297674	2.504838
33.16779	3.811861	6.573355	1.582668	10.40344	43.209
				108.2871	208.098

Hildreth-Lu Search Procedure

Step 1: Choose a grid of possible values (between -1 and +1) of ρ

Step 2: For each value of the grid, estimate the Generalized difference model and find the sum of square residuals

Step 3: The equation with min SS will be considered the best equation

Modified Step 1: Choose a grid of possible values (between -1 and +1) of ρ

We can select a simple grid like 0, 0.1, 0.2, 0.3 - - - - 0.9, 1. After repeating the previous process, select a further grid. For example if the first grid gives us 0.7, we start with another grid based on this information around 0.7 like 0.66, 0.67, 0.68, 0.69, 0.71, 0.72, 0.73, 0.74 and repeat the above procedure.

Problems:

- Time consuming and long process
- Grid values must be carefully selected to have a global minimum SS (not just a local min)

Comparison:

- Hildreth-Lu procedure is computer-time-intensive as compared to Cochrane-Orcutt procedure
- The Cochrane-Orcutt procedure iterates to a local minimum and may miss the global minimum SS or residuals

Lecture 23
Estimating Non-Linear equation by OLS

Estimating Quadratic Equation by OLS

Consider the relations of variables that provide U-shaped or inverted U-shaped curves. Some time we must use quadratic equations to capture the relationship. We introduce the square of the independent variable and include in regression. For U-shaped (convex) curves the coefficient of the quadratic term is positive. For Inverted-U-shaped (or concave) curves the coefficient of the quadratic term is negative.

Consider the quadratic equation (subscripts ignored)

$$Y = a + b\,X + c\,X^2 + e$$

Applying optimization (minimize the sum of squared residuals)

$$Min\,Z = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \hat{Y})^{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \boldsymbol{a} - \boldsymbol{bX} - \boldsymbol{cX^2})^{\boldsymbol{2}}$$

We have three parameters (a, b and c) so we need three partial derivatives set equal to zero

First partial derivative w.r.t. a

$$Z_a = \sum (Y - a - bX - cX^2)^{2-1} . \frac{\partial}{\partial a}(Y - a - bX - cX^2) = 0$$

$$\sum (Y - a - bX - cX^2)(-1) = 0$$

This gives the first normal equation as

$$\sum Y = na + b\sum X + c\sum X^2$$

Differentiating w.r.t. *b* & *c* we get the other normal equations

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

And

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

Hence for estimating a regression equation

We need to solve three normal equations

$$\sum Y = na + b\sum X + c\sum X^2$$

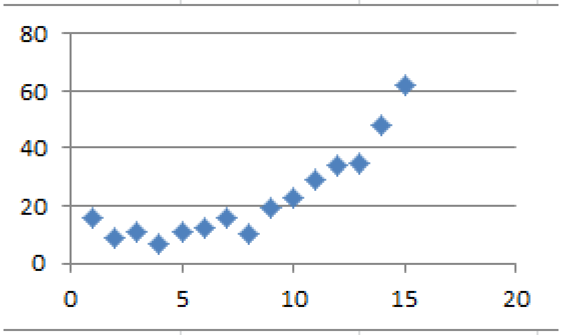
$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

For which we need all the columns and sums shown in the three normal equations. However, in the example, we will use LINEST function of Microsoft Excel.

Example I:

Y	X	X^2
16	1	1
9	2	4
11	3	9
7	4	16
11	5	25
12	6	36
16	7	49
10	8	64
19	9	81
23	10	100
29	11	121
34	12	144
35	13	169
48	14	196
62	15	225



Result of =LINEST() with quadratic term		
0.443842922	-4.019343891	18.26373626
0.048788309	0.802761669	2.791214341
0.967180725	3.133612459	
176.8193957	12	
3472.565676	**117.8343245**	
Regression line		
Y = 18.26 - 4.01934 X + 0.44 X^2		
SSR from linear equation:	930.5107	

Example I Modified: Centering X

Y	x=X-Xbar	x^2
16	-7	49
9	-6	36
11	-5	25
7	-4	16
11	-3	9
12	-2	4
16	-1	1
10	0	0
19	1	1
23	2	4
29	3	9
34	4	16
35	5	25
48	6	36
62	7	49

Old Result of =LINEST() with quadratic term		
0.443842922	-4.019343891	18.26373626
0.048788309	0.802761669	2.791214341
0.967180725	3.133612459	
176.8193957	12	
3472.565676	117.8343245	

Result of =LINEST()		
0.443842922	3.082142857	14.51493213
0.048788309	0.187269163	1.218210622
0.967180725	3.133612459	
176.8193957	12	
3472.565676	117.8343245	

Regression line

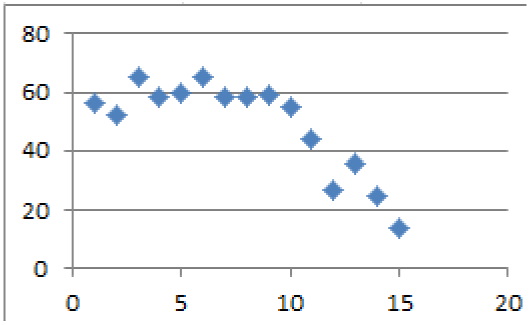
Y = 14.51 -3.082 X + 0.44 X^2

SSR from linear equation: 930.5107

. Y can be better predicted	Correl(X,X^2)= 0.97
. No correlation in expl vars	Correl(x,x^2)= 0

Example II:

Y	X	X^2
56	1	1
52	2	4
65	3	9
58	4	16
60	5	25
65	6	36
58	7	49
58	8	64
59	9	81
55	10	100
44	11	121
27	12	144
36	13	169
25	14	196
14	15	225



Result of =LINEST() with quadratic term		
-0.495475113	5.099030381	48.96703297
0.077250272	1.271074145	4.419543828
0.916745492	4.961689038	
66.06816929	12	
3252.979703	295.4202973	
Regression line		
Y = 48.97 + 5.099 X - 0.4955 X^2		
SSR from linear equation:		1308.171

Estimating Cubic Equation by OLS

Consider the relations of Cubic nature like the Total Cost function. We introduce the square and cube of the independent variable and include in regression. For Total Cost function, to get the required shape the coefficient of quadratic term is expected to be negative and that of cubic term should be non-zero. Consider the quadratic equation (subscripts ignored)

$$Y = a + b\,X + c\,X^2 + d\,X^3 + e$$

Applying optimization (minimize the sum of squared residuals)

$$Min\,Z = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{e_i^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y} - \hat{Y})^{\boldsymbol{2}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y - a - bX - cX^2 - dX^3})^{\boldsymbol{2}}$$

This time we have four parameters (a, b, c and d) so we need four partial derivatives set equal to zero. First partial derivative w.r.t. a

$$Z_a = \sum (Y - a - bX - cX^2 - dX^3)^{2-1} . \frac{\partial}{\partial a}(Y - a - bX - cX^2 - dX^3) = 0$$

$$\sum (Y - a - bX - cX^2 - dX^3)(-1) = 0$$

This gives the first normal equation as

$$\sum Y = na + b\sum X + c\sum X^2 + d\sum X^3$$

Differentiating w.r.t. *b* & *c* we get the other normal equations. Hence for estimating a regression equation $Y = a + b\,X + c\,X^2 + d\,X^3 + e$, We need to solve four normal equations

$$\sum Y = na + b\sum X + c\sum X^2 + d\sum X^3$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3 + d\sum X^4$$

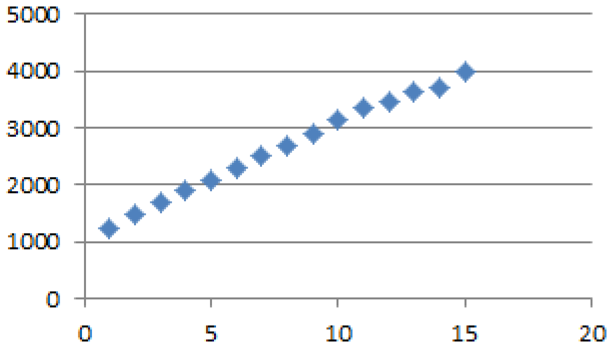
$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 + d\sum X^5$$

$$\sum X^3Y = a\sum X^3 + b\sum X^4 + c\sum X^5 + d\sum X^6$$

For which we need all the columns and sums shown in the three normal equations. However, in the example, we will use LINEST function of Microsoft Excel.

Consider the following example:

Y	X	X^2	X^3
1241	1	1	1
1470	2	4	64
1680	3	9	729
1900	4	16	4096
2100	5	25	15625
2290	6	36	46656
2500	7	49	117649
2680	8	64	262144
2900	9	81	531441
3150	10	100	1000000
3350	11	121	1771561
3450	12	144	2985984
3650	13	169	4826809
3700	14	196	7529536
4000	15	225	11390625



Result of =LINEST() with quadratic and cubic term			
-9.8497E-06	-1.305954071	221.8711228	1023.7397
1.33877E-05	1.611087348	18.97304895	48.58616
0.998021737	43.86210487		
1849.811606	11		
10676470.21	21162.72668		
Regresssion: Y = 1023.7 +221.87 X - 1.306 X^2 + 0.00000984 X^3			
SSR from linear equation: 45562.33			

Estimating Cobb-Douglas Production Function

Consider the Cobb-Douglas Production Function

$$Q = A\, l^{\alpha} k^{\beta}$$

Taking log on both sides

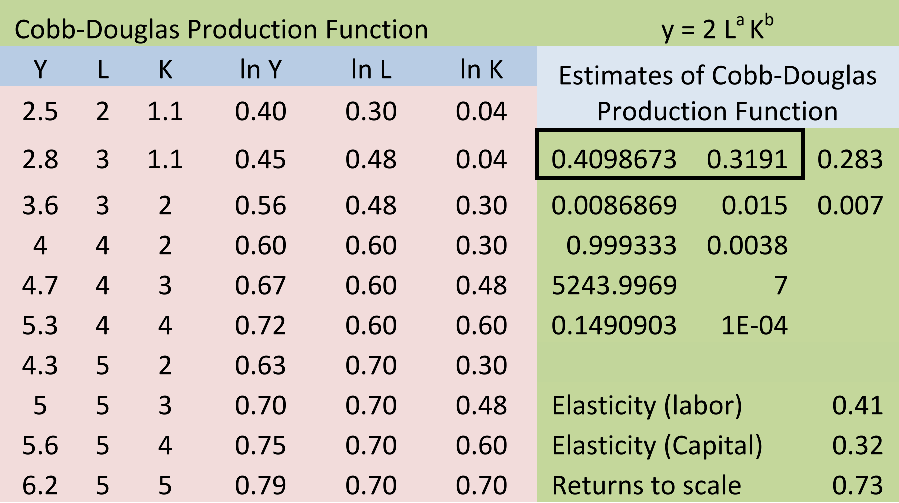
$$\ln Q = \ln A + \beta \ln l + \gamma \ln k$$

Let $Y = \ln Q,\ \alpha = \ln A\ and\ \ L = \ln l,\ \ K = \ln k$ then the above can be written as

$$Y = \alpha + \beta L + \gamma K$$

That can be estimated by OLS.

Consider the following example:



Cobb-Douglas Production Function						y = 2 $L^a K^b$		
Y	L	K	ln Y	ln L	ln K	Estimates of Cobb-Douglas Production Function		
2.5	2	1.1	0.40	0.30	0.04			
2.8	3	1.1	0.45	0.48	0.04	0.4098673	0.3191	0.283
3.6	3	2	0.56	0.48	0.30	0.0086869	0.015	0.007
4	4	2	0.60	0.60	0.30	0.999333	0.0038	
4.7	4	3	0.67	0.60	0.48	5243.9969	7	
5.3	4	4	0.72	0.60	0.60	0.1490903	1E-04	
4.3	5	2	0.63	0.70	0.30			
5	5	3	0.70	0.70	0.48	Elasticity (labor)		0.41
5.6	5	4	0.75	0.70	0.60	Elasticity (Capital)		0.32
6.2	5	5	0.79	0.70	0.70	Returns to scale		0.73

Transformation of models and Use of OLS

Method	Transformation	Regression equation
Standard linear	Not required	$y = b_0 + b_1 x$
Exponential model	Dependent variable = log(y)	$\log(y) = b_0 + b_1 x$
Logarithmic model	Independent variable = log(x)	$y = b_0 + b_1 \log(x)$
Double log	Dependent variable = log(y),	$\log(y) = b_0 + b_1 \log(x)$
Cobb Douglas	$Y = A\,L^{\alpha}K^{\beta}$	$\ln Y = \ln A + \alpha \ln L +$

Interpretation of different functional forms using OLS

Model	**Interpretation**	**Marginal Effect**	**Elasticity**
Linear in variable $Y = a + bX$	One unit change in X will cause, on the average, 'b' units of change in Y	b	$b\frac{X}{Y}$
Double log form (log-log)	One percent change in X will cause, on the average, 'b' % change in Y	$b\frac{Y}{X}$	b
Level-Log $Y = a + b\,ln\,X$	One percent change in X is expected to change Y by $\frac{100}{b}$ units	$\frac{b}{X}$	$\frac{b}{Y}$
Log-Level form $lnY = a + b\,X$	When X changes by one unit, Y will change by approximately (b*100)%	bY	bX
For interpretation, we assume that Gauss Markov assumption hold and parameters are significant			
Marginal effect of X is defined as the partial derivative of Y w.r.t. X			
marginal effect and elasticity ($\frac{dy}{dx}\frac{X}{Y}$) may be computed at mean values of X and Y			

Problems with OLS

We must take care of

- Outliers
- Non - linearity
- Wrong specification
- Missing Variables
- Multicollinearity
- Heteroskedasticity
- Autocorrelation

Taking care in use of OLS: 7 Questions

1. Are the explanatory variables helping the model?
2. Are the relationships what we expect?
3. Are any of the explanatory variables redundant?
4. Are the residuals normal?
5. Have all important variables been included?
6. How well the model explains the dependent variable?
7. Are the results free form MC, HSK & AC?

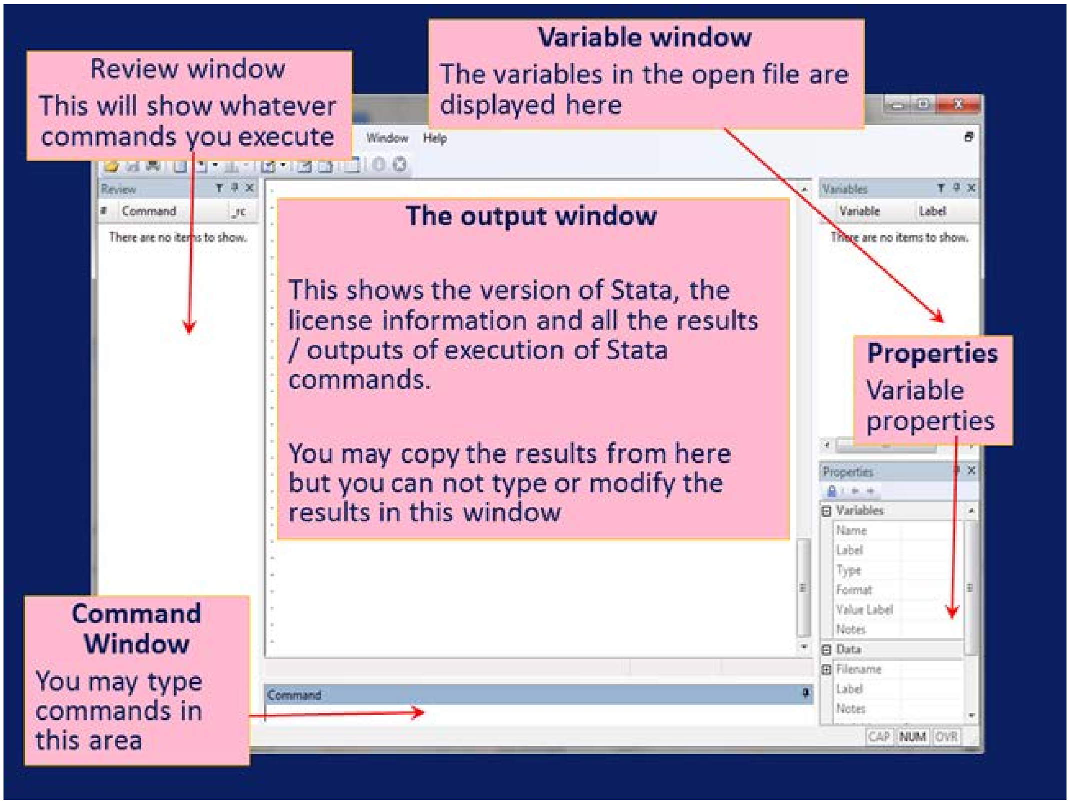
Remember that

- Explanatory variables can have categories
- We can use dummy variables for the above
- But if the dependent variable is categorical, do not use simple regression (Examples)
- Take care of origin and scale (magnitude of coefficients is better interpretable)
- Model not linear in parameters cannot be transformed for OLS

Lecture 24

Introduction to Stata- I

The stat environment looks as follows:



Review window

This will show whatever commands you execute

Variable window

The variables in the open file are displayed here

The output window

This shows the version of Stata, the license information and all the results / outputs of execution of Stata commands.

You may copy the results from here but you can not type or modify the results in this window

Properties

Variable properties

Command Window

You may type commands in this area

We have five windows visible. The main window is the output window in which you see all the output. Below that you will see a command window in which we can type and enter/execute the commands. Although stata has a good graphic user interface, we can efficiently use the command line for quick and efficient work. On the right you have a variable window in which all the variables will be displayed and below that there is a Properties window which will give the properties of all variables like their names, storage types, display format etc.

The review window on the left will give all the commands that have been executed in the current session. We can just click on any of them to execute them again instead of retyping the command. We have several type of files that we use in Stata. Data files have extension .dta although we can type file names without this extension and they will be treated as dta files. Log files can be also be used to record our session (.smcl or .log). We have also Do files that perform several operations at a time (.do). These files provide a list of commands that can be executed at one click.

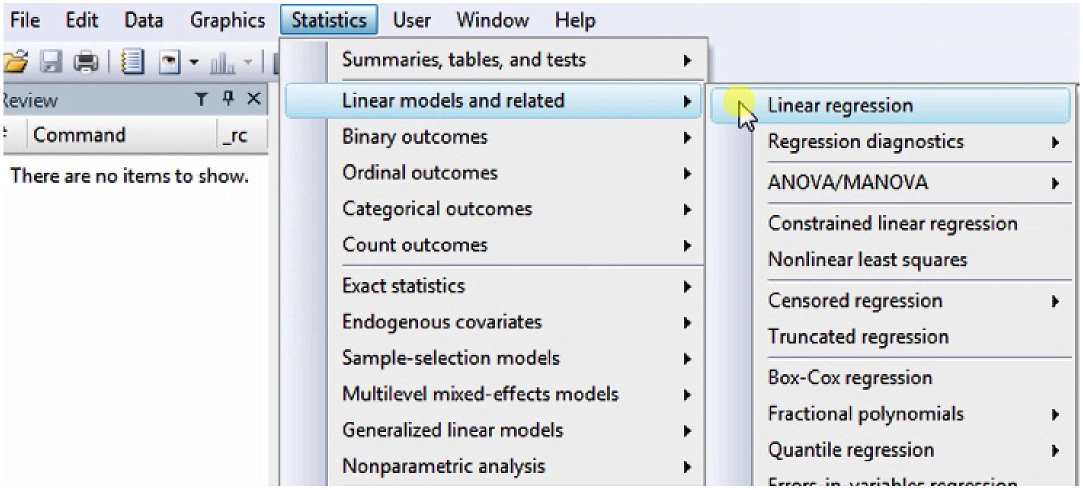
The data files (.dta) are of three types:

- Sample data files Shipped with Stata (opened by sysuse command)
- Files available on the web (using the webuse command)
- Files that you create and save on your hard disk (use command)

Other data files (like MS Excel) can be imported into stata.

The graphic use interface

A strong graphic user interface is available where we can perform operations by clicking on different icons. This is very user friendly. See some example below:



There are no items to show.

The Command window

We can use the command window to type and execute different commands. All that can be performed by the graphic user interface can be performed in the command window. This is undoubtedly very quick way of using stata.

This is useful in many ways

- Typing Commands is less time consuming
- You remember the commands in this way
- It may be helpful if you need to program in Stata

For the commands

- Use small letters; Stata is case sensitive
- Short names can be used (e.g. d / des instead of describe)
- Syntax can be viewed in the help window
- Underlined part in the syntax is the shortcut
- Commands have options (after typing comma ,)
- We can use conditions with commands (e.g. if, in etc.)

The 'help' command

The syntax is

help [command or topic] [,options]

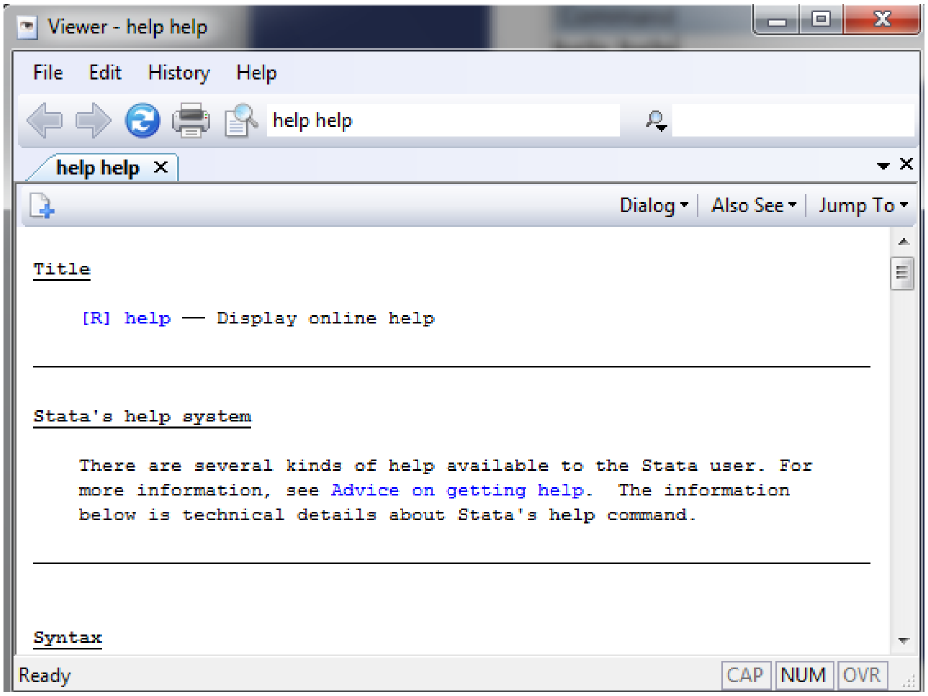
The underlined part is the minimum that you should write for help. If we just type *h describe* , we will get a window showing help for '*describe*'.

From the graphic user interface, we can click in this order

Help > Stata Command...

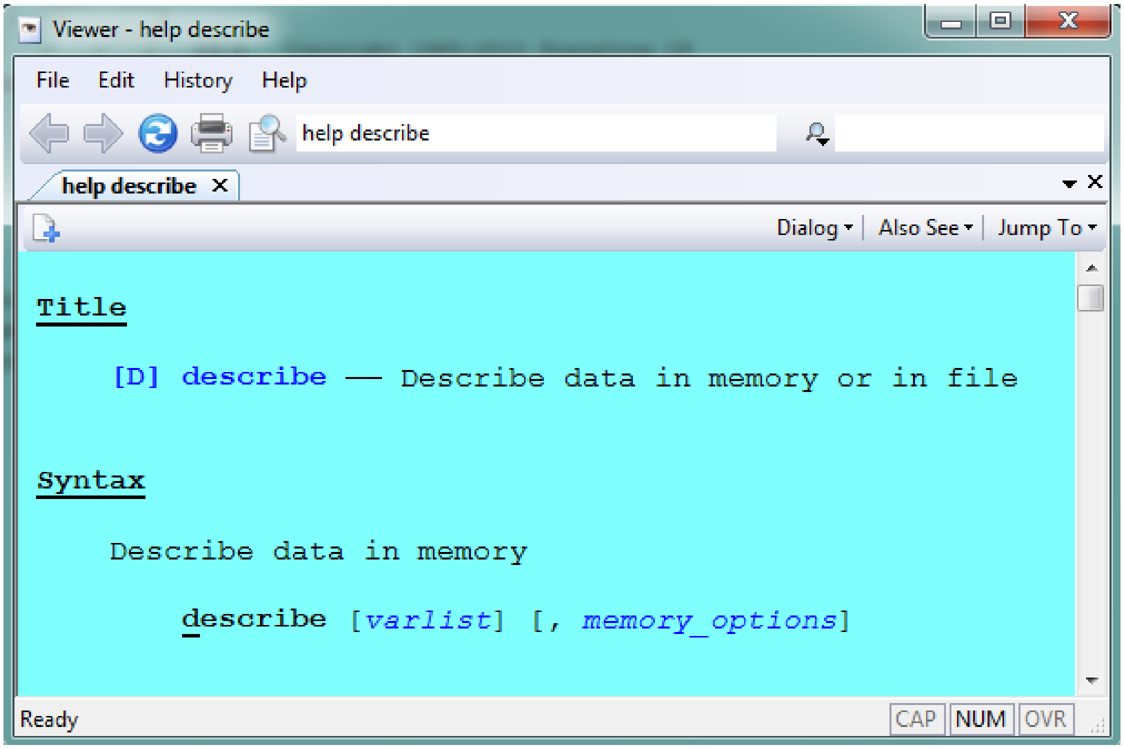
and write the command in the window that appears.

Typing '*help help*' in the command window opens the following window:



```
Title

    [R] help — Display online help

Stata's help system

    There are several kinds of help available to the Stata user. For
    more information, see Advice on getting help.  The information
    below is technical details about Stata's help command.

Syntax
```

Typing ‘help describe’ opens the following window:



```
Title

    [D] describe — Describe data in memory or in file

Syntax

    Describe data in memory

        describe [varlist] [, memory_options]
```

Almost all the help windows have the same structure. They show the command, its shortcut, its syntax, description and practical example of the use of the command.

Opening files in Stata

Sample data files Shipped with Stata (type in the command window: *sysuse filename*)

For the files on the web type *webuse filename*. On the hard disk of your computer (*use path and file name*)

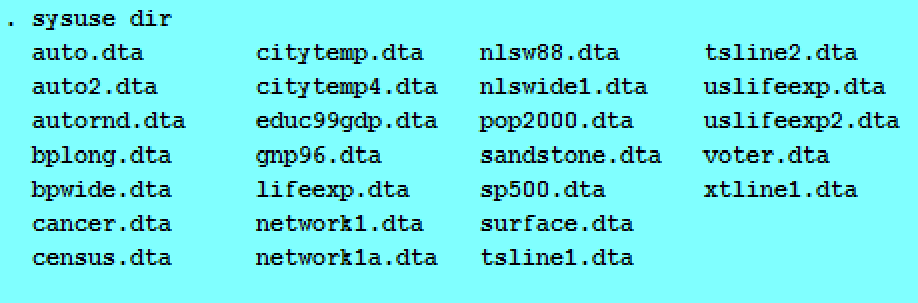
We can use *,clear* as an option.

To open file on your computer using the graphic user interface,

File > open (brows, find file and open)

Alternatively type: *use "path and file name"*or *use "path and file name", clear*

To know the files already available with stata type *sysuse dir*

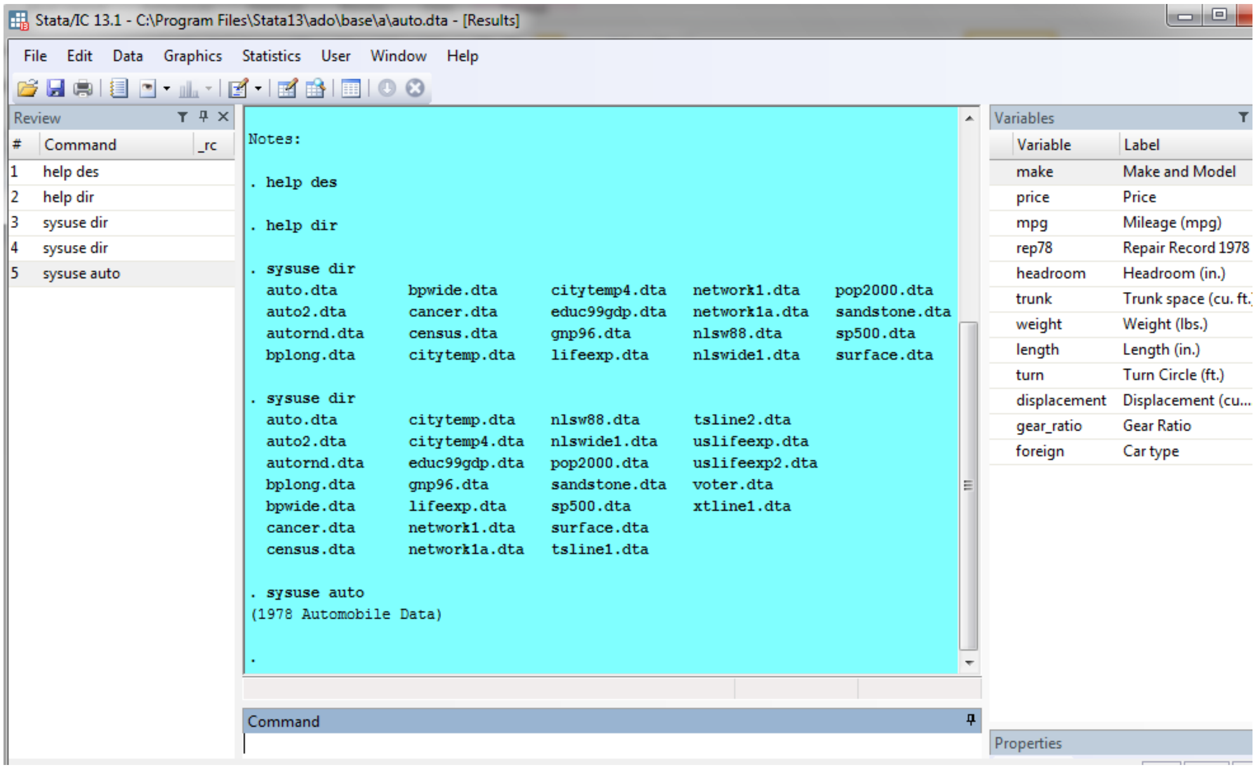


```
. sysuse dir
  auto.dta        citytemp.dta    nlsw88.dta      tsline2.dta
  auto2.dta       citytemp4.dta   nlswide1.dta    uslifeexp.dta
  autornd.dta     educ99gdp.dta   pop2000.dta     uslifeexp2.dta
  bplong.dta      gnp96.dta       sandstone.dta   voter.dta
  bpwide.dta      lifeexp.dta     sp500.dta       xtline1.dta
  cancer.dta      network1.dta    surface.dta
  census.dta      network1a.dta   tsline1.dta
```

A list of files will be seen in the output window. Any one of them can be opened.

For example, to open the first file type *sysuse auto*

File will open; a message will appear in the output window. Variables will be show in the variable window. Variable description also will be shown in the properties window. Note that all commands that you have been typing are visible in the review window on the left.



#	Command	_rc
1	help des	
2	help dir	
3	sysuse dir	
4	sysuse dir	
5	sysuse auto	

```
Notes:

. help des

. help dir

. sysuse dir
  auto.dta       bpwide.dta     citytemp4.dta   network1.dta    pop2000.dta
  auto2.dta      cancer.dta     educ99gdp.dta   network1a.dta   sandstone.dta
  autornd.dta    census.dta     gnp96.dta       nlsw88.dta      sp500.dta
  bplong.dta     citytemp.dta   lifeexp.dta     nlswide1.dta    surface.dta

. sysuse dir
  auto.dta       citytemp.dta   nlsw88.dta      tsline2.dta
  auto2.dta      citytemp4.dta  nlswide1.dta    uslifeexp.dta
  autornd.dta    educ99gdp.dta  pop2000.dta     uslifeexp2.dta
  bplong.dta     gnp96.dta      sandstone.dta   voter.dta
  bpwide.dta     lifeexp.dta    sp500.dta       xtline1.dta
  cancer.dta     network1.dta   surface.dta
  census.dta     network1a.dta  tsline1.dta

. sysuse auto
(1978 Automobile Data)

.
```

Variable	Label
make	Make and Model
price	Price
mpg	Mileage (mpg)
rep78	Repair Record 1978
headroom	Headroom (in.)
trunk	Trunk space (cu. ft.
weight	Weight (lbs.)
length	Length (in.)
turn	Turn Circle (ft.)
displacement	Displacement (cu...
gear_ratio	Gear Ratio
foreign	Car type

The *describe* command

Describe provides a description of data in the current file.

We use the 'describe' command for description of data in a file. The describe command lists the variables, labels, formats, storage type, number of observations, and date file was created

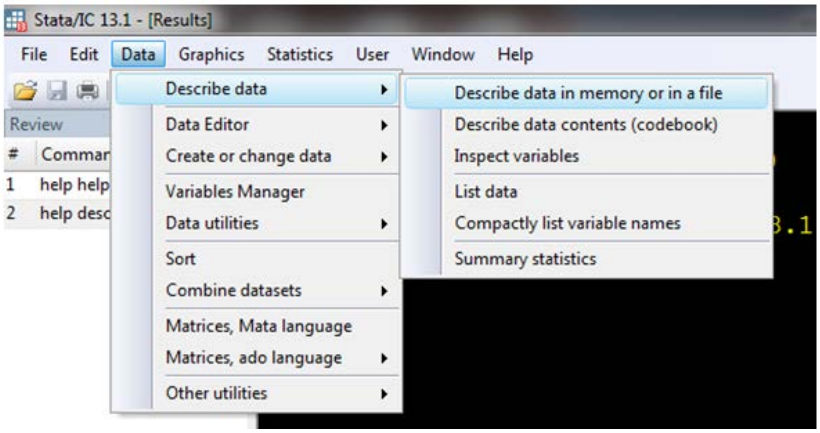
It can be used mainly in three ways

- *describe*: it will describe all the variables in memory
- *describe variable-names*: this will describe only the variables specified
- *describe variable1-variablen*: this will describe all the variables in range

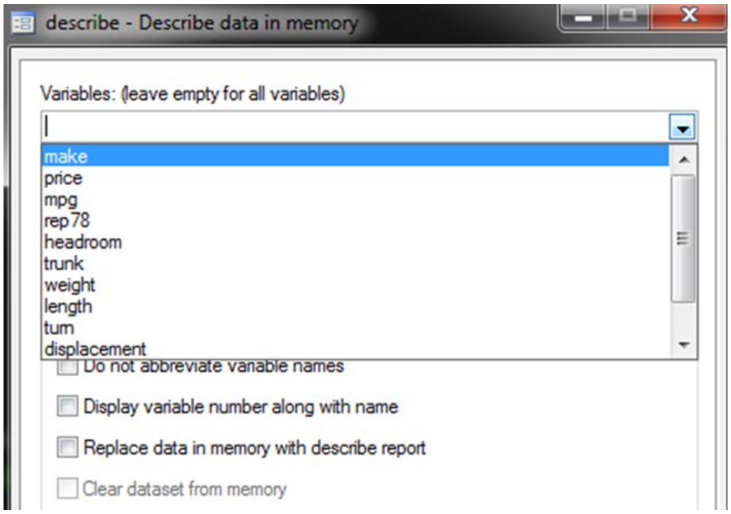
To use the graphic user interface click as follows:

Data > Describe data > Describe data in memory or in a file

A window will appear asking variable names etc. You need to use the pull-down menu to give the variable names (if file is open)



Here is what you may see:


describe - Describe data in memory

Variables: (leave empty for all variables)

make
price
mpg
rep78
headroom
trunk
weight
length
turn
displacement

□ Do not abbreviate variable names

□ Display variable number along with name

□ Replace data in memory with describe report

□ Clear dataset from memory

Write the variable names with the help of pull down menu and click OK.

EXAMPLES

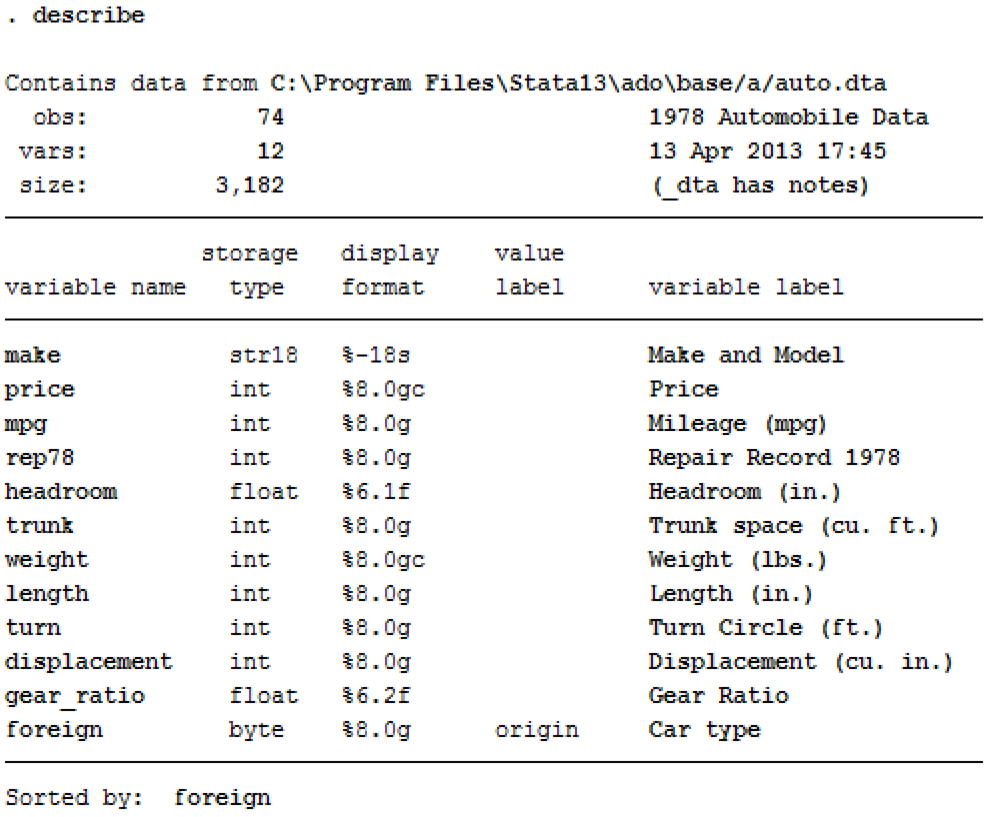
sysuse auto

describe

describe mpg price

describe price – length

Typing describe will give the following output:



```
. describe

Contains data from C:\Program Files\Stata13\ado\base/a/auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          13 Apr 2013 17:45
 size:         3,182                          (_dta has notes)
---------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
---------------------------------------------------------------------
make            str18  %-18s                 Make and Model
price           int    %8.0gc                Price
mpg             int    %8.0g                 Mileage (mpg)
rep78           int    %8.0g                 Repair Record 1978
headroom        float  %6.1f                 Headroom (in.)
trunk           int    %8.0g                 Trunk space (cu. ft.)
weight          int    %8.0gc                Weight (lbs.)
length          int    %8.0g                 Length (in.)
turn            int    %8.0g                 Turn Circle (ft.)
displacement    int    %8.0g                 Displacement (cu. in.)
gear_ratio      float  %6.2f                 Gear Ratio
foreign         byte   %8.0g      origin     Car type
---------------------------------------------------------------------
Sorted by:  foreign
```

This gives the variable names, the data type, the display type, and any labels if provided with the variables. We get to know the current file at a glance with all variables.

The *list* command

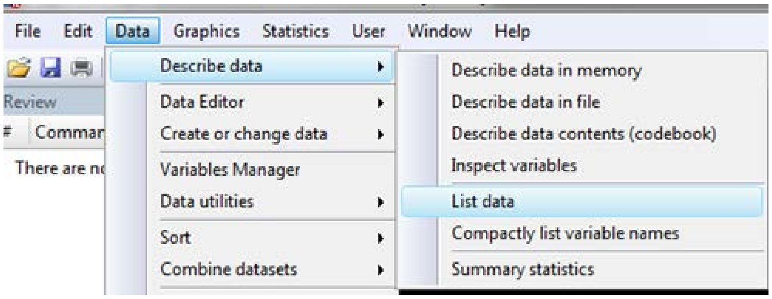
The list command lists rows and columns of the data file.

The list command provides the values of observations of data

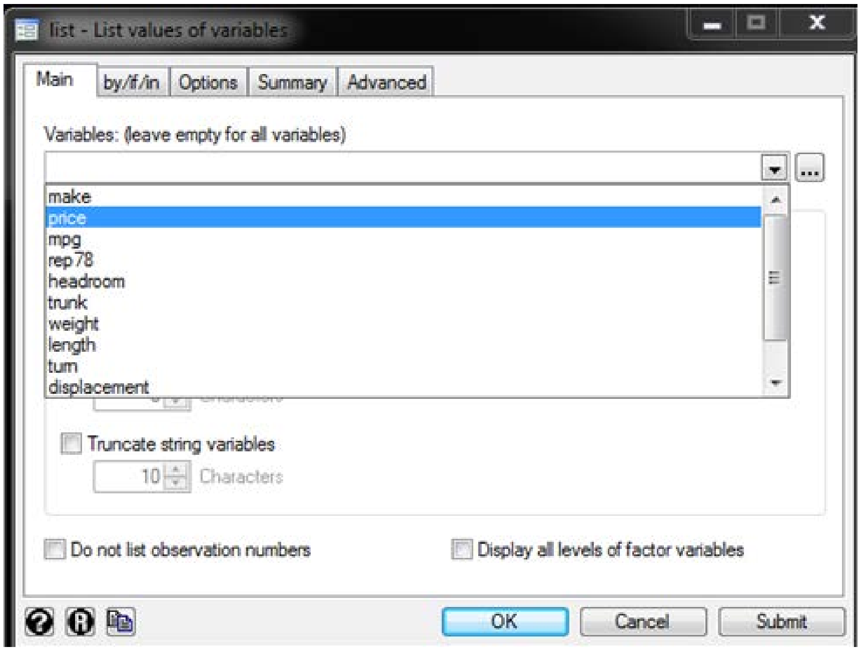
It can be used mainly in three ways (as in describe)

- *list*: it will list values of all the variables in memory
- *list variable-names*: this will list the values of only the variables specified
- *list variable1-variablen*: this will list values of all the variables in range

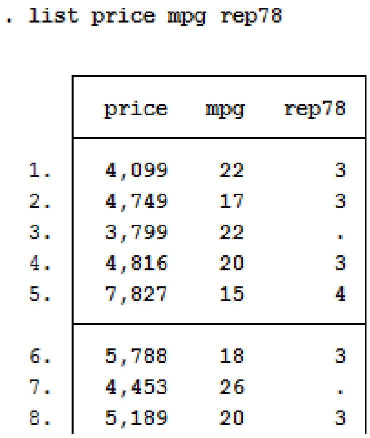
Form the Graphic interface:



A window will appear asking variable names etc. You need to use the pull-down menu to give the variable names (if file is open)


list - List values of variables
Main
by/if/in
Options
Summary
Advanced
Variables: (leave empty for all variables)
make
price
mpg
rep78
headroom
trunk
weight
length
turn
displacement
□ Truncate string variables
10 Characters
□ Do not list observation numbers
□ Display all levels of factor variables
OK
Cancel
Submit

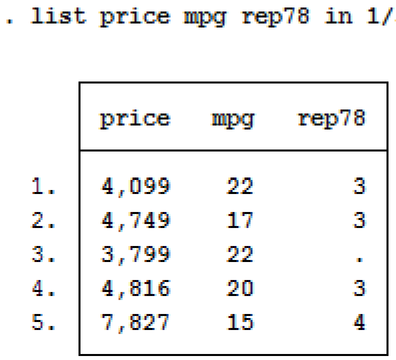
Select the variables and click OK



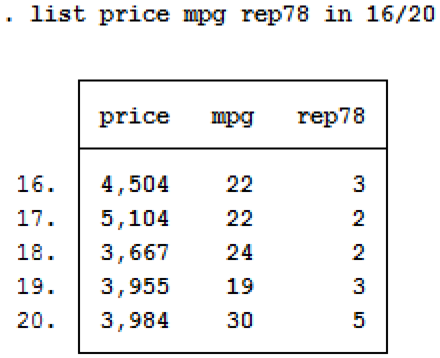
```
. list price mpg rep78

     +----------------------+
     | price   mpg   rep78  |
     |----------------------|
  1. | 4,099    22       3  |
  2. | 4,749    17       3  |
  3. | 3,799    22       .  |
  4. | 4,816    20       3  |
  5. | 7,827    15       4  |
     |----------------------|
  6. | 5,788    18       3  |
  7. | 4,453    26       .  |
  8. | 5,189    20       3  |
```

Using list with in and if

We can impose conditions for filtered data to be displayed (using ‘if’). We can specify rows of data to be displayed (data ‘in’ specific rows)

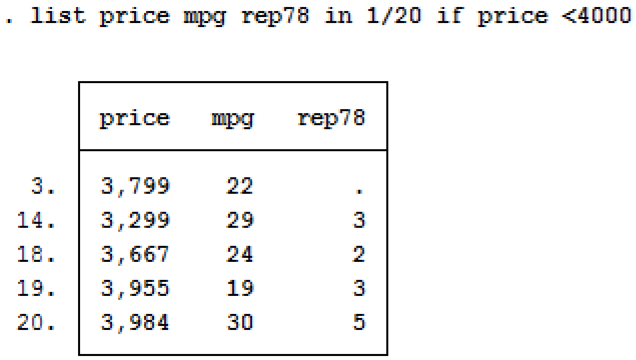


```
. list price mpg rep78 in 1/5

     +----------------------+
     | price   mpg   rep78  |
     |----------------------|
  1. | 4,099    22       3  |
  2. | 4,749    17       3  |
  3. | 3,799    22       .  |
  4. | 4,816    20       3  |
  5. | 7,827    15       4  |
     +----------------------+
```



```
. list price mpg rep78 in 16/20

     +----------------------+
     | price   mpg   rep78  |
     |----------------------|
 16. | 4,504    22       3  |
 17. | 5,104    22       2  |
 18. | 3,667    24       2  |
 19. | 3,955    19       3  |
 20. | 3,984    30       5  |
     +----------------------+
```

Conditions also may be attached



```
. list price mpg rep78 in 1/20 if price <4000

     +----------------------+
     | price   mpg   rep78  |
     |----------------------|
  3. | 3,799    22       .  |
 14. | 3,299    29       3  |
 18. | 3,667    24       2  |
 19. | 3,955    19       3  |
 20. | 3,984    30       5  |
     +----------------------+
```

Lecture 25

Introduction to Stata- II

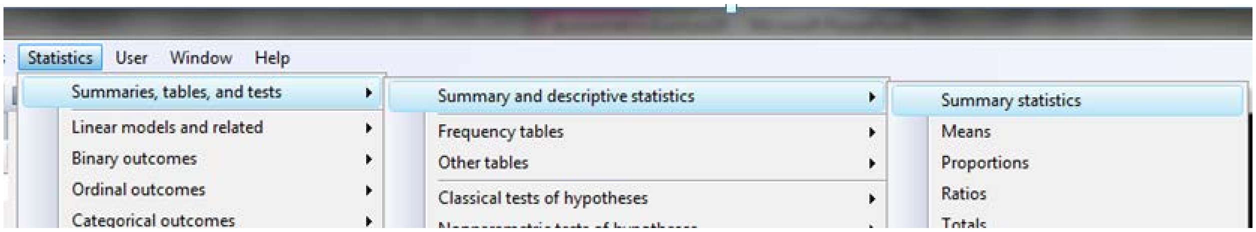
The *summarize* Command

The syntax is

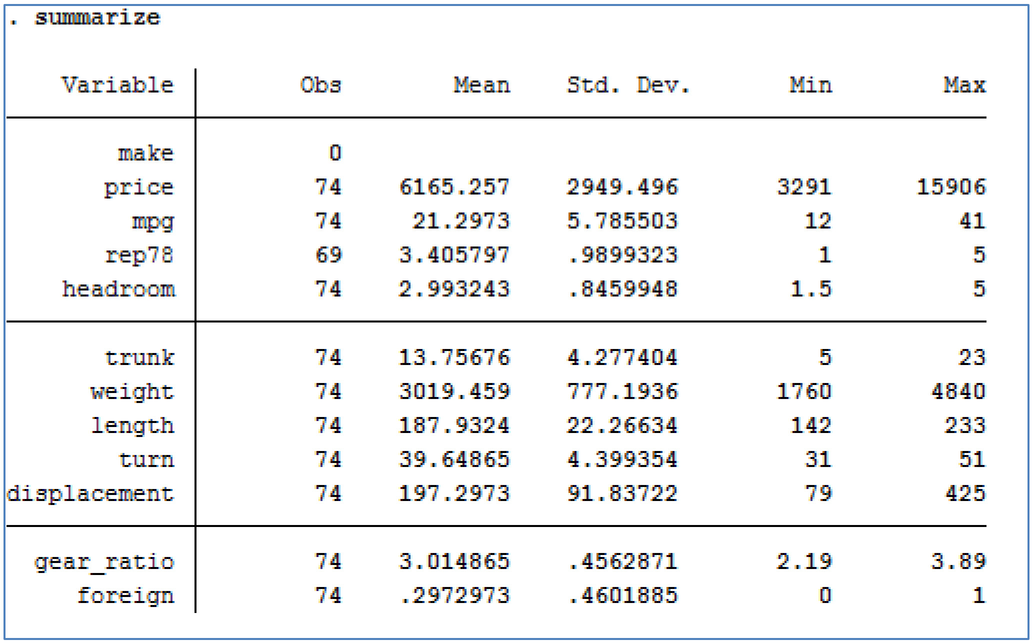
<u>Su</u>mmarize [varlist] [if] [in] [weight] [, options]

From the menu click:

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics



If the file auto is open (*sysuse auto*), summarize will give the following results containing the summary statistics of the variables in the currently open file.



```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        make |         0
       price |        74    6165.257    2949.496       3291      15906
         mpg |        74     21.2973    5.785503         12         41
       rep78 |        69    3.405797    .9899323          1          5
    headroom |        74    2.993243    .8459948        1.5          5
-------------+--------------------------------------------------------
       trunk |        74    13.75676    4.277404          5         23
      weight |        74    3019.459    777.1936       1760       4840
      length |        74    187.9324    22.26634        142        233
        turn |        74    39.64865    4.399354         31         51
displacement |        74    197.2973    91.83722         79        425
-------------+--------------------------------------------------------
  gear_ratio |        74    3.014865    .4562871       2.19       3.89
     foreign |        74    .2972973    .4601885          0          1
```

Like the previous commands, summarize can be used as standalone, with variables or with a range of variables.

EXAMPLES

sysuse auto, clear

summarize

summarize price mpg length

summarize price-length

The *tabulate* command

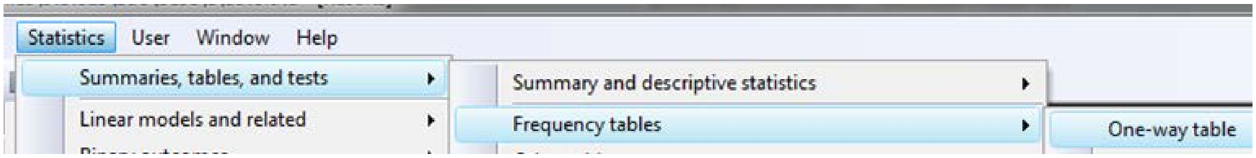
The command tabulate provides one way or two way frequency tables.

Syntax of the command is:

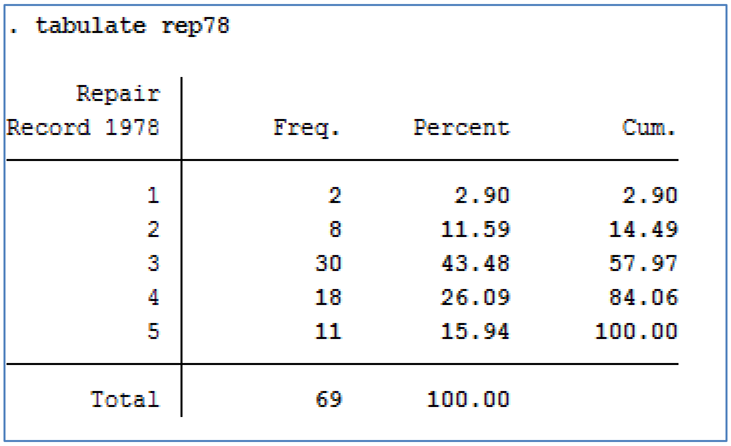
tabulate varname [if] [in] [weight] [, tabulate1_options]

From the menu, the command can be executed by clicking as:

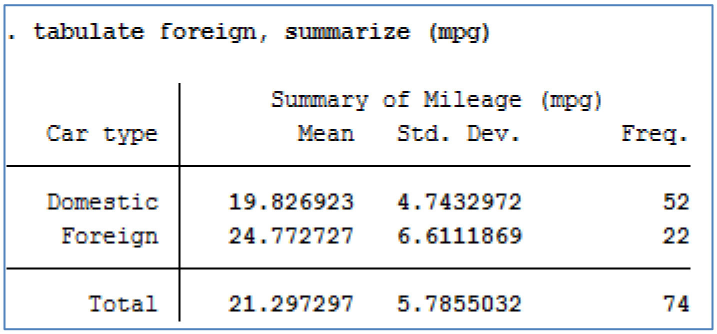
Statistics > Summaries, tables, and tests > Frequency tables > One-way table


Statistics
User
Window
Help
Summaries, tables, and tests
Linear models and related
Summary and descriptive statistics
Frequency tables
One-way table

The command ‘tabulate’ needs a variable name. See the following examples:



```
. tabulate rep78

     Repair │
Record 1978 │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
          1 │          2        2.90        2.90
          2 │          8       11.59       14.49
          3 │         30       43.48       57.97
          4 │         18       26.09       84.06
          5 │         11       15.94      100.00
────────────┼───────────────────────────────────
      Total │         69      100.00
```

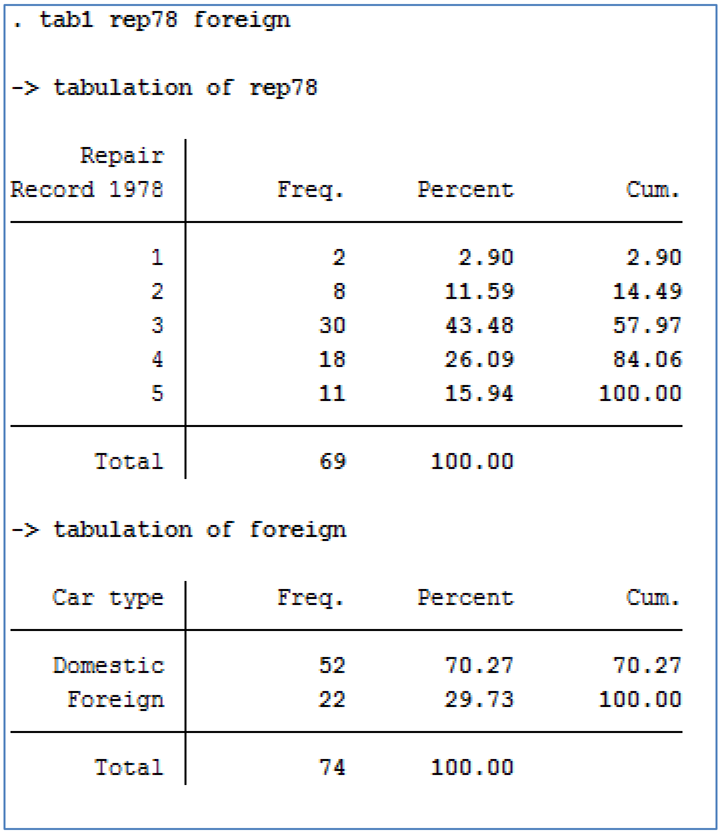


```
. tabulate foreign, summarize (mpg)

            │      Summary of Mileage (mpg)
   Car type │        Mean   Std. Dev.       Freq.
────────────┼────────────────────────────────────
   Domestic │   19.826923   4.7432972          52
    Foreign │   24.772727   6.6111869          22
────────────┼────────────────────────────────────
      Total │   21.297297   5.7855032          74
```

The above command provides summary statistics of the variable *mpg* according to categories in the variable *foreign*. In this way we use *summarize* as an option of the *tabulate* command. Remember that all options for any command in stata must be given after a comma (,).

In the tabulate command we can tabulate one variable at a time. For example if we need frequency tables for *rep78* and *foreign* (in the file auto), we need to type and execute the tabulate command separately for both the variables.

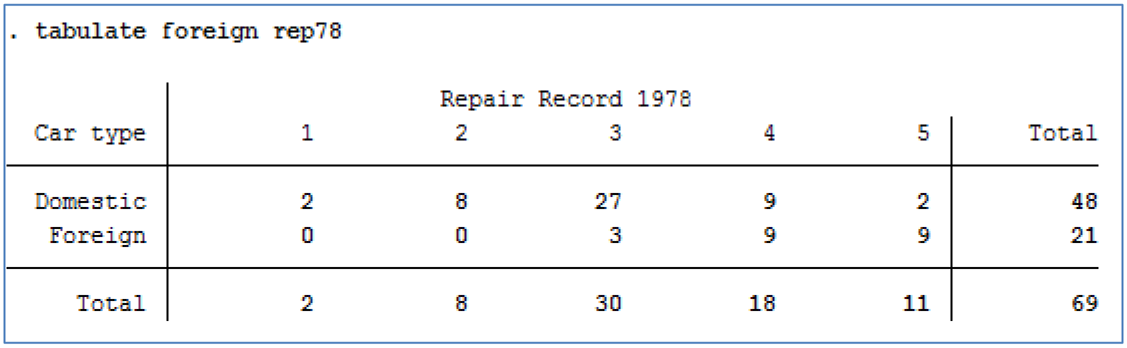
We have a command *tab1* that does this by specifying both the variable names in one go.



```
. tab1 rep78 foreign

-> tabulation of rep78

     Repair |
Record 1978 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2        2.90        2.90
          2 |          8       11.59       14.49
          3 |         30       43.48       57.97
          4 |         18       26.09       84.06
          5 |         11       15.94      100.00
------------+-----------------------------------
      Total |         69      100.00

-> tabulation of foreign

   Car type |      Freq.     Percent        Cum.
------------+-----------------------------------
   Domestic |         52       70.27       70.27
    Foreign |         22       29.73      100.00
------------+-----------------------------------
      Total |         74      100.00
```

Two way frequency tables

Two way tables can also be created. For this we should not have many categories for the variables (at least one). The following example provides a two way table.



```
. tabulate foreign rep78

           |                   Repair Record 1978
  Car type |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
  Domestic |         2          8         27          9          2 |        48
   Foreign |         0          0          3          9          9 |        21
-----------+-------------------------------------------------------+----------
     Total |         2          8         30         18         11 |        69
```

The *tabstat* command

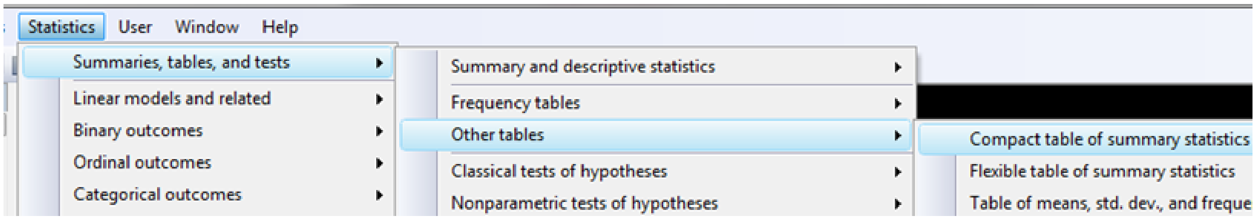
The command *tabstat* displays summary statistics for a series of numeric variables. It is a good alternative to ‘*summarize*’ as it allows us to specify the list of statistics to be displayed

The syntax is

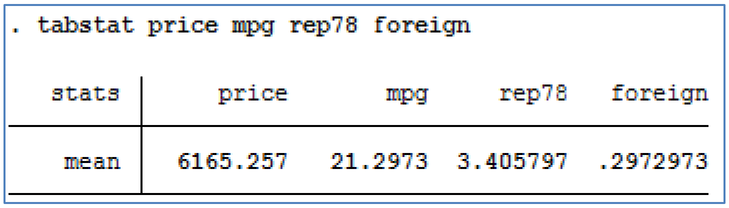
tabstat varlist [if] [in] [weight] [,options]

You can call this command from the menu by clicking:

Statistics > Summaries, tables, and tests > Other tables > Compact table of summary statistics

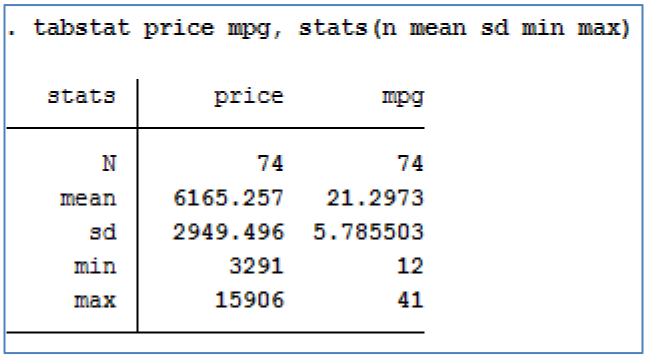

Statistics
User
Window
Help
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Summary and descriptive statistics
Frequency tables
Other tables
Classical tests of hypotheses
Nonparametric tests of hypotheses
Compact table of summary statistics
Flexible table of summary statistics
Table of means, std. dev., and freque

EXAMPLES



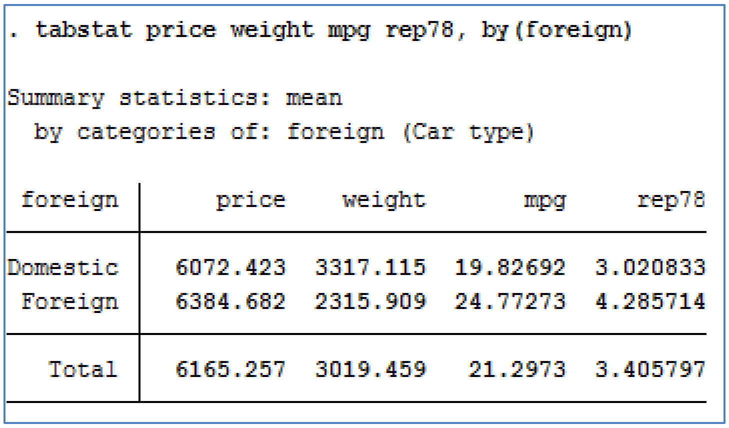
```
. tabstat price mpg rep78 foreign

   stats |    price       mpg     rep78   foreign
---------+----------------------------------------
    mean | 6165.257   21.2973  3.405797  .2972973
--------------------------------------------------
```

The above command (used without specifying any options) provides the means of the variables specified. If you need other statistics, you must use the stats option (after comma) as follows:



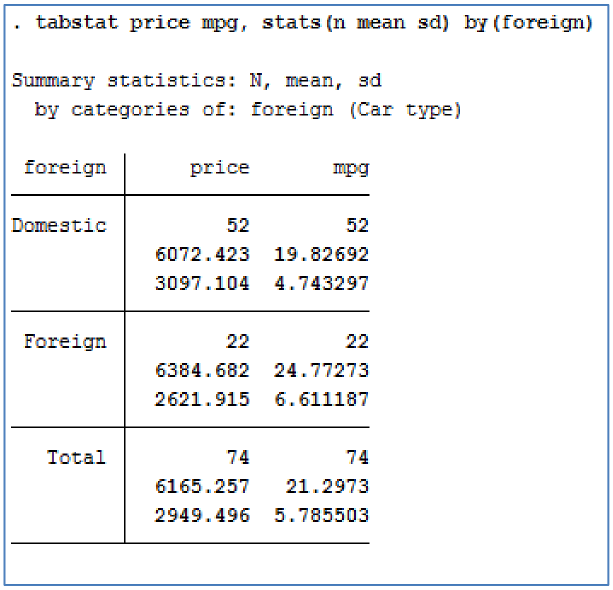
```
. tabstat price mpg, stats(n mean sd min max)

   stats |    price       mpg
---------+--------------------
       N |       74        74
    mean | 6165.257   21.2973
      sd | 2949.496  5.785503
     min |     3291        12
     max |    15906        41
------------------------------
```

You can do the same by any categorical variable (e.g. the variable *foreign* has two categories)



```
. tabstat price weight mpg rep78, by(foreign)

Summary statistics: mean
  by categories of: foreign (Car type)

 foreign |     price    weight       mpg     rep78
---------+----------------------------------------
Domestic |  6072.423  3317.115  19.82692  3.020833
 Foreign |  6384.682  2315.909  24.77273  4.285714
---------+----------------------------------------
   Total |  6165.257  3019.459   21.2973  3.405797
--------------------------------------------------
```

Bothe the above options can be used together as follows:



```
. tabstat price mpg, stats(n mean sd) by(foreign)

Summary statistics: N, mean, sd
  by categories of: foreign (Car type)

 foreign |     price       mpg
---------+--------------------
Domestic |        52        52
         |  6072.423  19.82692
         |  3097.104  4.743297
---------+--------------------
 Foreign |        22        22
         |  6384.682  24.77273
         |  2621.915  6.611187
---------+--------------------
   Total |        74        74
         |  6165.257   21.2973
         |  2949.496  5.785503
------------------------------
```

The *correlate* command

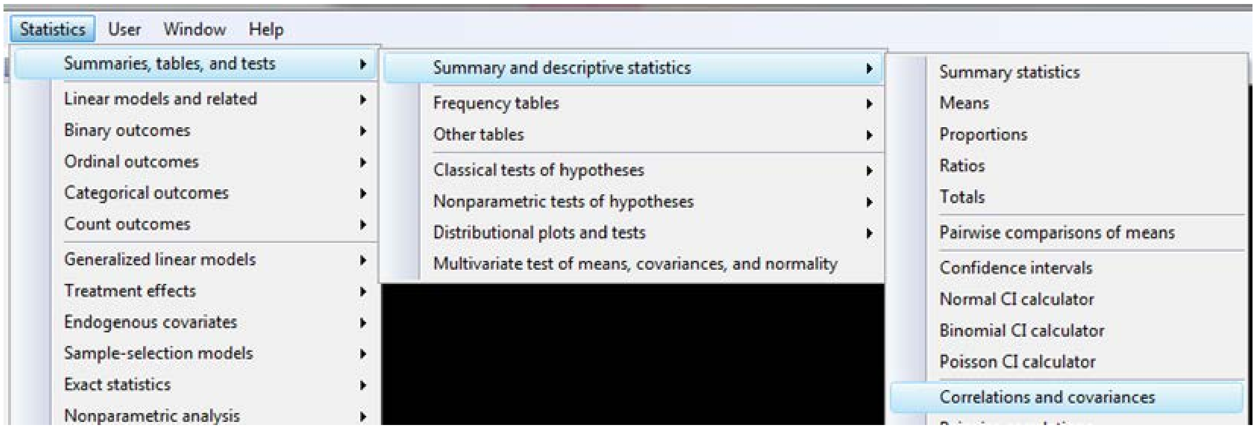
This command provides the correlation or covariance matrix of the variables specified.

The syntax is

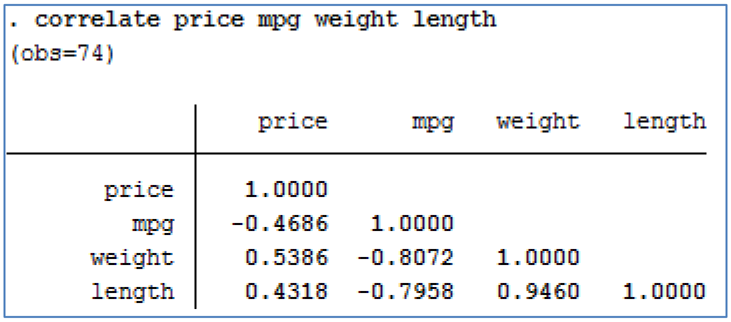
correlate [varlist] [if] [in] [weight] [,correlate_options]

If you want to use the graphic user interface instead, click on:

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Correlations and covariances

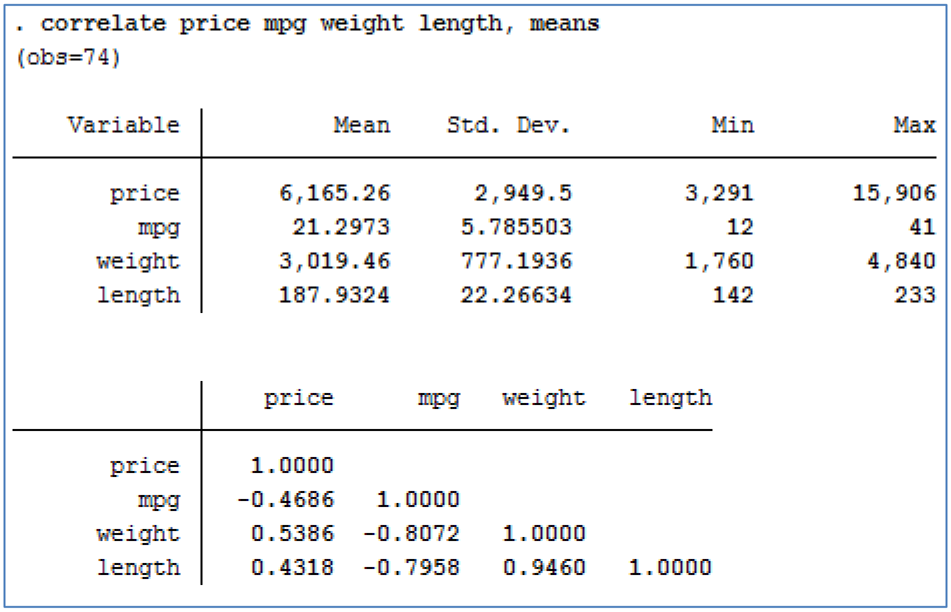


EXAMPLES



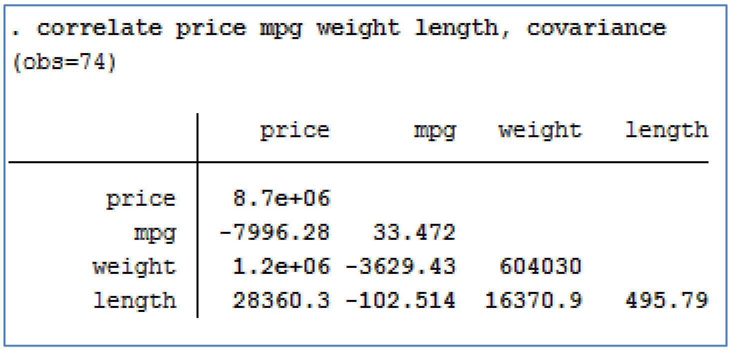
```
. correlate price mpg weight length
(obs=74)

             |    price      mpg   weight   length
-------------+------------------------------------
       price |   1.0000
         mpg |  -0.4686   1.0000
      weight |   0.5386  -0.8072   1.0000
      length |   0.4318  -0.7958   0.9460   1.0000
```

If you need summary statistics along with correlations, do the following



```
. correlate price mpg weight length, means
(obs=74)

    Variable |         Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
       price |     6,165.26      2,949.5        3,291       15,906
         mpg |      21.2973     5.785503           12           41
      weight |     3,019.46     777.1936        1,760        4,840
      length |     187.9324     22.26634          142          233


             |    price      mpg   weight   length
-------------+------------------------------------
       price |   1.0000
         mpg |  -0.4686   1.0000
      weight |   0.5386  -0.8072   1.0000
      length |   0.4318  -0.7958   0.9460   1.0000
```

Use the covariance option if you need the covariance matrix instead.



```
. correlate price mpg weight length, covariance
(obs=74)

             |    price       mpg    weight    length
-------------+----------------------------------------
       price |  8.7e+06
         mpg | -7996.28    33.472
      weight |  1.2e+06  -3629.43    604030
      length |  28360.3  -102.514   16370.9    495.79
```

Lecture 26
Introduction to Stata- III

Data Management in Stata

Importing Data

Data can be imported into stata from various types of files including Microsoft Excel files, raw data files, text files, csv files etc.

We can use the following commands in Stata to import files

- import excel: for importing .xls or .xlsx files

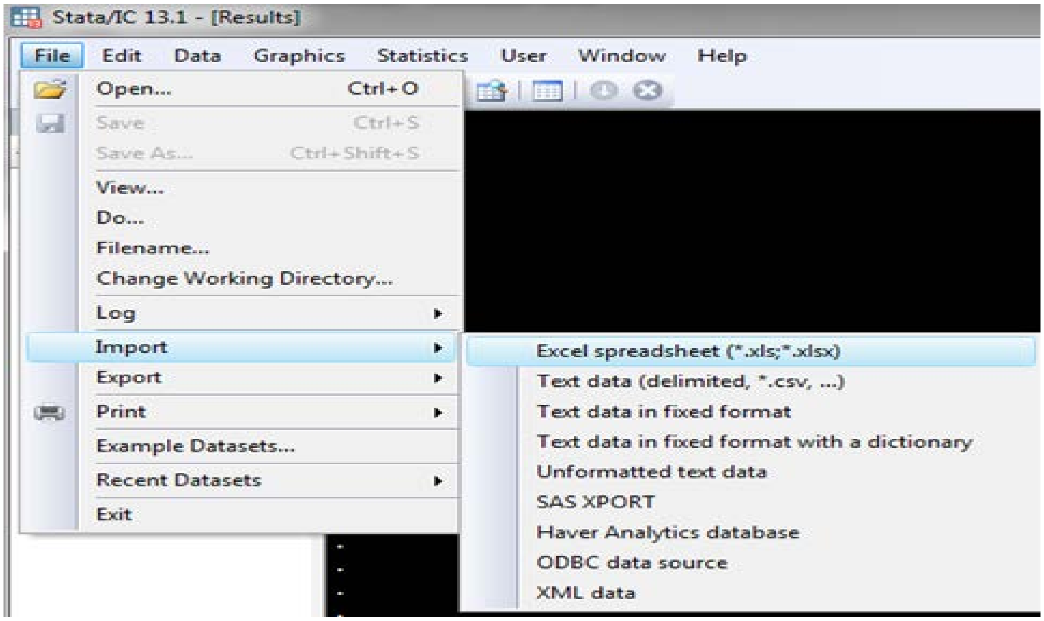
  it has the syntax: import excel [using] filename [, import_excel_options]

- infix: is used for fixed format files
- Infile: is used for free format files
- insheet: to import ASCII text data
- input: enter data from keyboard

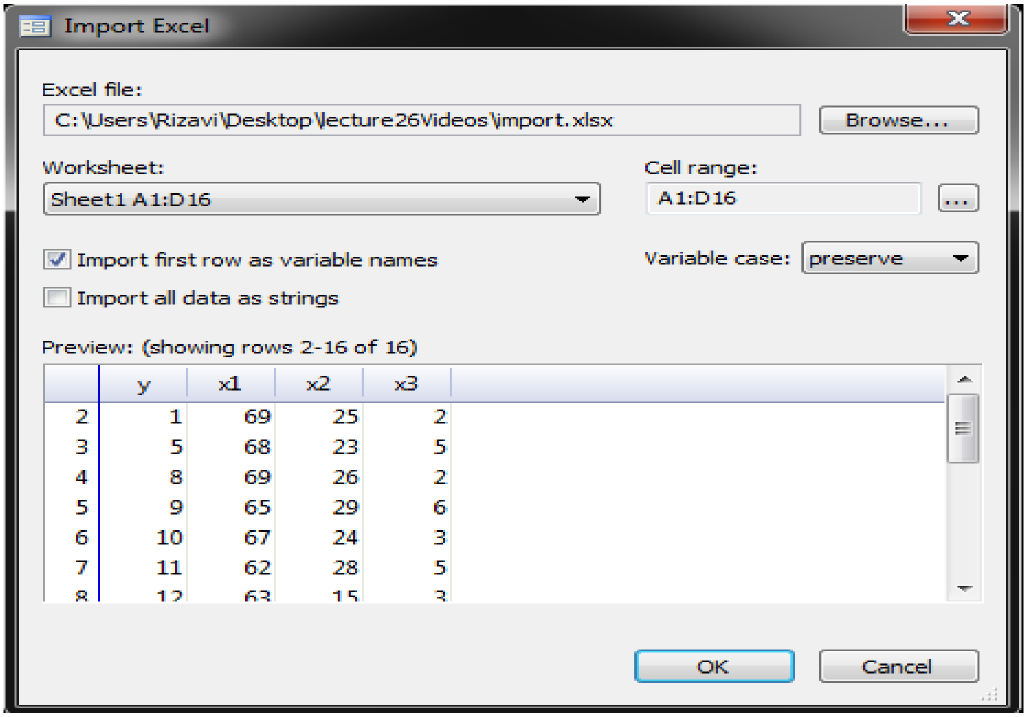
For example let us import an Excel file

From the menu click on:

File > Import > Excel spread sheet (or others)



Now brows, select your file and open


Import Excel

Excel file: C:\Users\Rizavi\Desktop\lecture26\videos\import.xlsx

Worksheet: Sheet1 A1:D16

Cell range: A1:D16

☑ Import first row as variable names

□ Import all data as strings

Variable case: preserve

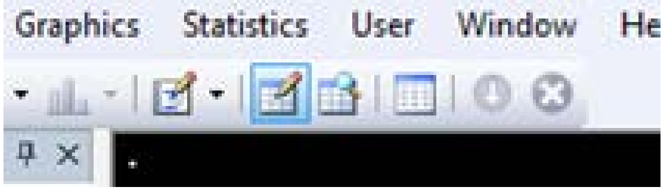
Preview: (showing rows 2-16 of 16)

	y	x1	x2	x3
2	1	69	25	2
3	5	68	23	5
4	8	69	26	2
5	9	65	29	6
6	10	67	24	3
7	11	62	28	5
8	12	63	15	3

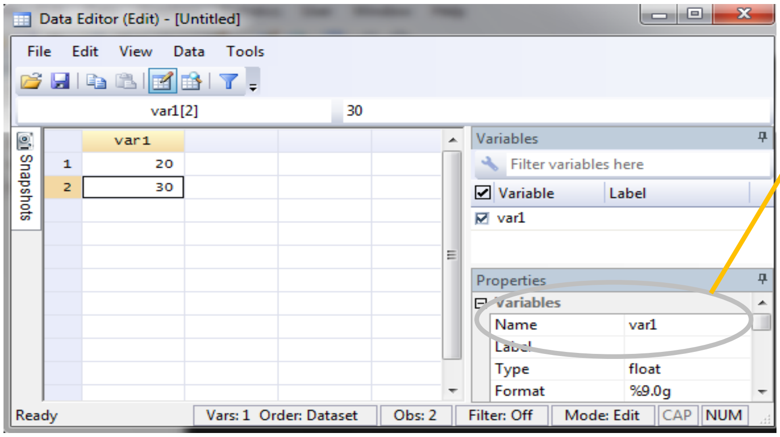
Select import variables names as first row and click OK

Entering Data

Click on the icon ‘data editor (edit)’



Start entering data (as in Excel). Variable names may be changed



var1[2] 30

	var1
1	20
2	30

☑ Variable	Label
☑ var1	

Variables	
Name	var1
Label	
Type	float
Format	%9.0g

Variable names can be changed here

Variable Names

Variable name in Stata should be according to some conventions:

- Up to 32 characters, 12 displayed
- 0 to 9, A to Z, _ ,
- First letter cannot be number
- Variable names are Case sensitive

Valid examples include

- mpg
- mpg2
- mpg_domestic
- _2014

Variable manager is used to change variable names. We can also use the *rename* command (*rename oldname newname*)

Creating Variables

Variables are created and/or modified using the *generate* and *replace* commands.

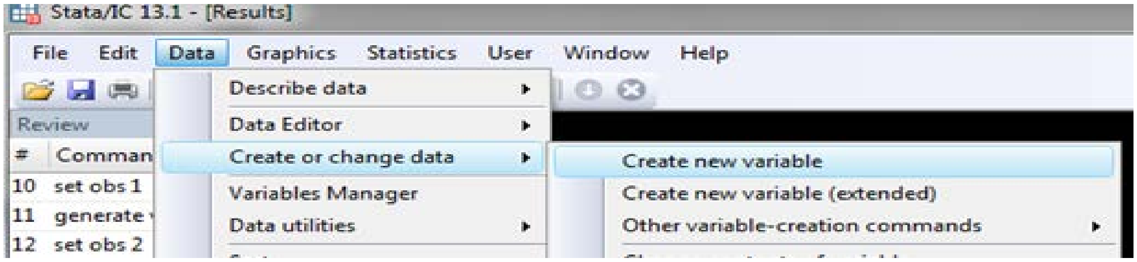
The command *generate* creates a new variable. The value of new variable is given by the =exp

It has a syntax:

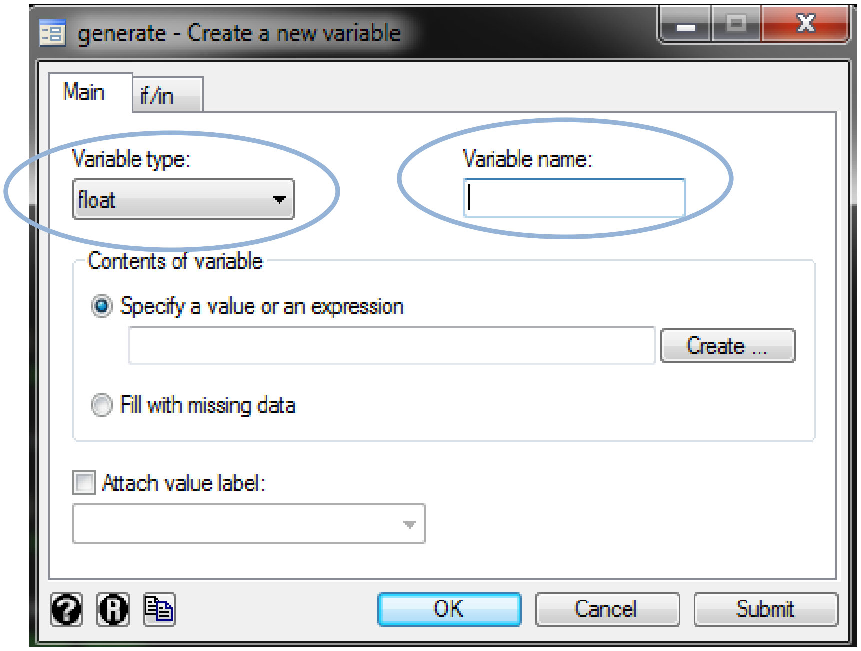
generate [type] newvar[:lblname] =exp [if] [in]

The command can also be executed using the menu:

Data > Create or change data > Create new variable



A window will open in which you need to fill the required information



Variable type: float

Variable name:

● Specify a value or an expression

○ Fill with missing data

□ Attach value label:

EXAMPLES

generate income = 0

gen income = 0

gen income = salary + bonus

gen int agesquare = age^2

gen lngdp = ln(gdp)

The replace command

The *replace* is used along with generate creates a new variable or modify existing variables

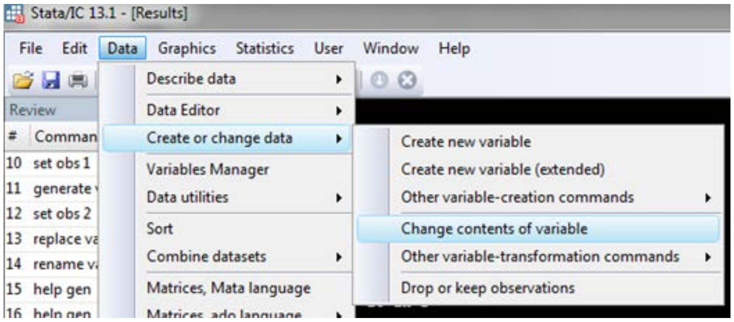
The use of the command 'in' is helpful .

It has the syntax:

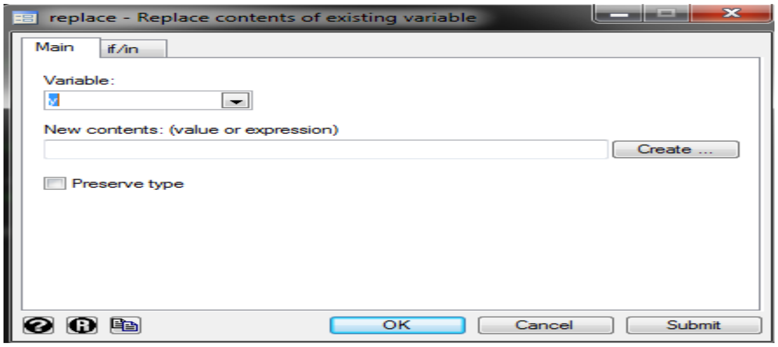
replace oldvar =exp [if] [in] [, nopromote]

From the menu click on:

Data > Create or change data > Change contents of variable



The following window will open


replace - Replace contents of existing variable
Main
if/in
Variable:
New contents: (value or expression)
Create ...
□ Preserve type
OK
Cancel
Submit

Fill in the variable name and then click on the if/in ribbon to type a condition to change the variable values and click OK.

EXAMPLE

The following are examples to use from the command window:

replace income = 1 in 5

replace income = 1000 if income == 0

replace highincome = 1 if income > 5000

gen age2 = 0

replace age2 = age^2

replace age2 = age^2 , nopromote

Operators in stata

The following usual operators are used in Stata

Arithmetic	Logical	Relational
+ add	! not (also ~ is used)	== equal
- subtract	| or	!= not equal (or ~ =)
* multiply	& and	< less than
/ divide		> greater than
^ raise to a power		<= less than or equal to
+ (In strings)		>= greater than or equal to

Functions

Type *help mathfun* to know all functions. Here are some examples

abs(x)	Absolute value of x
int(x)	Integer of x
log(x)	Logarithm
sqrt(x)	Square root of x
round(x)	Round x to the nearest integer

Assigning labels to variables

Variable labels are useful to understand what the variables contain. Label provides *data labels* and *variable labels*. They are helpful in understanding the data

We use the *label variable* command for assigning labels to variables.

To label the dataset as a whole use the following command

label data ["label"]

Example: *label data "1978 Automobile Data"*

To create variable labels use the following:

label variable varname ["label"]

Example: *label variable make "Make and Model"*

Creating Notes

The command 'note' places a note in the data. It can be general or attached to specific variable.

To create notes for a variable, use the following command:

notes [evarname]: text

Examples: Creating notes for variables

sysuse auto

note make: this is a note for variable make

To see the note type and enter:

note make

Examples: Creating general notes

sysuse auto

note: this version is from June 2014

To see the note type and enter:

notes or *notes _dta*

Dropping or Keeping Variables

The commands *drop* and *keep* are used to drop specific rows of data or entire variables.

Syntax

To Drop variables

drop varlist

To Drop observations

drop if exp

To Drop a range of observations

drop in range [if exp]

To Keep variables

keep varlist

To Keep observations that satisfy specified condition

keep if exp

To Keep a range of observations

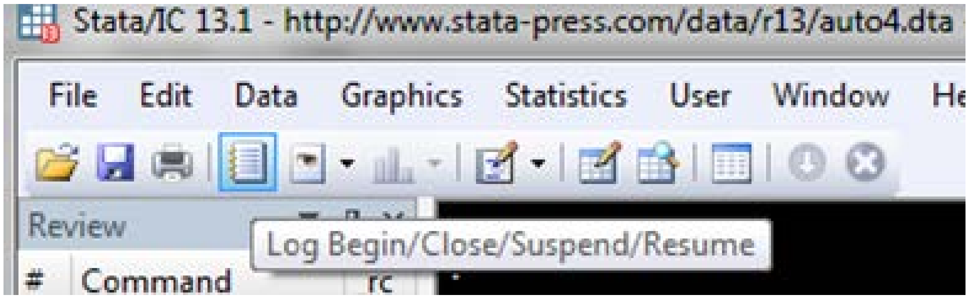
keep in range [if exp]

Using log files and do files

Log files are used to record your session. You can use .smcl or .log or .txt depending on needs

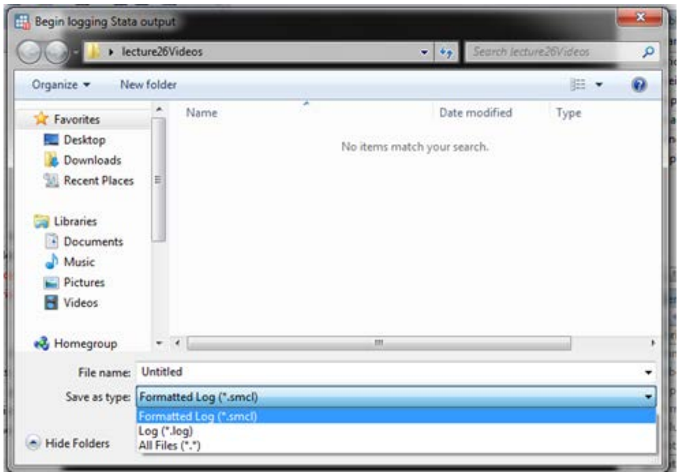
On the other hand, do files are used to run several commands at a time or replicate your results.

To open log files click on the icon similar to notepad on the menue



Stata/IC 13.1 - http://www.stata-press.com/data/r13/auto4.dta

You can create file of extension .smcl or .log (.log can be opened in notepad)



Begin logging Stata output

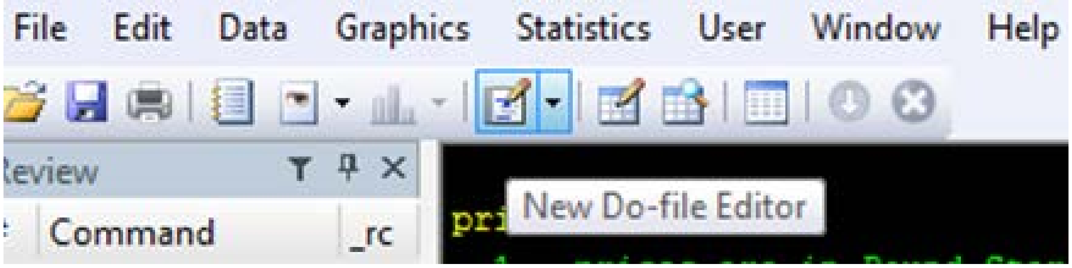
No items match your search.

File name: Untitled

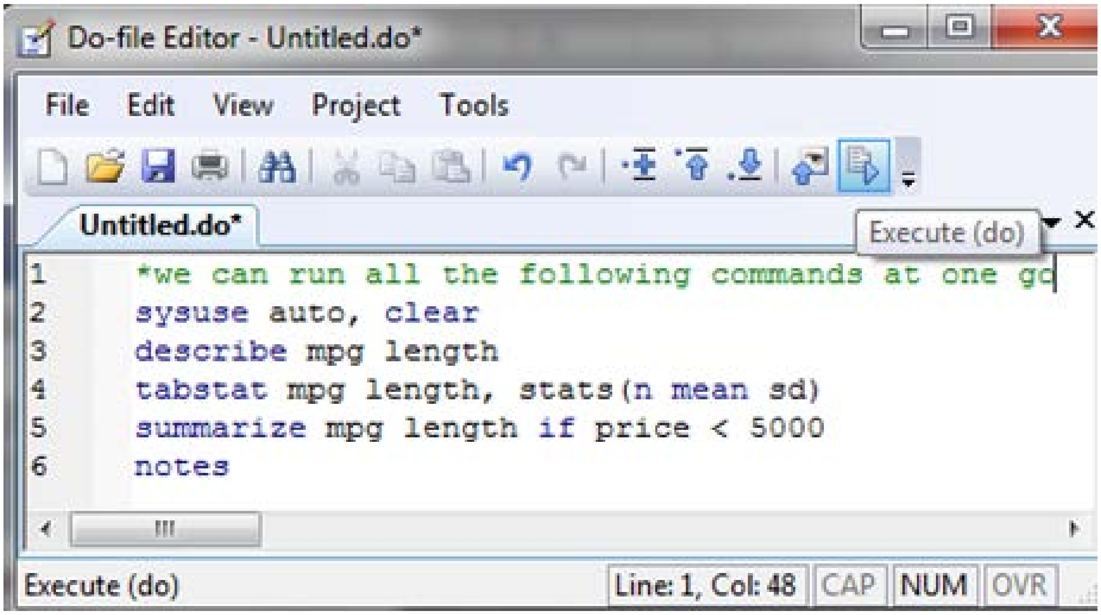
Save as type: Formatted Log (*.smcl)

Log files record all what you do. You can suspend log files in order to stop recording the session. The file can be resumed later. Finally we can close a log file to see it later.

The files of extension .do are text editor files that can execute several commands at a time. You can click on the following icon in order to open a do file. In the file you can type a list of command and execute those in one go.



Here is a snap of a do file editor with a list of commands.



```
*we can run all the following commands at one go
sysuse auto, clear
describe mpg length
tabstat mpg length, stats(n mean sd)
summarize mpg length if price < 5000
notes
```

Lecture 27

Introduction to Stata- IV

Video learning modules

Lecture27 is based on video tutorials for Stata. The students are required to consult the videos in lecture 27 in order to revise them.

Lecture 28

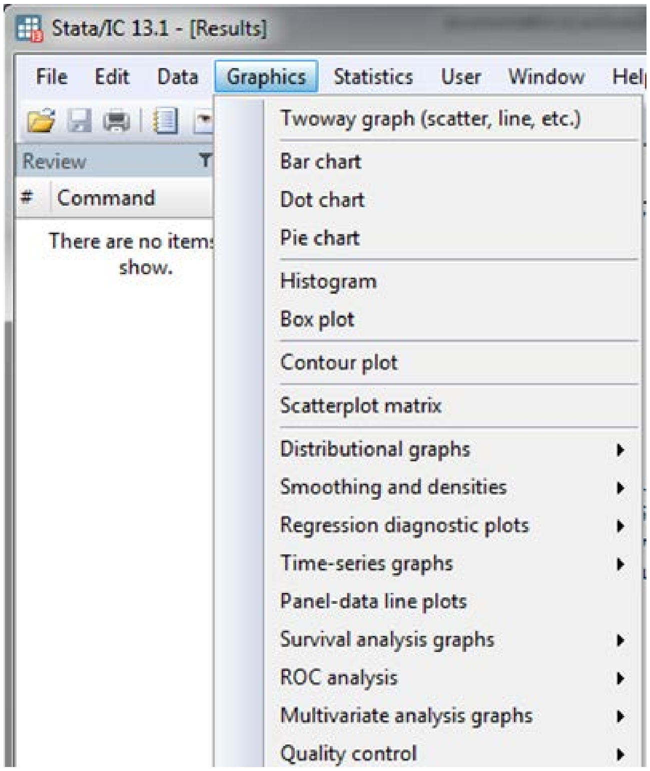
Introduction to Stata- V

Graphs in Stata

Stata has a very good capability to produce graphs with the following properties

- Rich set of tools
- Publishable quality of graphs
- GUI / Command line both
- Graph Editor
- Graphs can be copied/saved

You can create graphs by clicking on:



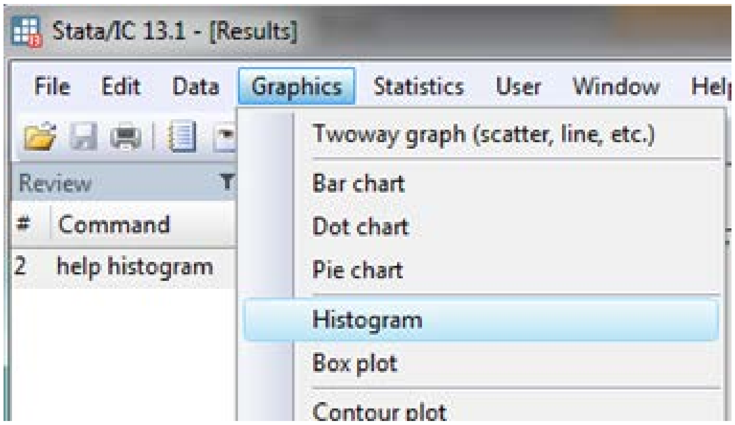
Examples

Histograms

We can create Histograms for both continuous and categorical variables

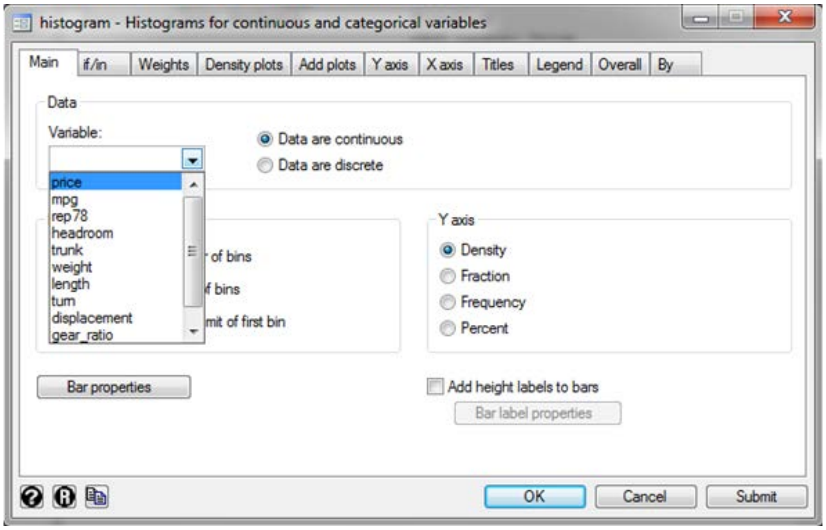
Syntax: *histogram varname [if] [in] [weight] [, [continuous_opts | discrete_opts] options]*

From the menu: *Graphics > Histogram*



```
2 help histogram
```

Specify the variable and click OK

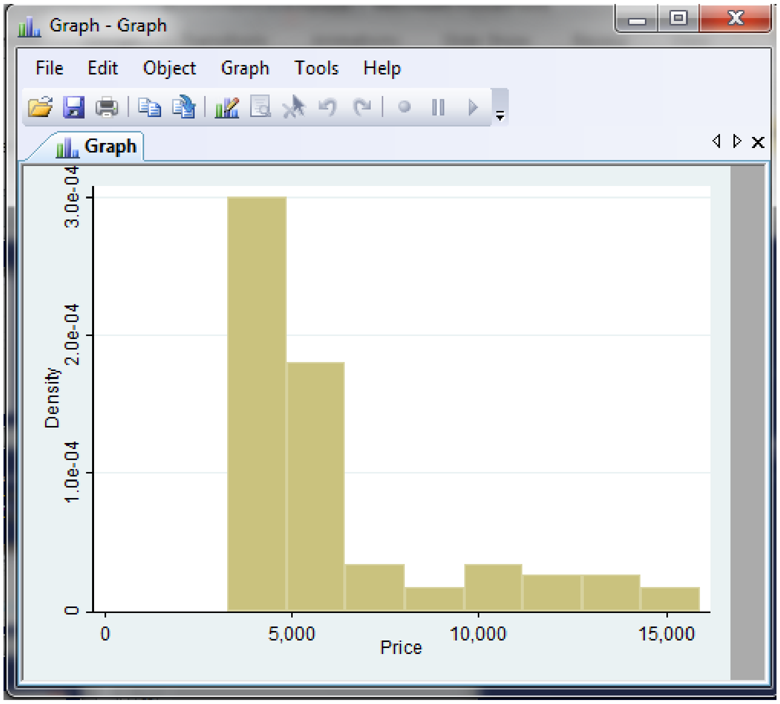


It would be better to use the command window as well to practice.

EXAMPLES:

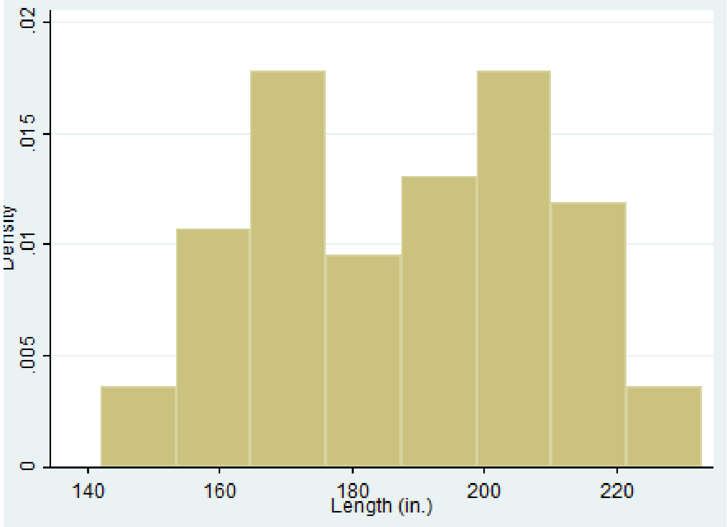
sysuse auto

histogram price



sysuse auto

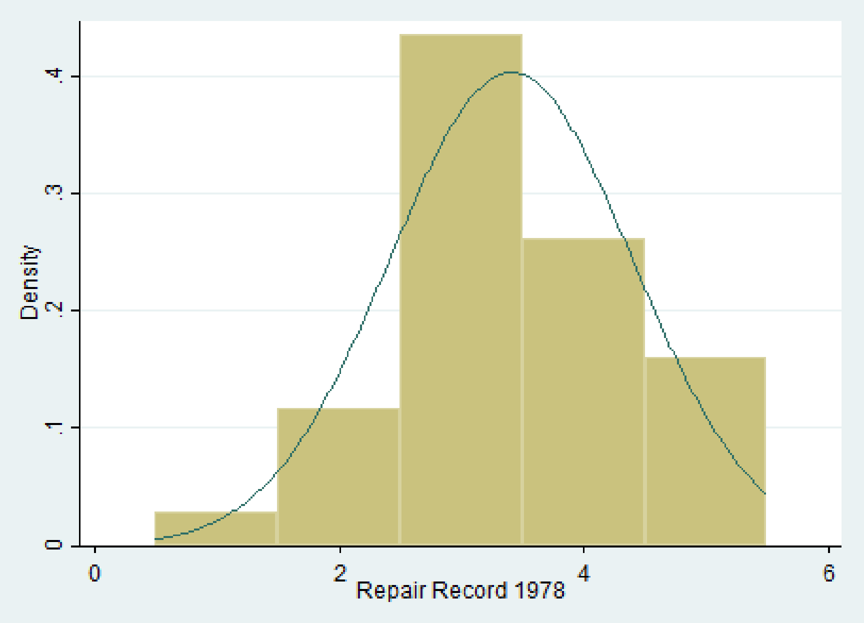
histogram length



The following example produces a normal curve as well (using the option 'normal')

sysuse auto

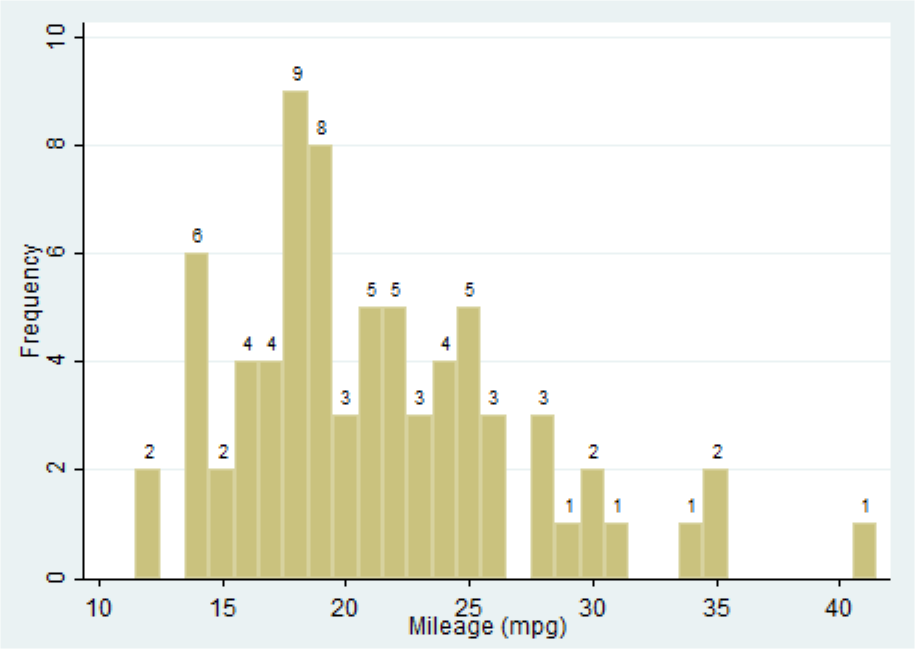
histogram rep78, discrete normal



The following example shows how to create histogram by specifying grids etc.

sysuse auto

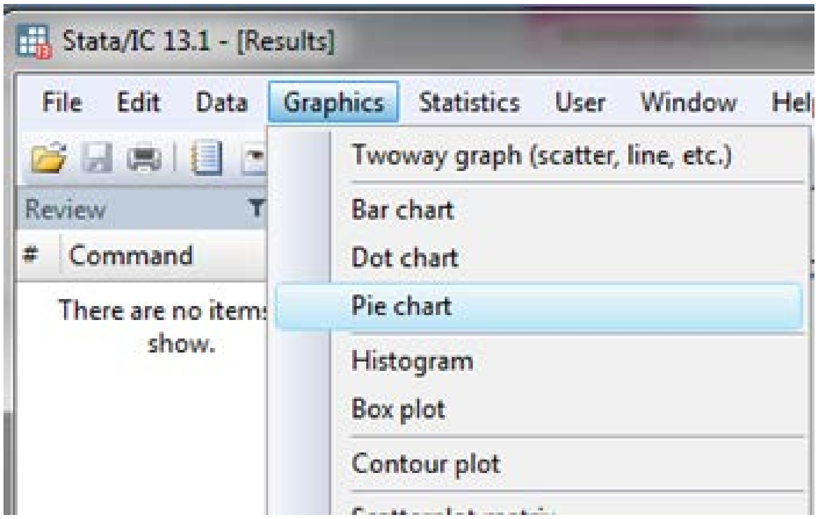
histogram mpg, discrete freq addlabels ylabel(,grid) xlabel(10(5)40)



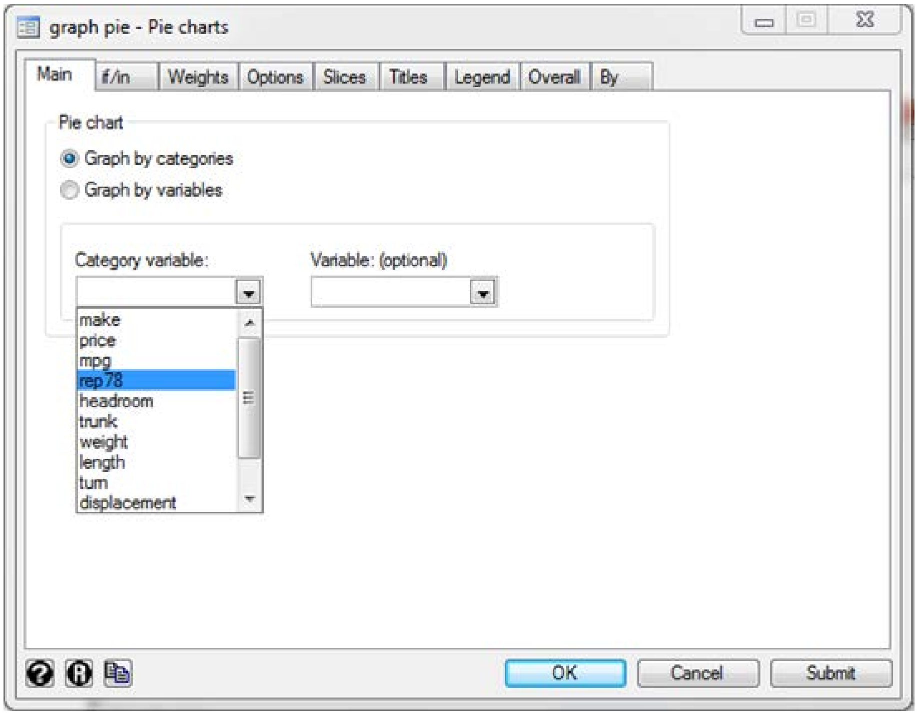
PIE Charts

Syntax: *graph pie varlist [if] [in] [weight] [, options]*

Menu: *Graph > Pie*



The following window will open. Fill the required information and click OK.



Pie chart

● Graph by categories

○ Graph by variables

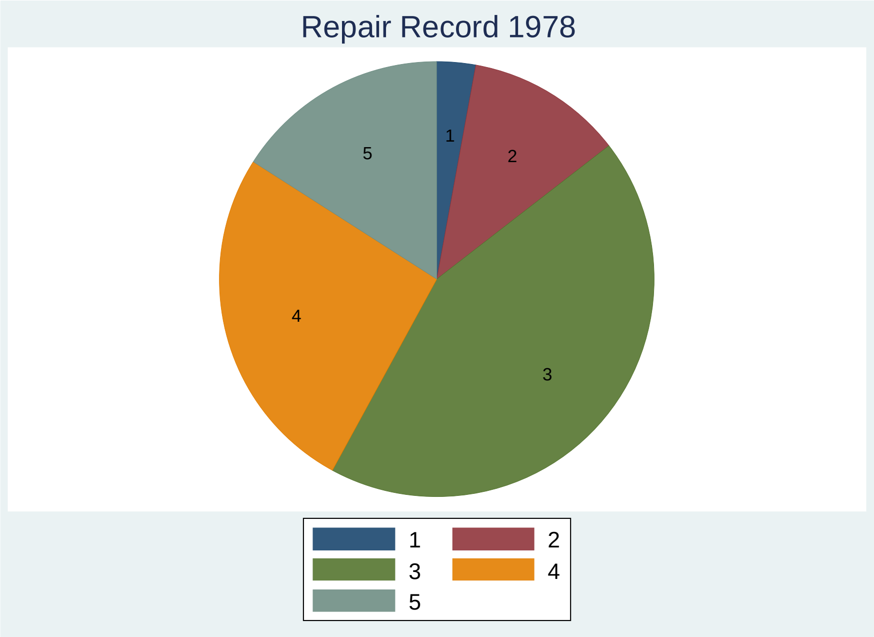
Category variable:

Variable: (optional)

EXAMPLES

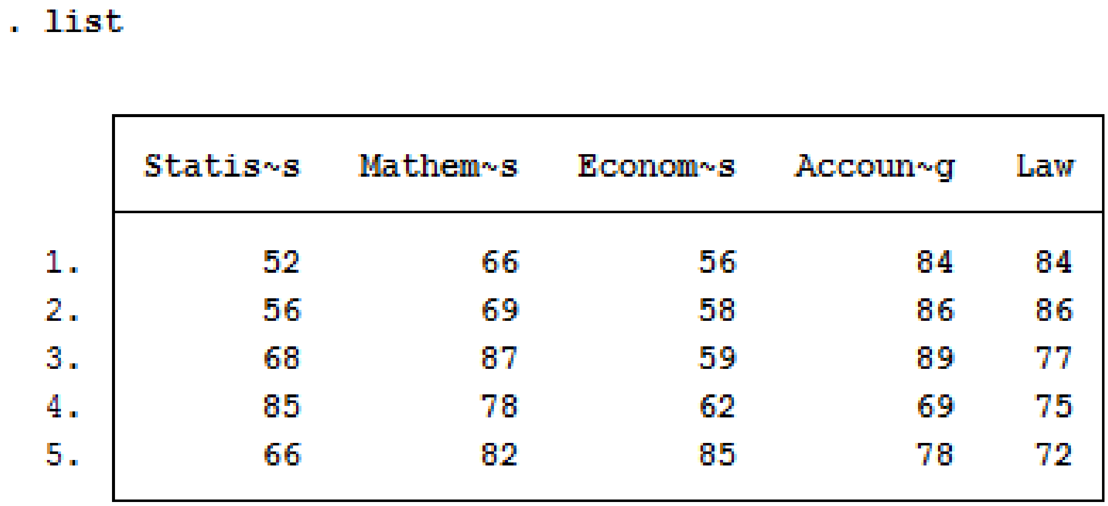
sysuse auto

graph pie, over(rep78) plabel(_all name) title("Repair Record 1978")



An Example of PIE chart

Enter the following data in a new file and use the commands *list* and *describe* and *summarize* to practice summary statistics.



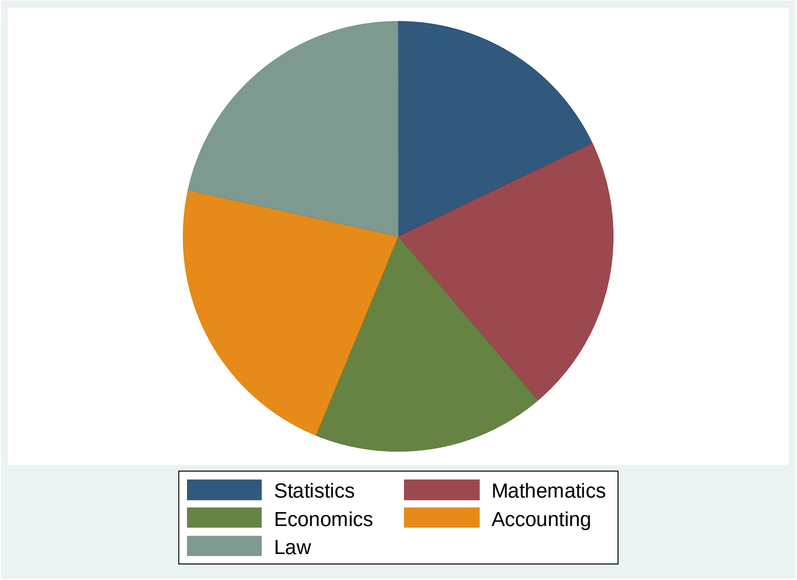
```
. list

     +------------------------------------------------------+
     | Statis~s   Mathem~s   Econom~s   Accoun~g   Law |
     |------------------------------------------------------|
  1. |       52         66         56         84    84 |
  2. |       56         69         58         86    86 |
  3. |       68         87         59         89    77 |
  4. |       85         78         62         69    75 |
  5. |       66         82         85         78    72 |
     +------------------------------------------------------+
```

Type and enter the following in the command window:

graph pie _all

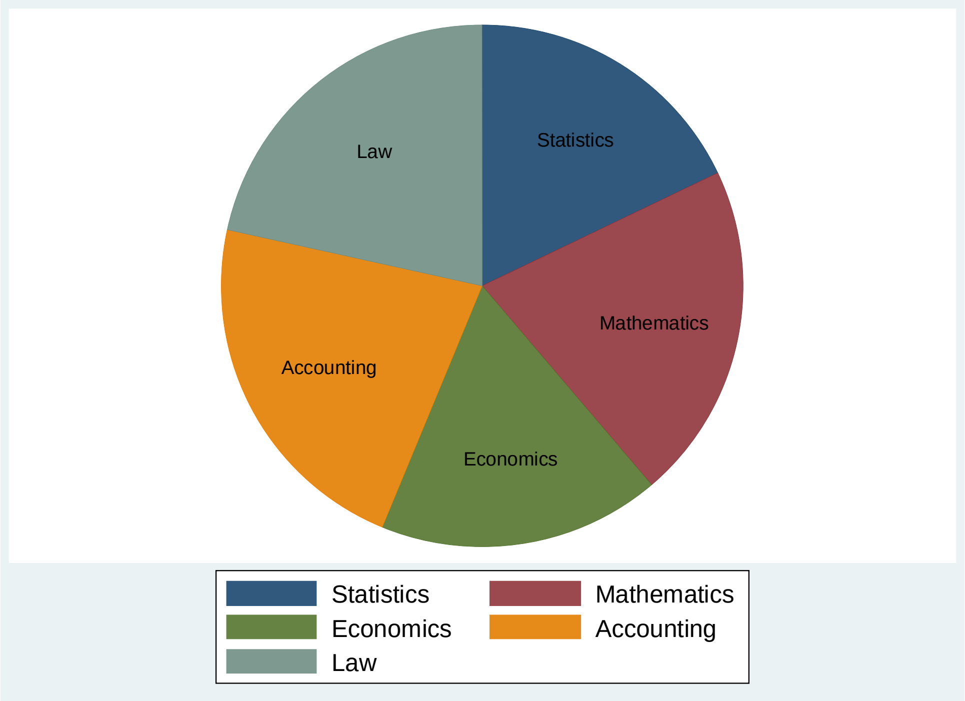
Or type

graph pie Statistics Mathematics Economics Accounting Law



Typing the following will produce a pied chart with labels:

graph pie _all , plabel(_all name) title("Distribution of marks")

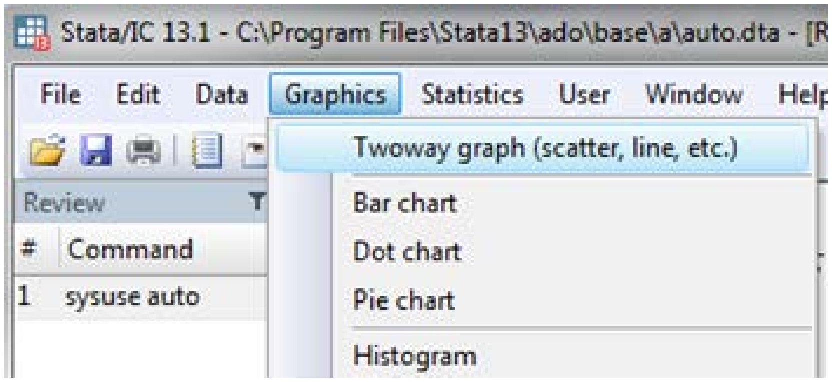


Two way graphs

When more than one variable is involved, we create twoway graphs like scatter plots

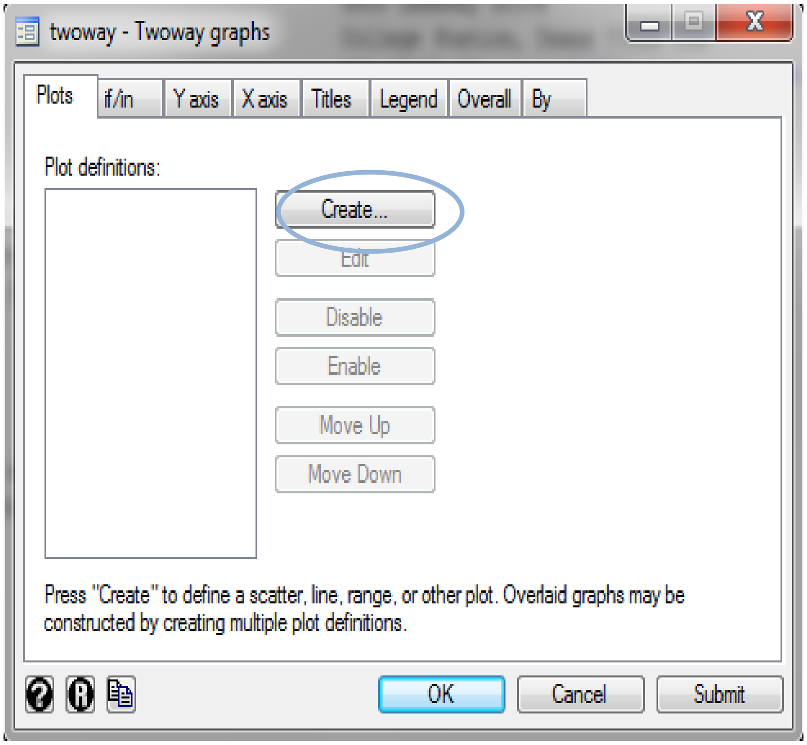
They can be created by using the menu as:

Graphics > Twoway graph


Stata/IC 13.1 - C:\Program Files\Stata13\ado\base\a\auto.dta - [R
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Twoway graph (scatter, line, etc.)
Bar chart
Dot chart
Pie chart
Histogram

#	Command
1	sysuse auto

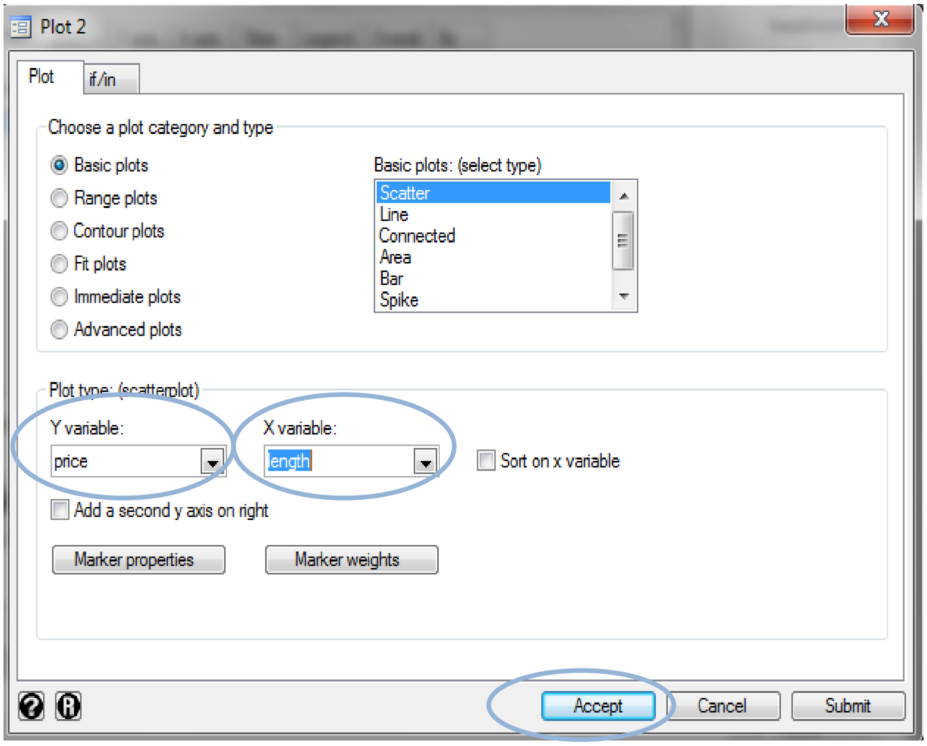
In the window that appears, click on create



Plot definitions:

Press "Create" to define a scatter, line, range, or other plot. Overlaid graphs may be constructed by creating multiple plot definitions.

The following will appear, select a scatter plot, specify variables and click *Accept*.



Choose a plot category and type

● Basic plots
○ Range plots
○ Contour plots
○ Fit plots
○ Immediate plots
○ Advanced plots

Basic plots: (select type)

Scatter (selected)
Line
Connected
Area
Bar
Spike

Plot type: (scatterplot)

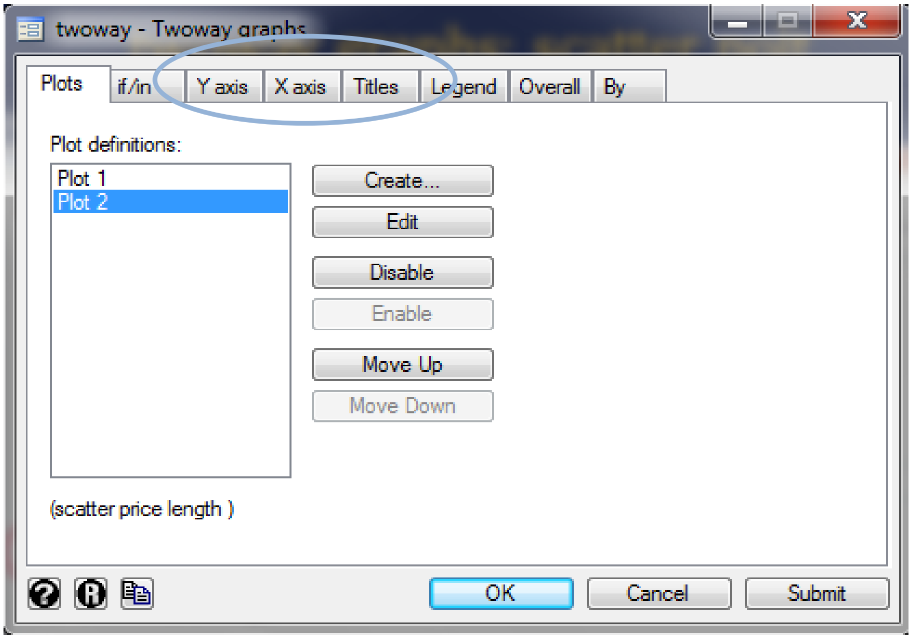
Y variable: price

X variable: length

□ Sort on x variable

□ Add a second y axis on right

To select titles of axis use the ribbons identified below:

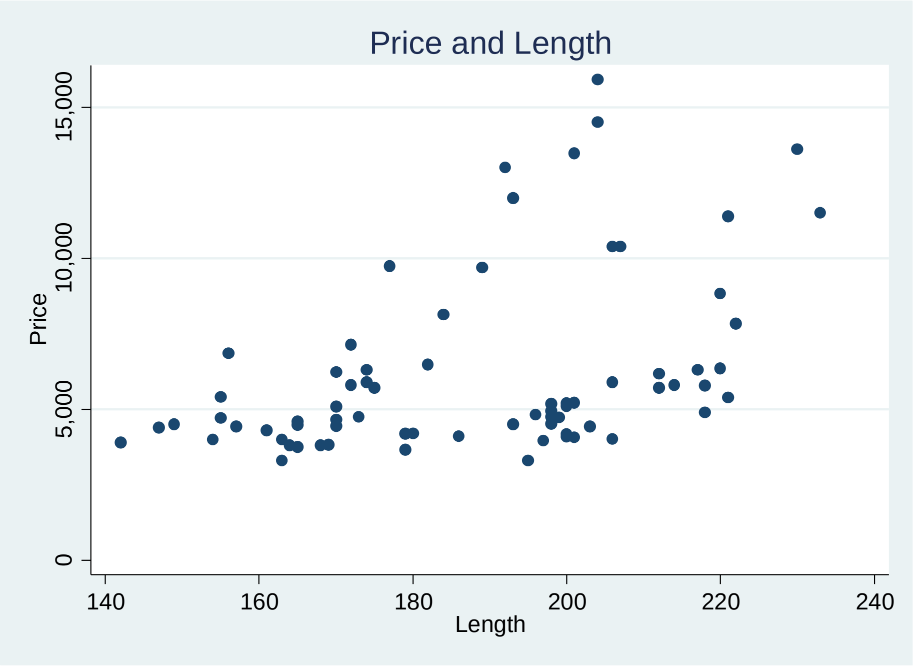


Plot definitions:

Plot 1
Plot 2

(scatter price length)

The result will be a scatter plot

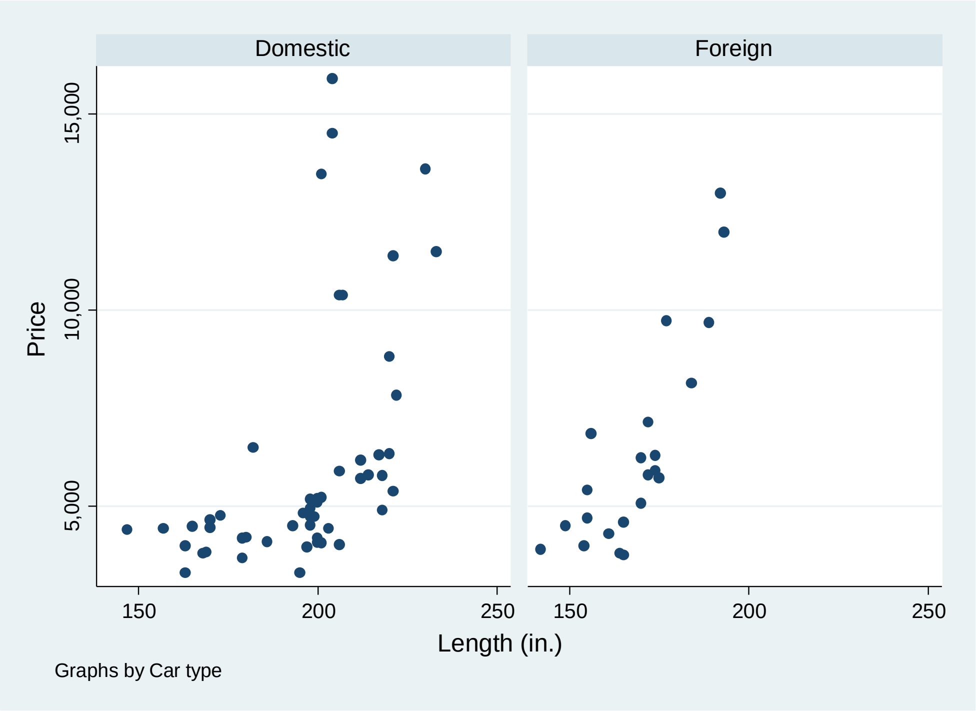


The above scatter plot can also be created by the following command:

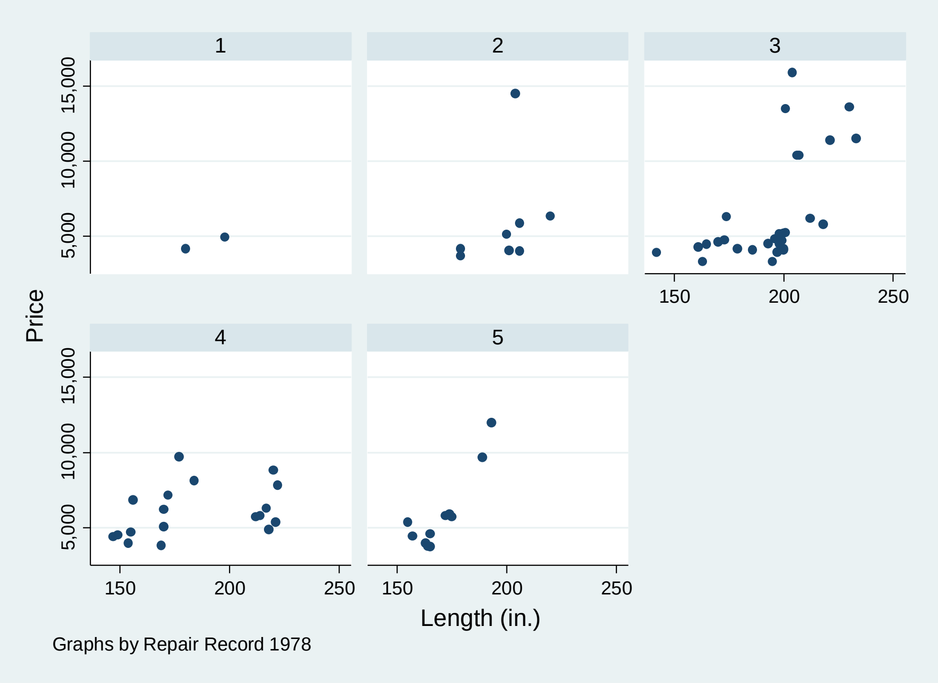
twoway (scatter price length), ytitle(Price) xtitle(Length) title(Price and Length)

Examples of Scatter Plots (type *sysuse auto, clear* to open the file auto.dta)

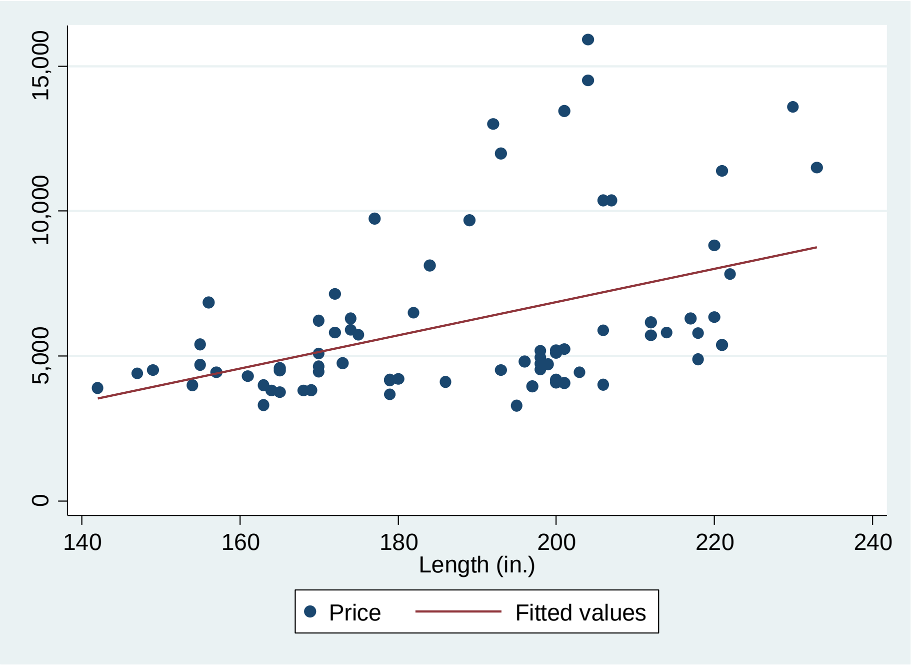
twoway (scatter price length), by(foreign)



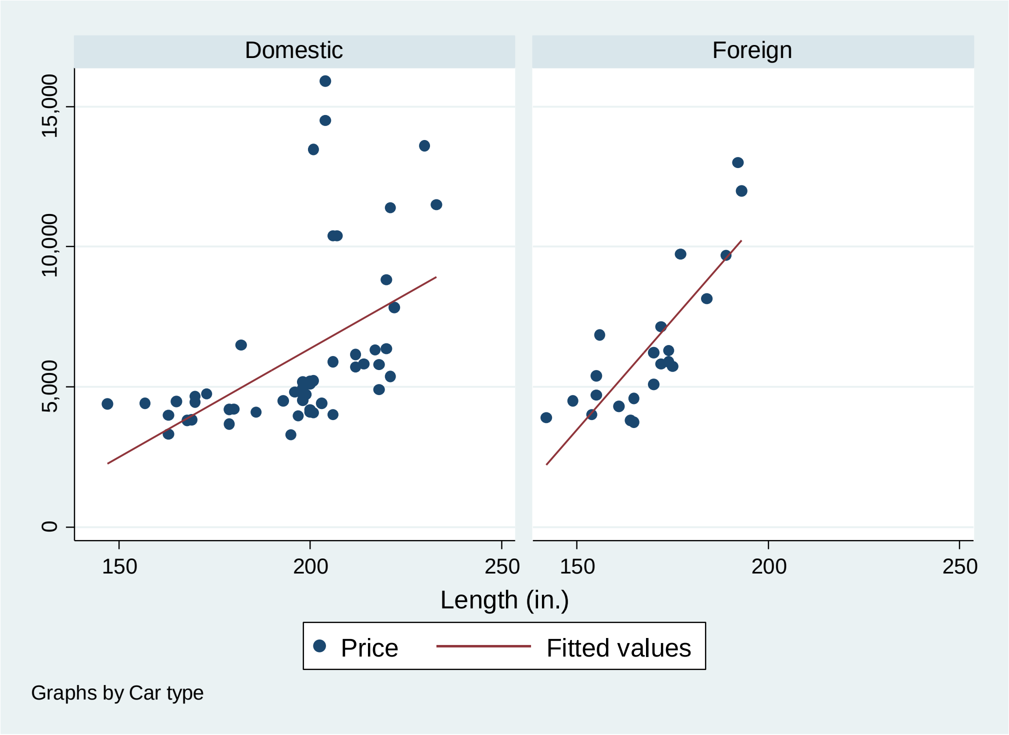
twoway (scatter price length), by(rep78)



twoway (scatter price length) (lfit price length)



twoway (scatter price length) (lfit price length), by(foreign)



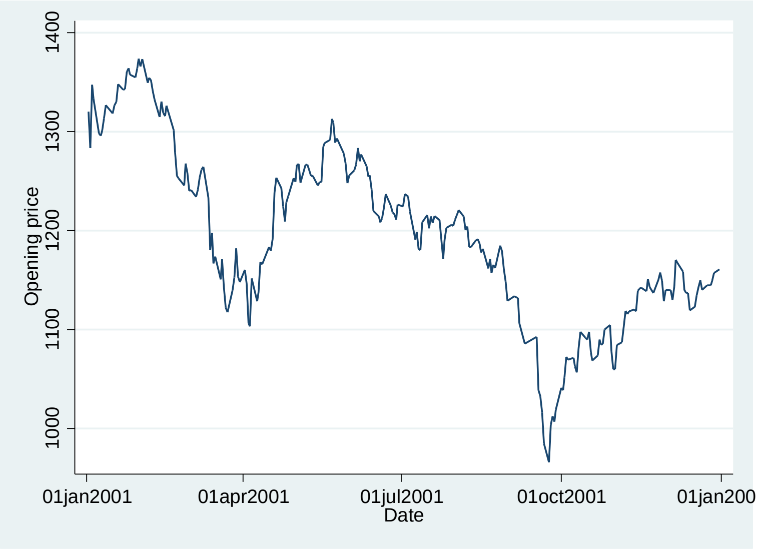
LINE plots

Using the menu, Line plots are created in the same way as scatter plots. Open the file sp500 which is a file available with stata.

Syntax: *[twoway] line varlist [if] [in] [, options]*

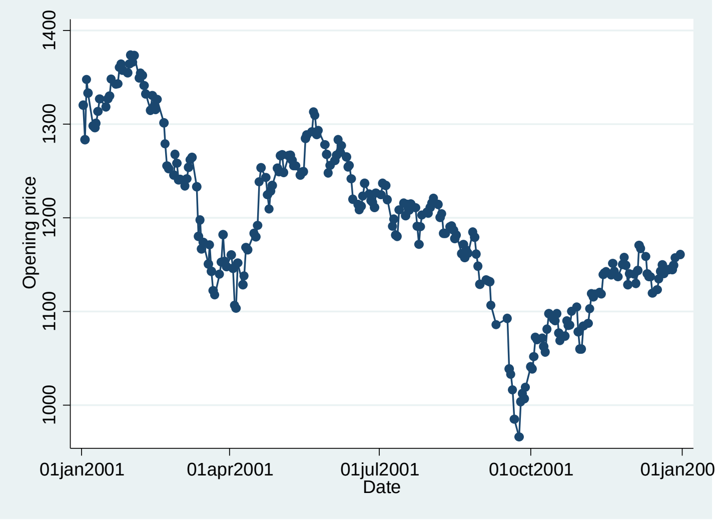
sysuse sp500

twoway (line open date)



sysuse sp500

twoway (connected close date)



Other twoway plots also can be produced in the same way.

Lecture 29

Introduction to Stata- VI

Regression with Stata

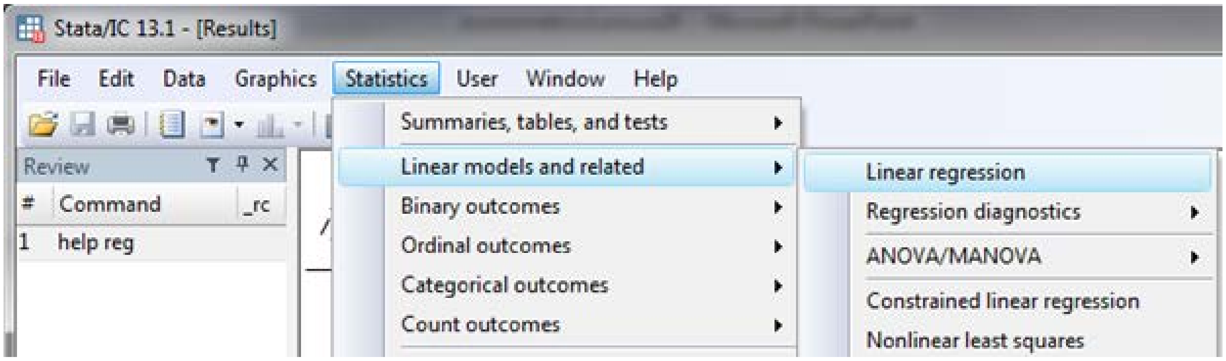
Regression analysis in stata is mainly done by the *regress* command.

Syntax:

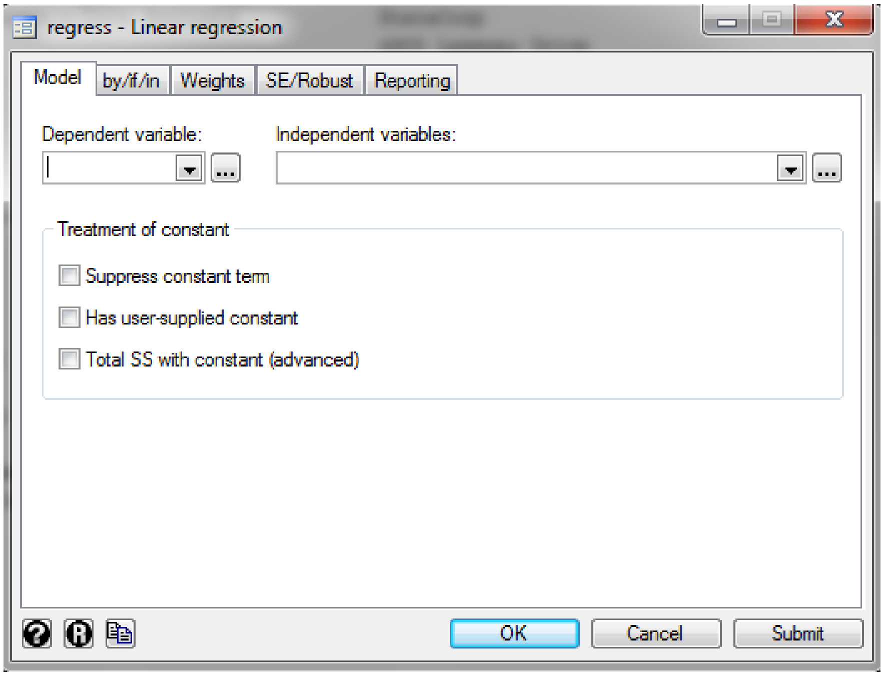
regress depvar [indepvars] [if] [in] [weight] [, options]

The menu also can be used by clicking on:

Statistics > Linear models and related > Linear regression



A window will appear that requires to specify the dependent and independent variables and other information.



Fill in the information and click OK.

STEP by STEP example of regression

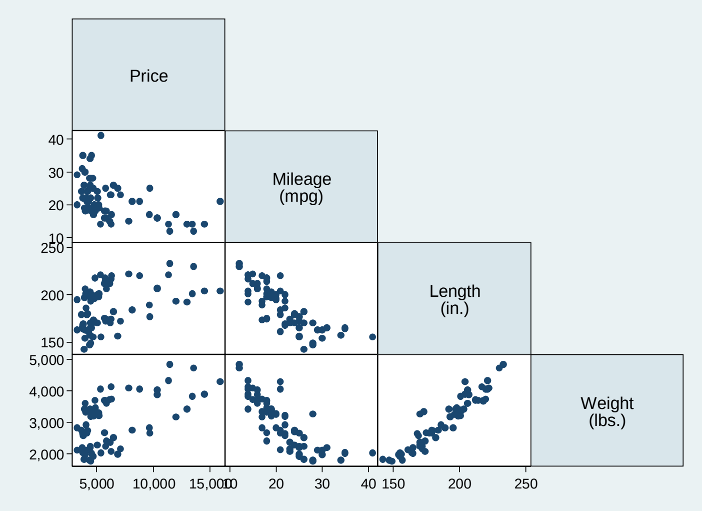
First open a file auto.dta

sysuse auto, clear

Summarize and analyze the data using related commands

Look at various two-way graphs

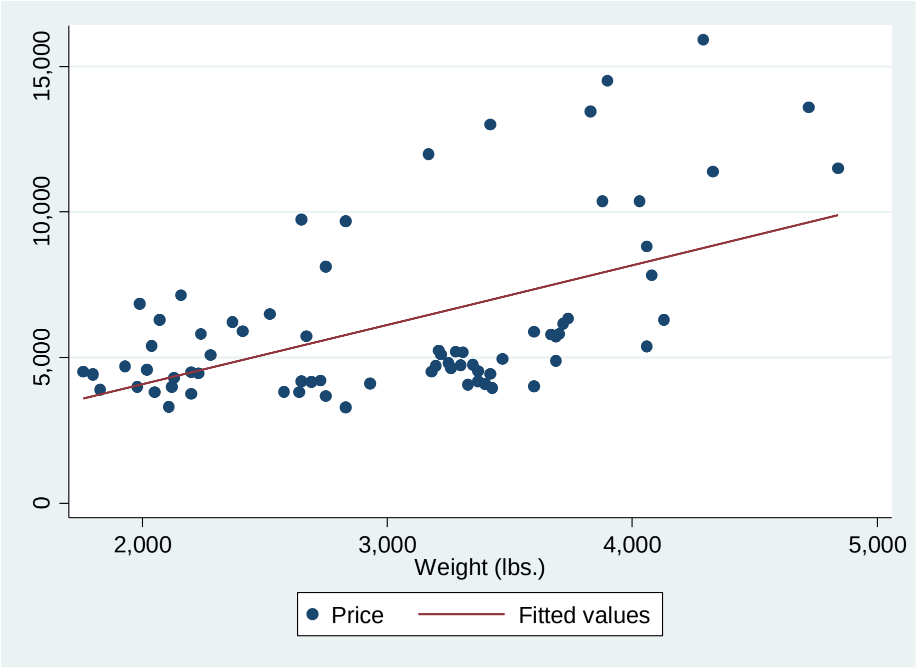
graph matrix price mpg length weight, half



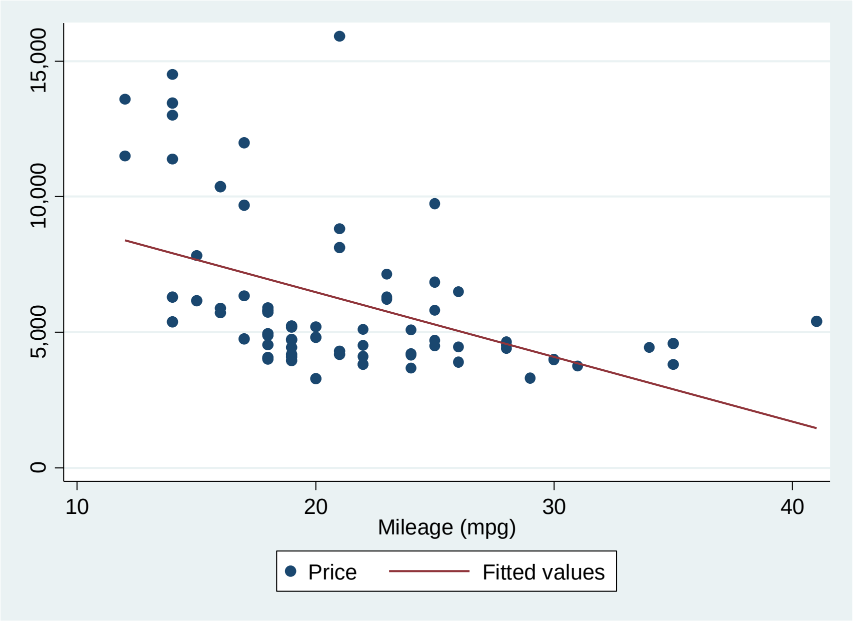
For several related variables perform the following:

To get a Scatter with a line fit:

twoway (scatter price weight) (lfit price weight)

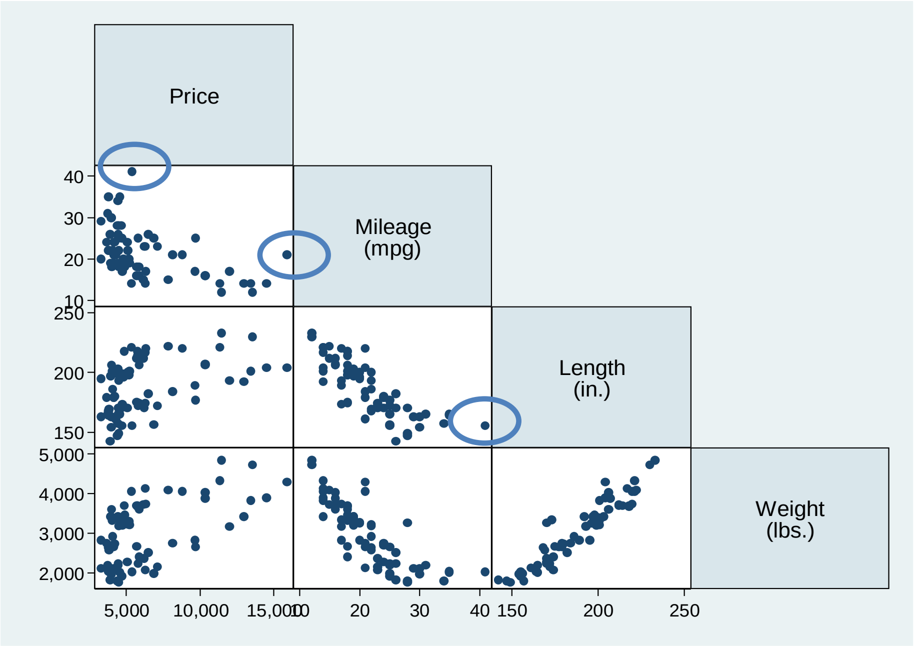


twoway (scatter price mpg) (lfit price mpg)



Check for outliers:

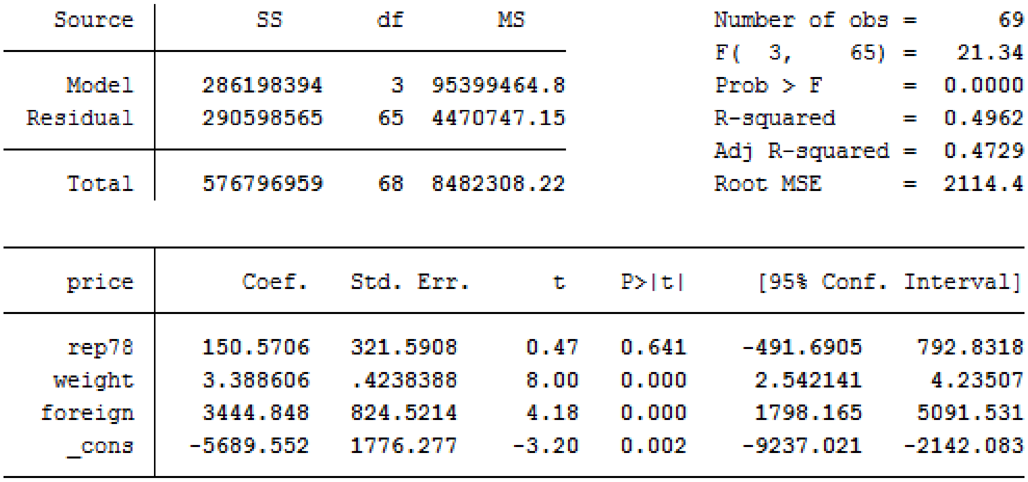
graph matrix price mpg length weight, half



Now let us perform a regression by using the regress command:

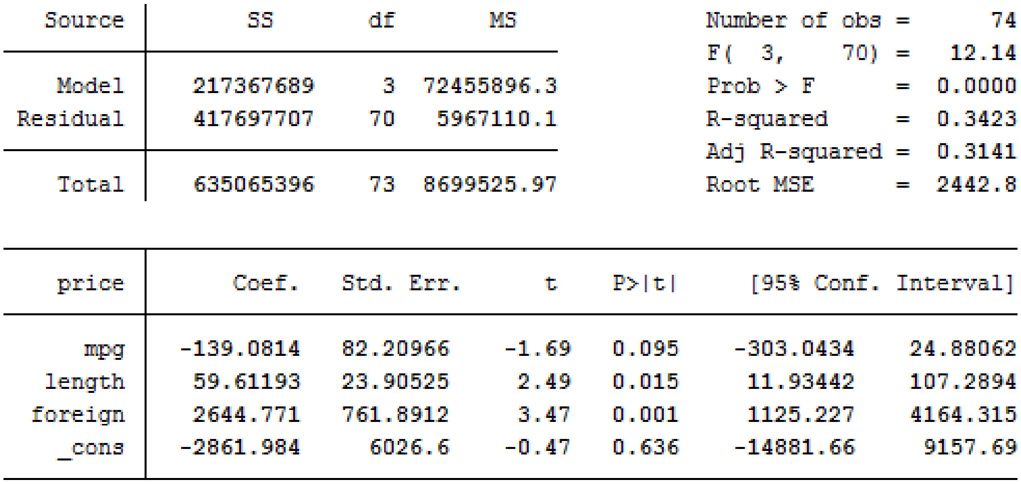
regress price rep78 weight foreign

You will get the following result that has the usual interpretation as you did with the regression in Microsoft Excel. (see the relevant lectures)



```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  3,    65) =   21.34
       Model |  286198394     3  95399464.8           Prob > F      =  0.0000
    Residual |  290598565    65  4470747.15           R-squared     =  0.4962
-------------+------------------------------           Adj R-squared =  0.4729
       Total |  576796959    68  8482308.22           Root MSE      =  2114.4

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rep78 |   150.5706   321.5908     0.47   0.641    -491.6905    792.8318
      weight |   3.388606   .4238388     8.00   0.000     2.542141     4.23507
     foreign |   3444.848   824.5214     4.18   0.000     1798.165    5091.531
       _cons |  -5689.552   1776.277    -3.20   0.002    -9237.021   -2142.083
------------------------------------------------------------------------------
```

regress price mpg length foreign



```
      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   12.14
       Model |  217367689     3  72455896.3            Prob > F      =  0.0000
    Residual |  417697707    70   5967110.1            R-squared     =  0.3423
-------------+------------------------------           Adj R-squared =  0.3141
       Total |  635065396    73  8699525.97            Root MSE      =  2442.8

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -139.0814   82.20966    -1.69   0.095    -303.0434    24.88062
      length |   59.61193   23.90525     2.49   0.015     11.93442    107.2894
     foreign |   2644.771   761.8912     3.47   0.001     1125.227    4164.315
       _cons |  -2861.984     6026.6    -0.47   0.636    -14881.66     9157.69
------------------------------------------------------------------------------
```

Post regression Analysis

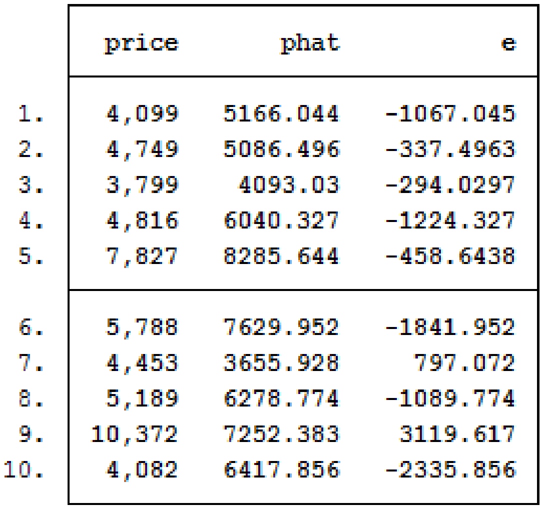
Use the following to get your trend values and residuals:

Predict phat

predict e, residual

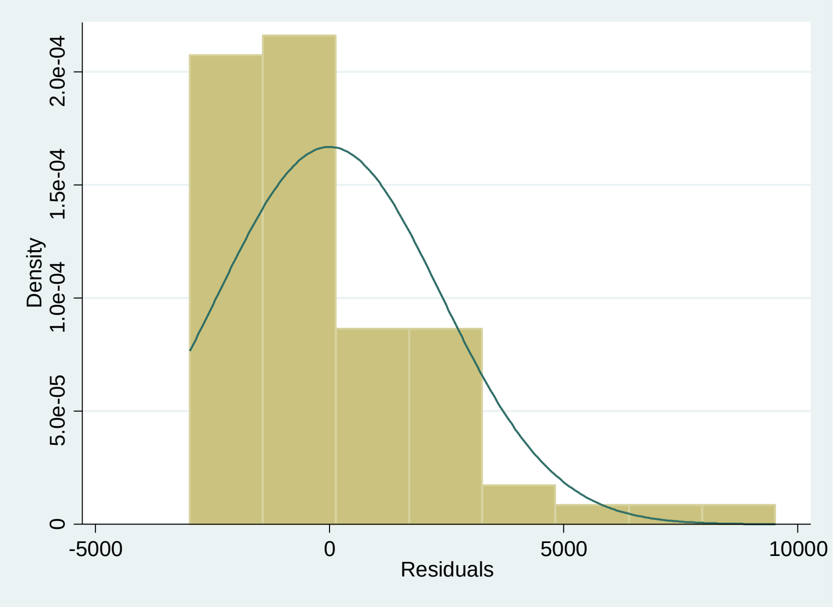
list price phat e in 1/20

The above will create new variables phat (the trend values through the predict command) and the variable 'e' for the residuals. You can use the list command to see the new columns of variables.



```
     +-------------------------------+
     |  price       phat           e |
     |-------------------------------|
  1. |  4,099   5166.044   -1067.045 |
  2. |  4,749   5086.496   -337.4963 |
  3. |  3,799    4093.03   -294.0297 |
  4. |  4,816   6040.327   -1224.327 |
  5. |  7,827   8285.644   -458.6438 |
     |-------------------------------|
  6. |  5,788   7629.952   -1841.952 |
  7. |  4,453   3655.928     797.072 |
  8. |  5,189   6278.774   -1089.774 |
  9. | 10,372   7252.383    3119.617 |
 10. |  4,082   6417.856   -2335.856 |
     +-------------------------------+
```

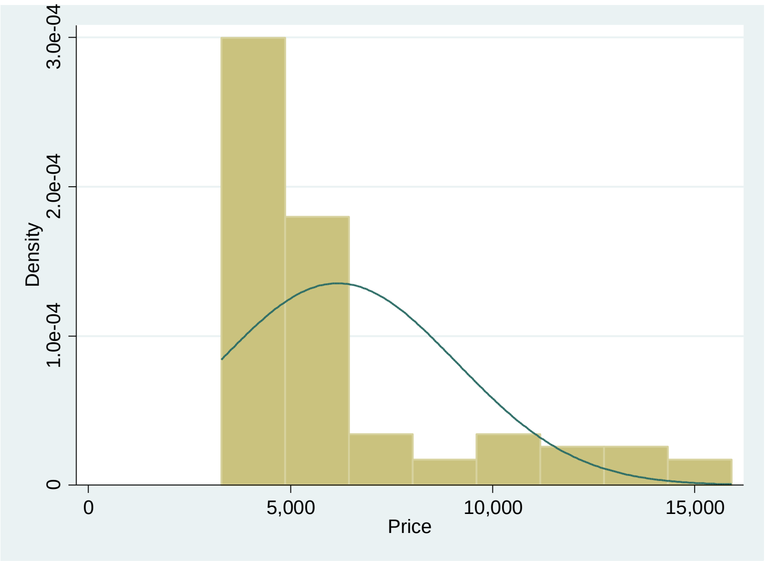
See if the residuals follow the assumption of normality *histogram e,normal*



The above shows that the residuals are not normally distributed.

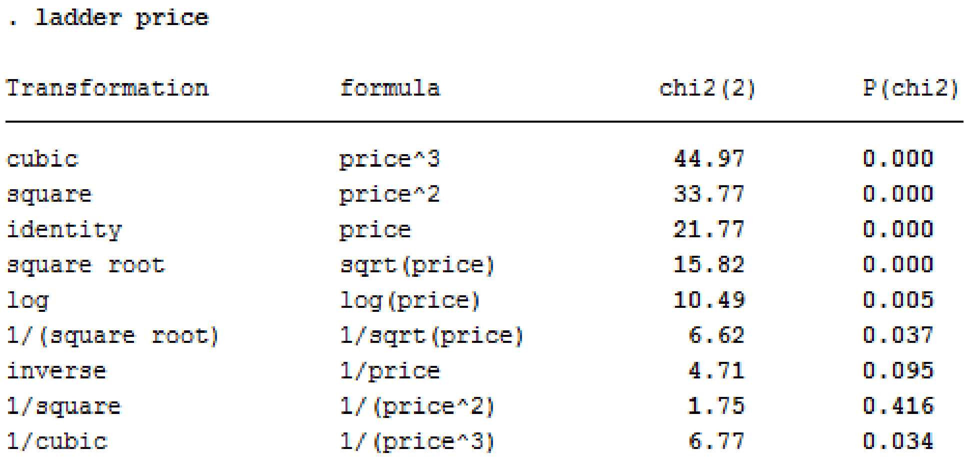
One reason of errors not being normally distributed is that the outcome (dependent) variable is not normally distributed.

histogram price, normal



We can transform the variable price.

Let us try to transform price to be normally distributed by trying taking log or square or other powers (*ladder* and *gladder* commands can help us)

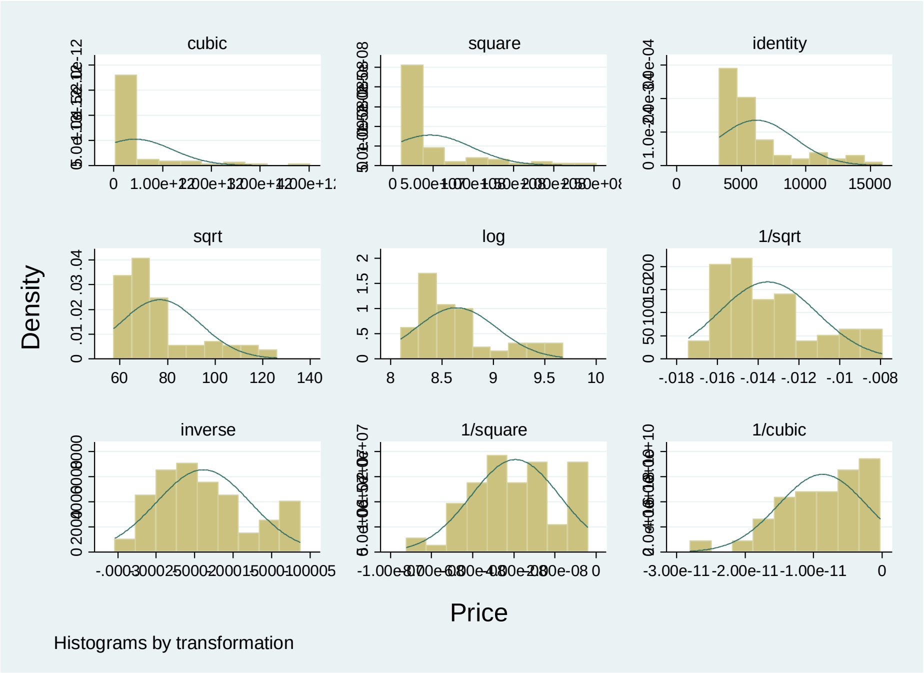


```
. ladder price

Transformation          formula                 chi2(2)       P(chi2)
---------------------------------------------------------------------
cubic                   price^3                   44.97         0.000
square                  price^2                   33.77         0.000
identity                price                     21.77         0.000
square root             sqrt(price)               15.82         0.000
log                     log(price)                10.49         0.005
1/(square root)         1/sqrt(price)              6.62         0.037
inverse                 1/price                    4.71         0.095
1/square                1/(price^2)                1.75         0.416
1/cubic                 1/(price^3)                6.77         0.034
```

The inverse and 1/square have the smallest chi-square but let us look at the graphs

The gladder gives a graph to analyze the situation (it may take some time to display the graph)

gladder price



Let us transform the variable price.

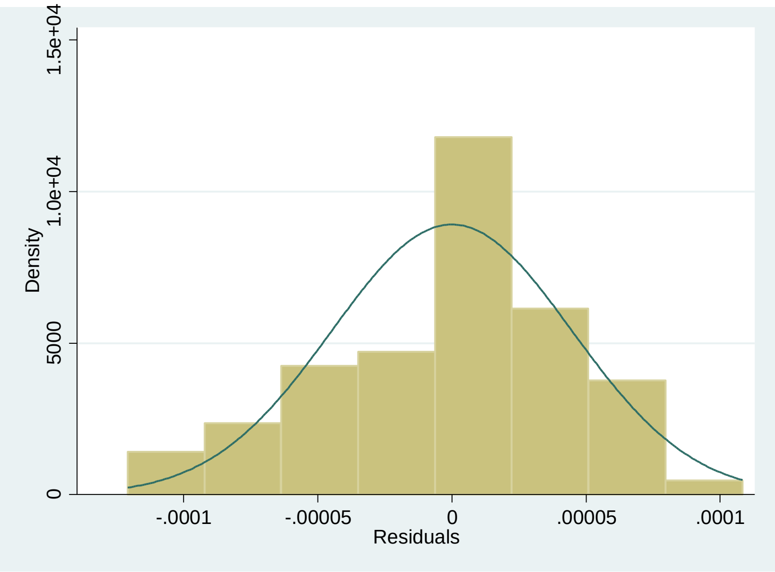
gen priceinv=1/price

reg priceinv mpg length foreign

predict e1, residual

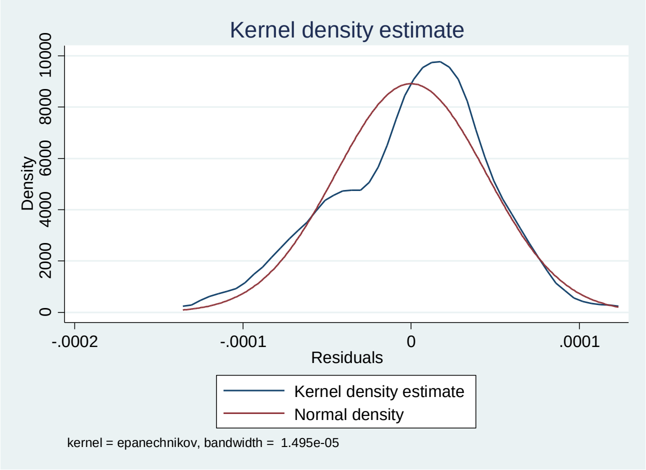
histogram e1, normal

First you will get the regression results, then the following diagram.



Now the residuals form regressing inverse of price looks somehow normal so we use it in our regression. To have the kernel density,type:

Kdensity e, normal

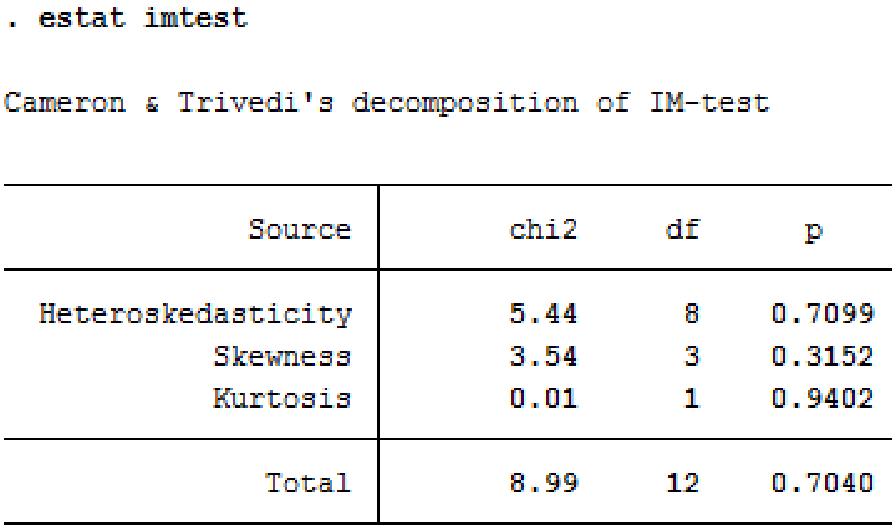


Post regression tests

To test for Multicollinearity and other problems, look at the following:

We have two different tests for Heteroskedasticity; the IM test and the hottest commands are used.

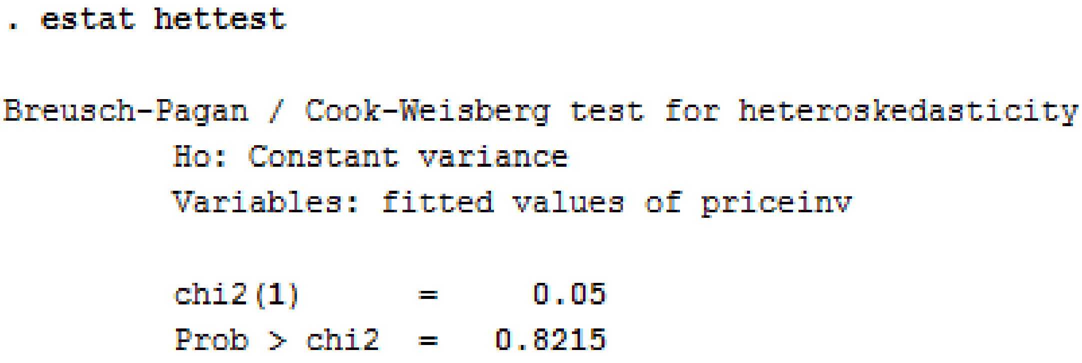
IM TEST (White Test)



```
. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       5.44      8    0.7099
            Skewness |       3.54      3    0.3152
            Kurtosis |       0.01      1    0.9402
---------------------+-----------------------------
               Total |       8.99     12    0.7040
---------------------------------------------------
```

This shows that there is heteroskedasticity.

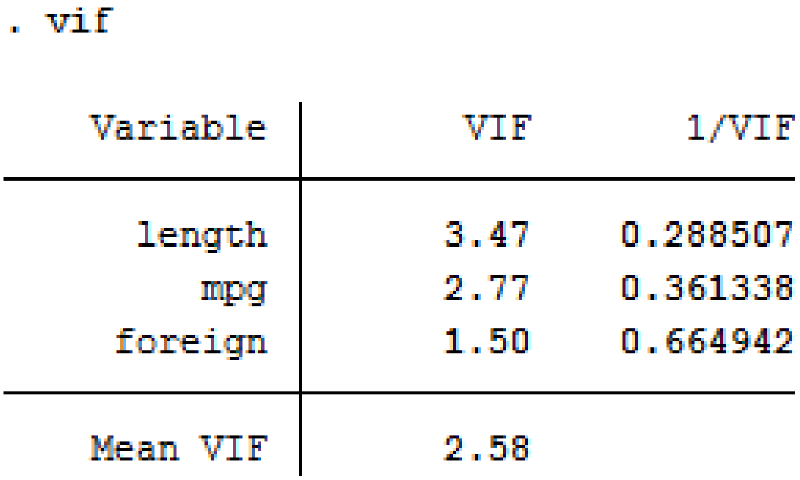
The second test is hettest



```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of priceinv

         chi2(1)      =     0.05
         Prob > chi2  =   0.8215
```

This also shows that there is heteroskedasticity.

To check for Multicollinearity use the VIF (variance inflation factor) command

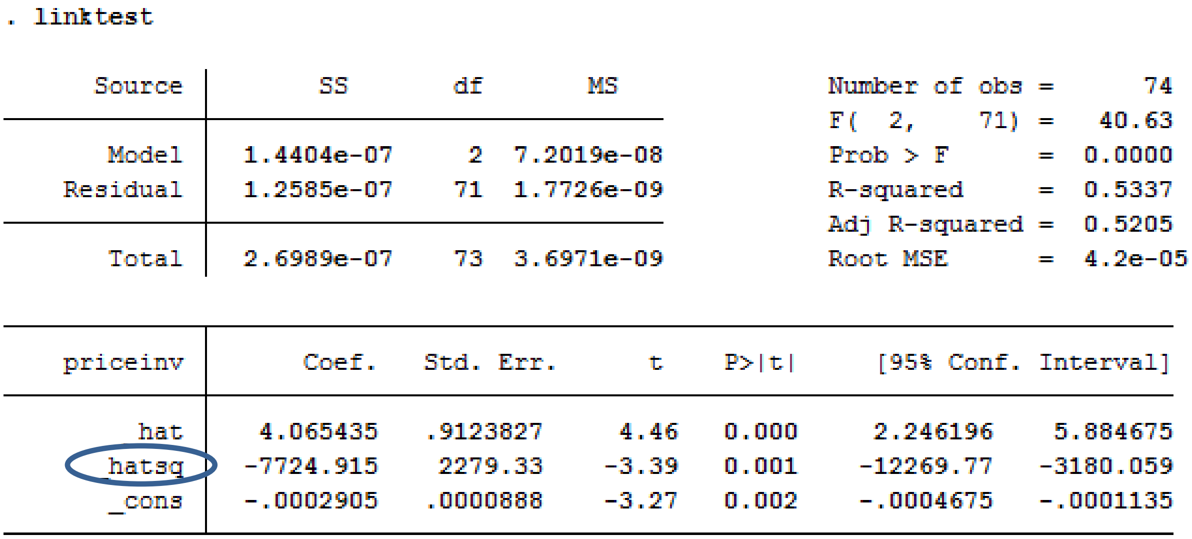


```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      length |      3.47    0.288507
         mpg |      2.77    0.361338
     foreign |      1.50    0.664942
-------------+----------------------
    Mean VIF |      2.58
```

As the values of VIF are less than 5, there is no multicollinearity detected.

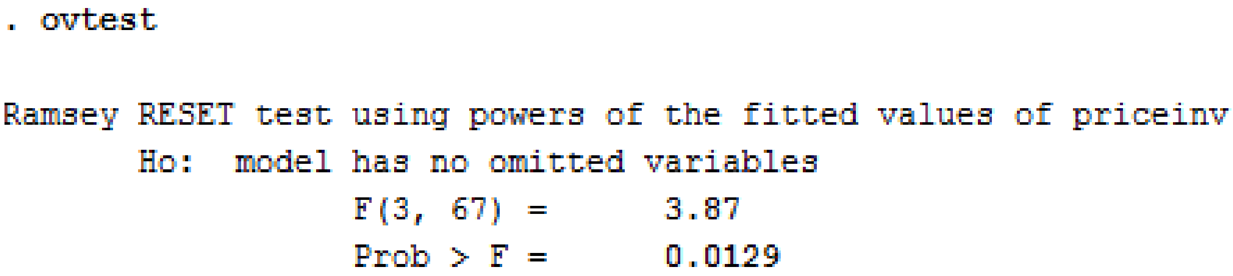
Another test the linktest is used. The linktest command performs a model specification test for single-equation models.

The null hypothesis is that the model is specified correctly. As the variable _hatsq is significant so the model is not specified correctly.



```
. linktest

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  2,    71) =   40.63
       Model |  1.4404e-07     2  7.2019e-08           Prob > F      =  0.0000
    Residual |  1.2585e-07    71  1.7726e-09           R-squared     =  0.5337
-------------+------------------------------           Adj R-squared =  0.5205
       Total |  2.6989e-07    73  3.6971e-09           Root MSE      =  4.2e-05

------------------------------------------------------------------------------
    priceinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   4.065435   .9123827     4.46   0.000     2.246196    5.884675
      _hatsq |  -7724.915    2279.33    -3.39   0.001    -12269.77   -3180.059
       _cons |  -.0002905   .0000888    -3.27   0.002    -.0004675   -.0001135
------------------------------------------------------------------------------
```

There is another test (missing variable). The command is ovtest. Here it shows that there may be some missing variables in the model that has been specified.



```
. ovtest

Ramsey RESET test using powers of the fitted values of priceinv
       Ho:  model has no omitted variables
                 F(3, 67) =      3.87
                 Prob > F =      0.0129
```

Testing for autocorrelation

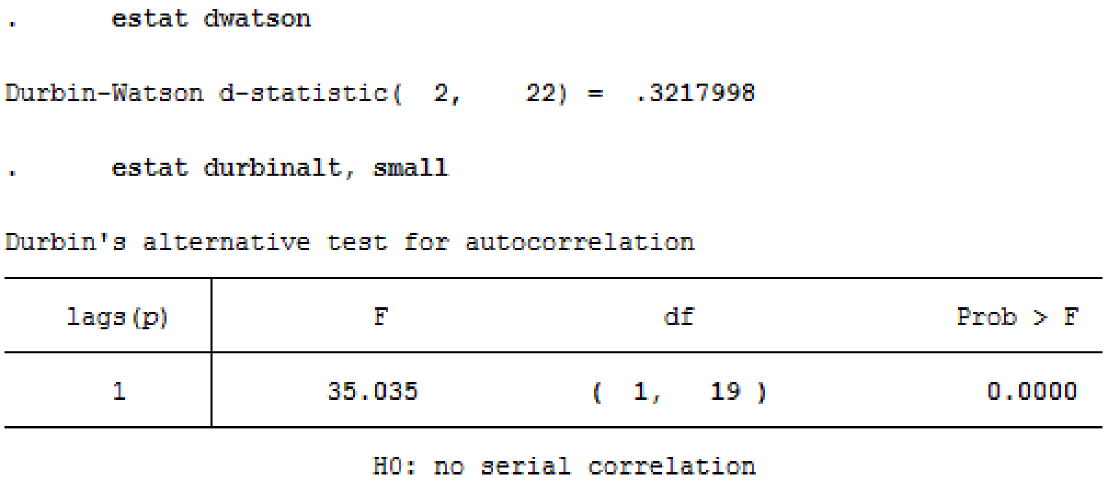
As autocorrelation is normally a time series phenomenon, we change our file an open a time series data file and use the Durbin Watson test. Type and enter the following:

webuse klein, clear

tsset yr

regress consump wagegovt

estat dwatson



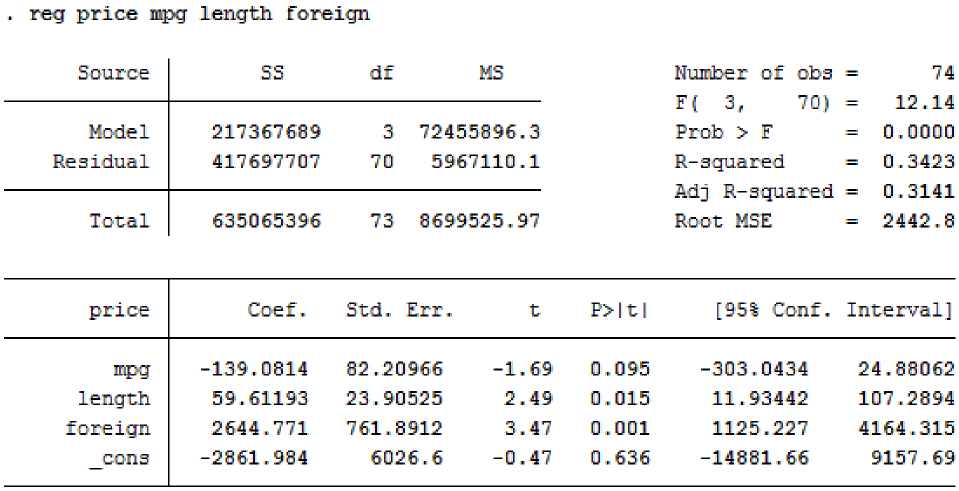
```
.      estat dwatson

Durbin-Watson d-statistic(  2,    22) =  .3217998

.      estat durbinalt, small

Durbin's alternative test for autocorrelation

    lags(p)  |          F                   df                      Prob > F
-------------+----------------------------------------------------------------
       1     |       35.035           (  1,    19 )                   0.0000
------------------------------------------------------------------------------
                             H0: no serial correlation
```

There is evidence that autocorrelation exists as we can reject the H0 of absence of autocorrelation (P value for F is less than 5%)

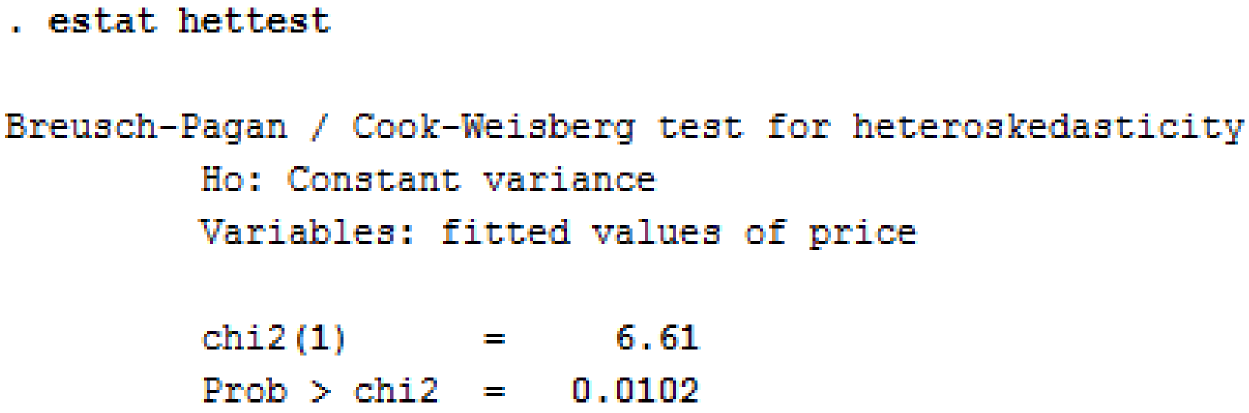
REMEDIES for Heteroskedasticity

White developed an estimator for standard errors that is robust to the presence of heteroskedasticity. Use the robust option with regression.

Remember we had transformed the outcome variable to 1/p. Now let us just use price. The errors will be heteroskedastic.



```
. reg price mpg length foreign

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   12.14
       Model |   217367689     3  72455896.3           Prob > F      =  0.0000
    Residual |   417697707    70   5967110.1           R-squared     =  0.3423
-------------+------------------------------           Adj R-squared =  0.3141
       Total |   635065396    73  8699525.97           Root MSE      =  2442.8

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -139.0814   82.20966    -1.69   0.095    -303.0434    24.88062
      length |   59.61193   23.90525     2.49   0.015     11.93442    107.2894
     foreign |   2644.771   761.8912     3.47   0.001     1125.227    4164.315
       _cons |  -2861.984     6026.6    -0.47   0.636    -14881.66     9157.69
------------------------------------------------------------------------------
```



```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =     6.61
         Prob > chi2  =   0.0102
```

This shows heteroskedasticity.

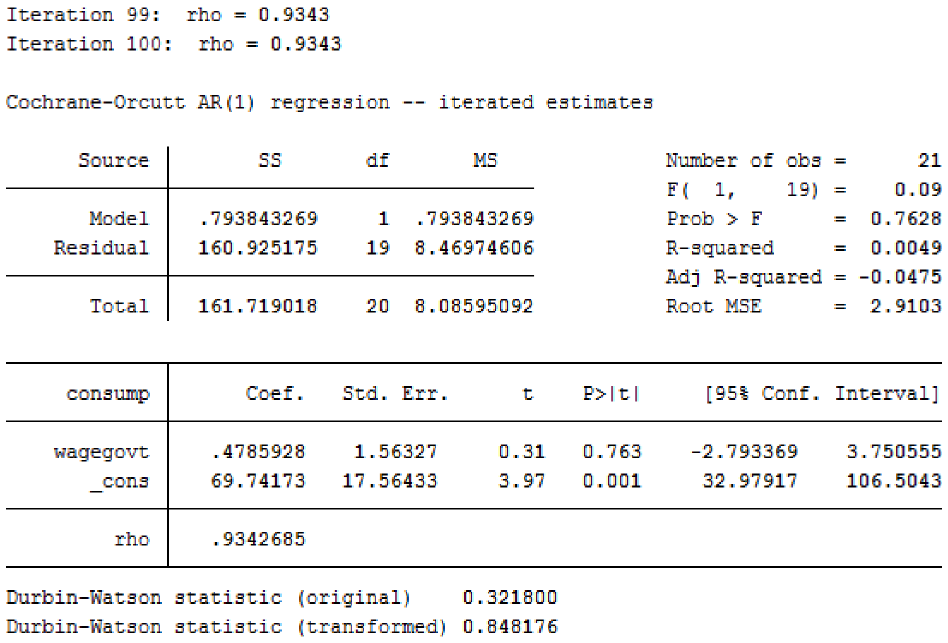
We can not run the regression with robust option

reg price mpg length foreign, robust

Standard errors will change, coefficients remain the same

Remedy for autocorrelation

Run the analysis with the Prais-Winston command, specifying the Cochran-Orcutt option.



```
Iteration 99:  rho = 0.9343
Iteration 100:  rho = 0.9343

Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =    0.09
       Model |  .793843269     1  .793843269           Prob > F      =  0.7628
    Residual |  160.925175    19  8.46974606           R-squared     =  0.0049
-------------+------------------------------           Adj R-squared = -0.0475
       Total |  161.719018    20  8.08595092           Root MSE      =  2.9103

------------------------------------------------------------------------------
     consump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    wagegovt |   .4785928    1.56327     0.31   0.763    -2.793369    3.750555
       _cons |   69.74173   17.56433     3.97   0.001     32.97917    106.5043
-------------+----------------------------------------------------------------
         rho |   .9342685
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    0.321800
Durbin-Watson statistic (transformed) 0.848176
```

After iterations dw has improved a bit. It is better to use models other than least square (AR1, ARIMA, lagged variable, GLS)

Lecture 30

Simultaneous Equation Models

OLS deals with single equations. We have a response variable and some explanatory variables. In fact, variables in real life may be a cause as well as an effect. The interdependence of variables may give rise to simultaneity bias if OLS is used. If we consider the demand and supply model or the national income model, we can quickly observe that variables may depend on each other. For example income is influenced by consumption level but consumption level itself depends on income levels.

To handle such relationships, we use simultaneous models with multiple equations where dependent variable in one equation may be the explanatory variable in some other equation. Consider the following examples of simultaneous models:

The Market Model

$$Q_d = \alpha_0 + \alpha_1 P + \alpha_2 I + e_1$$

$$Q_s = \beta_0 + \beta_1 P + \beta_2 T + e_2$$

$$Q_d = Q_s$$

Liquidity and Profitability

$$PROFITABILITY = \alpha_0 + \alpha_1 LIQUIDITY + \alpha_i F_i + e_1$$

$$LIQUIDITY = \beta_0 + \beta_1 PROFITABILITY + \alpha_i G_i + e_2$$

F_i AND G_i are matrices of determinants of prof and liq

The National Income (Macroeconomic) Model

$$Y_t = C_t + I_t + G_t$$

$$I_t = \alpha_0 + \alpha_1 Y_t$$

$$C_t = \beta_0 + \beta_1 Y_t + e_2$$

Endogenous and Exogenous Variables

Consider the following model

$$\boldsymbol{Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1}$$

$$\boldsymbol{Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_2 + e_2}$$

The number of unknowns is 4 and the number of equations is 2. To solve the model, the number of equations should be equal to No of unknowns. Note that $\boldsymbol{Y_1}$ and $\boldsymbol{Y_1}$ cause each other. Let us assume that $\boldsymbol{X_1}$and $\boldsymbol{X_1}$ are already given. Then we are left with two variables to be solved by two equations which is possible. The already given variables or the variables whose values externally (from outside the model) determined are called exogenous variables. The remaining variables whose values we seek by solving the model are called endogenous variables. Here $X_1$and X_2 are exogenous and Y_1 and Y_2 are endogenous variables.

Practice Question:

Consider

$$Q_d = \alpha_0 + \alpha_1 P + \alpha_2 I + e_1$$

$$Q_s = \beta_0 + \beta_1 P + \beta_2 T + e_2$$

$$Q_d = Q_s$$

Identify Exogenous and endogenous variables.

Example:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 Y_3 + \alpha_3 X_1 + e_1$$

$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 Y_3 + \beta_3 X_2 + e_2$$

$$Y_3 = \gamma_0 + \gamma_1 Y_1 + \gamma_2 Y_2 + \gamma_3 X_3 + e_3$$

- We have three equations so must have three endogenous variables
- Y_1, Y_2 and Y_3 are endogenous
- All remaining variables X_1, X_2 and X_3 are exogenous

Identification Problem

An equation (not model) can be Unidentified or Under-identified, exactly identified or Over identified. What do we need to do if OLS cannot be applied? we need to transform the model to reduced form.

Reduced form equation: equations that express endogenous variables as functions only of exogenous variables and disturbances. OLS can be applied to reduced-from equation (no endogenous variables on RHS). We need to solve and get reduced from equation for each Endogenous Variable.

No of endogenous variables in a model = no of reduced form equations

Example of simultaneous model is

$$\boldsymbol{Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1}$$

$$\boldsymbol{Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_2 + e_2}$$

After identifying the endogenous variables, we transform the model in to reduced form which may be like this

$$Y_1 = \pi_{10} + \pi_{11} X_1 + \pi_{12} X_2 + v_1$$

$$Y_2 = \pi_{20} + \pi_{21} X_1 + \pi_{22} X_2 + v_2$$

$\boldsymbol{\beta_i}$ are called structural parameters. π_{ij} are called reduced form coefficients.

- **Unidentified or Under-identified**: if we cannot express the structural parameters of an equation in terms of reduced form coefficients
- **Exactly identified**: if we CAN express the structural parameters of an equation in terms of reduced form coefficients in one way
- **Over identified**: if we CAN express the structural parameters of an equation in terms of reduced form coefficients in more than one way

Exclusion principle: consider a two equation system; For an equation to be identified, there should be at least one exogenous variable that is excluded from the equation i.e. the variable does not exist in the equation but must be found on the RHS of the other equation.

Example:

Consider the model

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1 \ldots\ldots\ldots\ldots (1)$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_2 + e_2 \ldots\ldots\ldots\ldots (2)$$

Substituting the value of Y_2 from equation 2 in equation 1 and solving for Y_1 gives

$$Y_1 = \frac{\alpha_0 + \alpha_1\beta_0}{1-\alpha_1\beta_1} + \frac{\alpha_2}{1-\alpha_1\beta_1} X_1 + \frac{\alpha_1\beta_2}{1-\alpha_1\beta_1} X_2 + u_1$$

Where u_1 contains expressions with residuals

The above can be written as

$$Y_1 = \pi_{10} + \pi_{11}X_1 + \pi_{12}X_2 + u_1$$

$$Y_1 = \pi_{10} + \pi_{11}X_1 + \pi_{12}X_2 + u_1$$

is a reduced from equation. Now substituting this expression for in equation 2, we get

$$Y_2 = (\beta_0 + \beta_1\pi_{11}) + \beta_1\pi_{11}X_1 + (\beta_1\pi_{12} + \beta_2)X_2 + \beta u_1 + e_2$$

or

$$Y_2 = \frac{\beta_0 + \alpha_0\beta_1}{1-\alpha_1\beta_1} + \frac{\beta_1\alpha_2}{1-\alpha_1\beta_1} X_1 + \frac{\beta_2}{1-\alpha_1\beta_1} X_2 + u_2$$

$$Y_2 = \pi_{20} + \pi_{21}X_1 + \pi_{22}X_2 + u_2$$

Results:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_2 + e_2$$

The above model is now transformed into reduced form equations

$$Y_1 = \pi_{10} + \pi_{11}X_1 + \pi_{12}X_2 + u_1$$

$$Y_2 = \pi_{20} + \pi_{21}X_1 + \pi_{22}X_2 + u_1$$

Where reduced form coefficients are symbols for some expression in the structural parameters

The reduced form coefficients

$$\pi_{10} = \frac{\alpha_0 + \alpha_1\beta_0}{1-\alpha_1\beta_1}, \pi_{11} = \frac{\alpha_2}{1-\alpha_1\beta_1}, \pi_{12} = \frac{\alpha_1\beta_2}{1-\alpha_1\beta_1}$$

$$\pi_{20} = \frac{\beta_0 + \alpha_0\beta_1}{1-\alpha_1\beta_1}, \pi_{21} = \frac{\beta_1\alpha_2}{1-\alpha_1\beta_1}, \pi_{22} = \frac{\beta_2}{1-\alpha_1\beta_1}$$

Solving the above, we can see that we can express the structural parameters (α_i and β_i) in the form of reduced form coefficients.

We can easily see that

$$\beta_1 = \frac{\pi_{21}}{\pi_{11}}, \quad \beta_0 = \pi_{20} - \frac{\pi_{21}}{\pi_{11}}\pi_{10}$$

$$and \;\; \beta_2 = \pi_{22} - \frac{\pi_{21}}{\pi_{11}}\pi_{12}$$

As the structural parameters of equation 2 can be expressed in one way in terms of reduced form coefficients, Equation 2 is exactly identified. Also

$$\alpha_1 = \frac{\pi_{12}}{\pi_{22}}, \quad \alpha_0 = \pi_{10} - \frac{\pi_{12}}{\pi_{22}}\pi_{20}$$

$$and \;\; \alpha_2 = \pi_{11} - \frac{\pi_{12}}{\pi_{22}}\pi_{21}$$

As the structural parameters of equation 1 can be expressed in one way in terms of reduced form coefficients, Equation 1 is also exactly identified

Lecture 31

Simultaneous Equation Models-II

Let us take another example of identification

Example:

Consider the market model

$$Q_d = \alpha_0 + \alpha_1 P + \alpha_2 I + e_1$$

$$Q_s = \beta_0 + \beta_1 P + \beta_2 T + e_2$$

$$Q_d = Q_s$$

This time it is not simple to identify the endogenous variable. We do that with the help of market equilibrium analysis

- Use equilibrium condition $Q_d = Q_s$
- Remember that in the market model we solve for price and quantity

Endogenous variables are Q and P and Exogenous variables are I and T

To solve, we use the equilibrium condition $Q_d = Q_s$

So

$$\boldsymbol{\alpha_0 + \alpha_1 P + \alpha_2 I + e_1 \qquad = \beta_0 + \beta_1 P + \beta_2 T + e_2}$$

Now solve for the endogenous variable P

$$\boldsymbol{P = \frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{\beta_2}{(\alpha_1 - \beta_1)}T - \frac{\alpha_2}{(\alpha_1 - \beta_1)}I + \frac{(e_2 - e_1)}{(\alpha_1 - \beta_1)}}$$

Labeling

$$P = \pi_{10} + \pi_{11}T + \pi_{12}I + v_1$$

$$P = \pi_{10} + \pi_{11}T + \pi_{12}I + v_1$$

Substituting in the demand equation

$$\boldsymbol{Q} = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1}(\pi_{10} + \pi_{11}T + \pi_{12}I + v_1) + \boldsymbol{\alpha_2 I} + \boldsymbol{e_1}$$

Or

$$\boldsymbol{= (\alpha_0 + \alpha_1 \pi_{10}) + \alpha_1 \pi_{11} T + (\alpha_1 \pi_{12} + \alpha_2) I + (e_1 + \alpha_1 v_1)}$$

Labeling, we get

$$Q\ = \boldsymbol{\pi}_{20} +\ \boldsymbol{\pi}_{21}T + \boldsymbol{\pi}_{22}I +\ v_2$$

Now we have both reduced forms

$$P =\ \pi_{10} +\ \pi_{11}T +\ \pi_{12}I +\ v_1$$

$$Q\ = \pi_{20} +\ \pi_{21}T + \pi_{22}I +\ v_2$$

Where $\pi_{10} = \frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)},\ \pi_{11} = \frac{\beta_2}{(\alpha_1 - \beta_1)},\ \pi_{12} = -\ \frac{\alpha_2}{(\alpha_1 - \beta_1)}$

And $\pi_{20} = \alpha_0 +\ \alpha_1\pi_{10}, \qquad \pi_{21} = \alpha_1\pi_{11},$

$$\pi_{22} = \alpha_1\pi_{12} +\ \alpha_2$$

All structural parameters may be expressed in terms of reduced form coefficients (solve to get the following)

$$\beta_1 = \frac{\pi_{22}}{\pi_{12}}, \beta_0 = \pi_{20} - \frac{\pi_{22}}{\pi_{12}}\pi_{10}, \beta_2 = \pi_{21} - \frac{\pi_{22}}{\pi_{12}}\pi_{11}$$

$$\alpha_1 = \frac{\pi_{21}}{\pi_{11}}, \alpha_2 = \pi_{22} - \frac{\pi_{21}}{\pi_{11}}\pi_{12}, \alpha_0 = \pi_{20} - \frac{\pi_{21}}{\pi_{11}}\pi_{10}$$

Hence both equations are exactly identified

Example: under-identified equation

Now let us see a model with one under-identified equation. Consider the model

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1$$

$$Y_2 =\ \beta_0 + \beta_1 Y_1 + e_2$$

Exogenous variables: X_1; Endogenous variables: $Y_1$and Y_2

Substituting value of Y_2 in equation 1

$$\boldsymbol{Y_1} = \frac{(\boldsymbol{\alpha}_0 +\ \boldsymbol{\alpha_1\beta}_0)}{(\mathbf{1} - \boldsymbol{\alpha_1\beta_1})} + \frac{\boldsymbol{\alpha_2}}{(\mathbf{1} - \boldsymbol{\alpha_1\beta_1})}\boldsymbol{X_1} + \frac{(\boldsymbol{\alpha_1 e_2} + \boldsymbol{e_1})}{(\mathbf{1} - \boldsymbol{\alpha_1\beta_1})}$$

Labeling

$$\boldsymbol{Y_1} =\ \boldsymbol{\pi_{10}} + \boldsymbol{\pi_{11}X_1} + \boldsymbol{V_1}$$

$$\boldsymbol{\pi_{10}} = \frac{(\boldsymbol{\alpha_0} + \boldsymbol{\alpha_1\beta_0})}{(\mathbf{1} - \boldsymbol{\alpha_1\beta_1})},\ \ \boldsymbol{\pi_{11}} = \frac{\boldsymbol{\alpha_2}}{(\mathbf{1} - \boldsymbol{\alpha_1\beta_1})}$$

Substituting reduced form of Y_1 in equation 2

$$\boldsymbol{Y_2} =\ \boldsymbol{\beta}_0 + \boldsymbol{\beta_1}(\boldsymbol{\pi_{10}} + \boldsymbol{\pi_{11}X_1} + \boldsymbol{V_1}) + \boldsymbol{e_2}$$

Or $\boldsymbol{Y_2} = (\boldsymbol{\beta}_0 + \boldsymbol{\beta_1\pi_{10}}) + \boldsymbol{\beta_1\pi_{11}X_1} + (\boldsymbol{\beta_1V_1} + \boldsymbol{e_2})$

Labeling

$$\boldsymbol{Y_2} = \boldsymbol{\pi_{20}} + \boldsymbol{\pi_{21}X_1} + \boldsymbol{V_2}$$

$$\boldsymbol{\pi_{20}}\text{=}\ \boldsymbol{\beta}_0 + \boldsymbol{\beta_1\pi_{10}}$$

$$\boldsymbol{\pi_{21}} = \boldsymbol{\beta_1\pi_{11}}$$

- We can solve for $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_0}$ but not for α_i so only equation 2 is exactly identified. Equation 1 is unidentified.
- The parameters of an unidentified equation have no interpretation, because you do not have enough information to obtain meaningful estimates

Example: One over identified equation

A model with one over identified equation

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + e_1$$

$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_1 + \beta_3 X_2 + e_2$$

Substituting the first equation in second and rearranging,

$$Y_2 = \frac{\beta_0 + \beta_1\alpha_0}{1 - \alpha_1\beta_1} + \frac{\beta_2}{1 - \alpha_1\beta_1}X_1 + \frac{\beta_3}{1 - \alpha_1\beta_1}X_2 + \beta_1 e_1 + e_2$$

Labeling, we get reduced form for Y_2

$$Y_2 = \pi_{20} + \pi_{21}X_1 + \pi_{22}X_2 + v_2$$

where

$$\pi_{20} = \frac{\beta_0 + \beta_1\alpha_0}{1 - \alpha_1\beta_1}$$

$$\pi_{21} = \frac{\beta_2}{1 - \alpha_1\beta_1}$$

$$\pi_{22} = \frac{\beta_3}{1 - \alpha_1\beta_1}$$

Substituting the reduced form for Y_2 in the first structural equation

$$Y_1 = \alpha_0 + \alpha_1(\pi_{20} + \pi_{21}X_1 + \pi_{22}X_2 + v_2) + e_1$$

$$Y_1 = \alpha_0 + \alpha_1\pi_{20} + \alpha_1\pi_{21}X_1 + \alpha_1\pi_{22}X_2 + \alpha_1 v_2 + e_1$$

$$Y_1 = \pi_{10} + \pi_{11}X_1 + \pi_{12}X_2 + e_1$$

Where

$$\pi_{10} = \alpha_0 + \alpha_1\pi_{20}$$

$$\pi_{11} = \alpha_1\pi_{21}$$

$$\pi_{12} = \alpha_1\pi_{22}$$

This means that we can find or express α_1 in terms of reduced from coefficients in two ways, as follows:

$$\alpha_1 = \frac{\pi_{11}}{\pi_{21}}$$

And in another way as,

$$\alpha_1 = \frac{\pi_{12}}{\pi_{22}}$$

Hence *equation 1 is over-identified.*

Equation 2 is under-identified.

Order Condition for Identification status

We can know the identification status by applying order condition without solving or using reduced forms. We compare two parameters (call them $P_1$and P_2).

- P_1: Number of exogenous variables excluded from the equation

(number of exogenous variables in the model – number of exogenous variables in the equation)

- P_2: Number of endogenous variable in the model minus 1

If $P_1 < P_2$, the equation is under-identified

If $P_1 = P_2$, the equation is exactly identified

If $P_1 > P_2$, the equation is over-identified

- If the equation is under-identified, we cannot find its parameters
- If the equation is exactly identified, we use ILS
- If the equation is over-identified, we may use 2SLS

Example:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_2 + e_2$$

Identification status of equation 1

$P_1 = 1$ (X_2 *is the only exog. variable excluded from equation 1)*

$P_2 = 2 - 1 = 1$ (Y_1 and Y_2 are two endogenous variables)

$P_1 = P_2$ So equation 1 is exactly identified *(apply ILS)*

Identification status of equation 2

$P_1 = 1$ (X_1 *is the only exog. variable excluded from equation 2)*

$P_2 = 2 - 1 = 1$ (Y_1 and Y_2 are two endogenous variables)

$P_1 = P_2$ So equation 1 is exactly identified *(apply ILS)*

Example:

$$Q_d = \alpha_0 + \alpha_1 P + \alpha_2 I + e_1$$

$$Q_s = \beta_0 + \beta_1 P + \beta_2 T + e_2$$

$$Q_d = Q_s$$

Identification status of equation 1

$P_1 = 1$ (T *is the only exog. variable excluded from equation 1)*

$P_2 = 2 - 1 = 1$ (Q and P are two endogenous variables)

$P_1 = P_2$ So equation 1 is exactly identified *(apply ILS)*

Identification status of equation 2

$P_1 = 1$ (I *is the only exog. variable excluded from equation 2)*

$P_2 = 2 - 1 = 1$ (Q and P are two endogenous variables)

$P_1 = P_2$ So equation 1 is exactly identified *(apply ILS)*

Example:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1$$

$$Y_2 = \beta_0 + \beta_1 Y_1 + e_2$$

Identification status of equation 1

$P_1 = 0$ ($No\ exogenous$ *variable is excluded from equation 1)*

$P_2 = 2 - 1 = 1$ (Y_1 and Y_2 are two endogenous variables)

$P_1 < P_2$ So equation 1 is under- identified *(no method can be applied)*

Identification status of equation 2

$P_1 = 1$ (X_1 *is the only exog. variable excluded from equation 2)*

$P_2 = 2 - 1 = 1$ (Y_1 and Y_2 are two endogenous variables)

$P_1 = P_2$ So equation 1 is exactly-identified *(apply ILS)*

Example:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + e_1$$

$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_1 + \beta_3 X_2 + e_2$$

Identification status of equation 1

$P_1 = 2$ ($No.\ of\ exogenous$ *variable excluded from equation 1)*

$P_2 = 2 - 1 = 1$ (Y_1 and Y_2 are two endogenous variables)

$P_1 > P_2$ So equation 1 is over- identified *(apply 2SLS)*

Identification status of equation 2

$P_1 = 0$ ($no\ exog.\ variable\ excluded\ from$ *equation 2)*

$P_2 = 2 - 1 = 1$ (Y_1 and Y_2 are two endogenous variables)

$P_1 < P_2$ So equation 1 is under-identified *(no method can be applied)*

Example:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 Y_3 + +\beta_2 X_1 + e_1$$

$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 Y_3 + \beta_3 X_2 + e_2$$

$$Y_3 = \gamma_0 + \gamma_1 Y_1 + \gamma_2 Y_2 + \gamma_3 X_3 + e_3$$

Identification status of equation 1:

$P_1 = 2$ *(two exogenous variables are excluded from equation 1, one is found in equation 2 and the other in equation 3)*

$P_2 = 3 - 1 = 2$ (Y_1, Y_2 and Y_3 are three endogenous variables)

$P_1 = P_2$ So equation 1 is exactly identified *(apply ILS)*

Identification status of equation 2 & 3:

Applying similar process, we can find that both equation 2 and 3 are exactly identified and we can apply ILS to find the parameters.

Lecture 32

Indirect Least Square (ILS)

Indirect least square may be applied when an equation is exactly identified. ILS will provide consistent estimates of the structural parameters. Applying OLS to individual equation of a simultaneous equation model does not provide consistent results (simultaneity bias).

STEPS

- Apply order condition
- Find the reduced from equations
- Express the structural parameters in terms of reduced form coefficients
- Estimate the reduced form coefficients by OLS
- Calculate the structural parameters

Example:

Consider:

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_2 + e_2$$

Applying order condition shows that both equations are exactly identified

Reduced form equations are

$$Y_1 = \pi_{10} + \pi_{11} X_1 + \pi_{12} X_2 + u_1$$

$$Y_2 = \pi_{20} + \pi_{21} X_1 + \pi_{22} X_2 + u_1$$

Expressing the structural parameters in terms of reduced form coefficients :

$$\beta_1 = \frac{\pi_{21}}{\pi_{11}}, \quad \beta_0 = \pi_{20} - \frac{\pi_{21}}{\pi_{11}} \pi_{10}$$

$$and \;\; \beta_2 = \pi_{22} - \frac{\pi_{21}}{\pi_{11}} \pi_{12}$$

and

$$\alpha_1 = \frac{\pi_{12}}{\pi_{22}}, \quad \alpha_0 = \pi_{10} - \frac{\pi_{12}}{\pi_{22}} \pi_{20}$$

$$and\ \ \alpha_2 = \pi_{11} - \frac{\pi_{12}}{\pi_{22}}\pi_{21}$$

Now consider the data (***ILS.xlsx***)

Y_1	Y_2	X_1	X_2
25	0.8	10	12
25	0.9	8	11
27	0.8	8	12
29	1.2	7	11
32	1.2	6	9
32	1.6	3	8
33	1.9	4	8
36	2.1	5	7
40	2	4	5
49	2	2	5

The OLS estimation of the reduced from equations are

Y_1=54.96 - 0.245 X_1 - 2.359 X_2 And Y_2=3.011 - 0.039 X_1 - 0.152 X_2

Comparing the results

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + e_1$$

Y_1=54.96 - 0.245 X_1 - 2.359 X_2

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_2 + e_2$$

Y_2=3.011 - 0.039 X_1 - 0.152 X_2

π_{10}	π_{11}	π_{12}
54.96086	-0.24518	-2.35947
π_{20}	π_{21}	π_{22}
3.010546	-0.03949	-0.15175

Using these estimations, we can compute the structural parameters indirectly.

Indirect estimation using the expression of structural parameters in terms of reduced form coefficients gives the following results.

$\alpha_{\text{0-ILS}}$	$\alpha_{\text{1-ILS}}$	$\alpha_{\text{2-ILS}}$
8.152606	15.54809	0.368888
$\beta_{\text{0-ILS}}$	$\beta_{\text{1-ILS}}$	$\beta_{\text{2-ILS}}$
-5.8429	0.161086	0.228326

The estimated equations can thus be written as:

$$Y_1 = 8.15 + 15.55\, Y_2 + 0.369\, X_1$$

$$Y_2 = -5.84 + 0.161\, Y_1 + 0.228\, X_2$$

If we would have estimated the structural equations individually by OLS we would have got different and inconsistent results

Example:

Consider the demand and supply equations:

$$Q_d \ = \ \alpha_0 + \ \alpha_1 P + \ \alpha_2 I + \ e_1$$

$$Q_s = \ \beta_0 + \ \beta_1 P + \ \beta_2 T + \ e_2$$

$$Q_d = Q_s$$

Applying order condition shows that both equations are exactly identified

Reduced form equations are

$$P = \ \pi_{10} + \ \pi_{11} T + \ \pi_{12} I + \ v_1$$

$$Q \ = \pi_{20} + \ \pi_{21} T + \pi_{22} I + \ v_2$$

The structural parameters expressed in terms of the reduced from coefficients as:

$$\beta_1 = \frac{\pi_{22}}{\pi_{12}}, \beta_0 = \pi_{20} - \frac{\pi_{22}}{\pi_{12}}\pi_{10}, \beta_2 = \pi_{21} - \frac{\pi_{22}}{\pi_{12}}\pi_{11}$$

$$\alpha_1 = \frac{\pi_{21}}{\pi_{11}}, \alpha_2 = \pi_{22} - \frac{\pi_{21}}{\pi_{11}}\pi_{12}, \alpha_0 = \pi_{20} - \frac{\pi_{21}}{\pi_{11}}\pi_{10}$$

Now consider the data (***ILS.xlsx***)

Quantity	Price	level of	level of technology
Q	P	I	T
5	3	5	0.2
6	5	10	0.5
7	2	6	0.5
4	6	7	0.2
6	10	15	0.2
9	5	8	0.5
11	6	10	0.9
7	6	8	0.5
6	12	12	0.2
12	12	16	0.9

The OLS estimation of the reduced from equations are

$$P = -0.82 - 2.62\,T \; + \; 0.9\,I$$

$$Q \; = \; 2.12 \; + \; 8.2\,T \; + \; 0.15\,I$$

Comparing the results

$$P = \; \pi_{10} + \; \pi_{11}T + \; \pi_{12}I + \; v_1$$

$$P = -0.82 - 2.62\,T \; + \; 0.9\,I$$

$$Q \; = \pi_{20} + \; \pi_{21}T + \pi_{22}I + \; v_2$$

$$Q \; = \; 2.12 \; + \; 8.2\,T \; + \; 0.15\,I$$

π_{10}	π_{11}	π_{12}
-0.82	-2.62	0.9
π_{20}	π_{21}	π_{22}
2.12	8.2	0.15

Using these estimations, we can compute the structural parameters indirectly

Indirect estimation using the expression of structural parameters in terms of reduced form coefficients gives the following results

$\alpha_{0\text{-ILS}}$	$\alpha_{1\text{-ILS}}$	$\alpha_{2\text{-ILS}}$
-0.45087	-3.12665	2.958705
$\beta_{0\text{-ILS}}$	$\beta_{1\text{-ILS}}$	$\beta_{2\text{-ILS}}$
2.250948	0.161495	8.623996

With these ILS estimations, the estimated equations can be written as

$$Q_d = -0.45 - 3.13\,P + 2.96\,I$$

$$Q_s = 2.25 + 0.16\,P + 8.62\,T$$

If we would have estimated the structural equations individually by OLS we would have got different and inconsistent results

ILS using Stata

- ***reg3*** can be used. reg3 is a command for 3SLS (option 2sls can be used)
- For exactly Identified equations: 2SLS Results identical to ILS
- Example II can be done as follows:

reg3 q p i, exog(i t) endog (q p) 2sls

and

reg3 q p t, exog(i t) endog (q p) 2sls

From the Interface

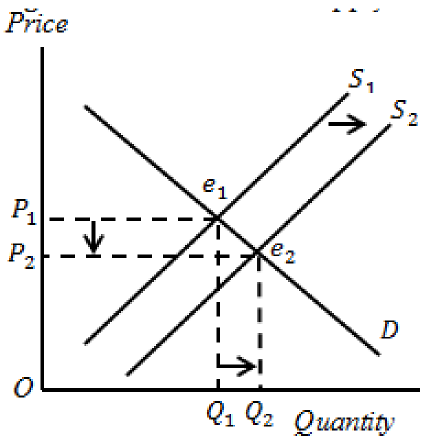
$$Statistics \rightarrow Linear\ Models \rightarrow Multiple\ equation\ models \rightarrow three\ stage\ least\ square$$

In a two equation system, at least one variable must be excluded from an equation to make it identified.

To identify the demand equation Technology must appear in the supply equation and Technology must NOT appear in the demand equation.

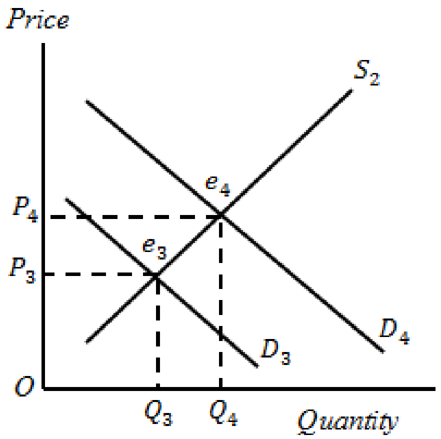
Understanding Identification: Exclusion Restriction

Consider two points on the same demand curve. Two points on the demand curve show different supply curves. Supply curve shifts through 'shift factors' or 'supply shifters'. Technology is a supply shifter.



The two points on the same demand curve. It must not appear in the demand equation because if it is then it will shift the demand as well. To have two points on the same demand curve, we need an exogenous variable that does not shift the demand curve but only the supply curve.

Now consider two points on the same supply curve. Two points on the supply curve show different demand curves. Demand curve shifts through 'shift factors' or 'demand shifters'. Income level is a demand shifter.



To show both points (different demand curves) Income level must appear in the demand equation. It must not appear in the supply equation because if it is then it will shift the supply as well. To have two points on the same supply curve, we need an exogenous variable that does not shift the supply curve but only the demand curve.

Some notes and assumption on ILS

We can directly write the reduced form equations without solving the model. We need to solve the model to derive the structural parameters from the reduced form coefficients. 2SLS also can be applied to exactly identified equations with the same results as ILS. ILS and 2SLS provide the same results when 2SLS is applied to a single equation. ILS and 3SLS may provide the same results when applied to a complete system. (Park, Canadian Journal of Statistics 01/1974)

Assumption of ILS:

- Equations must be exactly identified
- Error terms in all reduced from equations must satisfy all usual assumption of OLS estimation
- No Multicollinearity in exogenous variables
- Sample size not very small (ILS biased for small samples)

Lecture 33

Two Stage Least Square

We can use 2SLS to estimate the parameters of over-identified equations.

Stage I: Derive or write the reduced forms for all endogenous variables on the RHS of the over-identified equation and estimate them using OLS

Stage II: find trend value of the above estimated reduced form equations and replace with the variables in the over-identified equation and estimate it using OLS

Consider the model

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + e_1$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_1 + \beta_3 X_2 + e_2$$

Stage I: First equation is over-identified (we will use order condition to verify). As Y_2 is the endogenous variable on the RHS of the first (over-identified) equation, we need to estimate the reduced form for Y_2

Stage II: Replace Y_2 with $\widehat{Y_2}$(from the reduced form) and estimate it using OLS

Let us do it in detail

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + e_1$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_1 + \beta_3 X_2 + e_2$$

Order Condition: Equation 1

Number of exogenous variables excluded from the equation is 2 which is greater than the Number of endogenous variables in the model minus one (which gives 1) so equation 1 is over-identified. ILS cannot be used.

Stage 1

- We need the reduced form for the endogenous variable on the RHS of equation 1
- Y_2 is the variable

Reduced form equations contain only exogenous variables and they contain all the exogenous variables of the model.

$$Y_2 = \pi_{10} + \pi_{11} X_1 + \pi_{12} X_2 + v_1$$

For 2SLS We do not need to know what the π_i are equivalent to.

Now consider a small data set (2SLS.xlsx) to estimate

$$Y_2 = \pi_{10} + \pi_{11}X_1 + \pi_{12}X_2 + v_1$$

After estimating, we need to compute the trend values for Y_2

Y_1	Y_2	X_1	X_2
25	0.8	10	12
25	0.9	8	11
27	0.8	8	12
29	1.2	7	11
32	1.2	6	9
32	1.6	3	8
33	1.9	4	8
36	2.1	5	7
40	2	4	5
49	2	2	5

The estimates for $Y_2 = \pi_{10} + \pi_{11}X_1 + \pi_{12}X_2 + v_1$ are $Y_2 \ = \ 3.011 - 0.039\ X_1 - 0.152\ X_2$

Trend (Y_2)
0.794564
1.025306
0.873553
1.0648
1.407801
1.678038
1.638543
1.750802
2.093803
2.172792

This gives the trend values. Stage I is complete now.

In Stage II, we replace the variable Y2 in the original over-identified equation with the trend values from the reduced form of Y2 and run the regression.

The original (over-identified) equation was

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + e_1$$

After replacing Y2 with the estimated Y2, the equation becomes

$$Y_1 = \alpha_0 + \alpha_1 \widehat{Y_2} + e_1$$

Estimating the following

$$Y_1 = \alpha_0 + \alpha_1 \widehat{Y_2} + e_1$$

gives the following results

$$Y1 \; = \; 12.81 \; + \; 13.785\, Y2$$

12.81 and 13.785 are the 2SLS estimates of α_{0-2SLS} and α_{1-2SLS}. If we would have directly regressed $Y1$ on $Y2$ we would have got (not appropriate estimates). 16.03 and 11.6 as α_{0OLS} and α_{1OLS}.

Example:

Consider the model

$$Q_d \; = \; \alpha_0 + \; \alpha_1 P + \; e_1$$

$$Q_s = \; \beta_0 + \; \beta_1 P + \beta_2 C + \beta_3 T + \; e_2$$

$$Q_d = Q_s$$

Where

$$P \; = \; Price, \; Q = Quantity \; (demanded \; or \; supplied)$$

$C \; = \; Cost \; of \; production \; T = \; the \; level \; of \; technology.$

- 2SLS is needed for the first equation
- Stage I: First equation is over-identified. As P is the endogenous variable on the RHS of the first (over-identified) equation, we need to estimate the reduced form for P

- Stage II: Replace P with $\hat{P}$(from the reduced form) in the demand equation and estimate it using OLS

Let us do it in detail

$$Q_d = \alpha_0 + \alpha_1 P + e_1$$

$$Q_s = \beta_0 + \beta_1 P + \beta_2 C + \beta_3 T + e_2$$

$$Q_d = Q_s$$

Order Condition: Demand equation

Number of exogenous variables excluded from the equation is 2 which is greater than the Number of endogenous variables in the model minus one (which gives 1) so demand equation is over-identified. Here ILS cannot be used.

Stage 1

- We need the reduced form for the endogenous variable on the RHS of demand equation
- Q and P are endogenous
- C and T are exogenous variables
- P is the variable for which reduced form is needed

Reduced form equations have the following properties

- They contains only exogenous variables
- They contains all the exogenous variables of the model

So the reduced form for P can be written as

$$P = \pi_{10} + \pi_{11} C + \pi_{12} T + v_1$$

For 2SLS we do not need to know what the π_i are equivalent to

Now consider a small data set (2SLS.xlsx) to estimate

$$P = \pi_{10} + \pi_{11} C + \pi_{12} T + v_1$$

After estimating, we need to compute the trend values for P

Q	**P**	**C**	**T**
5	3	10	0.2
6	5	10	0.5
7	2	9	0.5
4	6	12	0.2
6	10	9	0.2
9	5	8	0.5
11	6	7	0.9
7	6	9	0.5
6	12	9	0.2
12	12	12	0.9

The estimates for

$$P = \pi_{10} + \pi_{11}C + \pi_{12}T + v_1$$

are

$$P = \ 0.868 \ + \ 0.548\, C \ + \ 1.3683\, T$$

This gives the trend values. Stage I is complete now.

Trend (P)
6.618038
7.028532
6.480933
7.713237
6.070439
5.933333
5.933059
6.480933
6.070439
8.671056

In Stage II, we replace the variable P in the original over-identified equation with the trend values from the reduced form of P and run the regression.

The original (over-identified) equation was

$$Q_d = \alpha_0 + \alpha_1 P + e_1$$

After replacing P with the estimated P, the equation becomes

$$Q = \alpha_0 + \alpha_1 \hat{P} + e_1$$

Estimating the following

$$Q = \alpha_0 + \alpha_1 \hat{P} + e_1$$

gives the following results (parameters written for original equation)

$$Q = 4.1735 + 0.4667\,P$$

4.1735 and 0.4667 are the 2SLS estimates of α_{0-2SLS} and α_{1-2SLS}. If we would have directly regressed Q on P we would have got (not appropriate estimates). 5.906 and 0.20799 as α_{0OLS} and α_{1OLS}

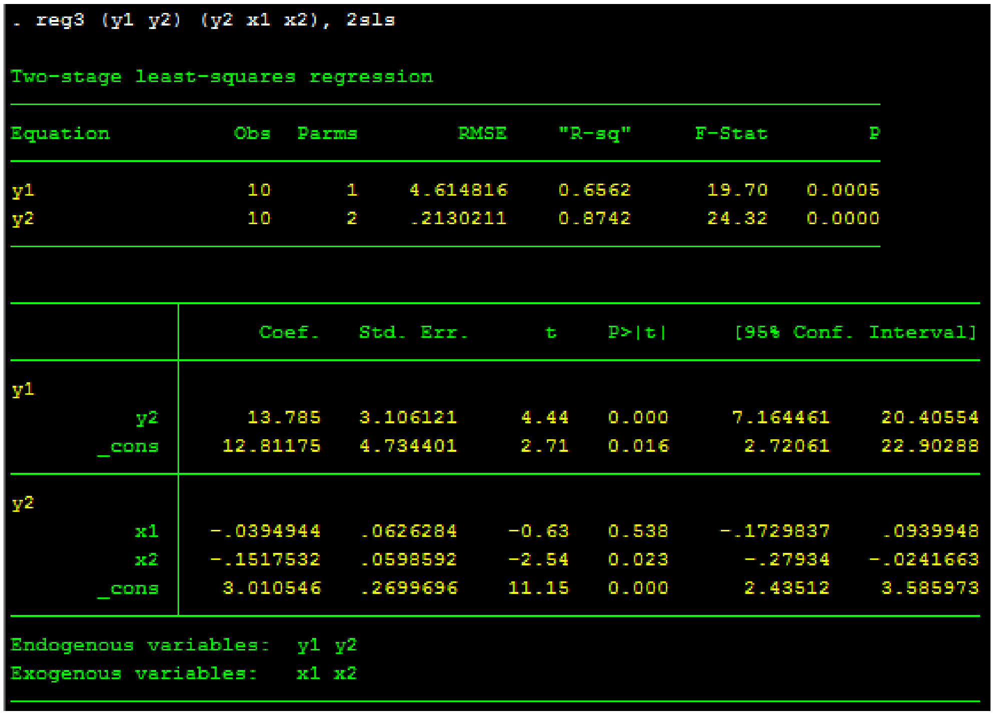
Using Stata for Two Stage Least Square (2SLS)

On the Menu: *Statistics > Endogenous covariates > Three-stage least squares*

In the window, after providing required information, we need to check the option 2SLS. In the command window (for the model in the first example) type:

$$reg3\ (y1\ y2)\ (y2\ x1\ x2),\ 2sls$$

That is, reg3 (first over-identified equation)(second reduced from equation), option-2SLS.

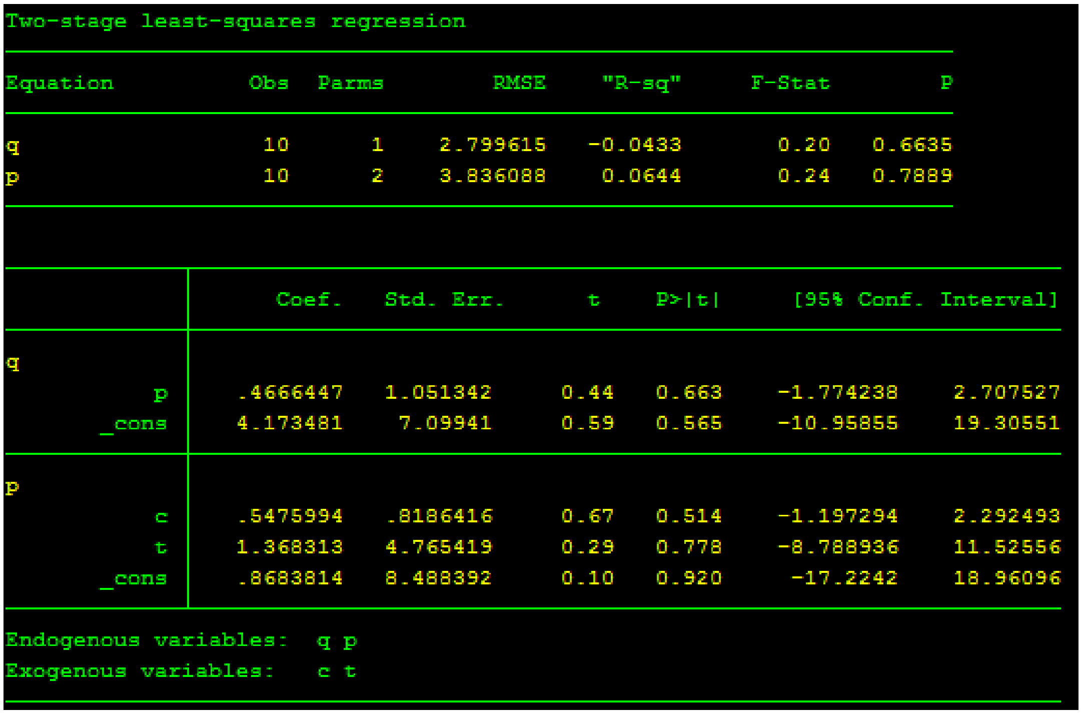


```
. reg3 (y1 y2) (y2 x1 x2), 2sls

Two-stage least-squares regression
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"     F-Stat        P
----------------------------------------------------------------------
y1                 10      1    4.614816    0.6562      19.70   0.0005
y2                 10      2    .2130211    0.8742      24.32   0.0000
----------------------------------------------------------------------


------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y1           |
          y2 |     13.785   3.106121     4.44   0.000     7.164461    20.40554
       _cons |   12.81175   4.734401     2.71   0.016      2.72061    22.90288
-------------+----------------------------------------------------------------
y2           |
          x1 |  -.0394944   .0626284    -0.63   0.538    -.1729837     .0939948
          x2 |  -.1517532   .0598592    -2.54   0.023      -.27934   -.0241663
       _cons |   3.010546   .2699696    11.15   0.000      2.43512    3.585973
------------------------------------------------------------------------------
Endogenous variables:  y1 y2
Exogenous variables:   x1 x2
------------------------------------------------------------------------------
```

For example II

In the command window (for the model in the first example) type:

$$reg3\ (q\ p)\ (p\ c\ t),\ 2sls$$



```
Two-stage least-squares regression
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"     F-Stat        P
----------------------------------------------------------------------
q                  10      1    2.799615   -0.0433       0.20   0.6635
p                  10      2    3.836088    0.0644       0.24   0.7889
----------------------------------------------------------------------


------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q            |
           p |   .4666447   1.051342     0.44   0.663    -1.774238    2.707527
       _cons |   4.173481    7.09941     0.59   0.565    -10.95855    19.30551
-------------+----------------------------------------------------------------
p            |
           c |   .5475994   .8186416     0.67   0.514    -1.197294    2.292493
           t |   1.368313   4.765419     0.29   0.778    -8.788936    11.52556
       _cons |   .8683814   8.488392     0.10   0.920     -17.2242    18.96096
------------------------------------------------------------------------------
Endogenous variables:  q p
Exogenous variables:   c t
------------------------------------------------------------------------------
```

Using *ivregress*

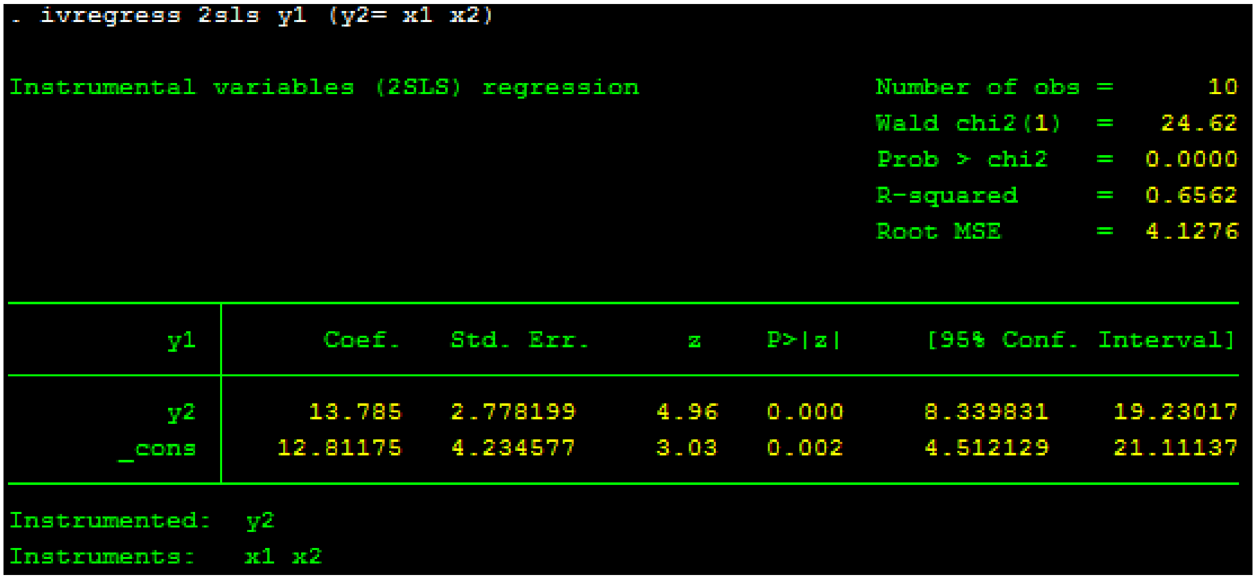
In some cases ivregress can be used but prefer reg3.

For example I

In the command window (for the model in the first example) type:

ivregress 2sls y1 (y2= x1 x2)

Ivregress 2sls dependent variable (instrumented variable=instruments)



```
. ivregress 2sls y1 (y2= x1 x2)

Instrumental variables (2SLS) regression               Number of obs =        10
                                                       Wald chi2(1)  =     24.62
                                                       Prob > chi2   =    0.0000
                                                       R-squared     =    0.6562
                                                       Root MSE      =    4.1276

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          y2 |     13.785   2.778199     4.96   0.000     8.339831    19.23017
       _cons |   12.81175   4.234577     3.03   0.002     4.512129    21.11137
------------------------------------------------------------------------------
Instrumented:  y2
Instruments:   x1 x2
```

We can use the option ‘first’ to see the stage I results

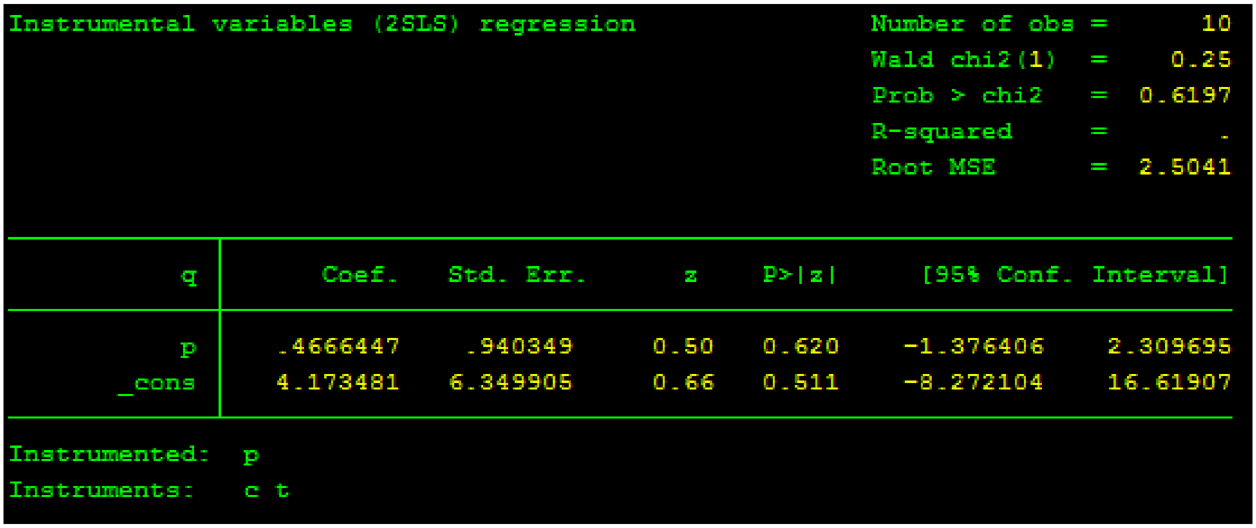
ivregress 2sls y1 (y2= x1 x2), first

For example II

In the command window (for the model in the first example) type:

ivregress 2sls q (p= c t)

iIvregress 2sls dependent variable (instrumented variable=instruments)



```
Instrumental variables (2SLS) regression               Number of obs =      10
                                                       Wald chi2(1)  =    0.25
                                                       Prob > chi2   =  0.6197
                                                       R-squared     =       .
                                                       Root MSE      =  2.5041

------------------------------------------------------------------------------
           q |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           p |   .4666447    .940349     0.50   0.620    -1.376406    2.309695
       _cons |   4.173481   6.349905     0.66   0.511    -8.272104    16.61907
------------------------------------------------------------------------------
Instrumented:  p
Instruments:   c t
```

We can use the option ‘first’ to see the stage I results: *ivregress 2sls q (p= c t), first*

Lecture 34
2SLS & 3SLS Models

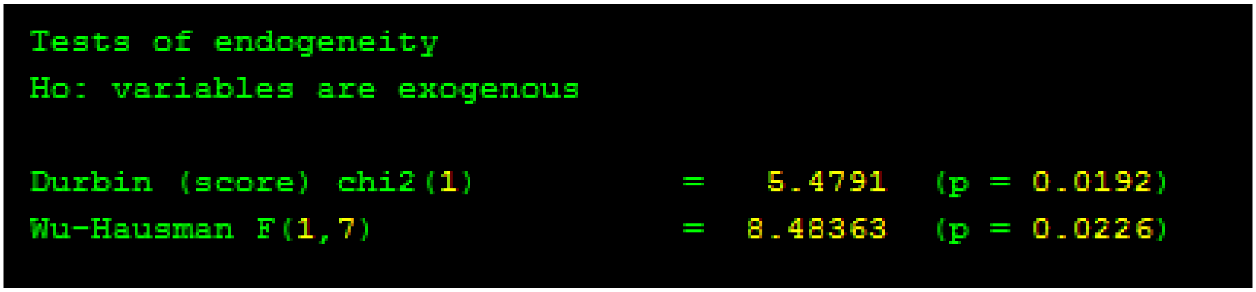
Why to use 2SLS?

OLS has only one endogenous variable (dependent). When we have system of equations, we may have several endogenous variables. When a variable is endogenous, it may be correlated with the disturbance term (biased OLS). 2sls goal is to find a proxy of the endogenous variable that is not correlated to e. (like $\hat{P}$). $\hat{P}$ should not be correlated to e_1.

Testing Endogeneity: Durban and Wu-Hausman Tests

After ivregress use the following post-estimation command:

estat endogeneity



```
Tests of endogeneity
Ho: variables are exogenous

Durbin (score) chi2(1)          =   5.4791   (p = 0.0192)
Wu-Hausman F(1,7)               =  8.48363   (p = 0.0226)
```

As p-values are less than 5%, we can reject H_0 of exogenous regressors and conclude that C and T can be used as endogenous variables.

Proof: 2SLS == ILS for exactly identified equations

consider the model

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + e_1$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_1 + e_2$$

The reduced form equations are

$$Y_1 = \pi_{10} + \pi_{11} X_1 + u_1$$

$$Y_2 = \pi_{20} + \pi_{21} X_1 + u_2$$

Where

$$\boldsymbol{\pi_{11}} = \frac{\boldsymbol{\alpha_1\beta_2}}{\boldsymbol{(1-\alpha_1\beta_1)}}, \qquad \boldsymbol{\pi_{21}} = \frac{\boldsymbol{\beta_2}}{\boldsymbol{(1-\alpha_1\beta_1)}}$$

$$\boldsymbol{\alpha_{1ILS} = \frac{\pi_{11}}{\pi_{21}}}$$

Applying ILS

We estimate the two reduced forms by OLS

$$Y_1 = \pi_{10} + \pi_{11}X_1 + u_1$$

$$Y_2 = \pi_{20} + \pi_{21}X_1 + u_2$$

Where

$$\boldsymbol{\pi_{11OLS} = \frac{\sum x_1 y_1}{\sum x_1^2}, \qquad \pi_{21OLS} = \frac{\sum x_1 y_2}{\sum x_1^2}}$$

$$\boldsymbol{\alpha_{1ILS} = \frac{\pi_{11}}{\pi_{21}} = \frac{\sum x_1 y_1}{\sum x_1 y_2}}$$

Applying 2LS

We estimate $\hat{Y}2$ and replace it in first equation

$$As\, Y_2 = \pi_{20} + \pi_{21}X_1 + u_2$$

$$\hat{Y}2 = \hat{\pi}_{21}X_1$$

Replacing in the original equation

$$Y_1 = \alpha_0 + \alpha_1\hat{Y}2 + e_1$$

$$\boldsymbol{\alpha_{1\,2SLS}} = \frac{\sum Y_{\mathbf{1}}\hat{Y}2}{\sum \hat{Y}2^{\mathbf{2}}} = \frac{\sum Y_{\mathbf{1}}\hat{\pi}_{21}X_1}{\sum(\hat{\pi}_{21}X_1)^{\mathbf{2}}}$$

$$= \frac{\hat{\pi}_{21}\sum X_1Y_{\mathbf{1}}}{(\hat{\pi}_{21})^2\sum(X_1)^{\mathbf{2}}}$$

$$= \frac{\hat{\pi}_{21}\sum X_1Y_{\mathbf{1}}}{(\hat{\pi}_{21})^2\sum(X_1)^{\mathbf{2}}}$$

$$= \frac{\sum X_1Y_{\mathbf{1}}}{\hat{\pi}_{21}\sum(X_1)^{\mathbf{2}}} = \frac{\sum(X_1)^{\mathbf{2}}}{\sum X_1Y_{\mathbf{2}}}\frac{\sum X_1Y_{\mathbf{1}}}{\sum(X_1)^{\mathbf{2}}}$$

$$\boldsymbol{\alpha_{1\,2SLS}} = \frac{\sum X_1Y_{\mathbf{1}}}{\sum X_1Y_{\mathbf{2}}} = \boldsymbol{\alpha_{1\,ILS}}$$

Numerical Example

Consider the model

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + e_1$$

$$Y_2 = \beta_{0+}\beta_1 Y_1 + \beta_2 X_1 + e_2$$

The reduced form equations are

$$Y_1 = \pi_{10} + \pi_{11} X_1 + u_1$$

$$Y_2 = \pi_{20} + \pi_{21} X_1 + u_2$$

We can estimate the above equations using data in ILS-2SLS.xlsx

Y_1	Y_2	X_1
25	0.8	10
25	0.9	8
27	0.8	8
29	1.2	7
32	1.2	6
32	1.6	3
33	1.9	4
36	2.1	5
40	2	4
49	2	2

Here $\boldsymbol{\pi_{11}} = \frac{\boldsymbol{\alpha_1\beta_2}}{\boldsymbol{(1-\alpha_1\beta_1)}}, \quad \boldsymbol{\pi_{21}} = \frac{\boldsymbol{\beta_2}}{\boldsymbol{(1-\alpha_1\beta_1)}}$

Using slope function in Excel

$$\boldsymbol{\pi_{11} = -2.45439}$$

$$\boldsymbol{\pi_{11} = -0.18158}$$

$$\boldsymbol{\alpha_{1ILS} = \frac{\pi_{11}}{\pi_{21}} = 13.51659}$$

Now 2SLS:

Y2 hat
0.669191
1.032358
1.032358
1.213941
1.395525
1.940275
1.758692
1.577108
1.758692
2.121859

Stage I: Y2 on Xi gives Y2hat

Stage II: replace y2hat in first equation

Estimate y1 on y2 haat

$$\boldsymbol{\alpha_{12LS} = 13.51659 = \alpha_{1ILS}}$$

Hence ILS and 2SLS provide identical results if applied to exactly identified equations.

Three Stage Least Square

Background

- OLS: inconsistent for systems with simultaneous equations
- 2SLS: ‘Cleans’ the endogenous regressors
  - I: regress endogenous variables against all predetermined variables of the system (reduced form). This gives ‘theoretical values’

  - II: Use OLS for theoretical values replacing the original values

In 2SLS

- Endogenous variable may be related to error term of another equation.

Possible Problem with 2SLS

- Focus on single equation in the system
- Correlations between error terms of various equations is ignored
- This may give inefficient estimates

Solution:

Use Three Stage Least Square (Zellner & Theil (1962))

- After 2SLS, we add a third stage to account for correlations of error terms.
- 3SLS uses the results of 2SLS to estimate 'ALL' coefficients of the system simultaneously
- 3SLS is more efficient as compared to 2SLS
- 3SLS is better when correlation of error terms is not small

3SLS using Stata

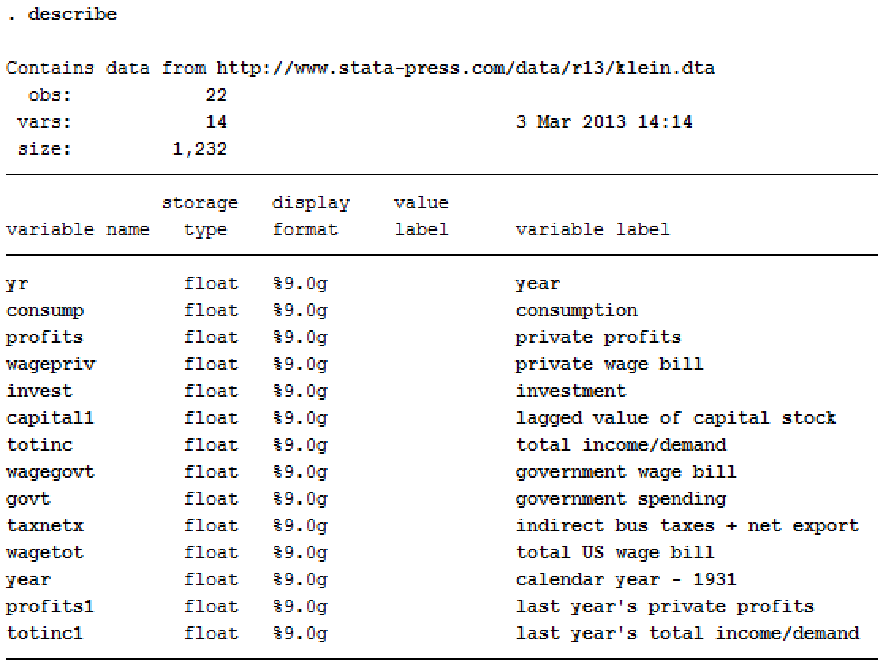
Syntax is *reg3 (depvar1 varlist1) (depvar2 varlist2) ……(depvarN varlistN) [if] [in] [weight]*

Menu: *Statistics > Endogenous Covariates > Three Stage Least Square*

EXAMPLE:

webuse klein

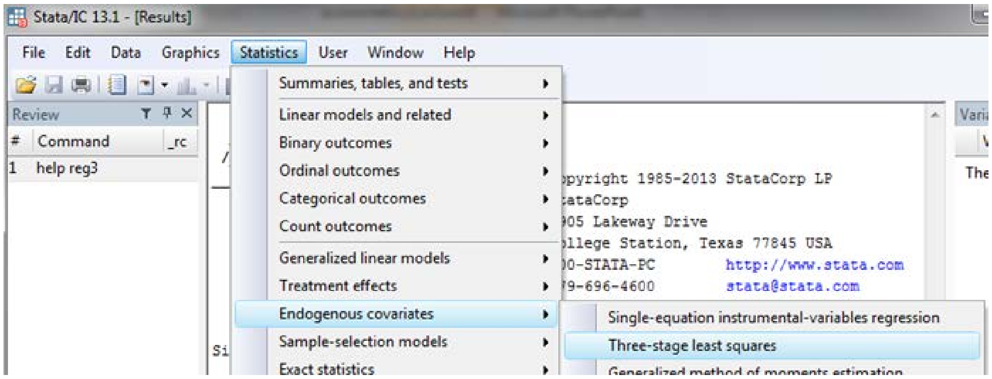
describe



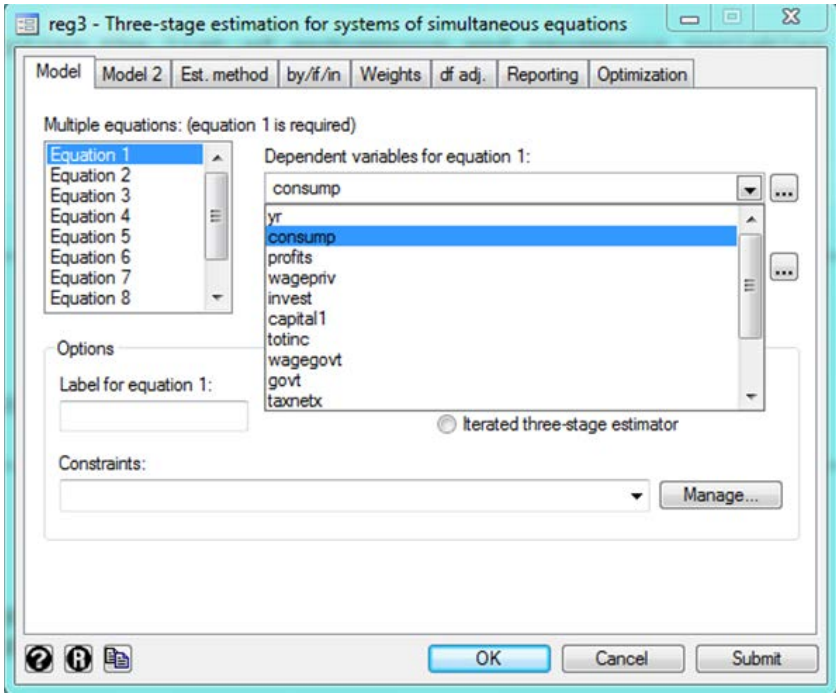
```
. describe

Contains data from http://www.stata-press.com/data/r13/klein.dta
  obs:            22
 vars:            14                          3 Mar 2013 14:14
 size:         1,232
-----------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-----------------------------------------------------------------------------
yr              float   %9.0g                 year
consump         float   %9.0g                 consumption
profits         float   %9.0g                 private profits
wagepriv        float   %9.0g                 private wage bill
invest          float   %9.0g                 investment
capital1        float   %9.0g                 lagged value of capital stock
totinc          float   %9.0g                 total income/demand
wagegovt        float   %9.0g                 government wage bill
govt            float   %9.0g                 government spending
taxnetx         float   %9.0g                 indirect bus taxes + net export
wagetot         float   %9.0g                 total US wage bill
year            float   %9.0g                 calendar year - 1931
profits1        float   %9.0g                 last year's private profits
totinc1         float   %9.0g                 last year's total income/demand
-----------------------------------------------------------------------------
```

EQ1: consump wagepriv wagegovt

EQ2: wagepriv consump govt capital1

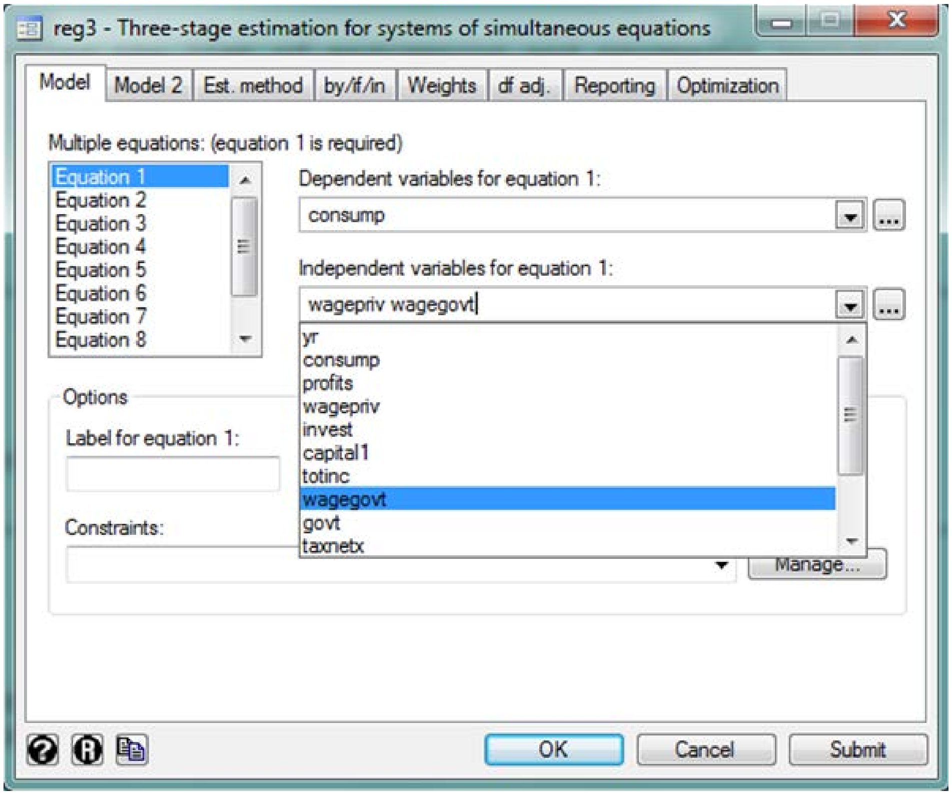


```
pyright 1985-2013 StataCorp LP
tataCorp
905 Lakeway Drive
ollege Station, Texas 77845 USA
00-STATA-PC          http://www.stata.com
79-696-4600          stata@stata.com
```



Dependent variables for equation 1: consump

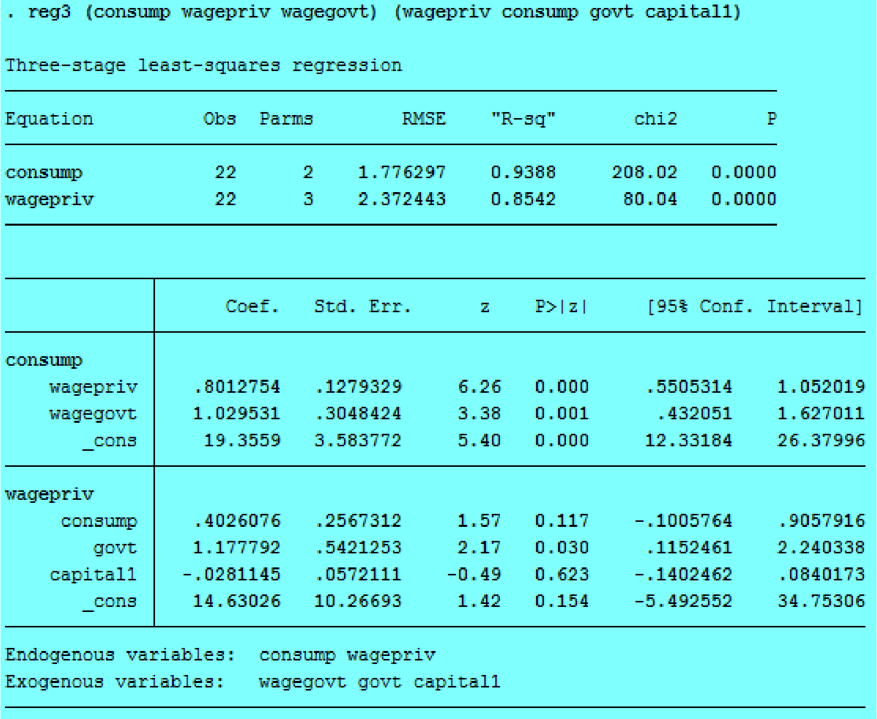
Select independent variable:



Dependent variables for equation 1: consump

Independent variables for equation 1: wagepriv wagegovt

Similarly now click on equation 2 and select variables. We can use various options. Click OK



```
. reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1)

Three-stage least-squares regression
----------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2        P
----------------------------------------------------------------------
consump            22      2    1.776297    0.9388     208.02   0.0000
wagepriv           22      3    2.372443    0.8542      80.04   0.0000
----------------------------------------------------------------------


------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
consump      |
    wagepriv |   .8012754   .1279329     6.26   0.000     .5505314    1.052019
    wagegovt |   1.029531   .3048424     3.38   0.001      .432051    1.627011
       _cons |    19.3559   3.583772     5.40   0.000     12.33184    26.37996
-------------+----------------------------------------------------------------
wagepriv     |
     consump |   .4026076   .2567312     1.57   0.117    -.1005764    .9057916
        govt |   1.177792   .5421253     2.17   0.030     .1152461    2.240338
    capital1 |  -.0281145   .0572111    -0.49   0.623    -.1402462    .0840173
       _cons |   14.63026   10.26693     1.42   0.154    -5.492552    34.75306
------------------------------------------------------------------------------
Endogenous variables:  consump wagepriv
Exogenous variables:   wagegovt govt capital1
------------------------------------------------------------------------------
```

Only important options of this stata procedure as an example

noconst

- omits constant term for an equation if specified in an equation
- omits constant term from instrument list (stage I) if specified overall

ireg3

- iterates over the estimated disturbance covariance matrix and parameter estimates until the parameter estimates converge

Sure

- performs a seemingly unrelated regression estimation of the system -even if dependent variables from some equations appear as regressors in other equations.

2sls

- performs equation-by-equation two stage least squares on the full system of equations.

first

- Requests display of first stage regression

Corr (Correlations)

- specifies the assumed form of the correlation structure of the equation disturbances; rarely requested

small

- Small sample statistics are also computed.

3sls

- Performs three stage least square; default

Lecture 35
Panel Data Methods

Types of Econometric Data

Time Series Data: different points of time

Cross Sectional Data: different entities

Pooled Data: mixture of time series and cross sectional data;

Panel /longitudinal Data/Repeated Cross Sections: type of panel data; same cross sectional entities over time

Definition of **Pooled Data**: Randomly sampled cross sections of individuals at different points in time

Example of Pooled Data

Respondent	**Year**	**Income (Rs.)**	**Expenditure (Rs.)**
Ali	2012	12000	9500
Arif	2012	15000	11500
Sakina	2012	11500	9000
Ali	2013	12500	10000
Jameel	2013	11000	.
Fatima	2013	14000	12500
Baqir	2013	9500	8500

Individuals may or may not be repeated

Different number of observations in different time periods

There may be missing value

Examples of pooled data include Women's Fertility over time, Population Survey, Labor Force Survey, Marketing Surveys, Consumer Surveys.

Definition of **Panel Data**: Observe cross sections of the same individuals at different points in time

Examples:

- Panel Survey of Income Dynamics (PSID)

- British Household Panel Survey (BHPS)
- US Time Use Surveys
- French Time Use Surveys
- German Socio-Economic Panel
- National Longitudinal Surveys (NLS)

Panel data is also a pooled data set. All pooled Data should not be called panels

Panel Dataset

Country	**Year**	**Y**	**X**
1	2012	6	9
1	2012	3	8
1	2012	5	11
2	2013	9	12
2	2013	5	7
2	2013	6	14
3	2014	11	14
3	2014	6	5
3	2014	8	16

Panel Data shows the behavior of individual entities across time

The same set of individuals are normally observed

There may be missing values

Types:

- Short and Long Panel
  - Short: Many entities, few time periods

- Balanced and Unbalanced Panel
  - Balanced: all entities have measurements in all time periods (no of observations is nT)
- Fixed and Rotating Panel
  - Fixed: individuals remain the same

Rotating panel may also be called Pseudo Panels

- Compact Panel: consecutive time periods

Panel Data benefits

- More degree of freedom (nT)
- Captures more complexities (blend both cross sectional and time series behavior)
- Better forecasts
- We can study heterogeneity and avoid omitted variable bias

Panel Data: Drawbacks

- Data collection issues
- Sampling design and coverage
- No response cases in micro panels
- Cross country correlation in case of macro panels (e.g. location or spacial correlation, gravity models)
- Attrition: dropping out of individuals leads to unbalance and/or uncompact panel

Rationale for using Panel data

- Unobserved Heterogeneity
  - Many individual characteristics are not observed; examples are: Risk taking behavior, Ability

They vary across individuals and are called unobserved heterogeneity. If they influence the response variable and are correlated to regressors, OLS results will be biased.

Examples of Unobserved Heterogeneity

- Returns to Education
    - Ability is not observed
    - Returns may be influenced due to ability
- Discrimination: Race, Gender, Religion
- Unobserved characteristics of groups
    - Attitude to risk
    - Social behavior
    - Working habits
- Macro Panels (e.g.country GDP)
- Countries have some unobserved characteristics

Pooled Regression

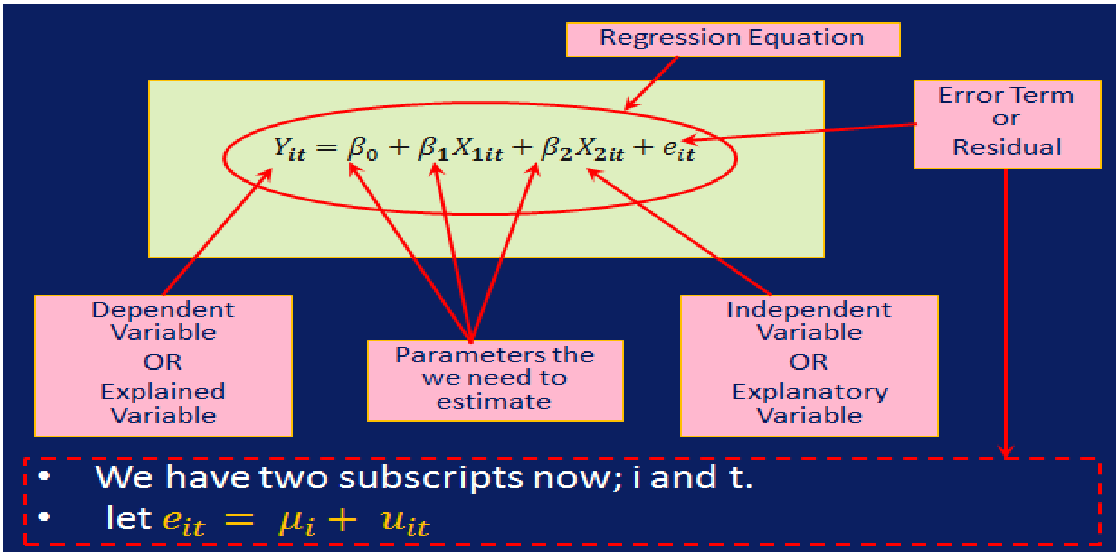
If we ignore the panel structure of the model and just run a regression, it is called a pooled regression or regression on pooled data. We have to assume that the error is

$$e_{it} \sim N(0,\ \sigma^2)$$

which may not be the case. We can pool the data with or without constraints on the disturbance term. We can use the Chow test (discussed with stata)

Least Square Dummy Variable Model

Dummy Variables can be used for different individual or time entities or both. Regression could be run using the dummy variables on the RHS of the regression equation. We can capture different intercepts for different groups. Problems include lack of degree of freedom or dummy variable trap.



Regression Equation

$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + e_{it}$

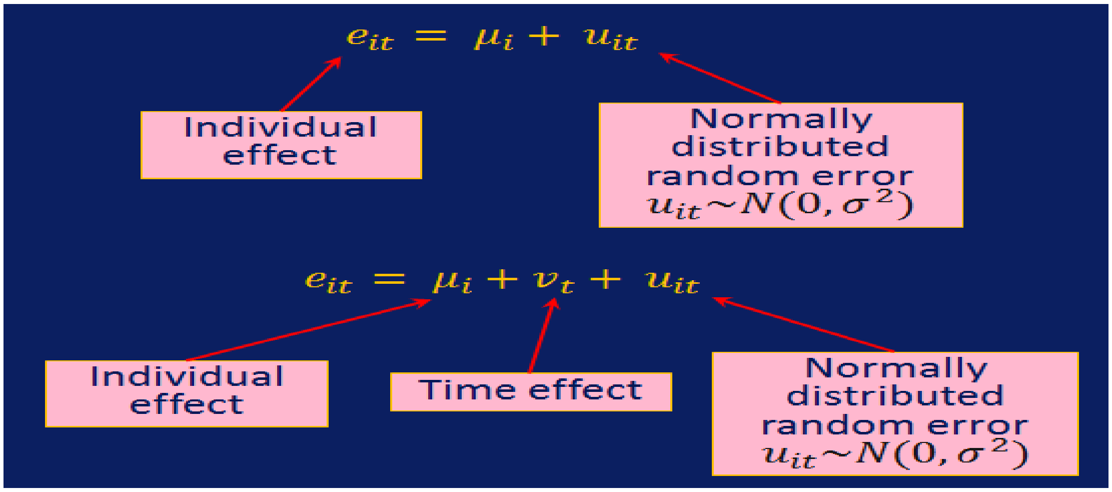
Error Term or Residual

Dependent Variable OR Explained Variable

Parameters the we need to estimate

Independent Variable OR Explanatory Variable

- We have two subscripts now; i and t.
- let $e_{it} = \mu_i + u_{it}$



$e_{it} = \mu_i + u_{it}$

Individual effect

Normally distributed random error $u_{it} \sim N(0, \sigma^2)$

$e_{it} = \mu_i + v_t + u_{it}$

Individual effect

Time effect

Normally distributed random error $u_{it} \sim N(0, \sigma^2)$

Error component: one way and two way

Consider the model with $e_{it} = \mu_i + u_{it}$

Fixed Effect: when we assume μ_i to be constant (for one individual, different for different individuals)

Random Effect: when we assume the μ_i is drawn independently from some probability distribution

Lecture 36
Panel Data Models-II

Fixed Effect Model

Consider the model

$$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + e_{it}}$$

with $e_{it} = \mu_i + u_{it}$

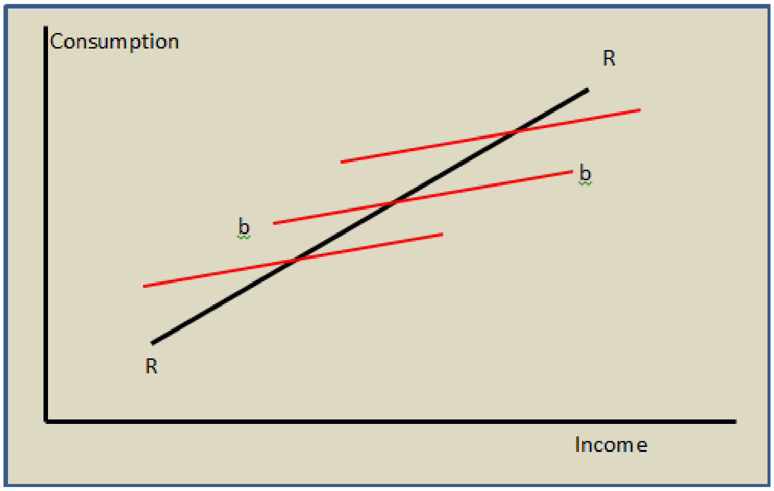
Then

$$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it}} + \mu_i + u_{it}$$

so

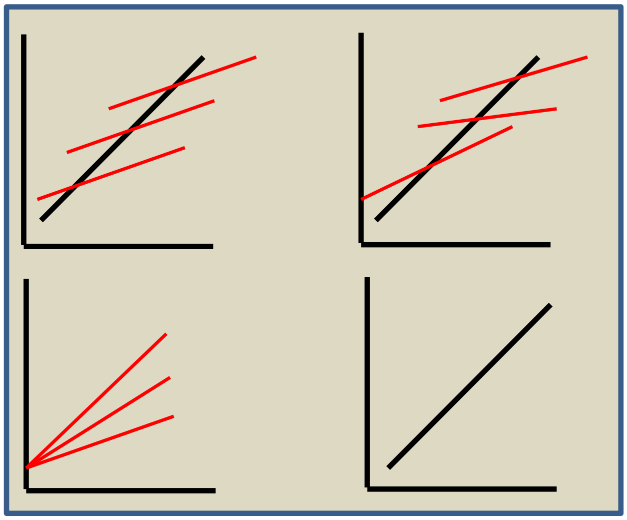
$$\boldsymbol{Y_{it}} = (\boldsymbol{\beta_0} + \mu_i) + \boldsymbol{\beta_1 X_{1it} + \beta_2 X_{2it}} + u_{it}$$

- μ_i is now a part of constant term but different for different individuals
- Individuals have different intercepts but common slope



Above diagrams shows three different classes with three different intercepts but a common slope as is the concept in fixed effect model.

Possible regressions with panel data are:



Fixed Effect Model Estimation: First Difference

Eliminating unobserved heterogeneity by first differencing

$$\boldsymbol{Y_{it}} = (\boldsymbol{\beta_0} + \mu_i) + \boldsymbol{\beta_1 X_{1it}} + \boldsymbol{\beta_2 X_{2it}} + u_{it}$$

With one time lag

$$\boldsymbol{Y_{it-1}} = (\boldsymbol{\beta_0} + \mu_i) + \boldsymbol{\beta_1 X_{1it-1}} + \boldsymbol{\beta_2 X_{2it-1}} + u_{it-1}$$

$$\boldsymbol{Y_{it}} - \boldsymbol{Y_{it-1}} = \boldsymbol{\beta_1}(\boldsymbol{X_{1it}} - \boldsymbol{X_{1it-1}}) + \boldsymbol{\beta_2}(\boldsymbol{X_{2it}} - \boldsymbol{X_{2it-1}}) + (u_{it} - u_{it-1})$$

$$\Delta \boldsymbol{Y_{it}} = \boldsymbol{\beta_1} \Delta \boldsymbol{X_{1it}} + \boldsymbol{\beta_2} \Delta \boldsymbol{X_{2it}} + \Delta u_{it}$$

- Which can be generalized for more variables
- Works fine for two time periods

Fixed Effect Model Estimation: Deviation from means (alternative approach)

We can use the deviations from means

$$\boldsymbol{Y_{it}} = (\boldsymbol{\beta_0} + \mu_i) + \boldsymbol{\beta_1 X_{1it}} + \boldsymbol{\beta_2 X_{2it}} + u_{it}$$

Taking deviations from mean, intercept will be eliminated

$$Y_{it-1} - \overline{Y_{\iota.}} = \beta_1(X_{1it} - \overline{X_{1\iota.}}) + \beta_2(X_{2it} - \overline{X_{2\iota.}}) + u_{it}$$

Constant part is eliminated again. Deviations from mean for each individual, averaged across all time periods. Model can be estimated by OLS, LSDV etc. This is called '*within*' estimator. This previous one is called '*within*' estimator. The 'between' estimator can be found by using the *individual averages*. The 'overall estimator' is the weighted average of within and between estimators.

Random Effect Model

If *correct weights* are used, combination of fixed (within) and between effect estimator is called random effect estimator

$$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + e_{it}}$$

$$e_{it} = \mu_i + u_{it}$$

when we assume the μ_i is drawn independently from some probability distribution we are dealing with the random effect model. When unobserved heterogeneity is uncorrelated with regressors, panel data techniques are not needed to produce a consistent estimator. We must correct for serial correlation between observations of the same entity (individual, firm, country etc.). If $E(Xit,\ eit) \neq 0$, Fixed Effect model may be used When $E(Xit,\ eit) = 0$, Random Effect model may be usedto overcome the serial correlation of panel data.

To estimate the random effect model, we assume that μ_i is part of the error term e_{it}. We evaluate the structure of the error and apply appropriate Generalized Least Square (with correct weights) to calculate efficient estimators. The following assumptions must hold:

$$E(\mu_i) = E(u_{it}) = 0$$

$E(\boldsymbol{u_{it}^2}) = \sigma_{\boldsymbol{u}}^{\boldsymbol{2}}$ **&** $E(\mu_{\boldsymbol{it}}^{\boldsymbol{2}}) = \sigma_\mu^{\boldsymbol{2}}$

$$E(u_{it}, \mu_i) = 0\ for\ all\ i\ and\ t$$

$$E(e_{it}^{\boldsymbol{2}}) = \sigma_{\boldsymbol{u}}^{\boldsymbol{2}} + \sigma_\mu^{\boldsymbol{2}}$$

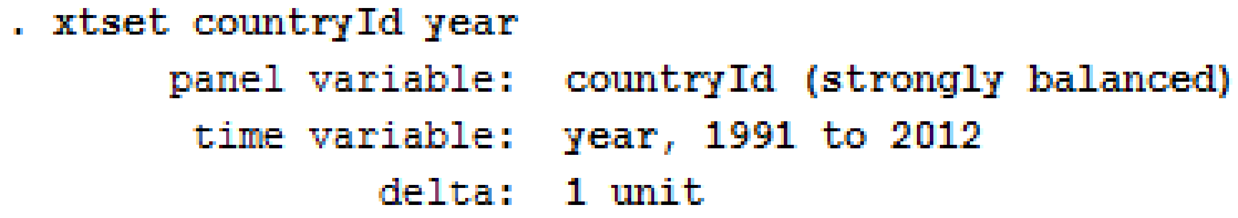
$$E(X_{kit}, \mu_i) = 0\ for\ all\ k,\ i\ and\ t$$

Panel Data Models using Stata

All panel data command will start with xt. We need to set the panel data first: xtset. If needed, we need to reshape our panel. We will use a file panel1.dta available with lecture notes. The following example uses data from world Development Indicators. Pakistan, India, SriLanka are included and data is from 1991 to 2012 which is a strongly balanced long panel.

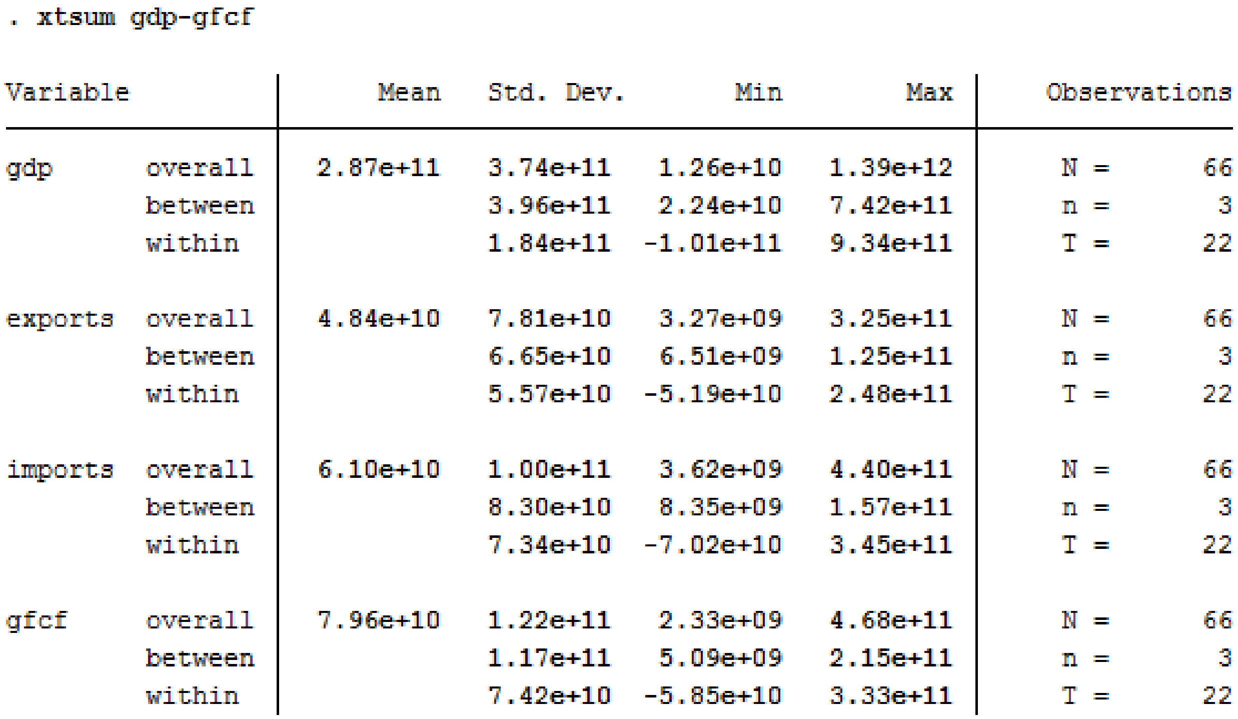
xtset command is used to set the data as panel data where both variables of cross sectional and time should be numeric. Country is a string variable, We need to create a numeric ID. We use the command

encode country, generate(countryId) to create a numeric variable.



```
. xtset countryId year
       panel variable:  countryId (strongly balanced)
        time variable:  year, 1991 to 2012
                delta:  1 unit
```

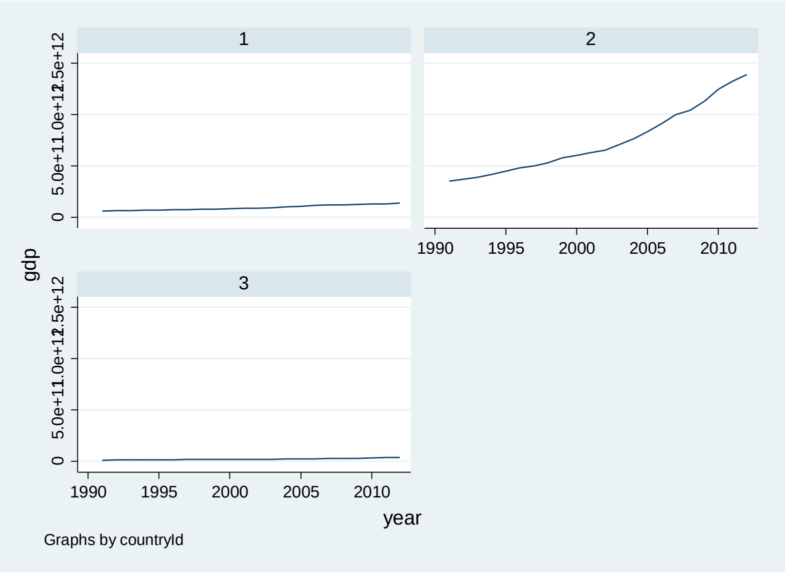
xtsummarize provides summary statistics



```
. xtsum gdp-gfcf

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
gdp      overall |  2.87e+11    3.74e+11   1.26e+10   1.39e+12 |     N =      66
         between |              3.96e+11   2.24e+10   7.42e+11 |     n =       3
         within  |              1.84e+11  -1.01e+11   9.34e+11 |     T =      22
                 |                                             |
exports  overall |  4.84e+10    7.81e+10   3.27e+09   3.25e+11 |     N =      66
         between |              6.65e+10   6.51e+09   1.25e+11 |     n =       3
         within  |              5.57e+10  -5.19e+10   2.48e+11 |     T =      22
                 |                                             |
imports  overall |  6.10e+10    1.00e+11   3.62e+09   4.40e+11 |     N =      66
         between |              8.30e+10   8.35e+09   1.57e+11 |     n =       3
         within  |              7.34e+10  -7.02e+10   3.45e+11 |     T =      22
                 |                                             |
gfcf     overall |  7.96e+10    1.22e+11   2.33e+09   4.68e+11 |     N =      66
         between |              1.17e+11   5.09e+09   2.15e+11 |     n =       3
         within  |              7.42e+10  -5.85e+10   3.33e+11 |     T =      22
```

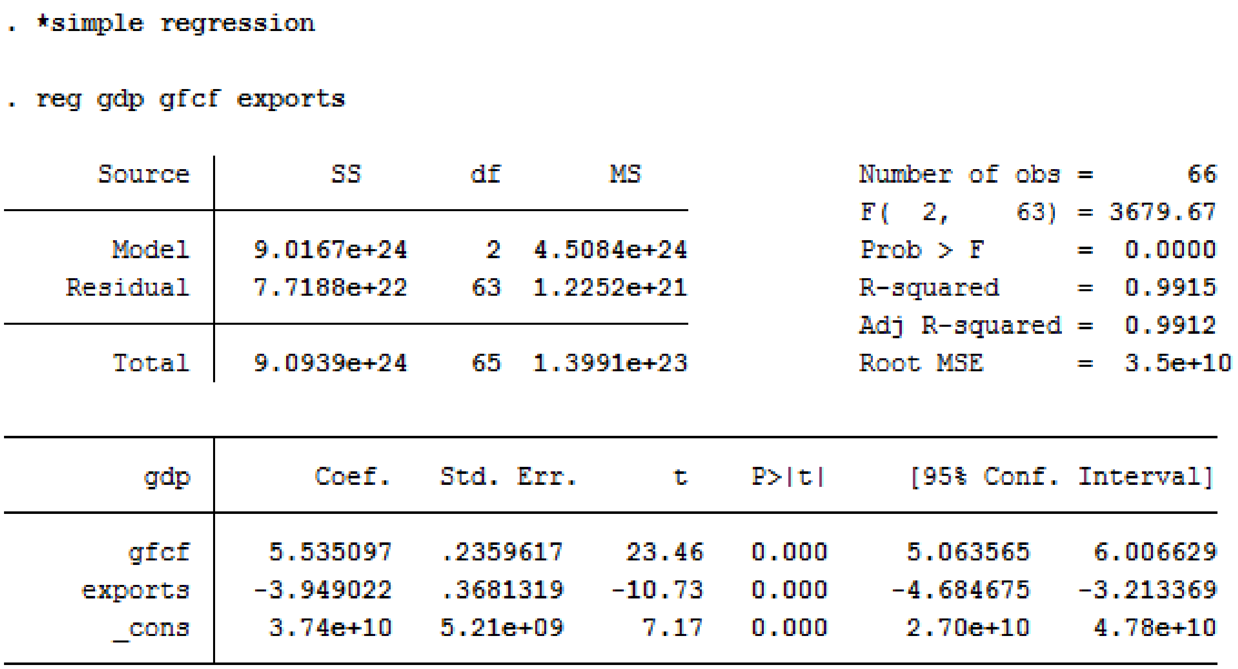
Let us think that gdp is a function of gfcf and exports and imports. Look at the dependent variable by country.

xtline gdp



Use fixed-effects (FE) whenever you are only interested in analyzing the impact of variables that vary over time. FE explores the relationship between predictor and outcome variables within an entity (country, person, company, etc.). Each entity has its own individual characteristics that may or may not influence the predictor variables (for example being a male or female could influence the opinion toward certain issue or the political system of a particular country could have some effect on trade or GDP or the business practices of a company may influence its stock price).

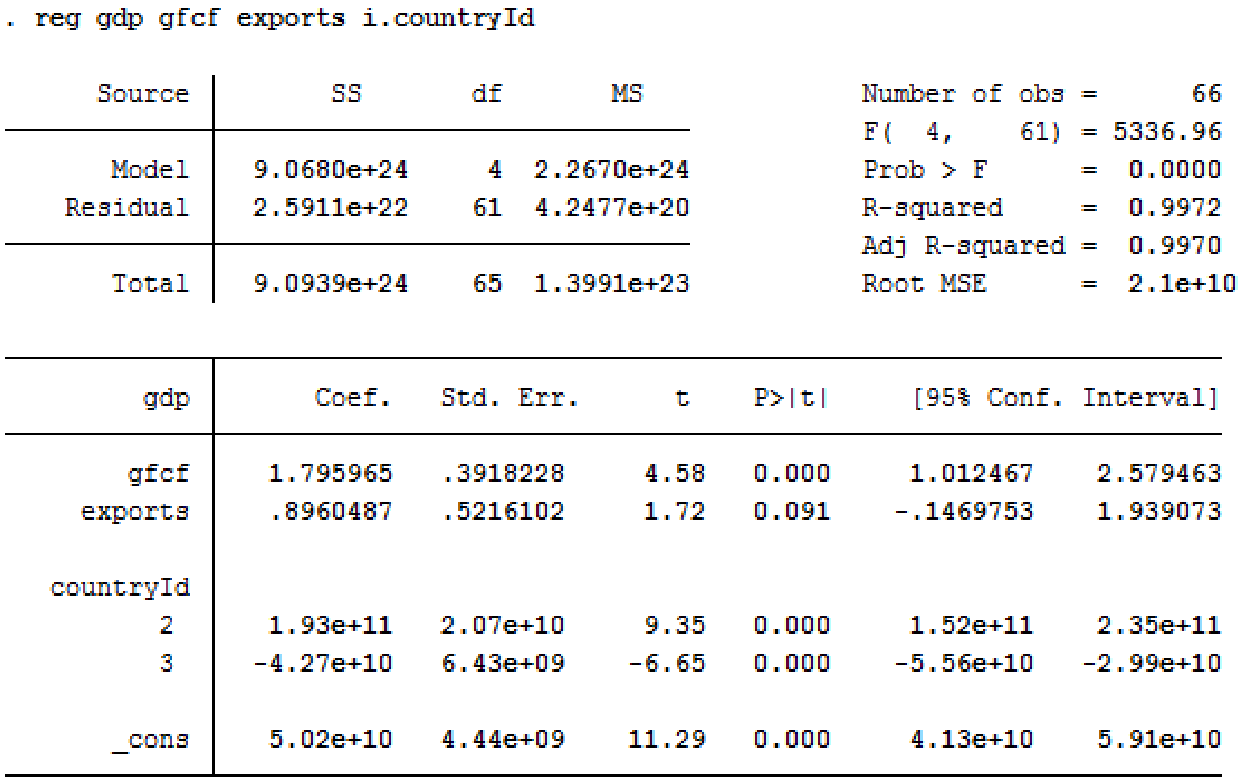
Simple OLS may provide results like this:



```
. *simple regression

. reg gdp gfcf exports

      Source |       SS       df       MS              Number of obs =      66
-------------+------------------------------           F(  2,    63) = 3679.67
       Model |  9.0167e+24     2  4.5084e+24           Prob > F      =  0.0000
    Residual |  7.7188e+22    63  1.2252e+21           R-squared     =  0.9915
-------------+------------------------------           Adj R-squared =  0.9912
       Total |  9.0939e+24    65  1.3991e+23           Root MSE      =  3.5e+10

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gfcf |   5.535097   .2359617    23.46   0.000     5.063565    6.006629
     exports |  -3.949022   .3681319   -10.73   0.000    -4.684675   -3.213369
       _cons |   3.74e+10   5.21e+09     7.17   0.000     2.70e+10    4.78e+10
------------------------------------------------------------------------------
```

When using FE we assume that something within the individual may impact or bias the predictor or outcome variables and we need to control for this. This is the rationale behind the assumption of the correlation between entity's error term and predictor variables. FE removes the effect of those time-invariant characteristics from the predictor variables so we can assess the predictors' net effect.

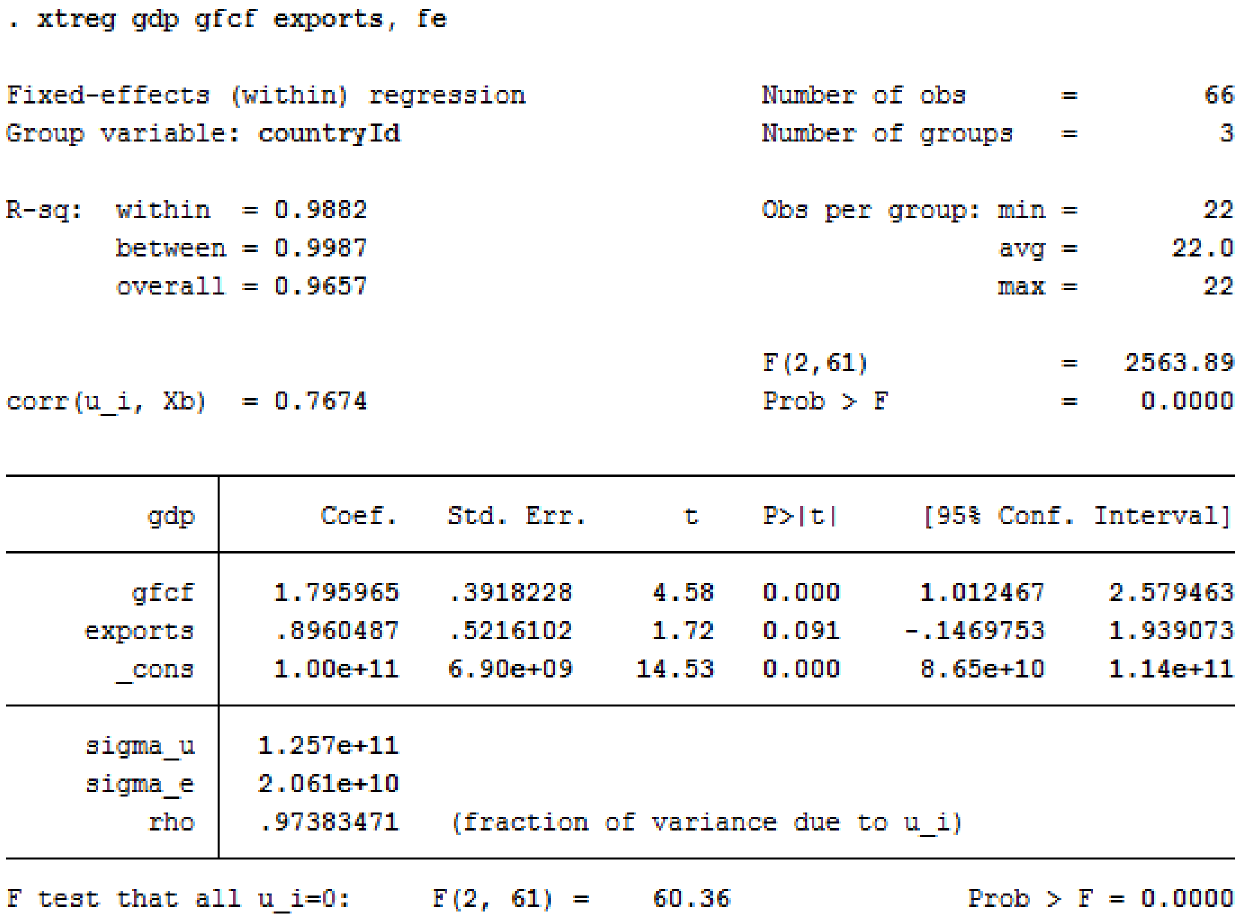
Least Square dummy variable model may provide the following result:



```
. reg gdp gfcf exports i.countryId

      Source |       SS       df       MS              Number of obs =      66
-------------+------------------------------           F(  4,    61) = 5336.96
       Model |  9.0680e+24     4  2.2670e+24           Prob > F      =  0.0000
    Residual |  2.5911e+22    61  4.2477e+20           R-squared     =  0.9972
-------------+------------------------------           Adj R-squared =  0.9970
       Total |  9.0939e+24    65  1.3991e+23           Root MSE      =  2.1e+10

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gfcf |   1.795965   .3918228     4.58   0.000     1.012467    2.579463
     exports |   .8960487   .5216102     1.72   0.091    -.1469753    1.939073
             |
   countryId |
          2  |   1.93e+11   2.07e+10     9.35   0.000     1.52e+11    2.35e+11
          3  |  -4.27e+10   6.43e+09    -6.65   0.000    -5.56e+10   -2.99e+10
             |
       _cons |   5.02e+10   4.44e+09    11.29   0.000     4.13e+10    5.91e+10
------------------------------------------------------------------------------
```

Another important assumption of the FE model is that those time-invariant characteristics are unique to the individual and should not be correlated with other individual characteristics. Each entity is different therefore the entity's error term and the constant (which captures individual characteristics) should not be correlated with the others. If the error terms are correlated then FE is no suitable since inferences may not be correct and you need to model that relationship (probably using random-effects), this is the main rationale for the Hausman test (presented later on in this document).

Fixed effect model may provide:



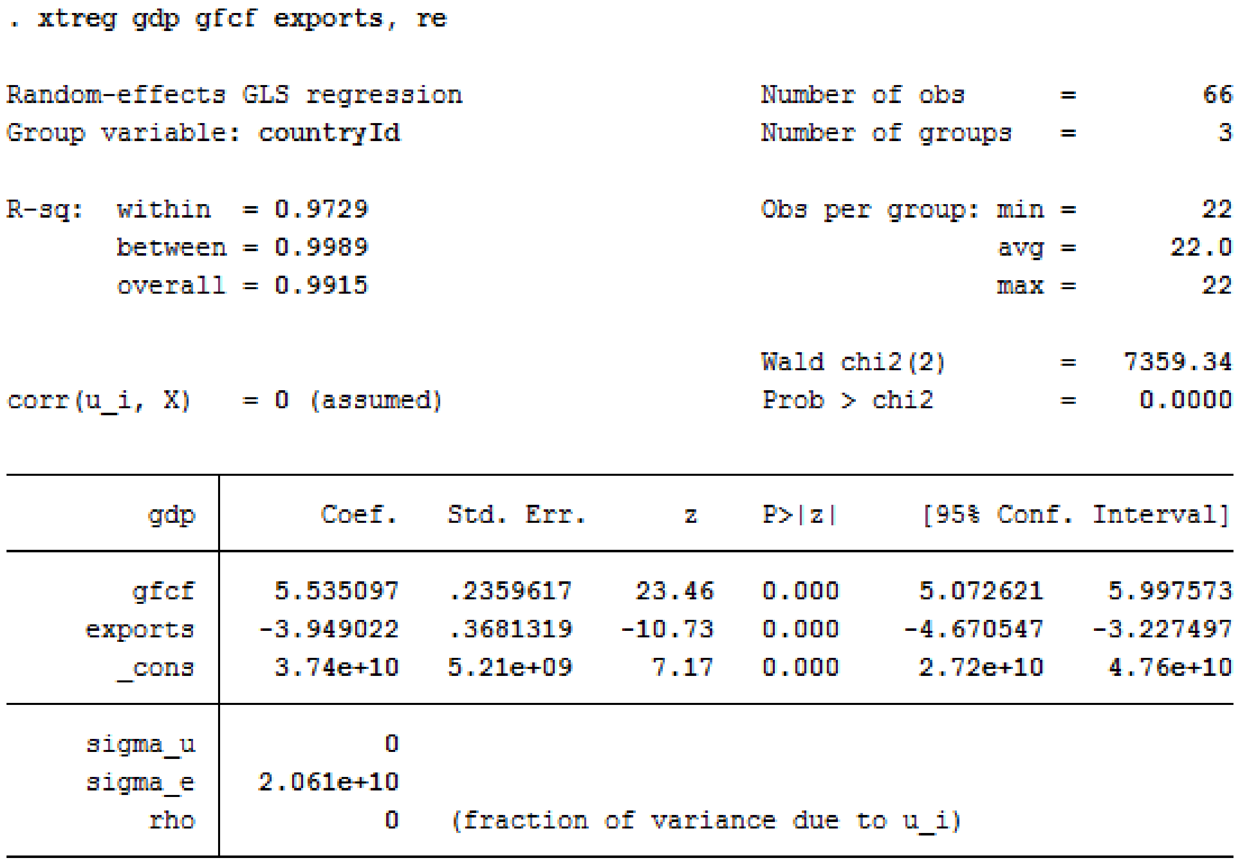
```
. xtreg gdp gfcf exports, fe

Fixed-effects (within) regression               Number of obs      =        66
Group variable: countryId                       Number of groups   =         3

R-sq:  within  = 0.9882                         Obs per group: min =        22
       between = 0.9987                                        avg =      22.0
       overall = 0.9657                                        max =        22

                                                F(2,61)            =   2563.89
corr(u_i, Xb)  = 0.7674                         Prob > F           =    0.0000

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gfcf |   1.795965   .3918228     4.58   0.000     1.012467    2.579463
     exports |   .8960487   .5216102     1.72   0.091    -.1469753    1.939073
       _cons |   1.00e+11   6.90e+09    14.53   0.000     8.65e+10    1.14e+11
-------------+----------------------------------------------------------------
     sigma_u |  1.257e+11
     sigma_e |  2.061e+10
         rho |  .97383471   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(2, 61) =    60.36               Prob > F = 0.0000
```

Random effect model

$$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + e_{it}}$$

Y = gdp, X_1 = gfcf, X_2 = exports

$$e_{it} = \mu_i + u_{it}$$



```
. xtreg gdp gfcf exports, re

Random-effects GLS regression                   Number of obs      =        66
Group variable: countryId                       Number of groups   =         3

R-sq:  within  = 0.9729                         Obs per group: min =        22
       between = 0.9989                                        avg =      22.0
       overall = 0.9915                                        max =        22

                                                Wald chi2(2)       =   7359.34
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        gfcf |   5.535097   .2359617    23.46   0.000     5.072621    5.997573
     exports |  -3.949022   .3681319   -10.73   0.000    -4.670547   -3.227497
       _cons |   3.74e+10   5.21e+09     7.17   0.000     2.72e+10    4.76e+10
-------------+----------------------------------------------------------------
     sigma_u |          0
     sigma_e |  2.061e+10
         rho |          0   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Fixed or Random? Hausman Test

Perform the following:

qui xtreg gdp gfcf exports, fe

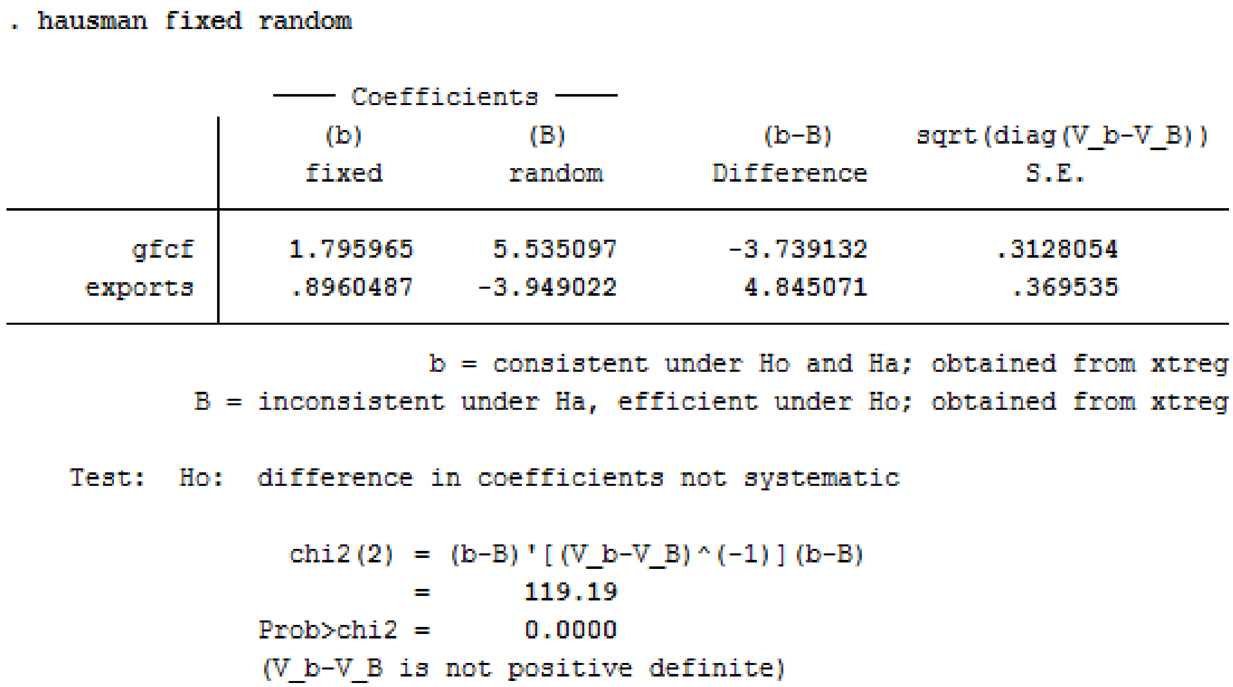
estimates store fixed

qui xtreg gdp gfcf exports, re

estimates store random

hausman fixed random

Rejecting H_0 means that fixed effects should be used



```
. hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+----------------------------------------------------------------
        gfcf |    1.795965     5.535097       -3.739132        .3128054
     exports |    .8960487    -3.949022        4.845071         .369535
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      119.19
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

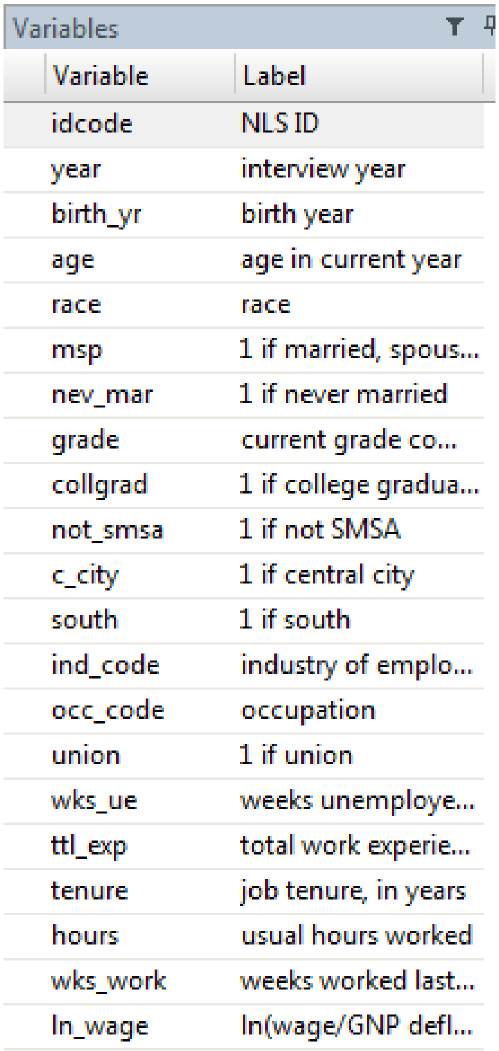
Lecture 37

Panel Data Methods & Post Estimation Tests

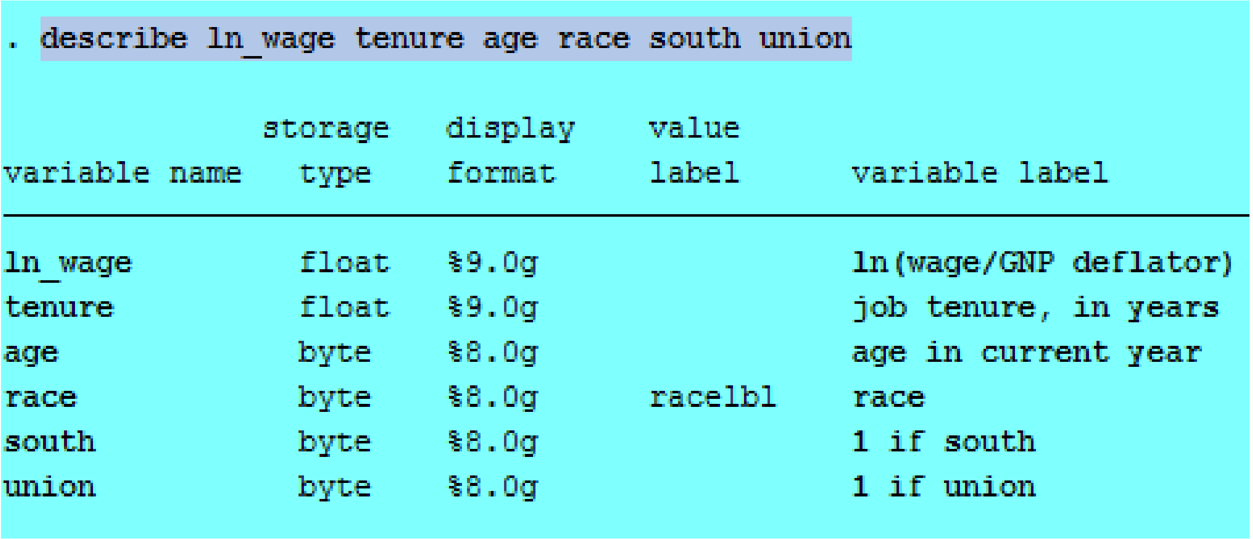
For practice download the file nlswork from the web. Perform the following operations:

webuse nlswork, clear

the variables are shown in the variable window

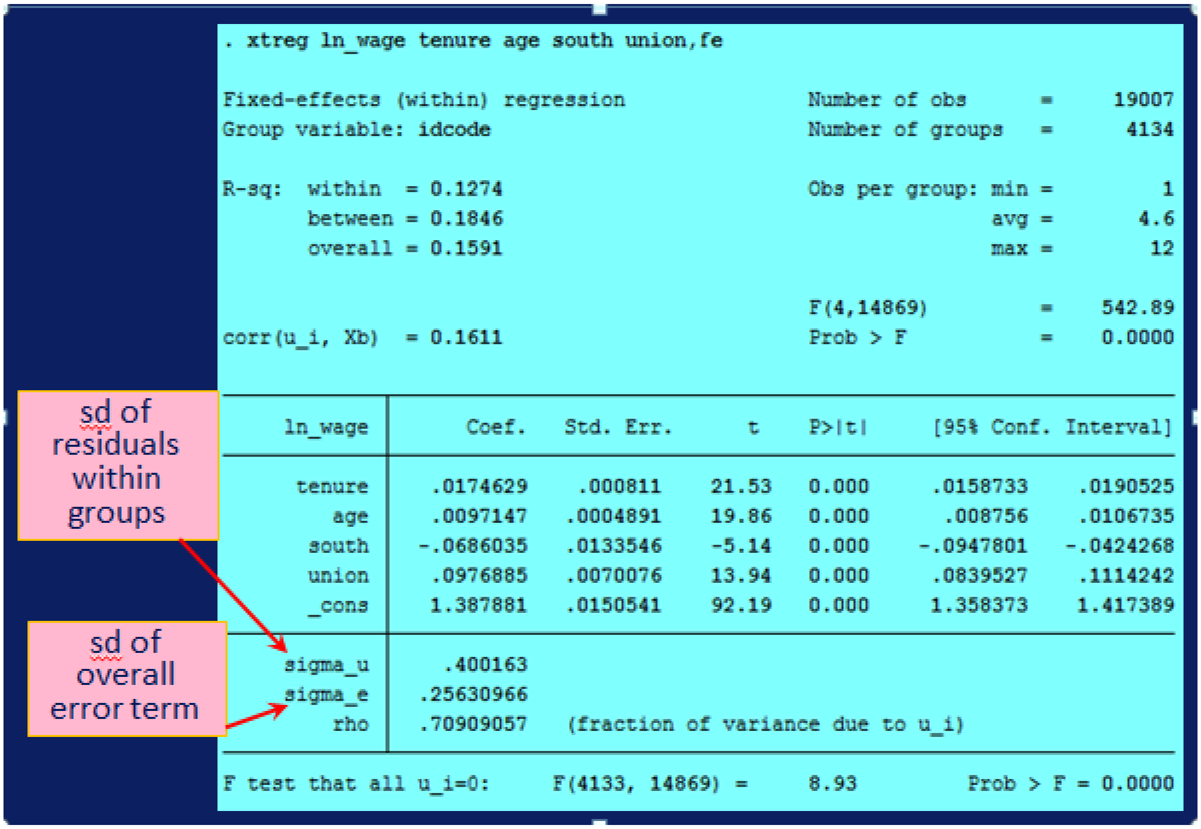


Variable	Label
idcode	NLS ID
year	interview year
birth_yr	birth year
age	age in current year
race	race
msp	1 if married, spous...
nev_mar	1 if never married
grade	current grade co...
collgrad	1 if college gradua...
not_smsa	1 if not SMSA
c_city	1 if central city
south	1 if south
ind_code	industry of emplo...
occ_code	occupation
union	1 if union
wks_ue	weeks unemploye...
ttl_exp	total work experie...
tenure	job tenure, in years
hours	usual hours worked
wks_work	weeks worked last...
ln_wage	ln(wage/GNP defl...



```
. describe ln_wage tenure age race south union

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------
ln_wage         float   %9.0g                 ln(wage/GNP deflator)
tenure          float   %9.0g                 job tenure, in years
age             byte    %8.0g                 age in current year
race            byte    %8.0g      racelbl    race
south           byte    %8.0g                 1 if south
union           byte    %8.0g                 1 if union
```

Fixed Effect Estimation

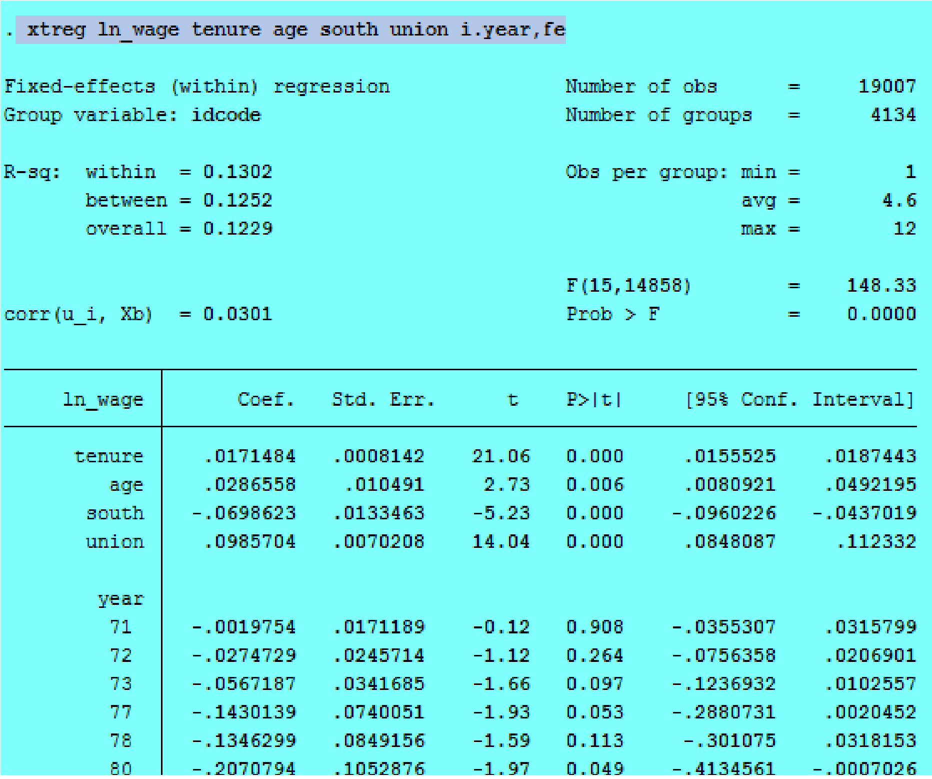


```
. xtreg ln_wage tenure age south union,fe

Fixed-effects (within) regression               Number of obs      =     19007
Group variable: idcode                          Number of groups   =      4134

R-sq:  within  = 0.1274                         Obs per group: min =         1
       between = 0.1846                                        avg =       4.6
       overall = 0.1591                                        max =        12

                                                F(4,14869)         =    542.89
corr(u_i, Xb)  = 0.1611                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .0174629    .000811    21.53   0.000     .0158733    .0190525
         age |   .0097147   .0004891    19.86   0.000      .008756    .0106735
       south |  -.0686035   .0133546    -5.14   0.000    -.0947801   -.0424268
       union |   .0976885   .0070076    13.94   0.000     .0839527    .1114242
       _cons |   1.387881   .0150541    92.19   0.000     1.358373    1.417389
-------------+----------------------------------------------------------------
     sigma_u |    .400163
     sigma_e |  .25630966
         rho |  .70909057   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4133, 14869) =     8.93         Prob > F = 0.0000
```

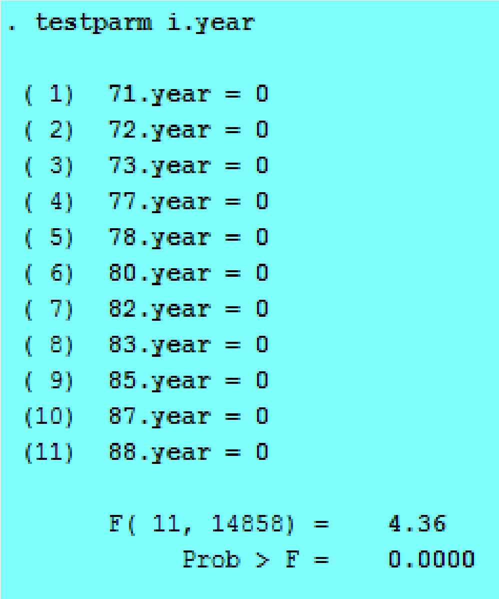
sd of residuals within groups

sd of overall error term

Testing Time Fixed Effects



```
. xtreg ln_wage tenure age south union i.year,fe

Fixed-effects (within) regression               Number of obs      =     19007
Group variable: idcode                          Number of groups   =      4134

R-sq:  within  = 0.1302                         Obs per group: min =         1
       between = 0.1252                                        avg =       4.6
       overall = 0.1229                                        max =        12

                                                F(15,14858)        =    148.33
corr(u_i, Xb)  = 0.0301                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .0171484   .0008142    21.06   0.000     .0155525    .0187443
         age |   .0286558    .010491     2.73   0.006     .0080921    .0492195
       south |  -.0698623   .0133463    -5.23   0.000    -.0960226   -.0437019
       union |   .0985704   .0070208    14.04   0.000     .0848087     .112332
             |
        year |
         71  |  -.0019754   .0171189    -0.12   0.908    -.0355307    .0315799
         72  |  -.0274729   .0245714    -1.12   0.264    -.0756358    .0206901
         73  |  -.0567187   .0341685    -1.66   0.097    -.1236932    .0102557
         77  |  -.1430139   .0740051    -1.93   0.053    -.2880731    .0020452
         78  |  -.1346299   .0849156    -1.59   0.113     -.301075    .0318153
         80  |  -.2070794   .1052876    -1.97   0.049    -.4134561   -.0007026
```



```
. testparm i.year

 ( 1)  71.year = 0
 ( 2)  72.year = 0
 ( 3)  73.year = 0
 ( 4)  77.year = 0
 ( 5)  78.year = 0
 ( 6)  80.year = 0
 ( 7)  82.year = 0
 ( 8)  83.year = 0
 ( 9)  85.year = 0
 (10)  87.year = 0
 (11)  88.year = 0

       F( 11, 14858) =    4.36
            Prob > F =    0.0000
```

We reject the null that all years coefficients are jointly equal to zero therefore time fixed effects are needed. Time dummies can be used.

Testing for Fixed Effects: Hausman test

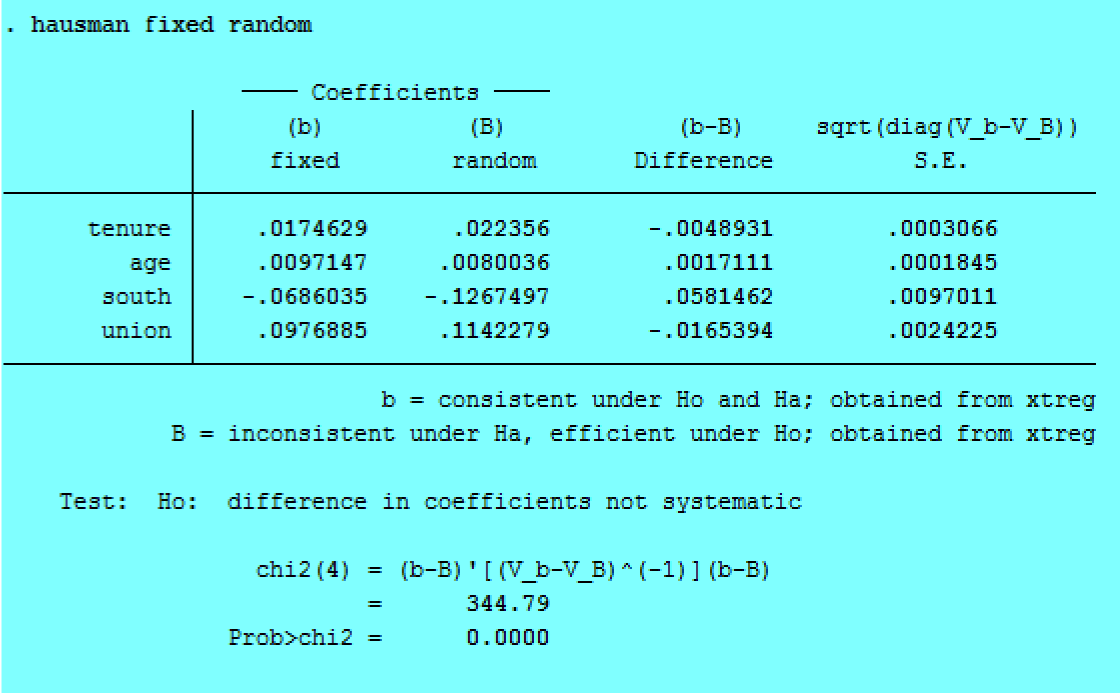
xtreg ln_wage tenure age south union,fe

estimates store fixed

xtreg ln_wage tenure age south union,re

estimates store random

hausman fixed random



```
. hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+----------------------------------------------------------------
      tenure |    .0174629      .022356       -.0048931        .0003066
         age |    .0097147     .0080036        .0017111        .0001845
       south |   -.0686035    -.1267497        .0581462        .0097011
       union |    .0976885     .1142279       -.0165394        .0024225
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      344.79
                Prob>chi2 =      0.0000
```

Fixed effect model is better (H_0 rejected)

Testing for random Effects: LM test (Breusch and Pagan LM test for random effects)

The LM test helps you decide between a random effects regression and a simple OLS regression.

Null hypothesis: variances across entities is zero / No significant difference across units / No panel effect.

Stata Command:

xttest0

Menu:

Statistics > Longitudinal/panel data > Linear models > Lagrange multiplier test for random effects

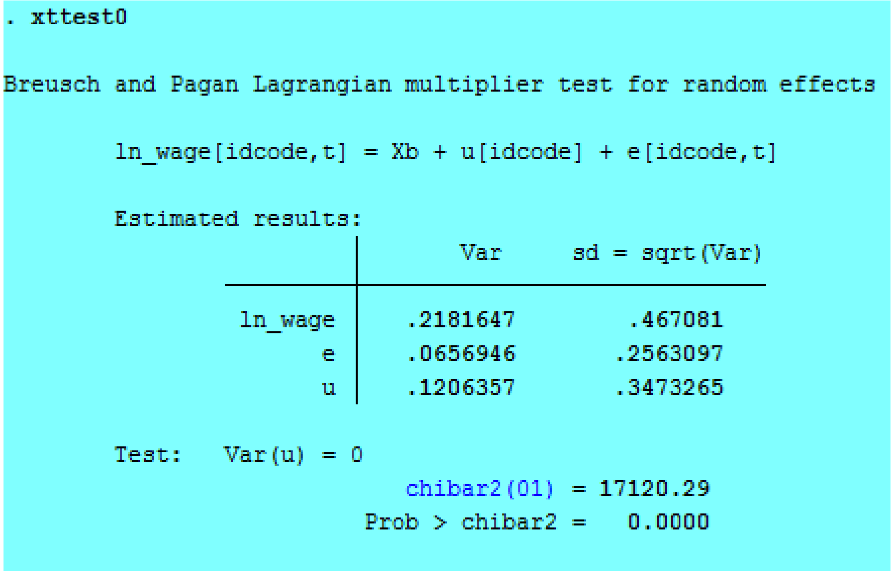
Prerequisite

- *xtset*
- Random effect model must be run first

Run the commands

xtreg ln_wage tenure age south union,re

xttest0



```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        ln_wage[idcode,t] = Xb + u[idcode] + e[idcode,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                 ln_wage |   .2181647        .467081
                       e |   .0656946       .2563097
                       u |   .1206357       .3473265

        Test:   Var(u) = 0
                             chibar2(01) = 17120.29
                          Prob > chibar2 =   0.0000
```

No random effects exist, simple OLS can be run

Breusch-Pagan LM test for cross-sectional correlation in fixed effects model (contemporaneous correlation / cross sectional dependence)

- This is a tests for cross-sectional independence in the residuals of a fixed effect regression model.
- This is a problem in macro panels with long time series

Null hypothesis:

- residuals across entities are not correlated

Stata Command:

xttset2

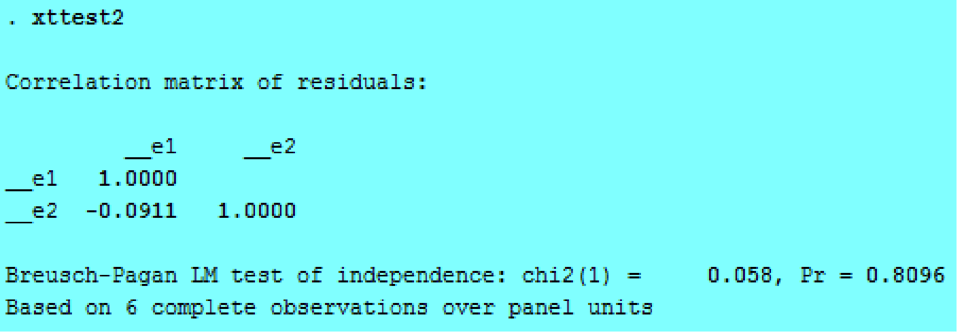
Prerequisite

xtset

Fixed effect model must be run first. Macro panel problem may exist so we reduce the entities.

xtreg ln_wage tenure age south union in 1/25,fe

xttest2



```
. xttest2

Correlation matrix of residuals:

          __e1      __e2
__e1   1.0000
__e2  -0.0911   1.0000

Breusch-Pagan LM test of independence: chi2(1) =     0.058, Pr = 0.8096
Based on 6 complete observations over panel units
```

No correlations found (H_0 accepted)

Testing for Heteroskedasticity

This calculates a modified Wald statistic for groupwise heteroskedasticity

Null hypothesis: Homoskedasticity (constant variance)

Stata Command: *xttest3*

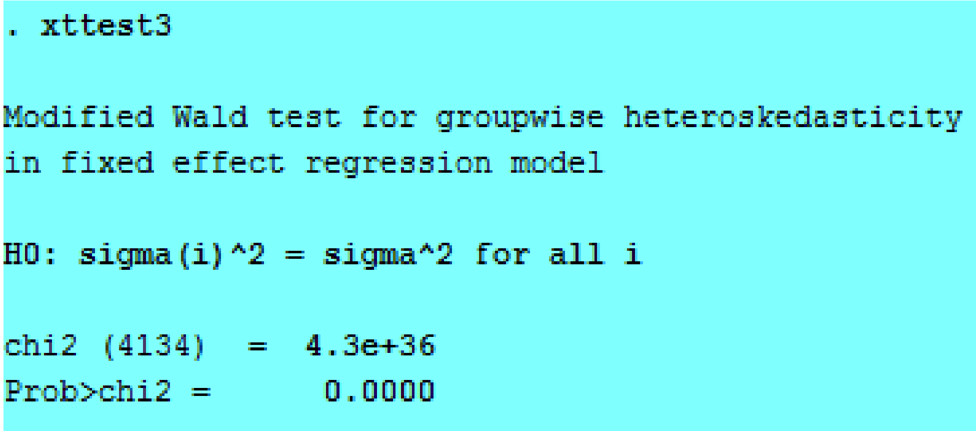
Prerequisite

xtset

Fixed effect model must be run first

xtreg ln_wage tenure age south union, fe

xttest3



```
. xttest3

Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (4134)  =  4.3e+36
Prob>chi2 =      0.0000
```

Conclusion

- Heteroskedasticity Exists
- Use robust option with regression (both fixed and random)

Testing for serial correlation:

Wooldridge test for serial correlation in panel-data models

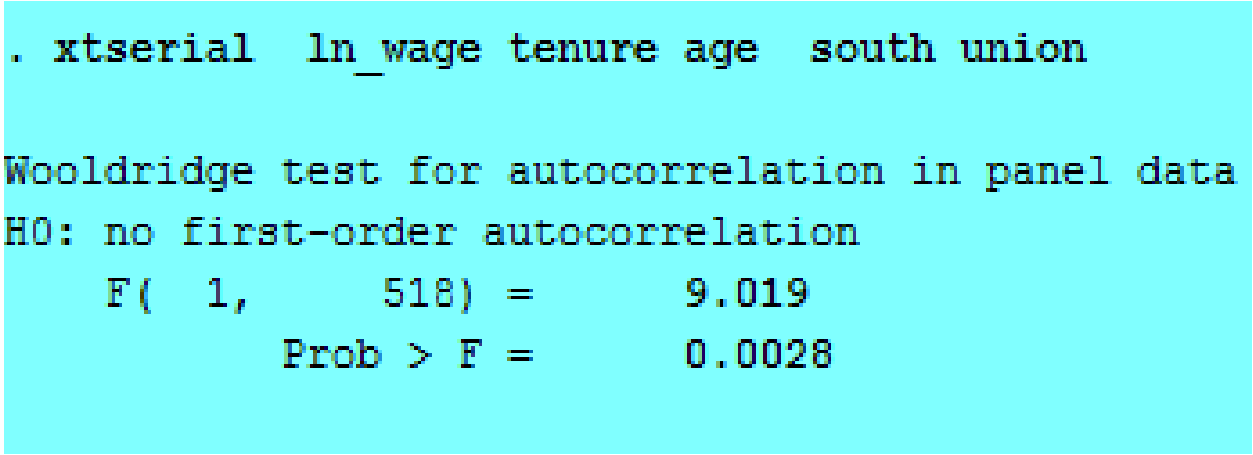
This implements a test for serial correlation in the idiosyncratic errors of a linear panel-data model discussed by Wooldridge (2002)

Null hypothesis: No Serial Correlation

Stata Command: *xtserial depvar [varlist] [if exp] [in range] [, output]*

Prerequisite: xtset

xtserial ln_wage tenure age south union



```
. xtserial  ln_wage tenure age  south union

Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,      518) =      9.019
           Prob > F =      0.0028
```

Conclusion:

- Serial Correlation Exists
- Use xtregar

Remedy for Serial Correlation: xtregar

xtregar fits regression models for panel data when the residual is first-order autoregressive.

xtregar offers a within estimator for fixed-effects models and a GLS estimator for random-effects models.

Syntax:

For Fixed-effects (FE) model

xtregar depvar [indepvars] [if] [in] [weight] , fe [options]

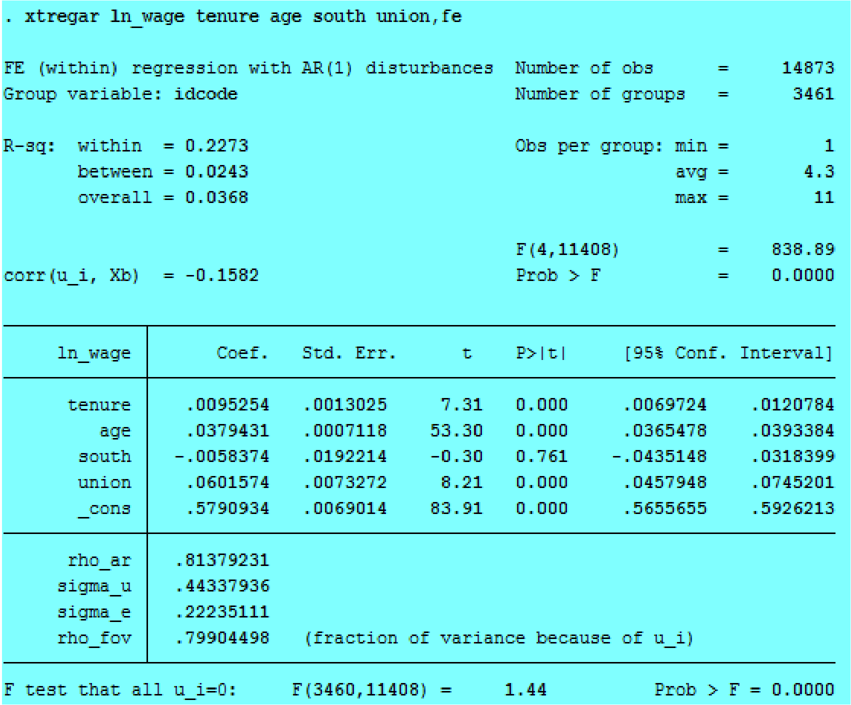
For GLS random-effects (RE) model

xtregar depvar [indepvars] [if] [in] [, re options]

Prerequisites: *xtset*

As serial correlation exists in our example so we can use xtregar

xtregar ln_wage tenure age south union, fe



```
. xtregar ln_wage tenure age south union,fe

FE (within) regression with AR(1) disturbances  Number of obs      =     14873
Group variable: idcode                          Number of groups   =      3461

R-sq:  within  = 0.2273                         Obs per group: min =         1
       between = 0.0243                                        avg =       4.3
       overall = 0.0368                                        max =        11

                                                F(4,11408)         =    838.89
corr(u_i, Xb)  = -0.1582                        Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   .0095254   .0013025     7.31   0.000     .0069724    .0120784
         age |   .0379431   .0007118    53.30   0.000     .0365478    .0393384
       south |  -.0058374   .0192214    -0.30   0.761    -.0435148    .0318399
       union |   .0601574   .0073272     8.21   0.000     .0457948    .0745201
       _cons |   .5790934   .0069014    83.91   0.000     .5655655    .5926213
-------------+----------------------------------------------------------------
      rho_ar |  .81379231
     sigma_u |  .44337936
     sigma_e |  .22235111
     rho_fov |  .79904498   (fraction of variance because of u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(3460,11408) =     1.44             Prob > F = 0.0000
```

Example of Panel Data

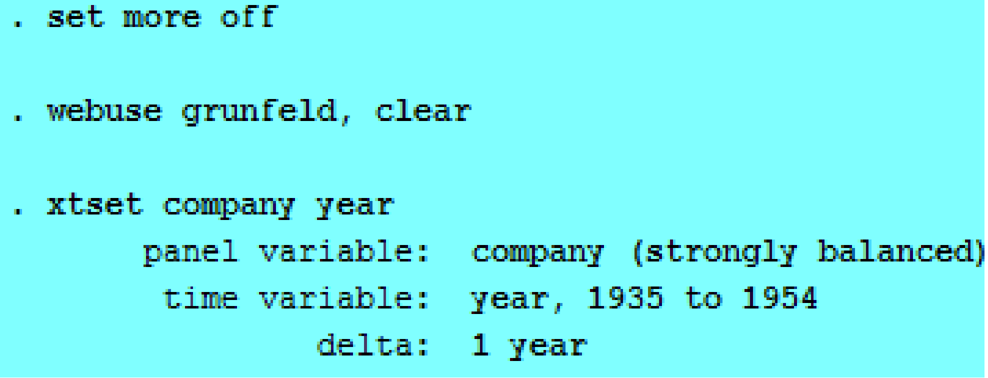
Using the file *grunfeld.dta* from the web perform various steps to apply panel data models:

See the following stat code (*panelExample.do*):

set more off

webuse grunfeld, clear

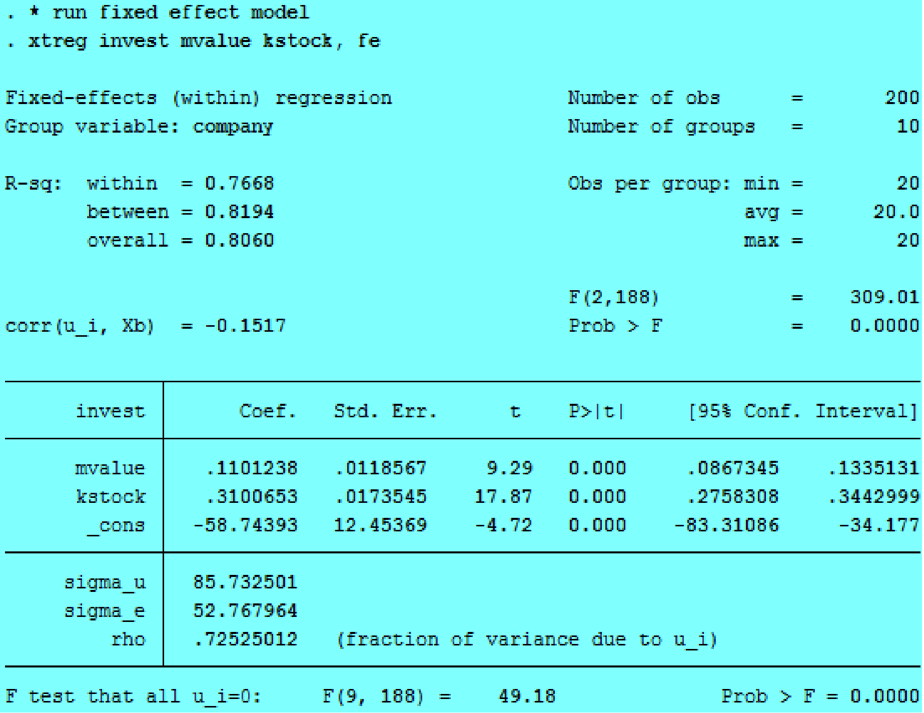
xtset company year



```
. set more off

. webuse grunfeld, clear

. xtset company year
       panel variable:  company (strongly balanced)
        time variable:  year, 1935 to 1954
                delta:  1 year
```

** run fixed effect model*

xtreg invest mvalue kstock, fe

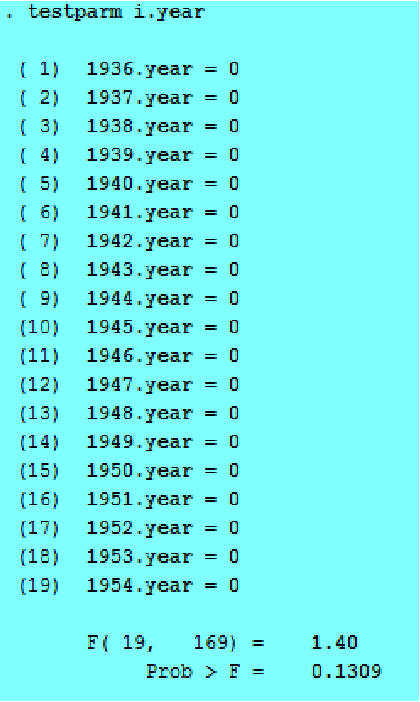


```
. * run fixed effect model
. xtreg invest mvalue kstock, fe

Fixed-effects (within) regression               Number of obs      =       200
Group variable: company                         Number of groups   =        10

R-sq:  within  = 0.7668                         Obs per group: min =        20
       between = 0.8194                                        avg =      20.0
       overall = 0.8060                                        max =        20

                                                F(2,188)           =    309.01
corr(u_i, Xb)  = -0.1517                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1101238   .0118567     9.29   0.000     .0867345    .1335131
      kstock |   .3100653   .0173545    17.87   0.000     .2758308    .3442999
       _cons |  -58.74393   12.45369    -4.72   0.000    -83.31086     -34.177
-------------+----------------------------------------------------------------
     sigma_u |  85.732501
     sigma_e |  52.767964
         rho |  .72525012   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(9, 188) =    49.18              Prob > F = 0.0000
```

To test time fixed effects:

qui xtreg invest mvalue kstock i.year, fe

testparm i.year



```
. testparm i.year

 ( 1)  1936.year = 0
 ( 2)  1937.year = 0
 ( 3)  1938.year = 0
 ( 4)  1939.year = 0
 ( 5)  1940.year = 0
 ( 6)  1941.year = 0
 ( 7)  1942.year = 0
 ( 8)  1943.year = 0
 ( 9)  1944.year = 0
 (10)  1945.year = 0
 (11)  1946.year = 0
 (12)  1947.year = 0
 (13)  1948.year = 0
 (14)  1949.year = 0
 (15)  1950.year = 0
 (16)  1951.year = 0
 (17)  1952.year = 0
 (18)  1953.year = 0
 (19)  1954.year = 0

       F( 19,   169) =    1.40
            Prob > F =    0.1309
```

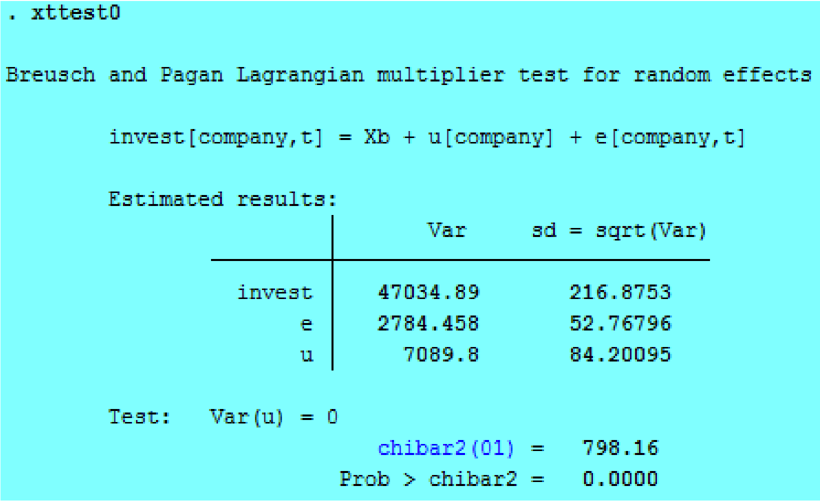
H_0: All year coefficients are zero

As p-value is greater than 5% or even 10%, we can not reject H0 so we accept that all year coefficients are zero and there are no time-FEs

To test for Random effects?

qui xtreg invest mvalue kstock, re

xttest0



```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        invest[company,t] = Xb + u[company] + e[company,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                  invest |   47034.89       216.8753
                       e |   2784.458       52.76796
                       u |     7089.8       84.20095

        Test:   Var(u) = 0
                             chibar2(01) =   798.16
                          Prob > chibar2 =   0.0000
```

As prob > chibar2, we reject H_0 and conclude that panel effects exist

Fixed of Random effects?

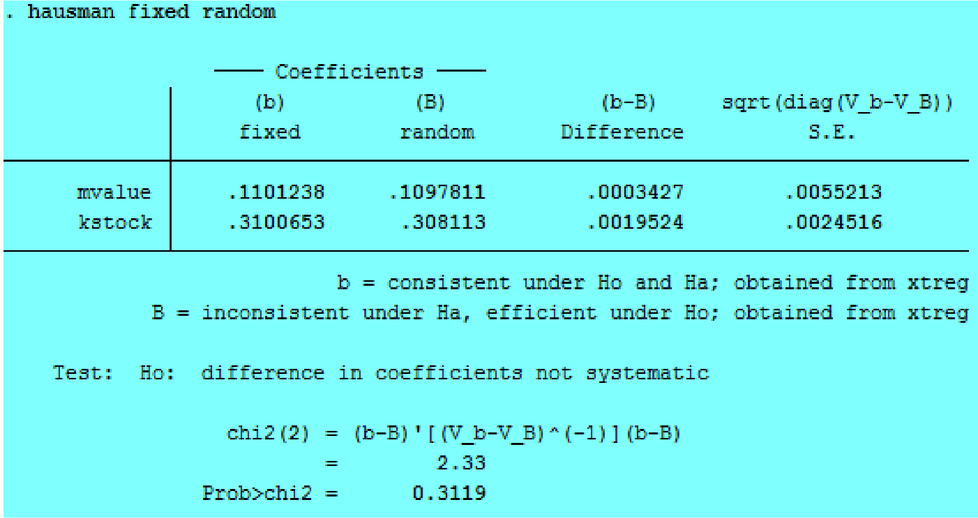
xtreg invest mvalue kstock, fe

estimates store fixed

xtreg invest mvalue kstock, re

estimates store random

hausman fixed random



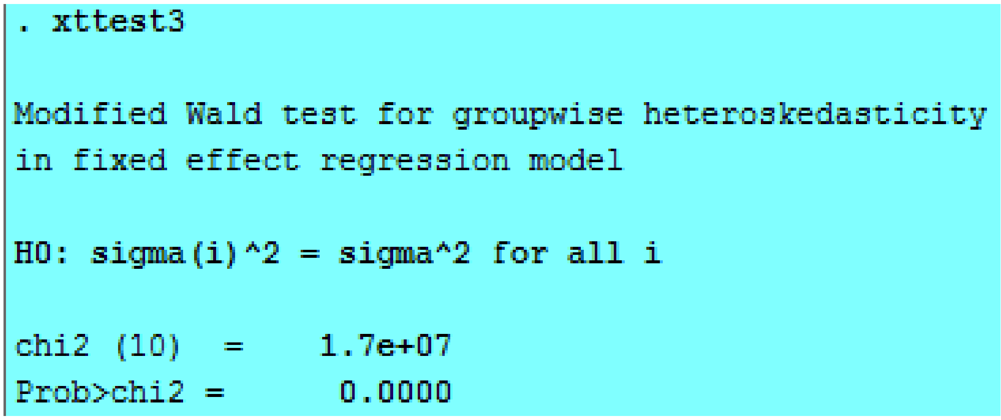
```
. hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+----------------------------------------------------------------
      mvalue |    .1101238     .1097811        .0003427        .0055213
      kstock |    .3100653      .308113        .0019524        .0024516
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from xtreg
           B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        2.33
                Prob>chi2 =      0.3119
```

Random effect model seems to be appropriate

Heteroskedasticity in FE

qui xtreg invest mvalue kstock, fe

xttest3



```
. xttest3

Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (10)  =    1.7e+07
Prob>chi2 =      0.0000
```

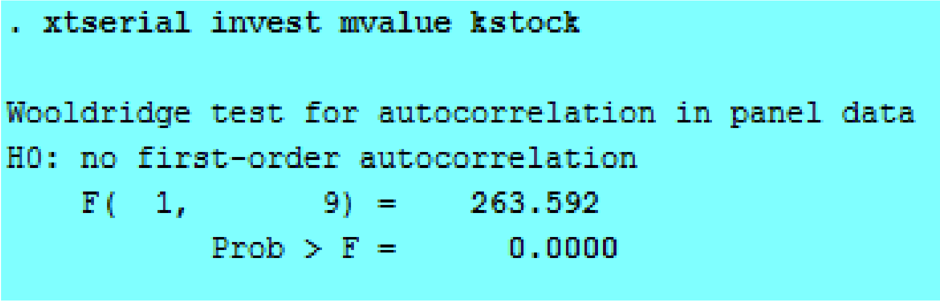
It is Required to run FE model before xttest3

H_0: homoscedasticity; we reject H0 and conclude that there is Heteroskedasticity. Use robust option for both FE & RE models

e.g. *xtreg invest mvalue kstock, fe robust*

Test for serial correlation

xtserial invest mvalue kstock



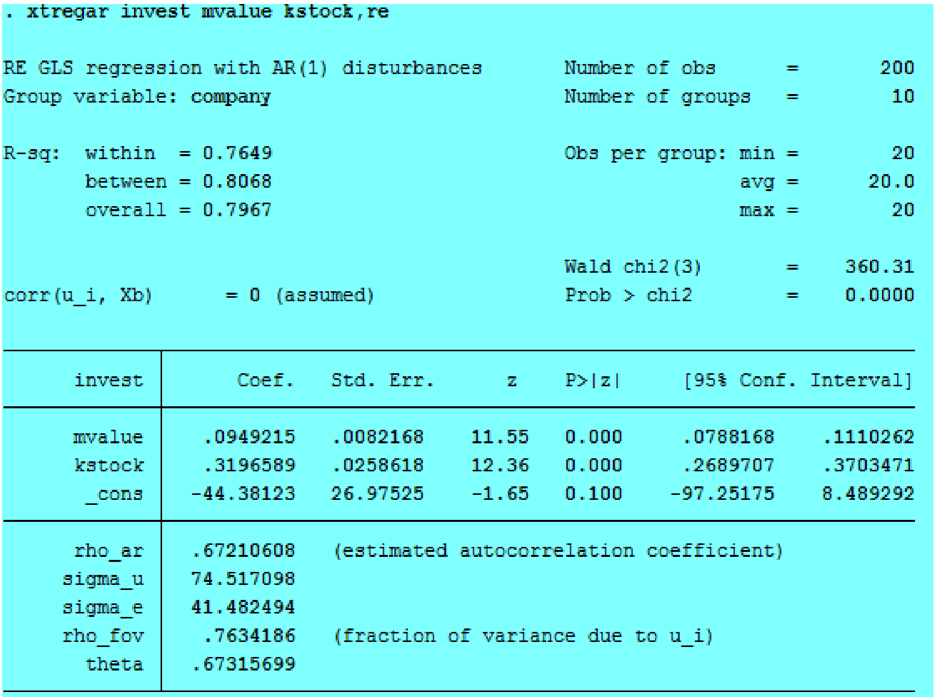
```
. xtserial invest mvalue kstock

Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,       9) =    263.592
           Prob > F =      0.0000
```

- reject H_0 , serial correlation exists
- xtregar may be used

Remedy for serial correlation

xtregar invest mvalue kstock,fe

xtregar invest mvalue kstock,re



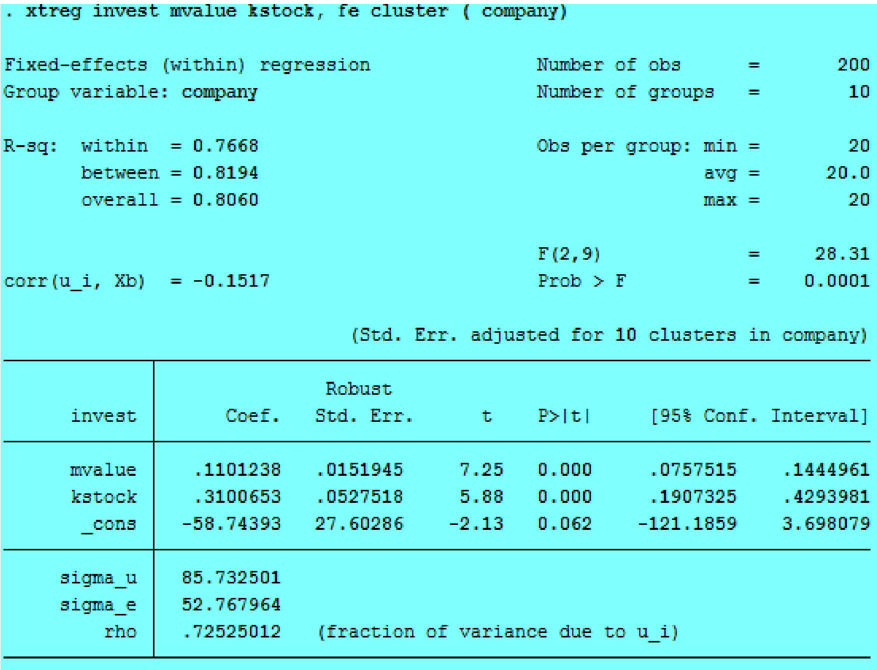
```
. xtregar invest mvalue kstock,re

RE GLS regression with AR(1) disturbances       Number of obs      =       200
Group variable: company                         Number of groups   =        10

R-sq:  within  = 0.7649                         Obs per group: min =        20
       between = 0.8068                                        avg =      20.0
       overall = 0.7967                                        max =        20

                                                Wald chi2(3)       =    360.31
corr(u_i, Xb)      = 0 (assumed)                Prob > chi2        =    0.0000

-----------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
      mvalue |   .0949215   .0082168    11.55   0.000     .0788168    .1110262
      kstock |   .3196589   .0258618    12.36   0.000     .2689707    .3703471
       _cons |  -44.38123   26.97525    -1.65   0.100    -97.25175    8.489292
-------------+---------------------------------------------------------------
      rho_ar |  .67210608   (estimated autocorrelation coefficient)
     sigma_u |  74.517098
     sigma_e |  41.482494
     rho_fov |   .7634186   (fraction of variance due to u_i)
       theta |  .67315699
-----------------------------------------------------------------------------
```

Only the random effect model results are shown above

Panel Data: both HSK and SC!

xtreg invest mvalue kstock, fe cluster (company)

Use cluster with individual ID variable as input (*cluster (company)*) option with xtreg



```
. xtreg invest mvalue kstock, fe cluster ( company)

Fixed-effects (within) regression               Number of obs      =       200
Group variable: company                         Number of groups   =        10

R-sq:  within  = 0.7668                         Obs per group: min =        20
       between = 0.8194                                        avg =      20.0
       overall = 0.8060                                        max =        20

                                                F(2,9)             =     28.31
corr(u_i, Xb)  = -0.1517                        Prob > F           =    0.0001

                               (Std. Err. adjusted for 10 clusters in company)
------------------------------------------------------------------------------
             |               Robust
      invest |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mvalue |   .1101238   .0151945     7.25   0.000     .0757515    .1444961
      kstock |   .3100653   .0527518     5.88   0.000     .1907325    .4293981
       _cons |  -58.74393   27.60286    -2.13   0.062    -121.1859    3.698079
-------------+----------------------------------------------------------------
     sigma_u |  85.732501
     sigma_e |  52.767964
         rho |  .72525012   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The resulting standard errors are completely robust to any kind of serial correlation and/or heteroskedasticity

Lecture 38
Qualitative and limited dependent variable models-I
(Categorical dependent variable models)

Variables with category:

- Variable that has categories
- Normally qualitative in nature
- binary, multivariate (ordinal, or nominal)

Examples

Gender, color of eyes, educational status, different outcomes of an event encoded in numbers are some examples. When dependent variable is categorical OLS is biased and inefficient.

Logistic Regression

A type of regression analysis used to predict the outcome of a categorical variable (binary or others). The dependent variable has a limited number of outcomes. The dependent variable is predicted by using a logistic function (e.g. logistic regression). Logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables. The minimum number of cases per independent variable is 10.

Logistic Function:

$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Where t is a linear function of the explanatory variables.

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X})}$$

Which can be interpreted as the probability of dependent variable equaling a ‘success’ or a ‘case’

The logistic function is:

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 X$$

This is the log of odds

$$g(x) = \ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 X$$

As a latent variable model this has an equivalent formulation. Let Y^*be a continuous latent variable (unobserved random variable) where $Y^* = \beta_1 X + \ \in$

Where ∈ $\sim Logistic(0,1)$

$$Y_i = \begin{cases} 1 & if\ Y^* > 0 \\ 0 & otherwise \end{cases}$$

Odd Ratio:

The logistic function is:

$$g(x) = \ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 X$$

This is the ln of odds

Where odd ratio is defined as : *ratio of probability of an event occurring to probability of an event not occurring*

Commands for logistic regression in Stata

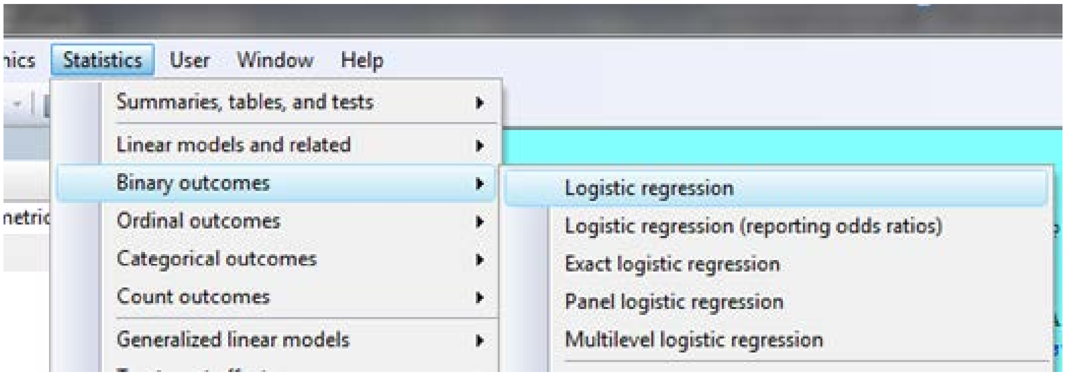
Binary logit: *logit*

Ordered logit: *ologit*

Multinomial logit: *mlogit*

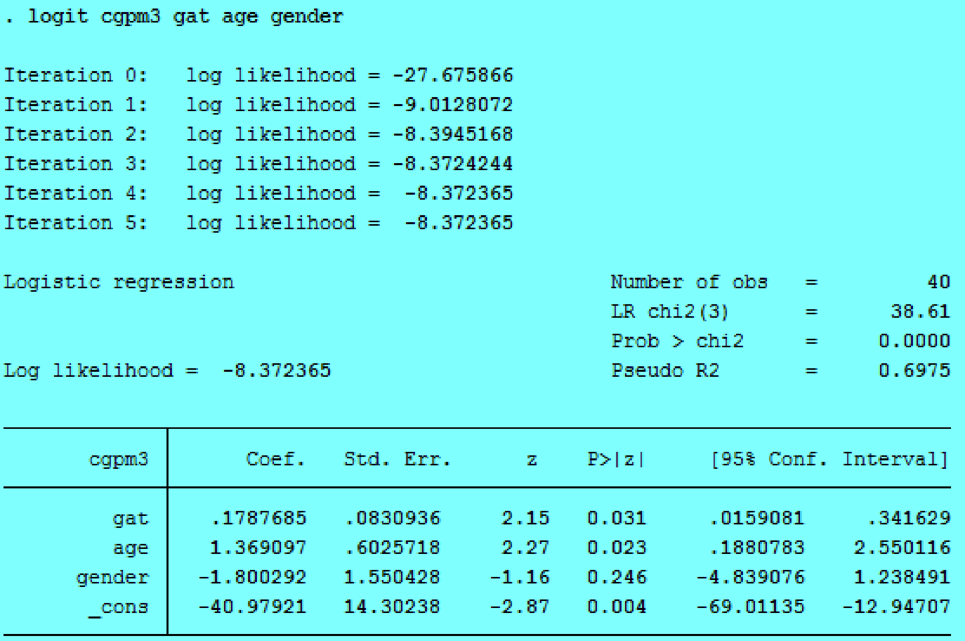
Syntax of the logit command is: ***logit depvar [indepvars] [if] [in] [weight] [, options]***

Menu: *Statistics > Binary outcomes > Logistic regression*


ics
Statistics
User
Window
Help
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Generalized linear models
netric
Logistic regression
Logistic regression (reporting odds ratios)
Exact logistic regression
Panel logistic regression
Multilevel logistic regression

Consider the file logitProbit.dta

Binary logistic regression



```
. logit cgpm3 gat age gender

Iteration 0:   log likelihood = -27.675866
Iteration 1:   log likelihood = -9.0128072
Iteration 2:   log likelihood = -8.3945168
Iteration 3:   log likelihood = -8.3724244
Iteration 4:   log likelihood =  -8.372365
Iteration 5:   log likelihood =  -8.372365

Logistic regression                               Number of obs   =         40
                                                  LR chi2(3)      =      38.61
                                                  Prob > chi2     =     0.0000
Log likelihood =  -8.372365                       Pseudo R2       =     0.6975

------------------------------------------------------------------------------
       cgpm3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gat |   .1787685   .0830936     2.15   0.031     .0159081     .341629
         age |   1.369097   .6025718     2.27   0.023     .1880783    2.550116
      gender |  -1.800292   1.550428    -1.16   0.246    -4.839076    1.238491
       _cons |  -40.97921   14.30238    -2.87   0.004    -69.01135   -12.94707
------------------------------------------------------------------------------
```

Interpreting Results

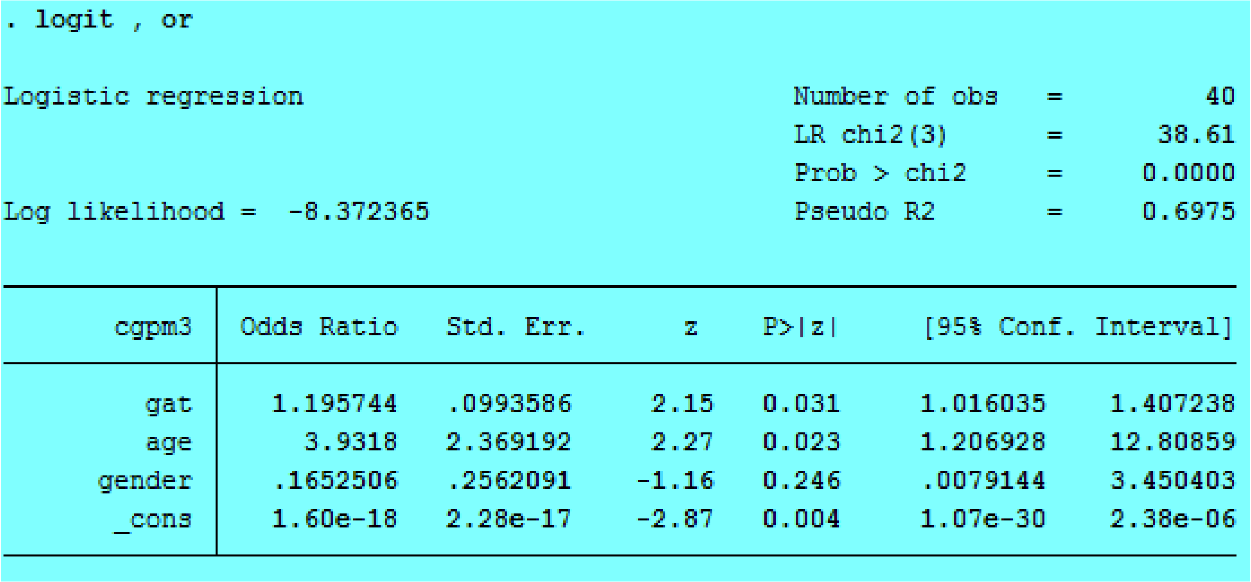
odds: odds mean ratio of favorable items to non-favorable items

Interpretation: The log of odds changes by the amount of the coefficient for one unit increase in the independent variable. Prob > Chi2 = 0.0000: our model as a whole fits significantly better than an empty model . we see the coefficients, their standard errors, the z-statistic, associated p-values, and the 95% confidence interval of the coefficients. For one unit increase in GAT score, the log odds of getting CGPA > 3 in Masters increases by 0.1787 (similar for other coefficients). Gender is not significant

Pseudo Rsquare

As equivalent statistic to R-squared does not exist because OLS not apply, pseudo R-squares are used because they look like R-squared (0 to 1). Higher values indicating better model fit different pseudo R-squars can give different values

To get odd ratios instead of log of odds



```
. logit , or

Logistic regression                               Number of obs   =         40
                                                  LR chi2(3)      =      38.61
                                                  Prob > chi2     =     0.0000
Log likelihood =  -8.372365                       Pseudo R2       =     0.6975

------------------------------------------------------------------------------
       cgpm3 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gat |   1.195744   .0993586     2.15   0.031     1.016035    1.407238
         age |     3.9318   2.369192     2.27   0.023     1.206928    12.80859
      gender |   .1652506   .2562091    -1.16   0.246     .0079144    3.450403
       _cons |   1.60e-18   2.28e-17    -2.87   0.004     1.07e-30    2.38e-06
------------------------------------------------------------------------------
```

Ordinal Logistic Regression

Ordinal categorical dependent variable examples:

- Encoding in age brackets
- Educational Grades (enccoded numerically)
- Income brackets encoded numerically
- size (small, medium, large or extra large)
- decesion (unlikely, somewhat likely, or very likely)
- Opinion (strongly agree , agree -- - - - - -)
- The actual values are irrelevant

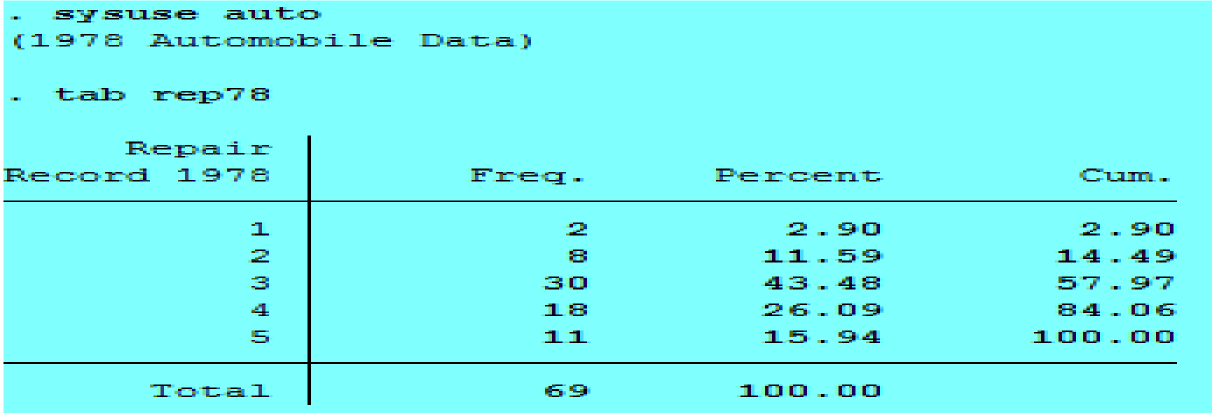
- Larger values are assumed to correspond to "higher" outcomes.

In Stata Command Syntax: *ologit depvar [indepvars] [if] [in] [weight] [, options]*

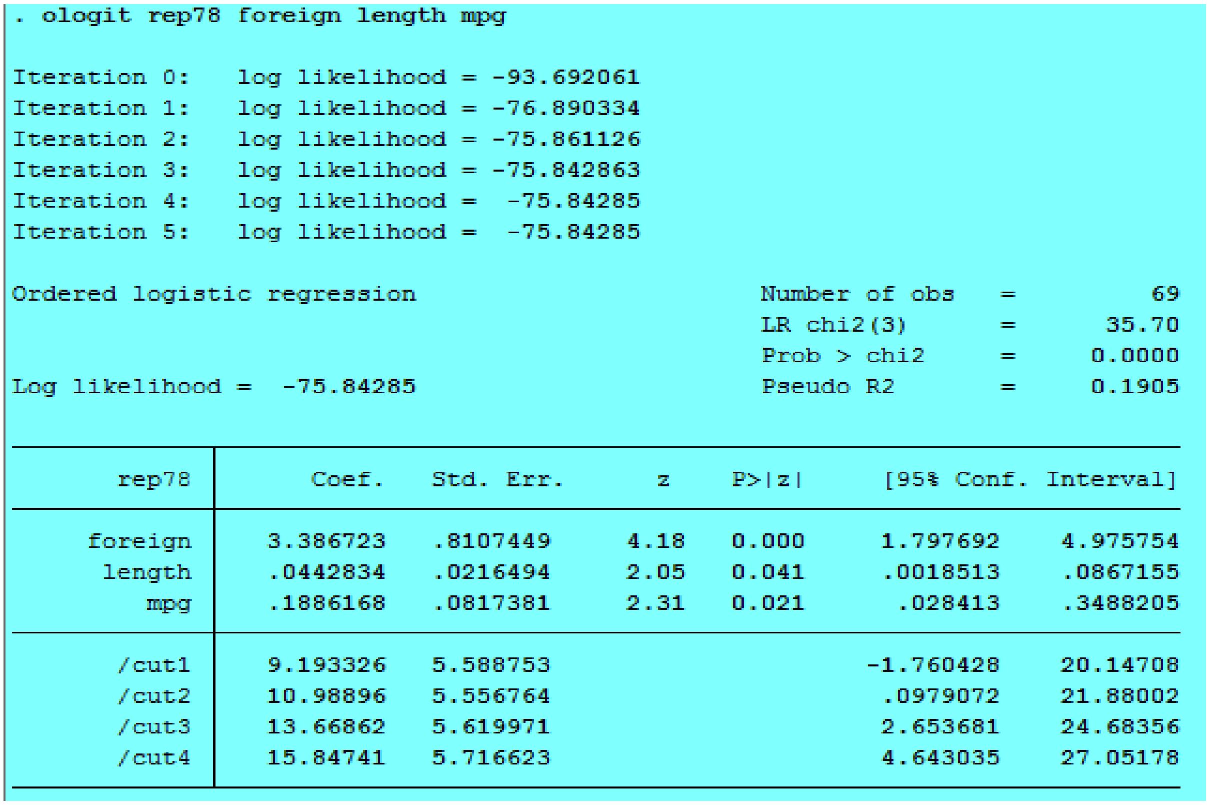
Menu:

Statistics > Ordinal outcomes > Ordered logistic regression

Follow the example given below:



```
. sysuse auto
(1978 Automobile Data)

. tab rep78

     Repair |
Record 1978 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          2        2.90        2.90
          2 |          8       11.59       14.49
          3 |         30       43.48       57.97
          4 |         18       26.09       84.06
          5 |         11       15.94      100.00
------------+-----------------------------------
      Total |         69      100.00
```



```
. ologit rep78 foreign length mpg

Iteration 0:   log likelihood = -93.692061
Iteration 1:   log likelihood = -76.890334
Iteration 2:   log likelihood = -75.861126
Iteration 3:   log likelihood = -75.842863
Iteration 4:   log likelihood =  -75.84285
Iteration 5:   log likelihood =  -75.84285

Ordered logistic regression                       Number of obs   =         69
                                                  LR chi2(3)      =      35.70
                                                  Prob > chi2     =     0.0000
Log likelihood =  -75.84285                       Pseudo R2       =     0.1905

------------------------------------------------------------------------------
       rep78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   3.386723   .8107449     4.18   0.000     1.797692    4.975754
      length |   .0442834   .0216494     2.05   0.041     .0018513    .0867155
         mpg |   .1886168   .0817381     2.31   0.021      .028413    .3488205
-------------+----------------------------------------------------------------
       /cut1 |   9.193326   5.588753                     -1.760428    20.14708
       /cut2 |   10.98896   5.556764                      .0979072    21.88002
       /cut3 |   13.66862   5.619971                      2.653681    24.68356
       /cut4 |   15.84741   5.716623                      4.643035    27.05178
------------------------------------------------------------------------------
```

Multinomial Logistic Regression

We use this model when no order to the categories of the outcome variable is found(i.e., the categories are nominal).

Examples

- Occupational choice (categories of job)
- Choice of specialization in Degrees (Accounting, Finance, marketing etc)

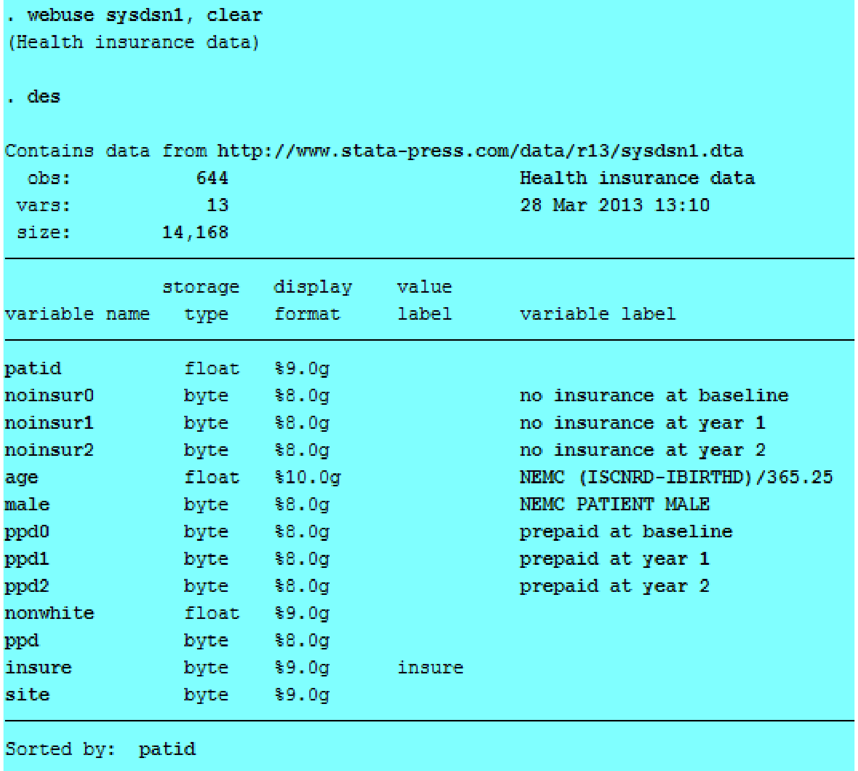
mlogit fits maximum-likelihood multinomial logit models, also known as polytomous logistic regression.

In Stata:

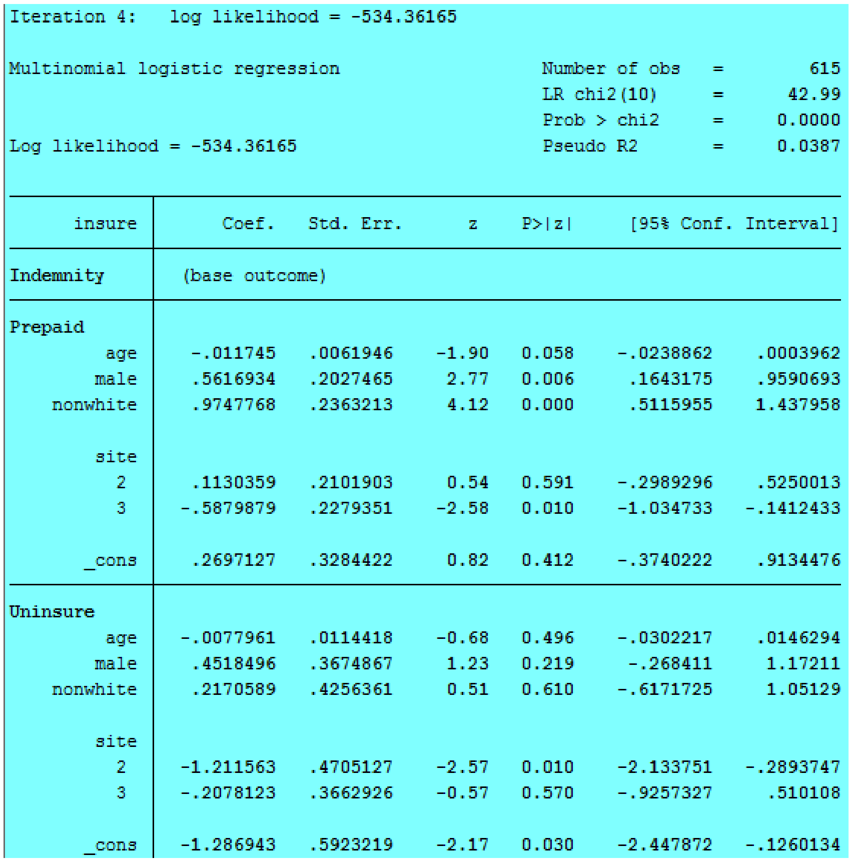
Menu: *Statistics > Categorical outcomes > Multinomial logistic regression*

Command Syntax: *mlogit depvar [indepvars] [if] [in] [weight] [, options]*

Please follow the example given below:



```
. webuse sysdsn1, clear
(Health insurance data)

. des

Contains data from http://www.stata-press.com/data/r13/sysdsn1.dta
  obs:           644                          Health insurance data
 vars:            13                          28 Mar 2013 13:10
 size:        14,168
-----------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
-----------------------------------------------------------------------------
patid           float  %9.0g
noinsur0        byte   %8.0g                 no insurance at baseline
noinsur1        byte   %8.0g                 no insurance at year 1
noinsur2        byte   %8.0g                 no insurance at year 2
age             float  %10.0g                NEMC (ISCNRD-IBIRTHD)/365.25
male            byte   %8.0g                 NEMC PATIENT MALE
ppd0            byte   %8.0g                 prepaid at baseline
ppd1            byte   %8.0g                 prepaid at year 1
ppd2            byte   %8.0g                 prepaid at year 2
nonwhite        float  %9.0g
ppd             byte   %8.0g
insure          byte   %9.0g      insure
site            byte   %9.0g
-----------------------------------------------------------------------------
Sorted by:  patid
```



```
Iteration 4:   log likelihood = -534.36165

Multinomial logistic regression                   Number of obs   =        615
                                                  LR chi2(10)     =      42.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -534.36165                       Pseudo R2       =     0.0387

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity    |  (base outcome)
-------------+----------------------------------------------------------------
Prepaid      |
         age |   -.011745   .0061946    -1.90   0.058    -.0238862    .0003962
        male |   .5616934   .2027465     2.77   0.006     .1643175    .9590693
    nonwhite |   .9747768   .2363213     4.12   0.000     .5115955    1.437958
             |
        site |
          2  |   .1130359   .2101903     0.54   0.591    -.2989296    .5250013
          3  |  -.5879879   .2279351    -2.58   0.010    -1.034733   -.1412433
             |
       _cons |   .2697127   .3284422     0.82   0.412    -.3740222    .9134476
-------------+----------------------------------------------------------------
Uninsure     |
         age |  -.0077961   .0114418    -0.68   0.496    -.0302217    .0146294
        male |   .4518496   .3674867     1.23   0.219     -.268411     1.17211
    nonwhite |   .2170589   .4256361     0.51   0.610    -.6171725     1.05129
             |
        site |
          2  |  -1.211563   .4705127    -2.57   0.010    -2.133751   -.2893747
          3  |  -.2078123   .3662926    -0.57   0.570    -.9257327     .510108
             |
       _cons |  -1.286943   .5923219    -2.17   0.030    -2.447872   -.1260134
```

The above is a result of using the *mlogit* command.

Lecture 39

Qualitative and limited dependent variable models-II

(Categorical dependent variable models)

Probit Regression

A type of regression analysis used to predict the Probability of a categorical variable (binary or others). Examples of categorical variables: as in logit. The dependent variable has a limited number of outcomes. The dependent variable is predicted by using a probit function. The minimum number of cases per independent variable is 10.

Probit Function:

Recall the standard normal distribution/ Z-scores

Given any Z-score , $\phi(Z)$ is the cumulative normal distribution function

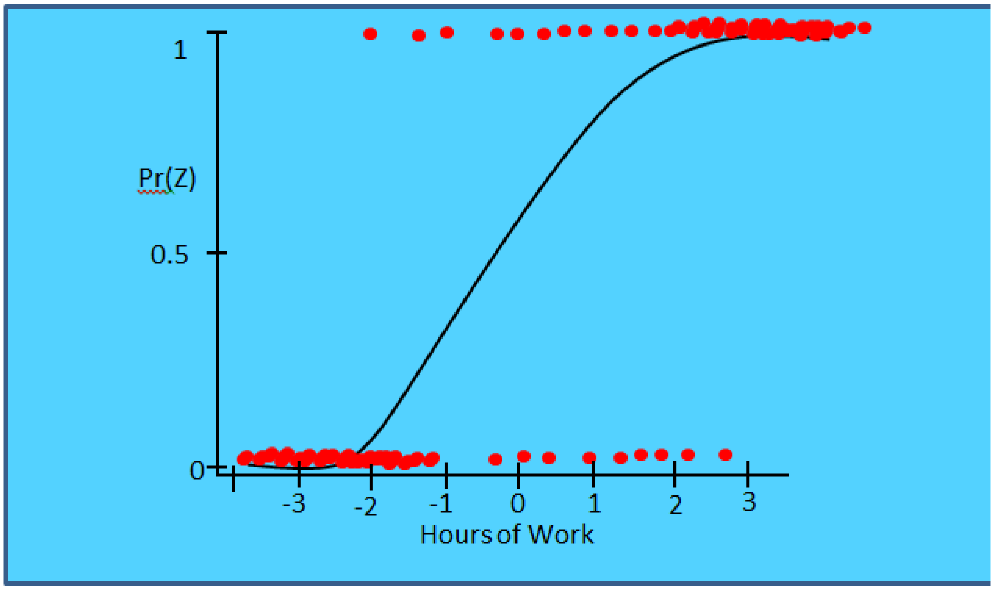
For any given Z, $\phi(Z) \in [0,1]$

$$Y = \phi(\boldsymbol{X}\beta + \epsilon) \text{ then } \phi^{-1}(Y) = \boldsymbol{X}\beta + \epsilon$$

$$F(Y) = \phi^{-1}(Y) \text{ is called the probit function}$$

Probit may be a short for ‘probability unit’ P(Y)=1

$\boldsymbol{X}\beta$ is the Z-value of a normal distribution. Higher the estimation value, more likely is the event to happen. A one unit change in the value of X bring β change in the Z-score of Y. Estimated Curve in S-shaped.



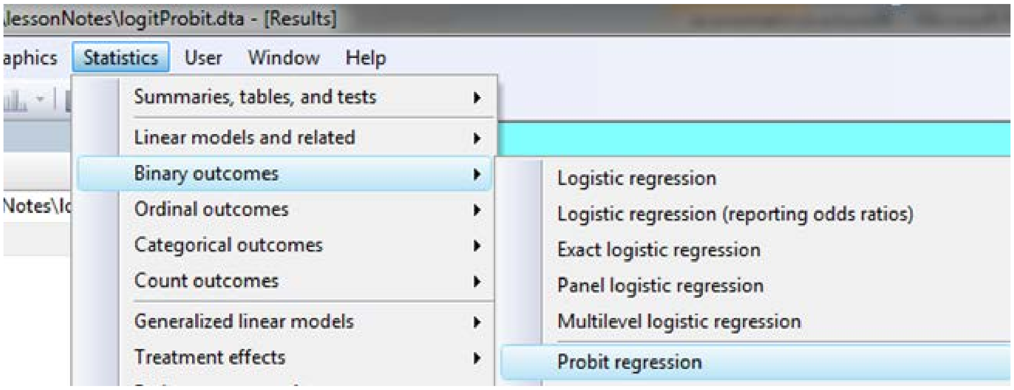
- Red dots are actual value of categorical variable (‘CGPA>3’ =1)
- The curve shows the probit

Example of binary probit

Binary probit command in Stata: *probit*

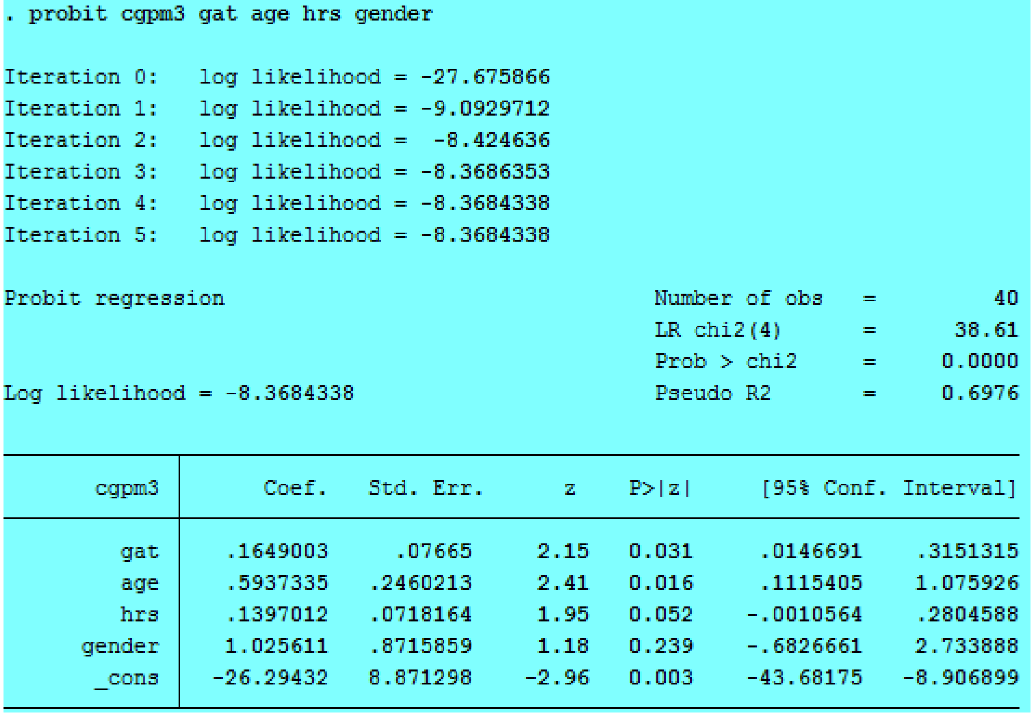
Sytax: *probit depvar [indepvars] [if] [in] [weight] [, options]*

Menu: *Statistics > Binary outcomes > Probit regression*


lessonNotes\logitProbit.dta - [Results]
aphics
Statistics
User
Window
Help
Notes\lo
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Generalized linear models
Treatment effects
Logistic regression
Logistic regression (reporting odds ratios)
Exact logistic regression
Panel logistic regression
Multilevel logistic regression
Probit regression

Consider the file logitProbit.dta

Run the Binary Probit regression



```
. probit cgpm3 gat age hrs gender

Iteration 0:   log likelihood = -27.675866
Iteration 1:   log likelihood = -9.0929712
Iteration 2:   log likelihood =  -8.424636
Iteration 3:   log likelihood = -8.3686353
Iteration 4:   log likelihood = -8.3684338
Iteration 5:   log likelihood = -8.3684338

Probit regression                                 Number of obs   =         40
                                                  LR chi2(4)      =      38.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -8.3684338                       Pseudo R2       =     0.6976

------------------------------------------------------------------------------
       cgpm3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gat |   .1649003     .07665     2.15   0.031     .0146691    .3151315
         age |   .5937335   .2460213     2.41   0.016     .1115405    1.075926
         hrs |   .1397012   .0718164     1.95   0.052    -.0010564    .2804588
      gender |   1.025611   .8715859     1.18   0.239    -.6826661    2.733888
       _cons |  -26.29432   8.871298    -2.96   0.003    -43.68175   -8.906899
------------------------------------------------------------------------------
```

Interpreting Results

- Significance is almost as in logit
- Model, as a whole is statistically significant (Pr > chi2) < 0.05
- Coefficients: Change in Z-score or probit due to change in one unit of independent variable (they are less than one)

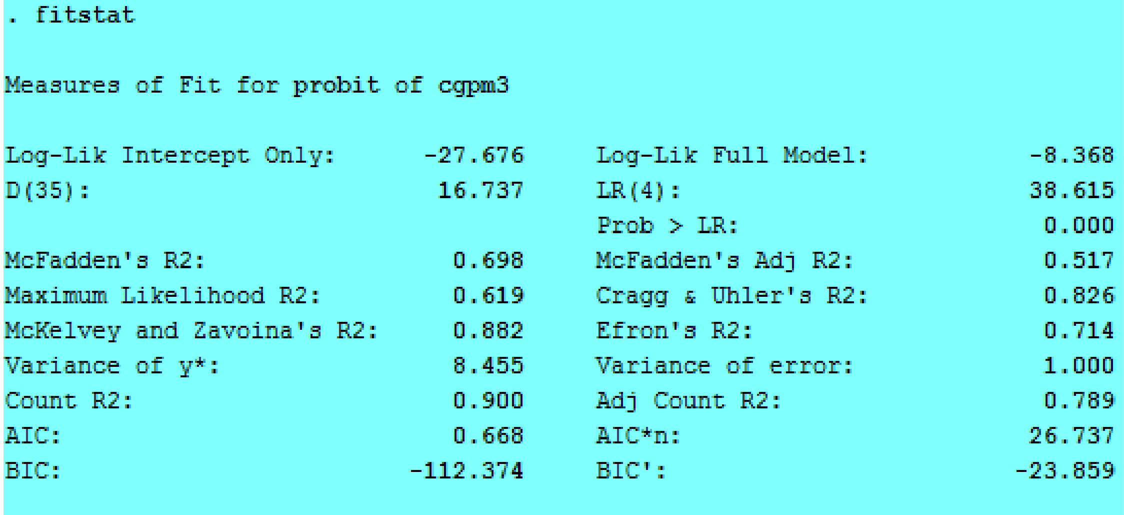
Interpretation of the coefficients:

- Not as in OLS

- Not slope coefficient or marginal effects
- The increase in probability due to a one-unit increase in a given independent variable/predictor depends both on the values of the other predictors and the starting value of the given predictors (different for different rows of observation)
- The predicted probability of CGPA>3 can be calculated as
- $F\ (\beta_0 + \beta_1 GAT + \beta_2 AGE + \beta_3 HRS)$
- Where $\beta_i$are the coefficients generated by the probit regression.
- Effect of change in one variable on the probability of getting CGPA>3 depends on the current value of a variable under consideration and also on the GIVEN/CONSTANT values of the other variables

Tests after probit (fitness, different pseudo R squares)

If fitstat is not installed try *ssc install fitstat*



```
. fitstat

Measures of Fit for probit of cgpm3

Log-Lik Intercept Only:      -27.676    Log-Lik Full Model:           -8.368
D(35):                        16.737    LR(4):                        38.615
                                        Prob > LR:                     0.000
McFadden's R2:                 0.698    McFadden's Adj R2:             0.517
Maximum Likelihood R2:         0.619    Cragg & Uhler's R2:            0.826
McKelvey and Zavoina's R2:     0.882    Efron's R2:                    0.714
Variance of y*:                8.455    Variance of error:             1.000
Count R2:                      0.900    Adj Count R2:                  0.789
AIC:                           0.668    AIC*n:                        26.737
BIC:                        -112.374    BIC':                        -23.859
```

Ordinal Probit Regression

Ordinal categorical dependent variable examples:

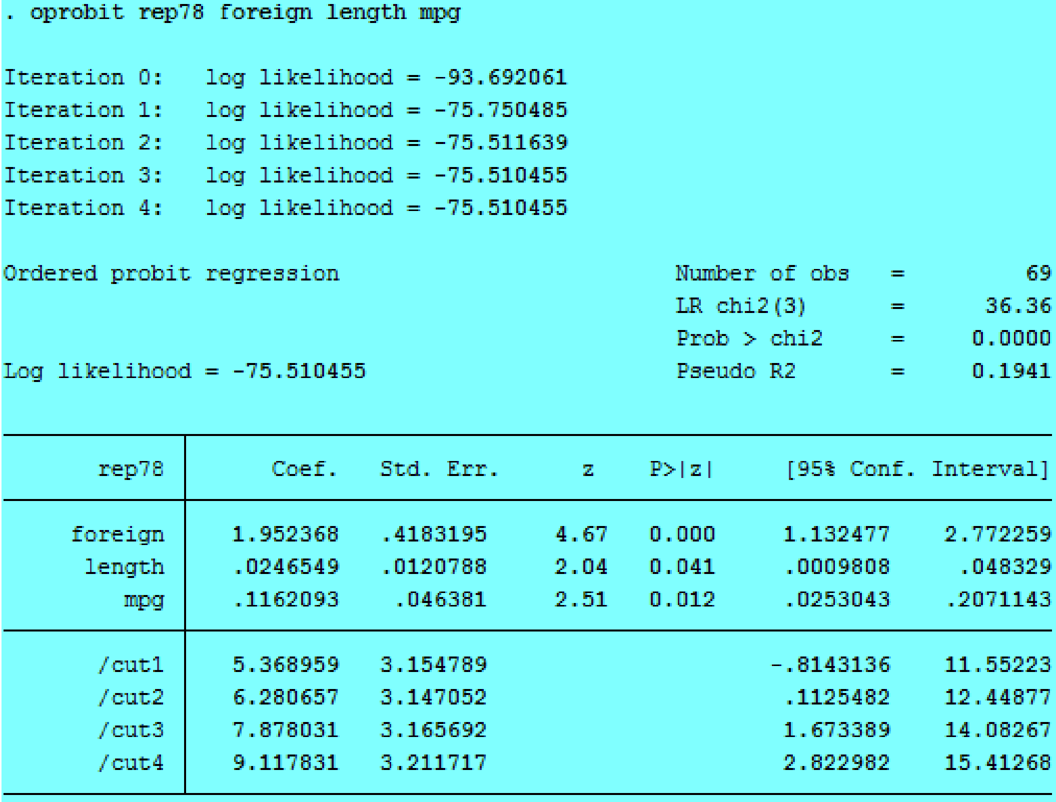
- Encoding in age brackets
- Educational Grades (enccoded numerically)
- Income brackets encoded numerically
- size (small, medium, large or extra large)

- decesion (unlikely, somewhat likely, or very likely)
- Opinion (strongly agree , agree -- - - - - -)
- The actual values are irrelevant
- Larger values are assumed to correspond to "higher" outcomes.

In Stata

Menu: *Statistics > Ordinal outcomes > Ordered probit regression*

Command Syntax: *oprobit depvar [indepvars] [if] [in] [weight] [, options]*



```
. oprobit rep78 foreign length mpg

Iteration 0:   log likelihood = -93.692061
Iteration 1:   log likelihood = -75.750485
Iteration 2:   log likelihood = -75.511639
Iteration 3:   log likelihood = -75.510455
Iteration 4:   log likelihood = -75.510455

Ordered probit regression                         Number of obs   =         69
                                                  LR chi2(3)      =      36.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -75.510455                       Pseudo R2       =     0.1941

------------------------------------------------------------------------------
       rep78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   1.952368   .4183195     4.67   0.000     1.132477    2.772259
      length |   .0246549   .0120788     2.04   0.041     .0009808     .048329
         mpg |   .1162093    .046381     2.51   0.012     .0253043    .2071143
-------------+----------------------------------------------------------------
       /cut1 |   5.368959   3.154789                     -.8143136    11.55223
       /cut2 |   6.280657   3.147052                      .1125482    12.44877
       /cut3 |   7.878031   3.165692                      1.673389    14.08267
       /cut4 |   9.117831   3.211717                      2.822982    15.41268
------------------------------------------------------------------------------
```

Magnitude of Coefficients differs from probit by a scale factor. Magnitude of Coefficients is not interpreted. The sign will matter.

Multinomial Probit Regression

No order to the categories of the outcome variable is found (i.e., the categories are nominal).

Examples

- Occupational choice (categories of job)

- Choice of specialization in Degrees (Accounting, Finance, marketing etc)

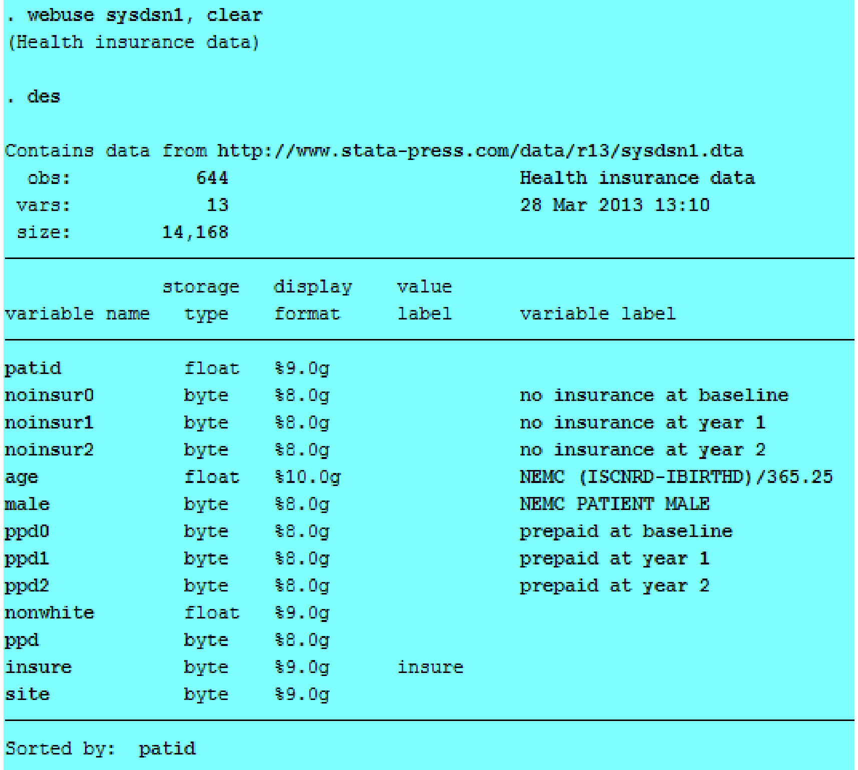
mprobit fits maximum-likelihood multinomial probit models, also known as polytomous probit regression.

In Stata:

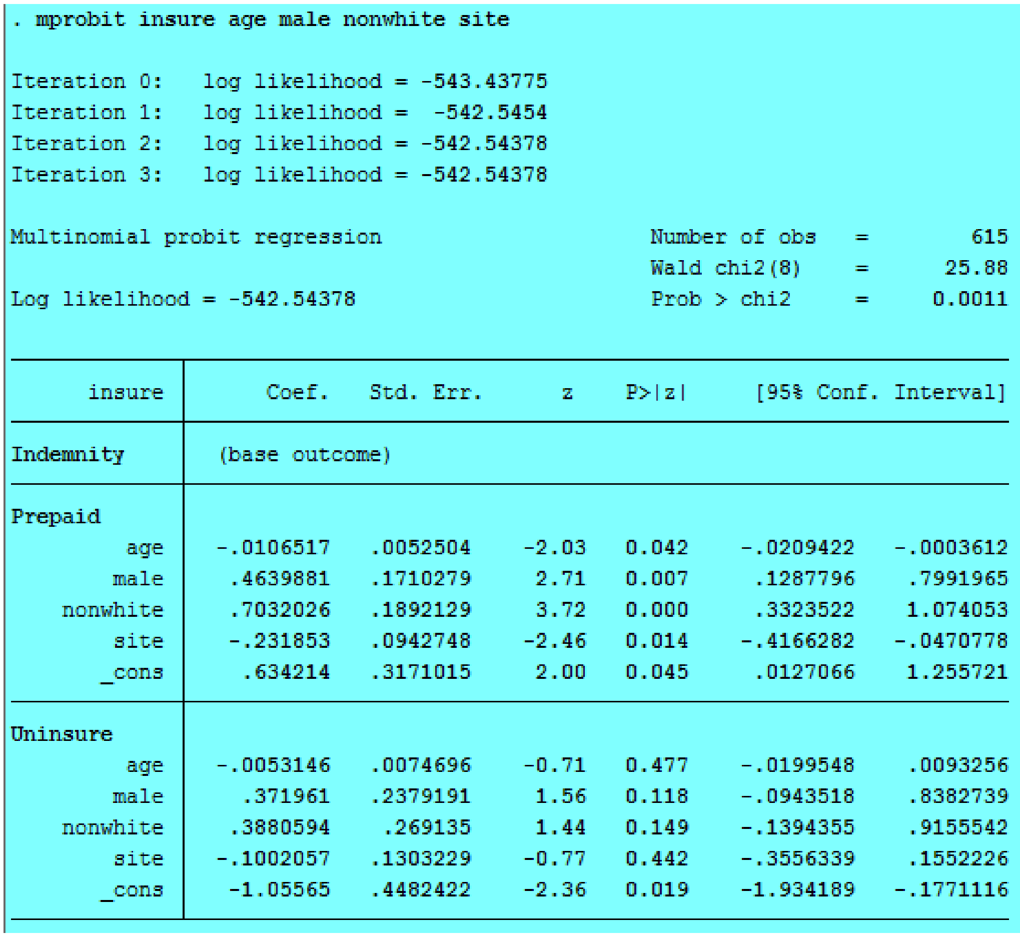
Menu: *Statistics > Categorical outcomes > Independent multinomial probit*

Command Syntax: *mprobit depvar [indepvars] [if] [in] [weight], [options]*

Follow the example below:



```
. webuse sysdsn1, clear
(Health insurance data)

. des

Contains data from http://www.stata-press.com/data/r13/sysdsn1.dta
  obs:           644                          Health insurance data
 vars:            13                          28 Mar 2013 13:10
 size:        14,168
------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
------------------------------------------------------------------------------
patid           float  %9.0g
noinsur0        byte   %8.0g                 no insurance at baseline
noinsur1        byte   %8.0g                 no insurance at year 1
noinsur2        byte   %8.0g                 no insurance at year 2
age             float  %10.0g                NEMC (ISCNRD-IBIRTHD)/365.25
male            byte   %8.0g                 NEMC PATIENT MALE
ppd0            byte   %8.0g                 prepaid at baseline
ppd1            byte   %8.0g                 prepaid at year 1
ppd2            byte   %8.0g                 prepaid at year 2
nonwhite        float  %9.0g
ppd             byte   %8.0g
insure          byte   %9.0g      insure
site            byte   %9.0g
------------------------------------------------------------------------------
Sorted by:  patid
```



```
. mprobit insure age male nonwhite site

Iteration 0:   log likelihood = -543.43775
Iteration 1:   log likelihood =  -542.5454
Iteration 2:   log likelihood = -542.54378
Iteration 3:   log likelihood = -542.54378

Multinomial probit regression                     Number of obs   =        615
                                                  Wald chi2(8)    =      25.88
Log likelihood = -542.54378                       Prob > chi2     =     0.0011

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity    |  (base outcome)
-------------+----------------------------------------------------------------
Prepaid      |
         age |  -.0106517   .0052504    -2.03   0.042    -.0209422   -.0003612
        male |   .4639881   .1710279     2.71   0.007     .1287796    .7991965
    nonwhite |   .7032026   .1892129     3.72   0.000     .3323522    1.074053
        site |   -.231853   .0942748    -2.46   0.014    -.4166282   -.0470778
       _cons |    .634214   .3171015     2.00   0.045     .0127066    1.255721
-------------+----------------------------------------------------------------
Uninsure     |
         age |  -.0053146   .0074696    -0.71   0.477    -.0199548    .0093256
        male |    .371961   .2379191     1.56   0.118    -.0943518    .8382739
    nonwhite |   .3880594    .269135     1.44   0.149    -.1394355    .9155542
        site |  -.1002057   .1303229    -0.77   0.442    -.3556339    .1552226
       _cons |   -1.05565   .4482422    -2.36   0.019    -1.934189   -.1771116
------------------------------------------------------------------------------
```

Lecture 40
Qualitative and limited dependent variable models-III
&
Censored Regression Models (Tobit Model)

Probability Vs. Odd Ratio

The following table maps the odd ratio and probability concepts.

Probability	Odd Ratio	Ln (odds)	Odds
0	**0**	**-**	**-**
1/5	**1/4**	**Negative**	**Against (1 to 4)**
1/4	1/3	Negative	Against (1 to 3)
1/3	1/2	Negative	Against (1 to 2)
1/2	**1**	**Zero**	**Even**
2/3	**2**	**Positive**	**Favor (2 to one)**
3/4	**3**	**Positive**	**Favor (3 to one)**
1	∞	**Positive/∞**	**-**
Range= 0 to 1	**Range= 0 to ∞**	**-∞ to ∞**	

Note that Probability ranges from 0 to 1 while Odds range from 0 to infinity and Log (odds) ranges from $-\infty\ to\ \infty$. Also:

$$Odd\ ratio = \frac{Probability\ of\ x}{1 - probability\ of\ x}$$

$$Probability = \frac{Odd\ Ratio}{1 + Odd\ ratio}$$

Remember that the coefficients are the CHANGE in probability or odd ratios, not the probability or Odds. Above formula are not for the coefficients.

Marginal Effects

For simple regression, slopes are the derivative of dependent variable w.r.t. independent variable. This is not the case in probit models where

$$\frac{\partial \mathbb{P}\left[Y_i = 1 \middle| X_{1i},, \;\ldots\ldots X_{ki}; \beta_{0,\ldots\ldots}\beta_k\right]}{\partial X_{ki}} = \beta_k \Phi\left(\beta_0 + \sum_{\mathbf{1}}^{\boldsymbol{k}} \beta_k X_{ki}\right)$$

Here $\phi(\cdot)$ is the standard normal probability density function.

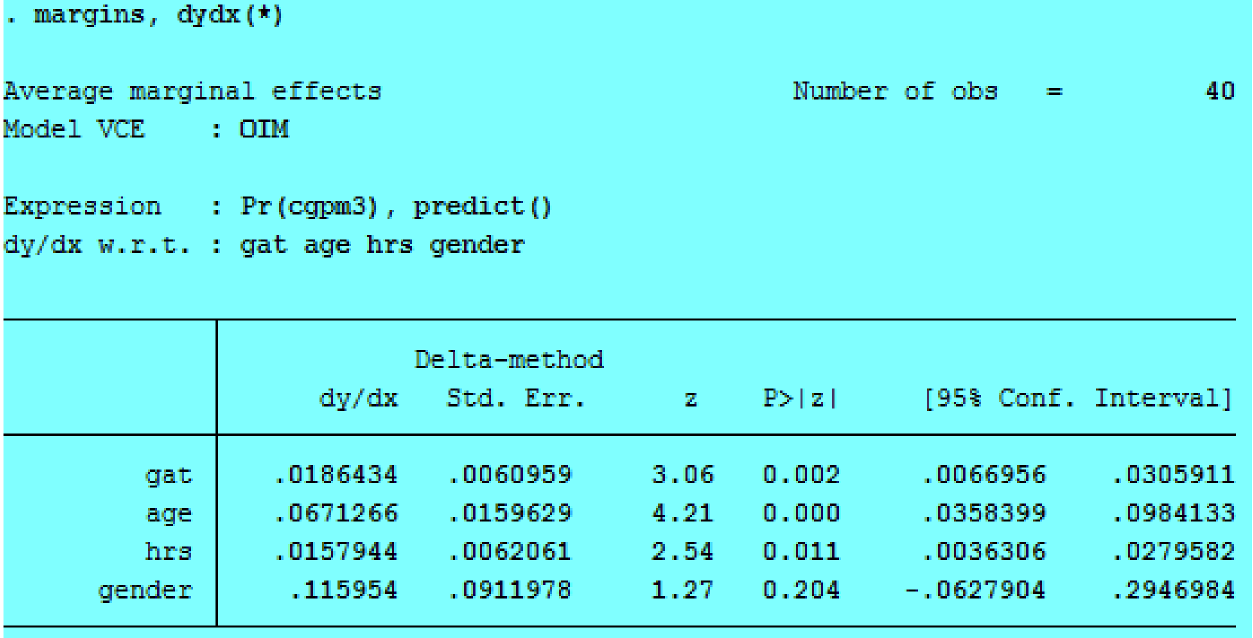
The increase in probability due to a one-unit increase in a given independent variable/predictor depends both on the values of the other predictors and the starting value of the given predictors (different for different rows of observation). Effect of change in one variable on the probability of getting CGPA>3 depends on the current value of a variable under consideration and also on the GIVEN/CONSTANT values of the other variables

Marginal Effects in Stata

Open the file **logitProbit.dta** and run the following Probit regression

probit cgpm3 gat age hrs gender

margins, dydx(*)



```
. margins, dydx(*)

Average marginal effects                          Number of obs   =         40
Model VCE    : OIM

Expression   : Pr(cgpm3), predict()
dy/dx w.r.t. : gat age hrs gender

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gat |   .0186434   .0060959     3.06   0.002     .0066956    .0305911
         age |   .0671266   .0159629     4.21   0.000     .0358399    .0984133
         hrs |   .0157944   .0062061     2.54   0.011     .0036306    .0279582
      gender |    .115954   .0911978     1.27   0.204    -.0627904    .2946984
------------------------------------------------------------------------------
```

One unit increase in gat increases the probability of CGPA>3 by 0.0128985.

Tests after logit/probit models

Likelihood ratio test

This test is used both for logit and probit models. It estimates two models and compares them (e.g. empty vs full OR restricted model vs. unrestricted model). This statistic is distributed chi-squared with degrees of freedom equal to the difference in the number of degrees of freedom between the two models. Significance means that the model improves.

Open **logitProbit.dta** and run the following commands

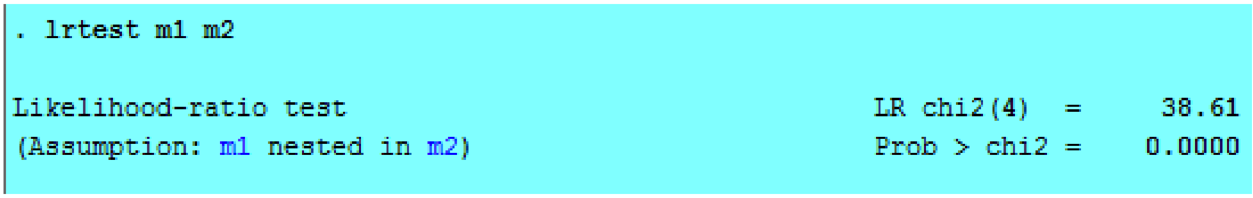
probit cgpm3

estimates store m1

probit cgpm3 gat age hrs gender

estimates store m2

lrtest m1 m2



```
. lrtest m1 m2

Likelihood-ratio test                                 LR chi2(4)  =     38.61
(Assumption: m1 nested in m2)                         Prob > chi2 =    0.0000
```

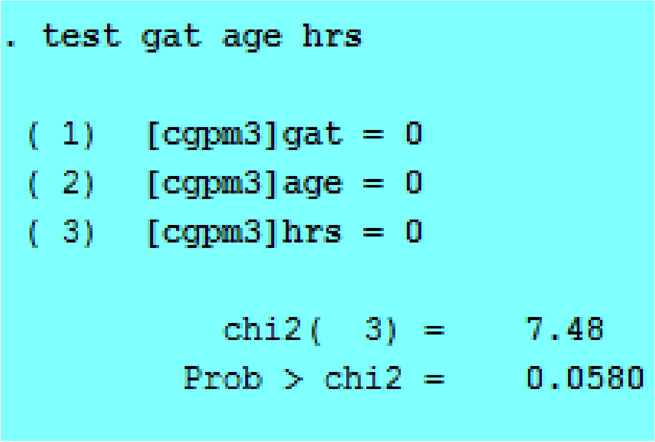
Here the last line gives the result above. As the p-value is less than 0.01, including all variable together improves the model. We can use this test for two models where the first one is nested in the second one.

Wald Test

Wald test estimates the *lrtest* but is better as we need to run only one model. It tests if the parameters of interest are jointly equal to zero. For the file logitProbit.dta, run the following (Other combination of variables may also be tested):

probit cgpm3 gat age hrs

test gat age hrs



```
. test gat age hrs

 ( 1)  [cgpm3]gat = 0
 ( 2)  [cgpm3]age = 0
 ( 3)  [cgpm3]hrs = 0

           chi2(  3) =    7.48
         Prob > chi2 =    0.0580
```

The above result shows that H_0 is rejected at 10%: all coefficients of variables (gat, age, hrs) are not jointly zero.

Censored Data Regression Models: The Tobit Model

OLS is inefficient when The dependent variable is incompletely observed or when the dependent variable is observed completely but the selected sample is not representative of the population.

Truncated and Censored Data

Truncated Data: observations on both dependent and independent variable are lost

Example: small firms only included, low income respondents only etc.

Censored Data: when observations on dependent variable only are lost (or limited)

Examples: Income top-coded to 50,000; time bottom-coded to 10 minutes chunks etc.

Tobit Model: Structure

Suppose $y^*$is a latent variable that is observed partially for values greater than θ

$$y^* = X_i\beta + \epsilon_i$$

And

$$\epsilon_i \sim N(0,\ \sigma^2)$$

Then the observed y is defined as

$$y = \begin{cases} y^* & if y^* > \theta \\ \theta_y & if y^* \leq \theta \end{cases}$$

As the data is usually censured at zero, the model becomes

$$y = \begin{cases} y^* & if y^* > 0 \\ 0 & if y^* \leq 0 \end{cases}$$

Examples where the Tobit model can be applied may include:

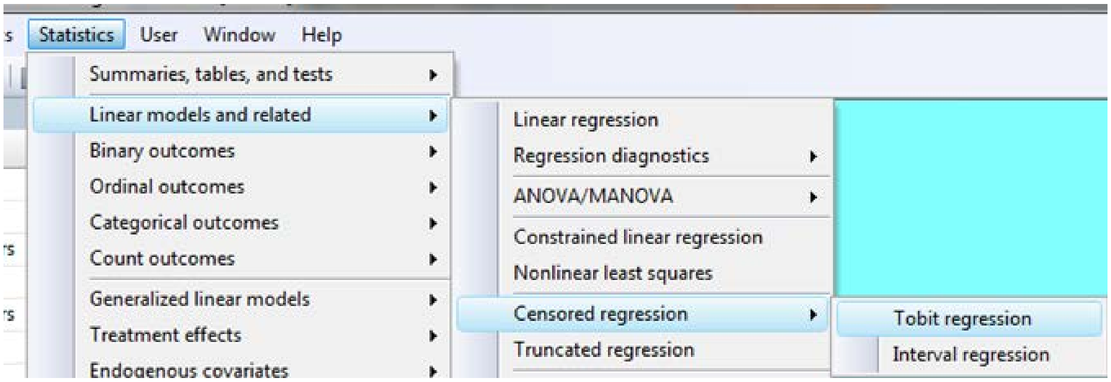
- Time use surveys: 10 minutes time chunks
- Glucometers to detect blood sugar give a reading Error or not more than a specific value (like, for example, 500 mg/dL)
- GAT (Gen): students answering all 100 questions may have different IQ levels and/or aptitude.

Tobit Model using Stata

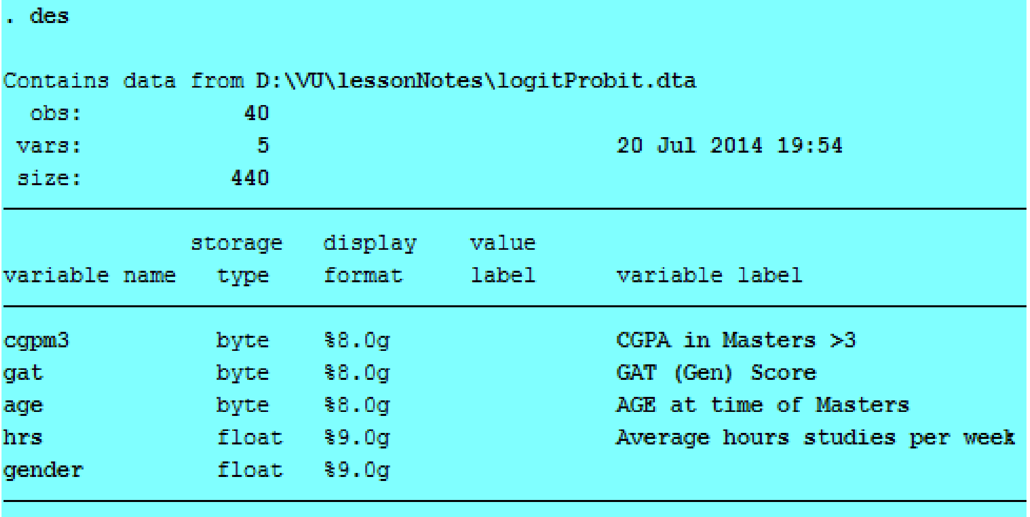
Tobit Model Stata Command: *tobit*

Syntax: *tobit depvar [indepvars] [if] [in] [weight] , ll[(#)] ul[(#)] [options]*

Menu: *Statistics > Linear models and related > Censored regression > Tobit regression*



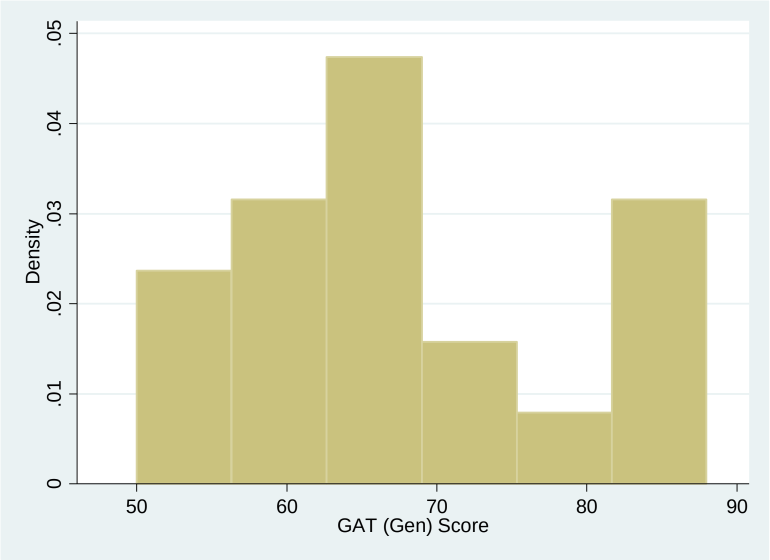
Open the file logitProbit.dta, use usual commands to know your data (sum, des, etc.)



```
. des

Contains data from D:\VU\lessonNotes\logitProbit.dta
  obs:            40
 vars:             5                          20 Jul 2014 19:54
 size:           440

              storage  display    value
variable name   type   format     label      variable label

cgpm3           byte   %8.0g                 CGPA in Masters >3
gat             byte   %8.0g                 GAT (Gen) Score
age             byte   %8.0g                 AGE at time of Masters
hrs             float  %9.0g                 Average hours studies per week
gender          float  %9.0g
```

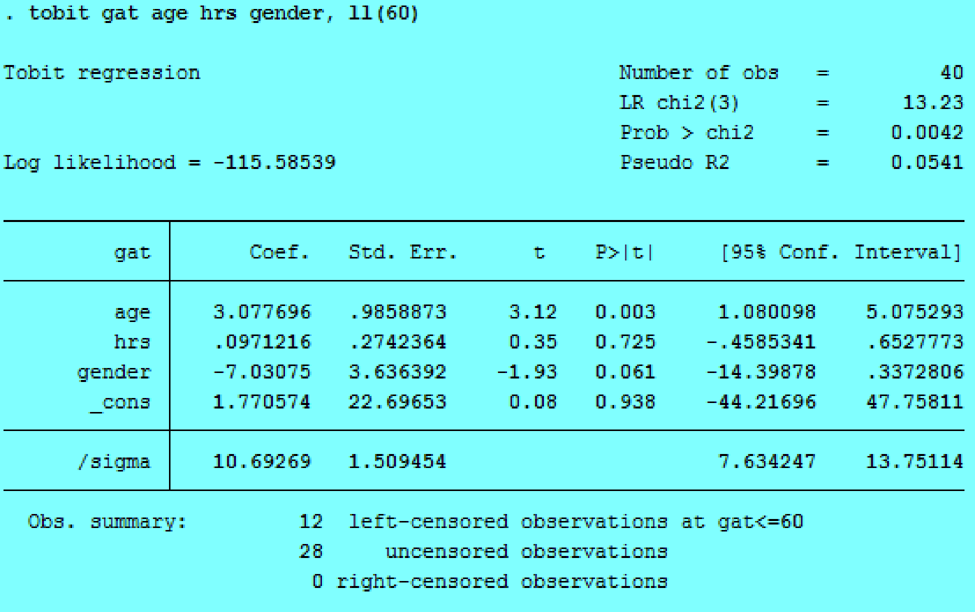
Consider the variable 'gat'.

histogram gat



Just for example, let us consider censoring data at different levels.

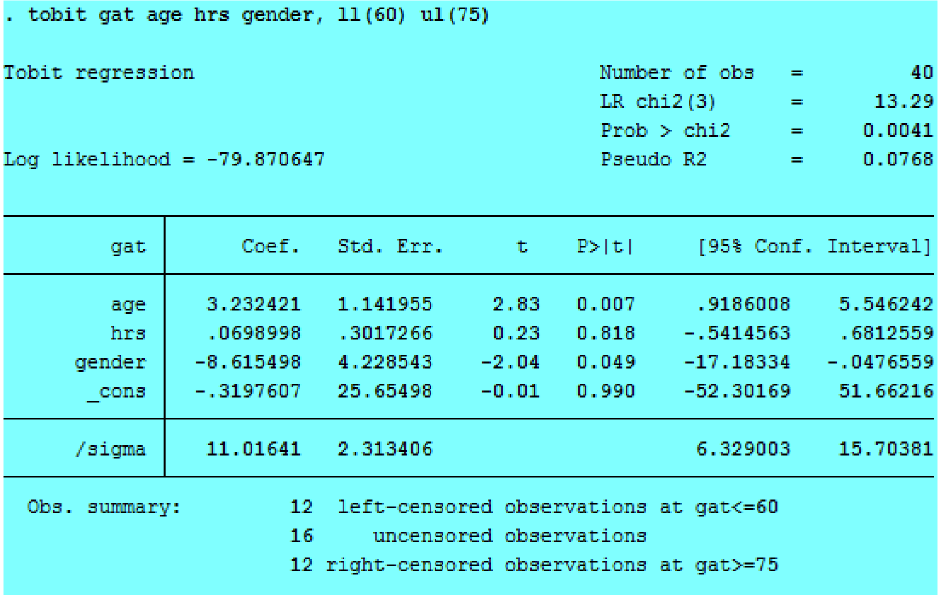
First let us consider left censoring at 60. There were 40 observations with 10 with gat<60. Even after censoring there are 40!



```
. tobit gat age hrs gender, ll(60)

Tobit regression                                  Number of obs   =         40
                                                  LR chi2(3)      =      13.23
                                                  Prob > chi2     =     0.0042
Log likelihood = -115.58539                       Pseudo R2       =     0.0541

------------------------------------------------------------------------------
         gat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   3.077696   .9858873     3.12   0.003     1.080098    5.075293
         hrs |   .0971216   .2742364     0.35   0.725    -.4585341    .6527773
      gender |   -7.03075   3.636392    -1.93   0.061    -14.39878    .3372806
       _cons |   1.770574   22.69653     0.08   0.938    -44.21696    47.75811
-------------+----------------------------------------------------------------
      /sigma |   10.69269   1.509454                      7.634247    13.75114
------------------------------------------------------------------------------
  Obs. summary:         12  left-censored observations at gat<=60
                        28     uncensored observations
                         0 right-censored observations
```

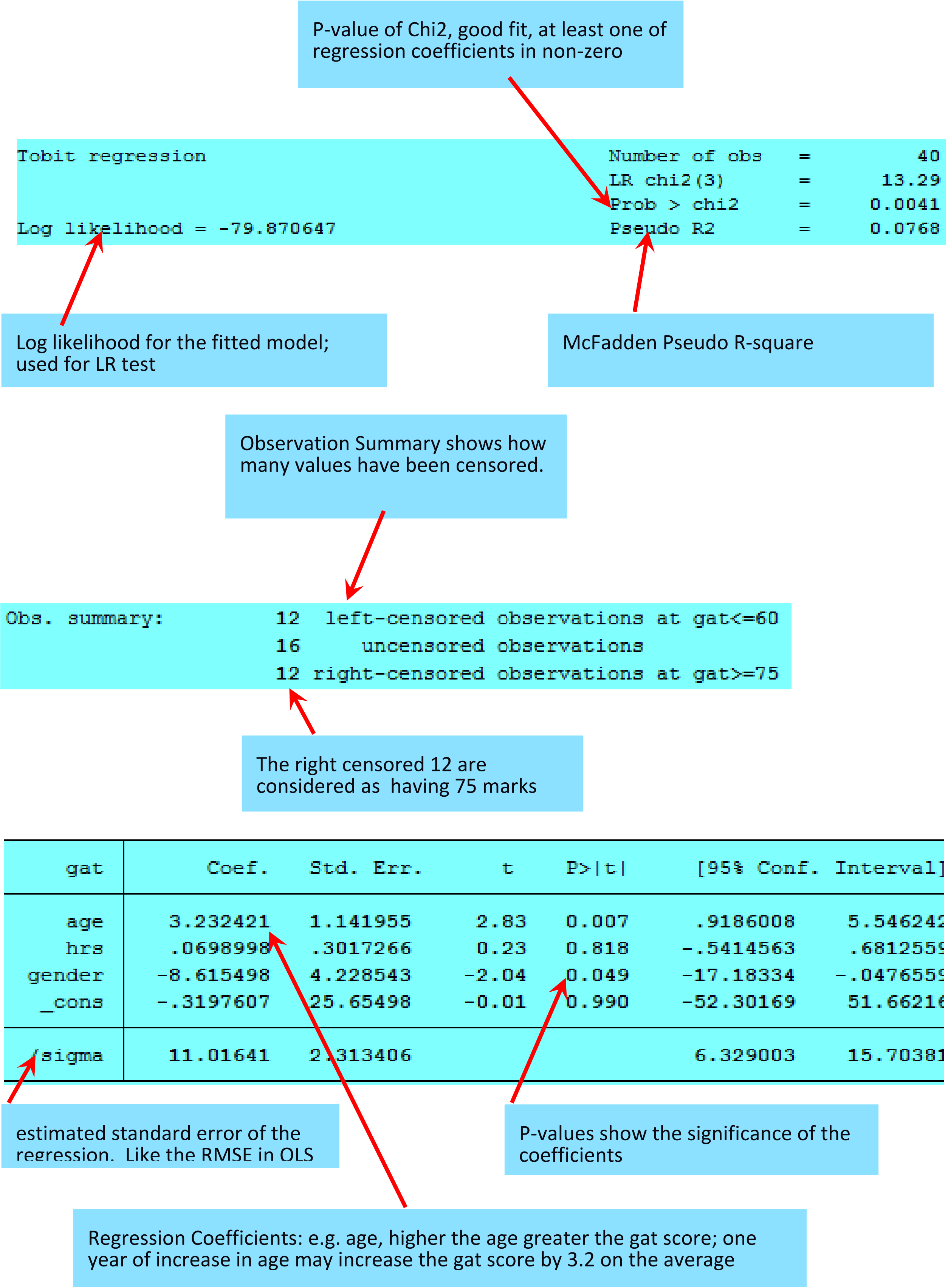
Here lower/left censoring is used at 60

We can also use both left and right censor. Even after censoring there are 40 observations!



```
. tobit gat age hrs gender, ll(60) ul(75)

Tobit regression                                  Number of obs   =         40
                                                  LR chi2(3)      =      13.29
                                                  Prob > chi2     =     0.0041
Log likelihood = -79.870647                       Pseudo R2       =     0.0768

------------------------------------------------------------------------------
         gat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   3.232421   1.141955     2.83   0.007     .9186008    5.546242
         hrs |   .0698998   .3017266     0.23   0.818    -.5414563    .6812559
      gender |  -8.615498   4.228543    -2.04   0.049    -17.18334   -.0476559
       _cons |  -.3197607   25.65498    -0.01   0.990    -52.30169    51.66216
-------------+----------------------------------------------------------------
      /sigma |   11.01641   2.313406                      6.329003    15.70381
------------------------------------------------------------------------------
  Obs. summary:         12  left-censored observations at gat<=60
                        16     uncensored observations
                        12 right-censored observations at gat>=75
```

Now let us look at various parts of the result using the following diagramatic representation.



P-value of Chi2, good fit, at least one of regression coefficients in non-zero

```
Tobit regression                                  Number of obs   =         40
                                                  LR chi2(3)      =      13.29
                                                  Prob > chi2     =     0.0041
Log likelihood = -79.870647                       Pseudo R2       =     0.0768
```

Log likelihood for the fitted model; used for LR test

McFadden Pseudo R-square

Observation Summary shows how many values have been censored.

```
Obs. summary:          12  left-censored observations at gat<=60
                       16     uncensored observations
                       12 right-censored observations at gat>=75
```

The right censored 12 are considered as having 75 marks

```
         gat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   3.232421   1.141955     2.83   0.007     .9186008    5.546242
         hrs |   .0698998   .3017266     0.23   0.818    -.5414563    .6812559
      gender |  -8.615498   4.228543    -2.04   0.049    -17.18334   -.0476559
       _cons |  -.3197607   25.65498    -0.01   0.990    -52.30169    51.66216
-------------+----------------------------------------------------------------
      /sigma |   11.01641   2.313406                      6.329003    15.70381
```

estimated standard error of the regression. Like the RMSE in OLS

P-values show the significance of the coefficients

Regression Coefficients: e.g. age, higher the age greater the gat score; one year of increase in age may increase the gat score by 3.2 on the average

Lecture 41
Forecasting-I

Meaning and Types of Forecasting

- Variance of Unconditional and Conditional Forecasts
- Forecasting Performance

Meaning

Prediction: inference from laws of nature

Forecasting is more 'probabilistic'

Forecast is any statement about the future. Econometric forecast has some basis; A systematic procedure to predict future events.

In Econometrics forecasting can be defined as "Predicting a known or unknown value of a dependent variable based on known or unknown values of the independent variables by using an econometric model.

Fore – "in front" or "in advance"

Cast – dice, lots, spells, horoscopes are all cast

Good Forecasting requires that

- There are some patterns or regularities or trends to be captured
- These patterns or regularities or trends provide some information about the future
- The econometric model used can capture the regular trends

However

- Irregular variations and shocks cannot be captured or predicted or forecasted.

Methods of Forecasting

- Guess using some rules of thumbs
- Use leading indicators

- Extrapolation
- Time series models
- Econometric forecasting models

Forecasting: Never 100% accurate

There are things we know but we may not Incorporated in the model

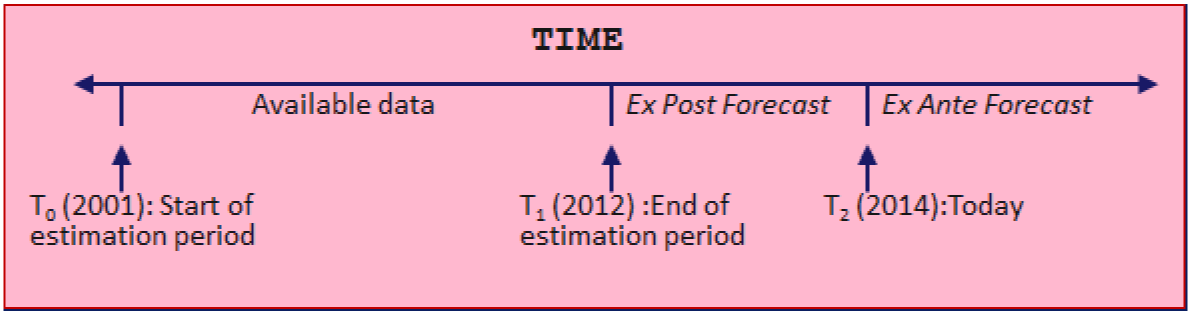
There are things we do not know which can be classified as

- *What we know that we do not know*
- *What we do not know that we do not know*

Another problem is using static models for dynamic world. Data problem also may influence the quality of forecast.

Types of Forecast

Ex Ante and Ex post forecast


TIME
Available data
Ex Post Forecast
Ex Ante Forecast
T_0 (2001): Start of estimation period
T_1 (2012) :End of estimation period
T_2 (2014):Today

Mean Forecast and Individual Forecast

- Individual Forecast: Forecasting the individual value of the dependent variable
- Mean Forecast: Forecasting the expected value of the dependent variable

Conditional and unconditional forecast

- Unconditional Forecast: Forecasting the dependent variable when the values of explanatory variables are known.
- Conditional Forecast: Forecasting dependent variable when the all values of explanatory variables are not known.

- Ex post forecasts are always unconditional
- Ex Ante forecasts may or may not be unconditional

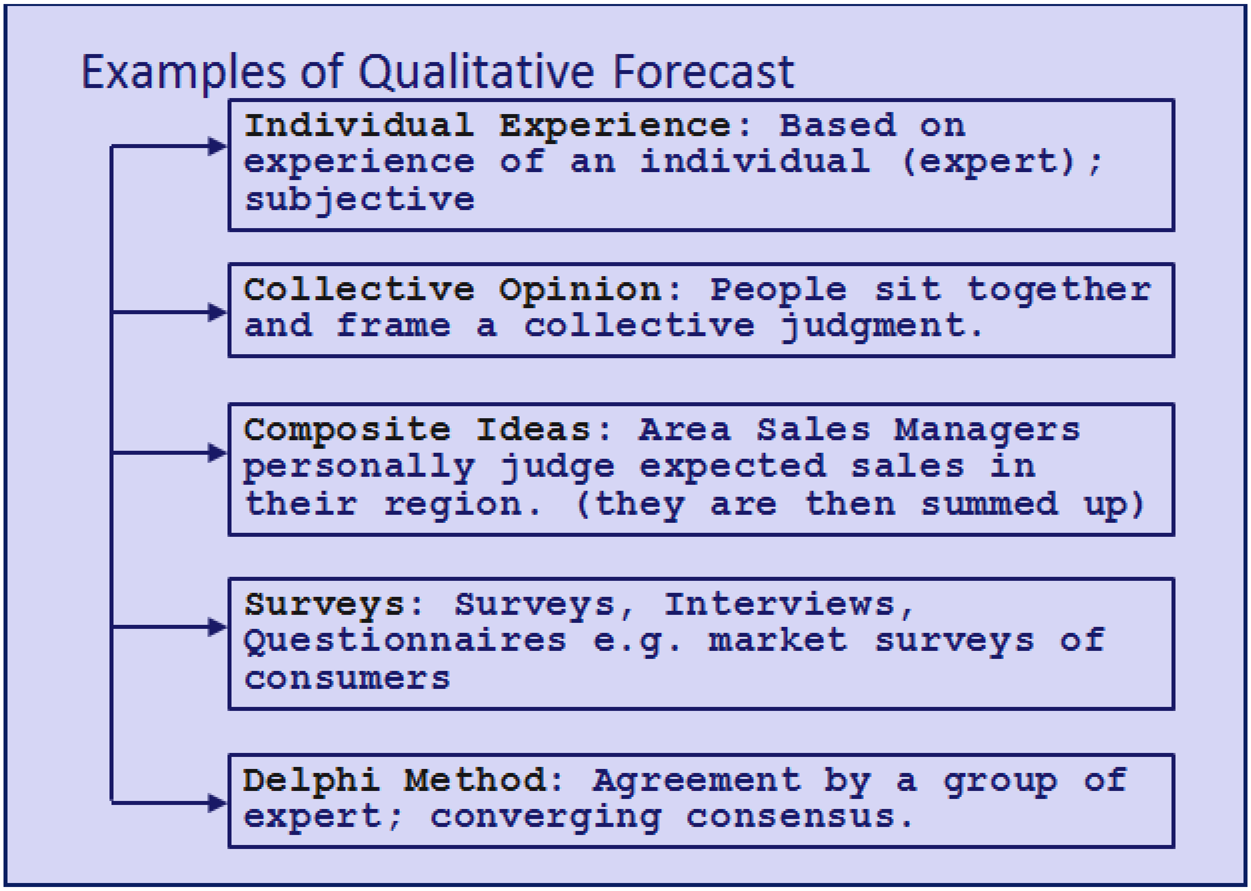
$$Y_t = f(X_{t-3}, X_{t-4})$$

Qualitative and Quantitative Forecast

Qualitative Forecast is based on opinion, emotion, personal experience, surveys etc. which is Subjective in nature while Quantitative Forecast: is based on Mathematical Models

Examples of forecasts

Qualitative Forecasts



Examples of Qualitative Forecast

Individual Experience: Based on experience of an individual (expert); subjective

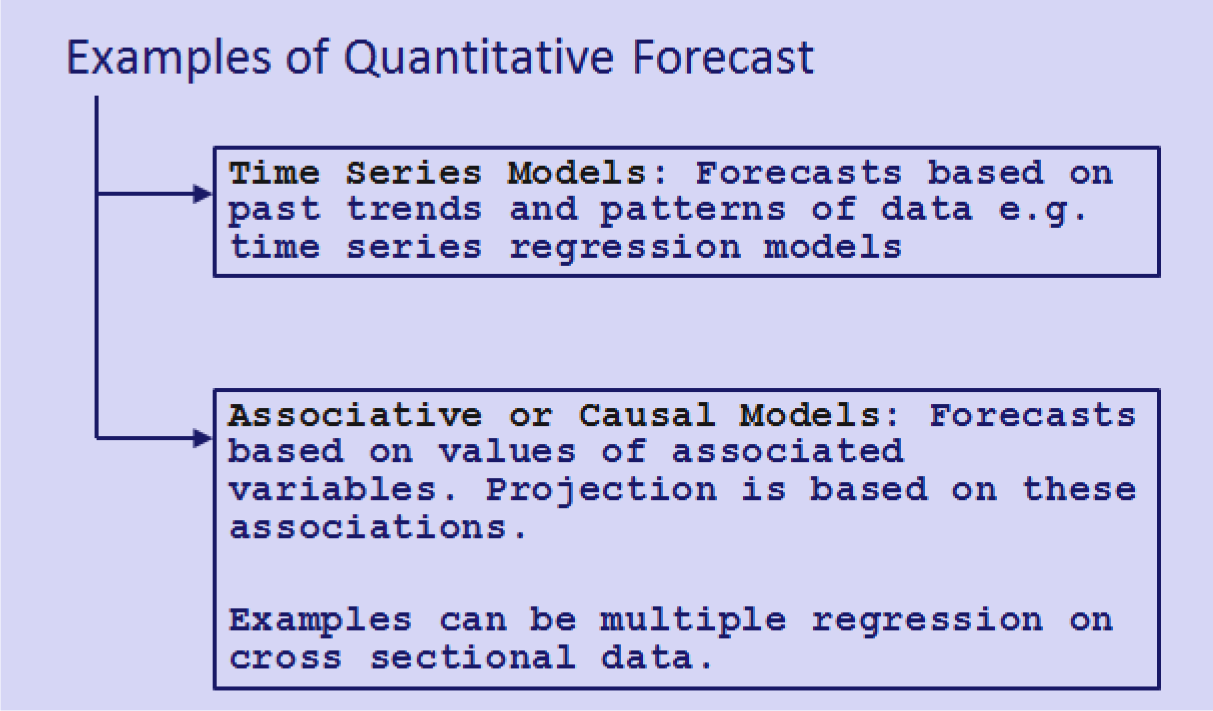
Collective Opinion: People sit together and frame a collective judgment.

Composite Ideas: Area Sales Managers personally judge expected sales in their region. (they are then summed up)

Surveys: Surveys, Interviews, Questionnaires e.g. market surveys of consumers

Delphi Method: Agreement by a group of expert; converging consensus.

Quantitative Forecast



Examples of Quantitative Forecast

Time Series Models: Forecasts based on past trends and patterns of data e.g. time series regression models

Associative or Causal Models: Forecasts based on values of associated variables. Projection is based on these associations.

Examples can be multiple regression on cross sectional data.

Patterns in Time Series Data

Trend: Steady growth or decline over time in the long run.

Seasonal Variations: Variations in data normally in different parts of the year (short run)

Cyclical Variation: upward or downward movement in the data over the medium term. (e.g. Business Cycle)

Irregular or Random Variations or Shocks: Unpredictable variation without any pattern (e.g. due to war, natural disasters etc.)

Time Series Models for forecasting

a) Last Period: Last period value as a forecast

b) Simple or Weighted Average: simple of weighted average (e.g. Arithmetic Mean) as a forecast

c) Simple or weighted Moving Averages: moving averages (e.g. three years, five years, four years centered etc.)

d) Seasonal Indices: Uses seasonal variation or seasonal patterns existing in the data

e) Exponential Smoothing: Weighted technique; more weight for recent observation

f) Trends: Uses least square method to fit a regression line to forecast

The examples for these models are given below.

Examples of Forecasting Techniques

Use data given in ***forecast.xlsx*** for the purpose of all examples. It is Demand Data (hypothetical)

Year	Quarter	Time Q	Demand	Price
2006	1	1	12	9.9
	2	2	18	9.2
	3	3	17	9.65
	4	4	14	9.8
2007	1	5	13	9.75
	2	6	18	9.4
	3	7	17	9.55
	4	8	14	9.4
Data continues till 2013				

with Quarterly Time series from 2006 to 2013 but We convert data to annual.

year	time	Demand	price (sums)	price (average)
2006	1	61	38.55	9.6375
2007	2	62	38.1	9.525
2008	3	66	38.8	9.7
2009	4	68	38.5	9.625
2010	5	74	38.2	9.55
2011	6	74	38.9	9.725
2012	7	83	37.45	9.3625
2013	8	85	37.55	9.3875
Reshaped / Transformed Data				

Last Period Method considers the value of the previous time period as a forecast for the current period.

Last Period Method

Description: We simply use the value of the last period as a forecast

year	time	Demand	Forecasted Demand$_{(t-1)}$	Remarks
2006	1	61	-	
2007	2	62	61	previous value
2008	3	66	62	
2009	4	68	66	
2010	5	74	68	
2011	6	74	74	
2012	7	83	74	
2013	8	85	83	
2014		Forecast	85	

Simple Average Method uses the average of previous time periods as a forecast. Starting from the second period i.e. 2007, the first value is just the previous one. Next time the average of 2006 and 2007 is used as a forecast for 2008. For 2009 the average of 2006 to 2008 is used and this continues in the same way.

Simple Average Method

Description:the forecast for the next period is the average of the previous demands

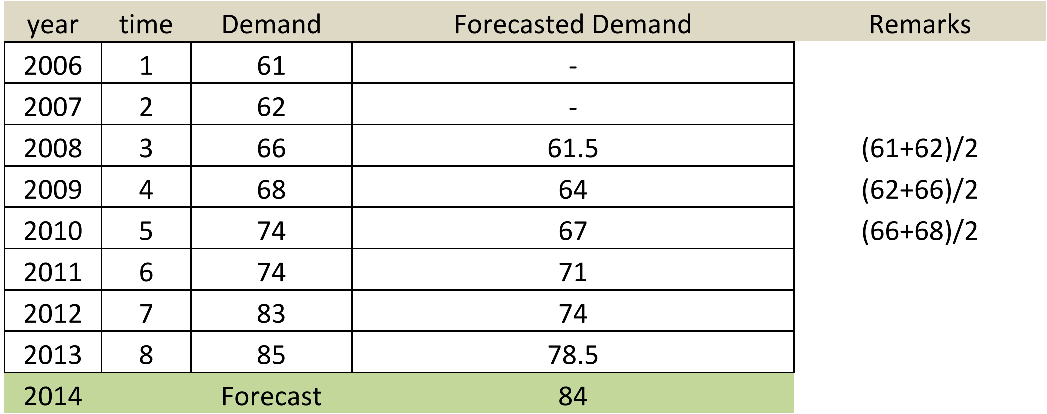
year	time	Demand	Forecasted Demand	Remarks
2006	1	61	-	
2007	2	62	61	
2008	3	66	61.5	(61+62)/2
2009	4	68	63	(61+62+66)/3
2010	5	74	64.25	(61+62+66+68)/4
2011	6	74	66.2	
2012	7	83	67.5	
2013	8	85	69.71428571	
2014		Forecast	71.625	

Moving Average Method uses the moving average of previous (fixed e.g. 2) time periods. For example in 2008 the average demand of 2006 and 2007 is used. The years change dynamically e.g. for 2009, the average demand of 2007 and 2008 is used as a forecast.

Moving Average Method

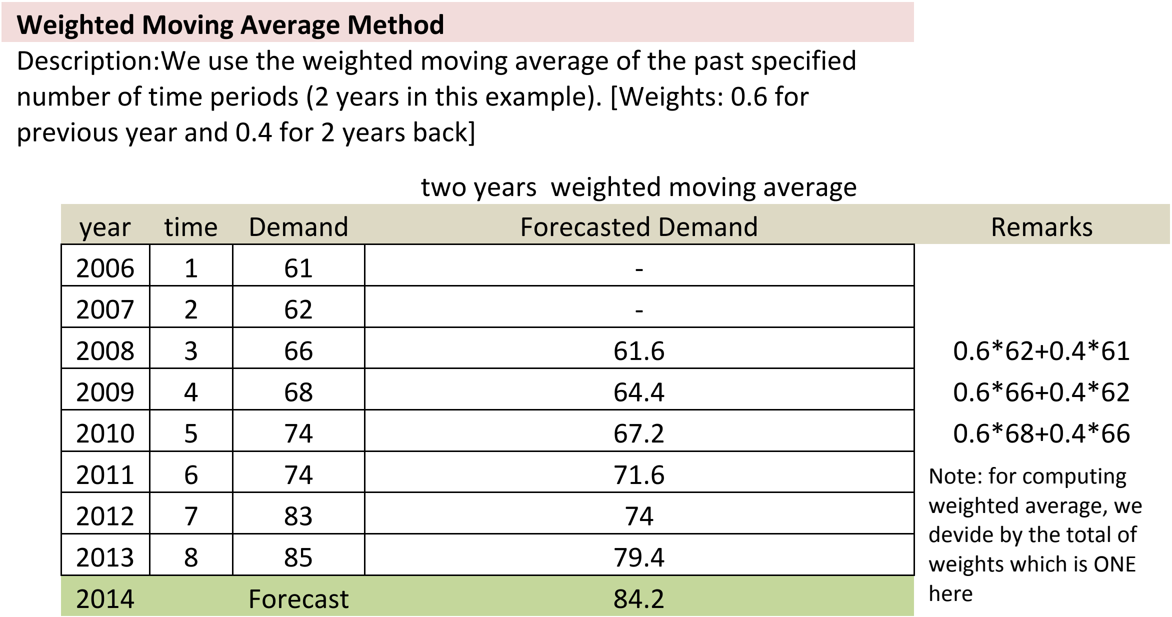
Description:We use the moving average of the past (2 in this example) periods of time

two years moving average



year	time	Demand	Forecasted Demand	Remarks
2006	1	61	-	
2007	2	62	-	
2008	3	66	61.5	(61+62)/2
2009	4	68	64	(62+66)/2
2010	5	74	67	(66+68)/2
2011	6	74	71	
2012	7	83	74	
2013	8	85	78.5	
2014		Forecast	84	

Weighted Moving Average method is similar to the moving average method except that some weights are assigned to the values. In the example below, out of the two years that we consider for averaging, a weight of 0.6 is assigned to the first year and 0.4 to the second year.



Weighted Moving Average Method

Description:We use the weighted moving average of the past specified number of time periods (2 years in this example). [Weights: 0.6 for previous year and 0.4 for 2 years back]

two years weighted moving average

year	time	Demand	Forecasted Demand	Remarks
2006	1	61	-	
2007	2	62	-	
2008	3	66	61.6	0.6*62+0.4*61
2009	4	68	64.4	0.6*66+0.4*62
2010	5	74	67.2	0.6*68+0.4*66
2011	6	74	71.6	Note: for computing weighted average, we devide by the total of weights which is ONE here
2012	7	83	74	
2013	8	85	79.4	
2014		Forecast	84.2	

Exponential Smoothing Method may be selected by minimizing the mean square error or any other technique. This is a forecasting technique and is different form the exponential smoothing. Do not be confused!

Exponential Smoothing Method

Description: The forecast of a period is calculated as the last periods forecast plus a factor multiplied by the difference of the last periods actual value and the forecasted value

$$F_t = F_{t-1} + \alpha(Y_{t-1} - F_{t-1})$$

Where alpha is a smoothing coefficient having values between zero and one

year	time	Demand	Forecasted Demand	Remarks
2006	1	61	-	
2007	2	62	61	first forecast is prev. value
2008	3	66	61.4	61+0.1(62-61)
2009	4	68	63.24	61.4+0.1(66-61.4)
2010	5	74	65.144	
2011	6	74	68.6864	
2012	7	83	70.81184	
2013	8	85	75.687104	
2014		Forecast	79.4122624	

smoothing constant may be selected by minimizing the mean square error or any other technique

This is a forecasting technique and is different form the exponential smoothing. Do not be confused!

Forecasting by Trend Method uses regression equation to forecast the values as we estimate trend values.

Forecasting by Trend

Description: use regression on the time trend to forecast

Y = a + b T gives **Y= 55.393 + 3.0607 T, use time value 9 for 2014**

			Trend	
year	time	Demand	Forecasted Demand	Remarks
2006	1	61	59	intercept
2007	2	62	62.60714286	55.39285714
2008	3	66	66.21428571	slope
2009	4	68	69.82142857	3.607142857
2010	5	74	73.42857143	
2011	6	74	77.03571429	
2012	7	83	80.64285714	
2013	8	85	84.25	
2014		Forecast	87.85714286	55.393 + 3.607 (9)

Associative Causal Forecast depends on the values of the associated variables. It is similar to the trend method but here the demand may depend on various variables in the system instead of being dependent on time.

Associative or Causal Forecast

Description: Forecast depends on the values of the associated variables e.g a regression line where demand may depend on price. This is possible in cross section as well.

Demand = a + b Price; this gives 500.49 - 44.84 Price

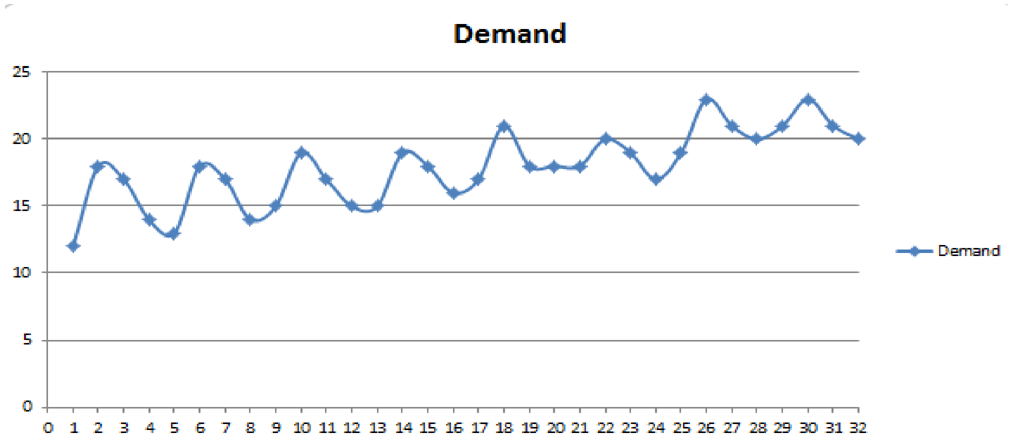
Demand = a + b Price

year	time	Demand	price (average)	Trend	Remarks
2006	1	61	9.6375	68.33200062	intercept
2007	2	62	9.525	73.37659542	500.485622
2008	3	66	9.7	65.52944795	
2009	4	68	9.625	68.89251115	slope
2010	5	74	9.55	72.25557435	-44.84084269
2011	6	74	9.725	64.40842688	
2012	7	83	9.3625	80.66323235	
2013	8	85	9.3875	79.54221129	price(2014) = 9.3
2014			Forecast	83.46578502	500.49 - 44.84 (9.3)

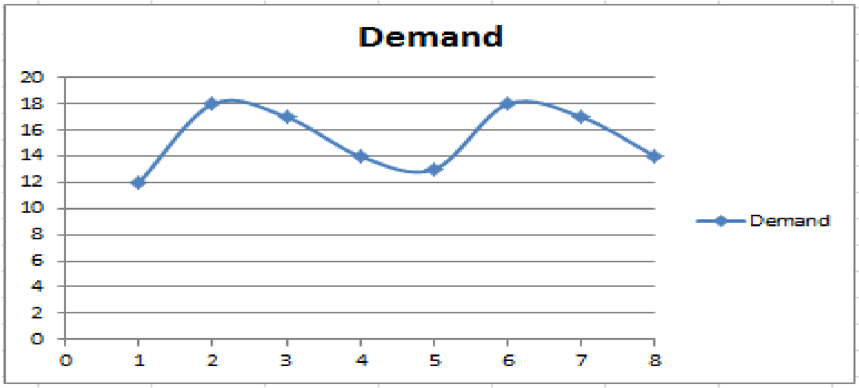
Lecture 42

Forecasting-II

Seasonal Variation and forecasting



- The plot shows a clear demonstration of seasonal variation
- Values are high in the second quarter



- Values are high in the second quarter in general
- Values are least in the first quarter and slightly higher in third and fourth quarter and the highest in the second quarter

Calculating Seasonal Index

The following example shows how to calculate seasonal indices.

First we need to reshape the data

Quarterly Demand

	Q1	Q2	Q3	Q4	Annual Demand
2006	12	18	17	14	61
2007	13	18	17	14	62
2008	15	19	17	15	66
2009	15	19	18	16	68
2010	17	21	18	18	74
2011	18	20	19	17	74
2012	19	23	21	20	83
2013	21	23	21	20	85
qaurterly averages					573
	16.25	20	18.5	16.75	71.625
				AVG	71.625/4=17.90625
Seasonal index	devide each quarterly average by average of the four averages				
	0.9075044	1.1	1.0331588	0.9354276	4

Average of annual demands is equal to total of the quarterly averages. Sum of seasonal index equals FOUR

Now we need to adjust the forecast of each quarter multiplying by the relevant index. Any forecasting method may be used

Example

Run a regression Demand on Time, Forecast the demand for four quarters of 2014, Adjust the forecasts by multiplying them to the relevant seasonal index

Regression of Demand on time is

$$Demand \ = \ 14.24395 \ + \ 0.221957 \ (Time)$$

Now we need to encoded time values for the four quarters of 2014 are 33,34,35 and 36

Unadjusted Forecasts		Adjusted forecasts		**computation**
2014-1	21.568531	2014-1	19.573536	21.56853 (0.907504)
2014-2	21.790488	2014-2	24.490531	21.79049 (1.123909)
2014-3	22.012445	2014-3	22.742352	22.01245 (1.033159)
2014-4	22.234402	2014-4	20.798673	22.2344 (0.935428)

Measuring Forecasting Performance

To see which forecasting technique works the best, we use various criteria. We look at the success of each method using performance measures.

Almost all the performance measures start by looking at the difference on the actual and forecasted values $Y_t - F_t$

This is called the **Forecast Error**.

This enables us to see the reliability of different econometric models.

Different methods or measures are commonly used.

These methods relate to the *Ex Post* Forecast as we know the actual values of the dependent variable.

Some of the measures are listed below:

1. Mean Forecast Error (MFE)
2. Mean Absolute Error (MAE) or Mean Absolute Deviation (MAD)
3. Mean Absolute Percent Error (MAPE)
4. Root Mean Square Error (RMSE) [its square is Mean Square Error (MSE)]
5. Root Mean Square Percent Error (RMSPE)
6. Theil's Inequality Coefficient (TIC)

Let us look at them one by one considering the following (taken from previous data)

We are considering (just as an example) the values of 2008-2013 from two methods: weighted moving average to illustrate the calculations and later we will compare this with the forecasting performance of the projection by trend.

Year	Actual	Forecast	Forecast Error
	Y	F	Y - F
2008	66	61.6	4.4
2009	68	64.4	3.6

2010	74	67.2	6.8
2011	74	71.6	2.4
2012	83	74	9
2013	85	79.4	5.6

Let T denote the number of observations (or number of forecasted values)

1. Mean Forecast Error (MFE)

It is the arithmetic mean of the forecast error. It may be sometimes misleading as negative and positive errors cancel each other. This is why it may be better to used MEA.

$$MFA = \frac{\sum(Y_t - F_t)}{T}$$

2. Mean Absolute Error (MAE) or Mean Absolute Deviation (MAD)

$$MAD = \frac{\sum|Y_t - F_t|}{T}$$

3. Mean Absolute Percent Error (MAPE)

$$MAPE = \frac{\sum(\frac{|Y_t - F_t|}{Y_t})}{T}$$

4. Root Mean Square Error (RMSE) [its square is Mean Square Error (MSE)]
   This gives more weight to larger observations. It is commonly used. The weak point is that its value will be larger if values of the dependent variable are large. In such cases it is better to use RMSPE.

$$RMSE = \sqrt{\frac{\sum(Y_t - F_t)^2}{T}}$$

5. Root Mean Square Percent Error (RMSPE)

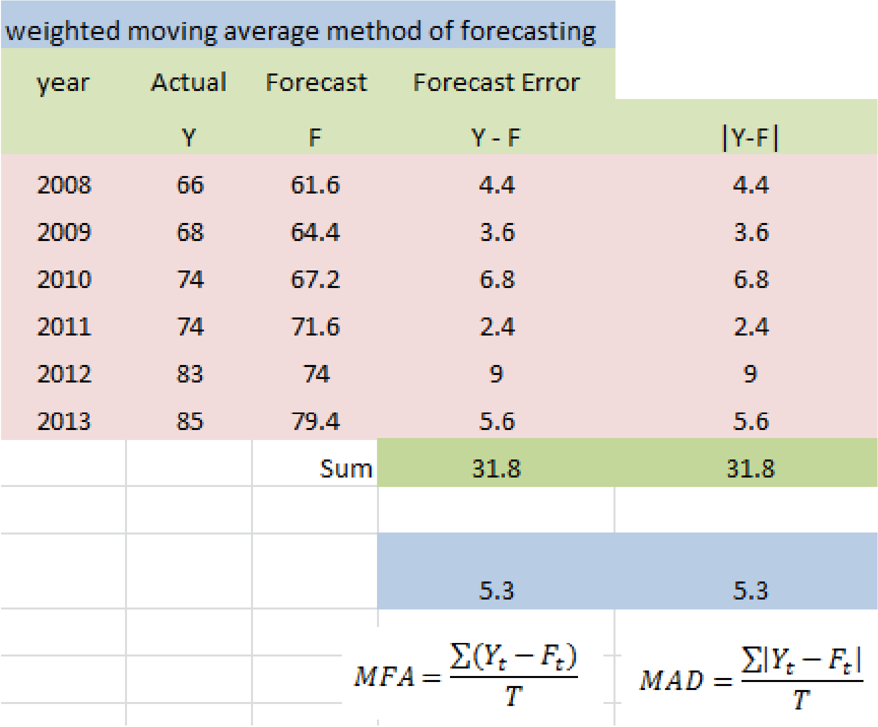
$$RMSPE = \sqrt{\frac{\sum(\frac{Y_t - F_t}{Y_t})^2}{T}}$$

6. Theil's Inequality Coefficient (TIC)

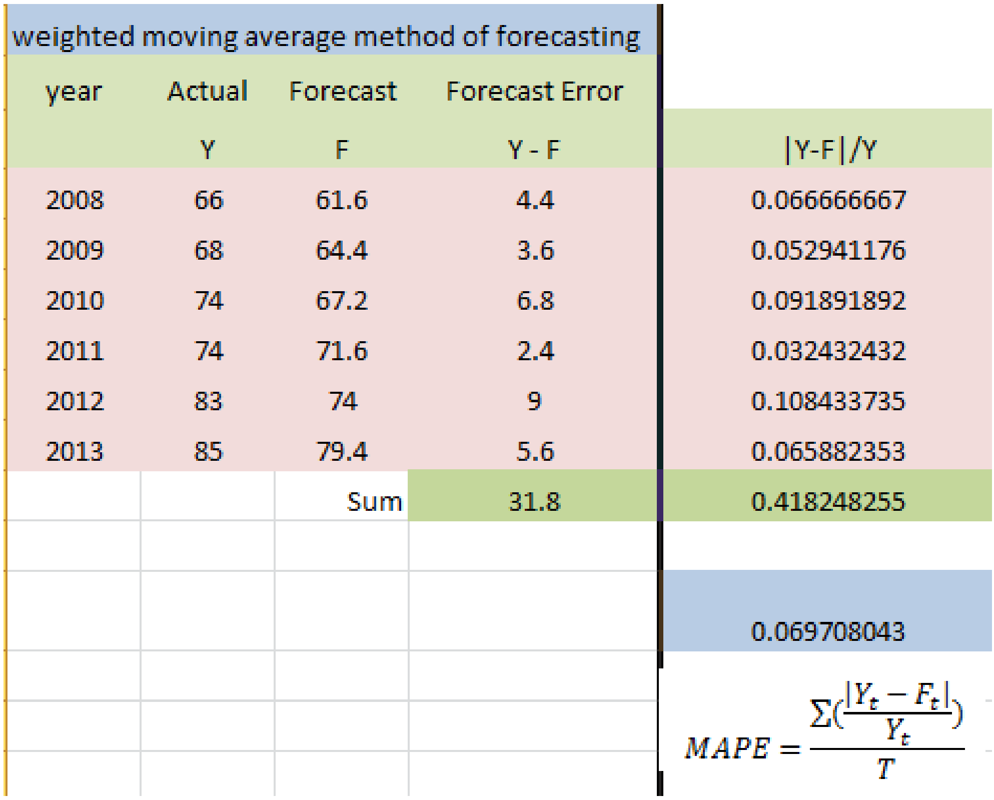
Concept was given by Theil in 1961.

$$U = \frac{\sqrt{\frac{1}{T}\sum(Y_t - F_t)^2}}{\sqrt{\frac{1}{T}\sum Y_t{}^2} + \sqrt{\frac{1}{T}\sum F_t{}^2}} = \frac{RMSE}{\sqrt{\frac{1}{T}\sum Y_t{}^2} + \sqrt{\frac{1}{T}\sum F_t{}^2}}$$

Example: Measuring Forecast Performance



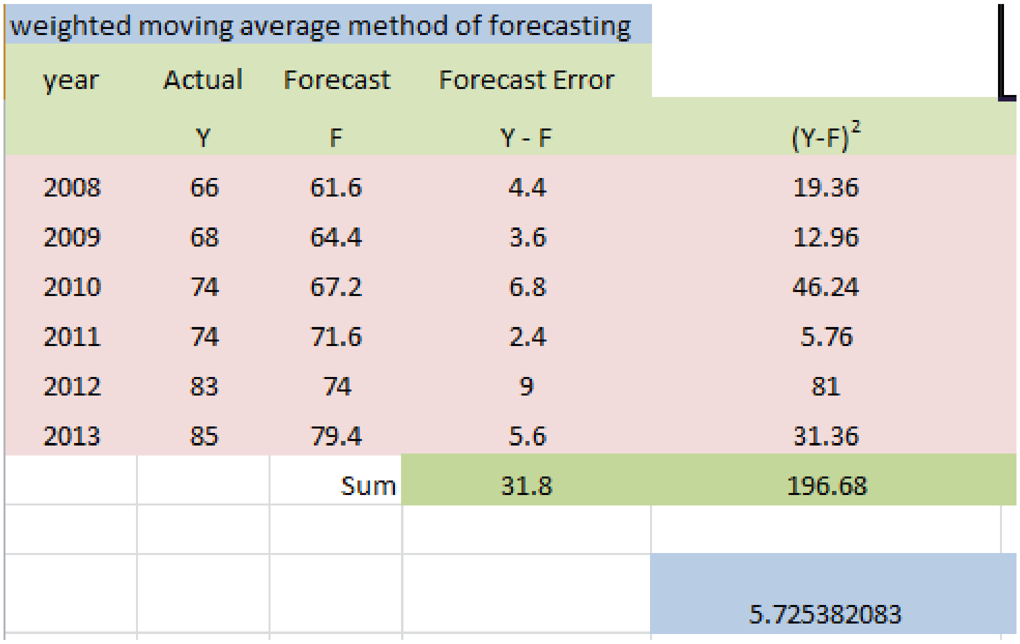
weighted moving average method of forecasting				
year	Actual	Forecast	Forecast Error	
	Y	F	Y - F	|Y-F|
2008	66	61.6	4.4	4.4
2009	68	64.4	3.6	3.6
2010	74	67.2	6.8	6.8
2011	74	71.6	2.4	2.4
2012	83	74	9	9
2013	85	79.4	5.6	5.6
		Sum	31.8	31.8
			5.3	5.3
			$MFA = \frac{\sum(Y_t - F_t)}{T}$	$MAD = \frac{\sum\lvert Y_t - F_t\rvert}{T}$



weighted moving average method of forecasting

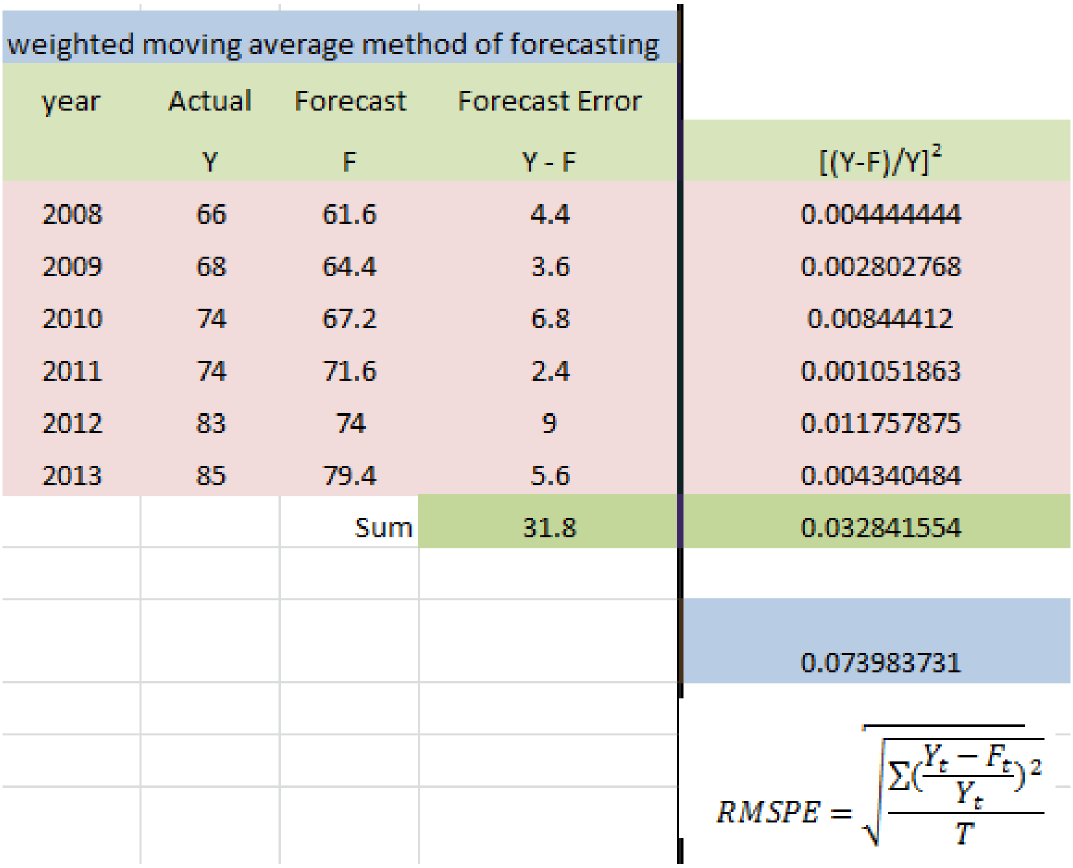
year	Actual	Forecast	Forecast Error	
	Y	F	Y - F	|Y-F|/Y
2008	66	61.6	4.4	0.066666667
2009	68	64.4	3.6	0.052941176
2010	74	67.2	6.8	0.091891892
2011	74	71.6	2.4	0.032432432
2012	83	74	9	0.108433735
2013	85	79.4	5.6	0.065882353
		Sum	31.8	0.418248255
				0.069708043

$$MAPE = \frac{\sum\left(\frac{|Y_t - F_t|}{Y_t}\right)}{T}$$



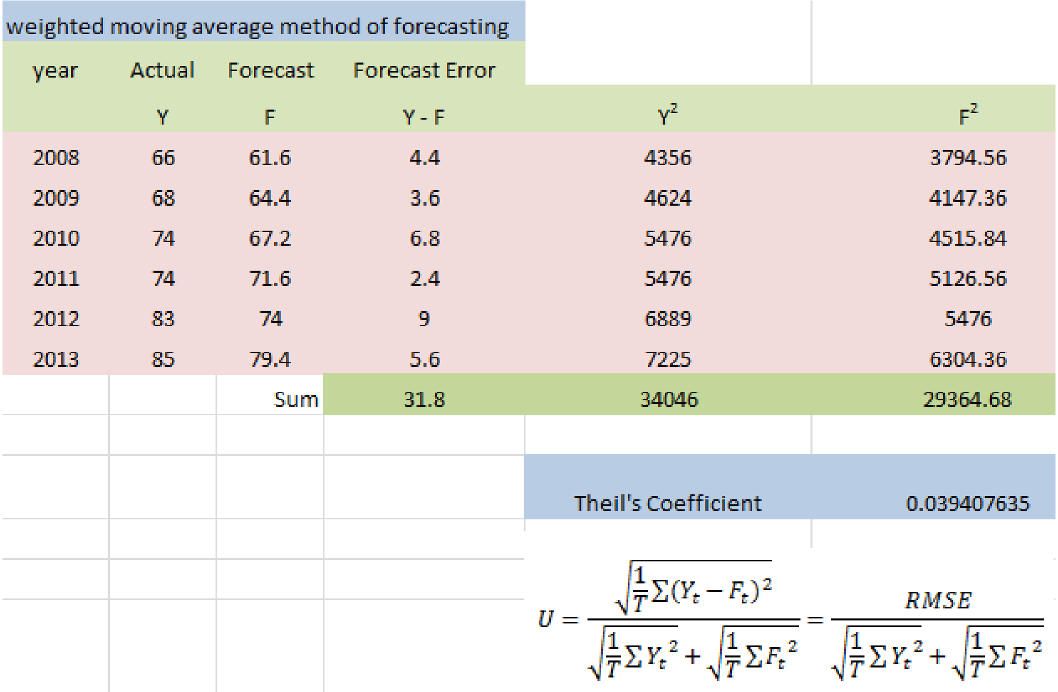
weighted moving average method of forecasting

year	Actual	Forecast	Forecast Error	
	Y	F	Y - F	$(Y-F)^2$
2008	66	61.6	4.4	19.36
2009	68	64.4	3.6	12.96
2010	74	67.2	6.8	46.24
2011	74	71.6	2.4	5.76
2012	83	74	9	81
2013	85	79.4	5.6	31.36
		Sum	31.8	196.68
				5.725382083



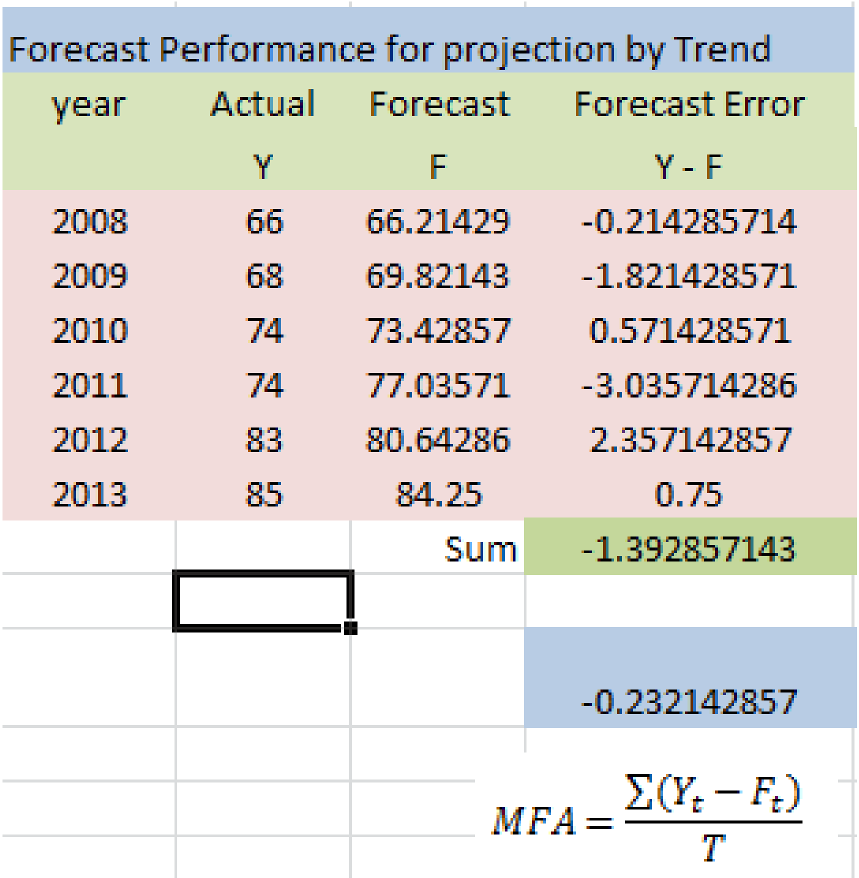
weighted moving average method of forecasting				
year	Actual	Forecast	Forecast Error	
	Y	F	Y - F	$[(Y-F)/Y]^2$
2008	66	61.6	4.4	0.004444444
2009	68	64.4	3.6	0.002802768
2010	74	67.2	6.8	0.00844412
2011	74	71.6	2.4	0.001051863
2012	83	74	9	0.011757875
2013	85	79.4	5.6	0.004340484
		Sum	31.8	0.032841554
				0.073983731

$$RMSPE = \sqrt{\frac{\Sigma(\frac{Y_t - F_t}{Y_t})^2}{T}}$$



weighted moving average method of forecasting					
year	Actual	Forecast	Forecast Error		
	Y	F	Y - F	Y^2	F^2
2008	66	61.6	4.4	4356	3794.56
2009	68	64.4	3.6	4624	4147.36
2010	74	67.2	6.8	5476	4515.84
2011	74	71.6	2.4	5476	5126.56
2012	83	74	9	6889	5476
2013	85	79.4	5.6	7225	6304.36
		Sum	31.8	34046	29364.68
				Theil's Coefficient	0.039407635

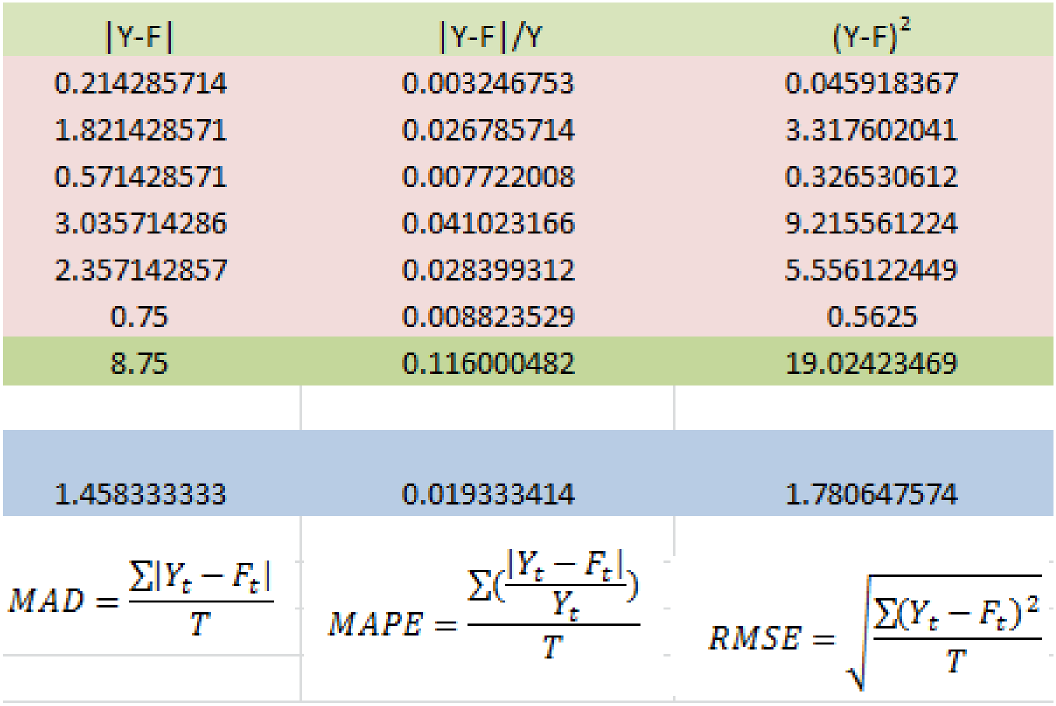
$$U = \frac{\sqrt{\frac{1}{T}\Sigma(Y_t - F_t)^2}}{\sqrt{\frac{1}{T}\Sigma Y_t{}^2} + \sqrt{\frac{1}{T}\Sigma F_t{}^2}} = \frac{RMSE}{\sqrt{\frac{1}{T}\Sigma Y_t{}^2} + \sqrt{\frac{1}{T}\Sigma F_t{}^2}}$$



Forecast Performance for projection by Trend

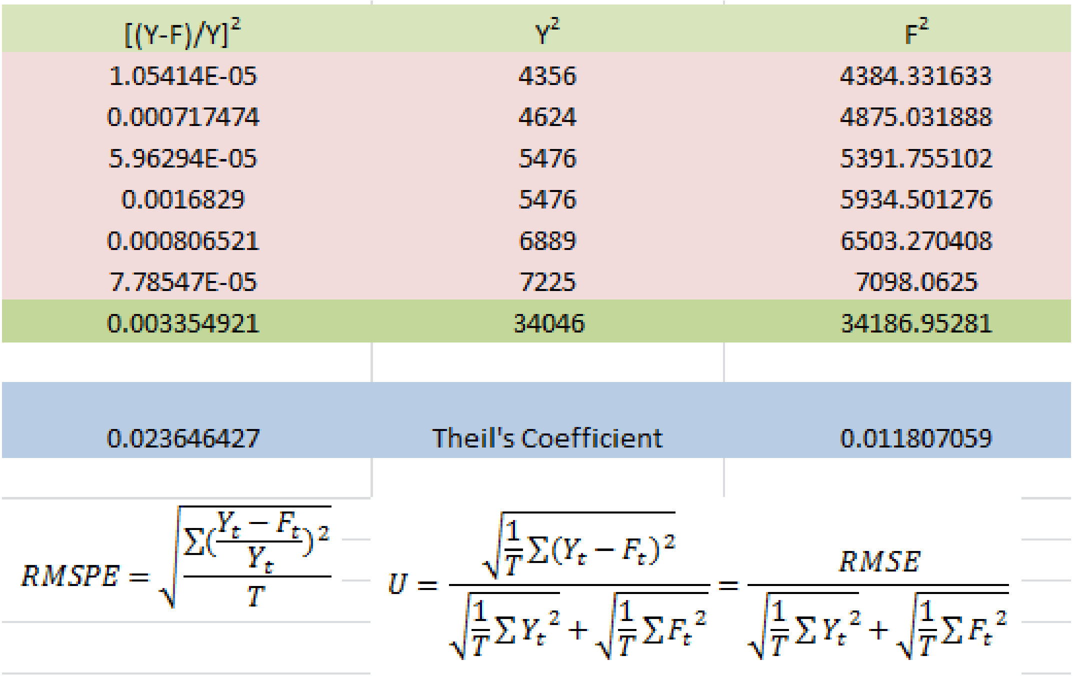
year	Actual	Forecast	Forecast Error
	Y	F	Y - F
2008	66	66.21429	-0.214285714
2009	68	69.82143	-1.821428571
2010	74	73.42857	0.571428571
2011	74	77.03571	-3.035714286
2012	83	80.64286	2.357142857
2013	85	84.25	0.75
		Sum	-1.392857143
			-0.232142857

$$MFA = \frac{\sum(Y_t - F_t)}{T}$$



|Y-F|	|Y-F|/Y	$(Y\text{-}F)^2$
0.214285714	0.003246753	0.045918367
1.821428571	0.026785714	3.317602041
0.571428571	0.007722008	0.326530612
3.035714286	0.041023166	9.215561224
2.357142857	0.028399312	5.556122449
0.75	0.008823529	0.5625
8.75	0.116000482	19.02423469
1.458333333	0.019333414	1.780647574

$$MAD = \frac{\sum|Y_t - F_t|}{T} \qquad MAPE = \frac{\sum(\frac{|Y_t - F_t|}{Y_t})}{T} \qquad RMSE = \sqrt{\frac{\sum(Y_t - F_t)^2}{T}}$$

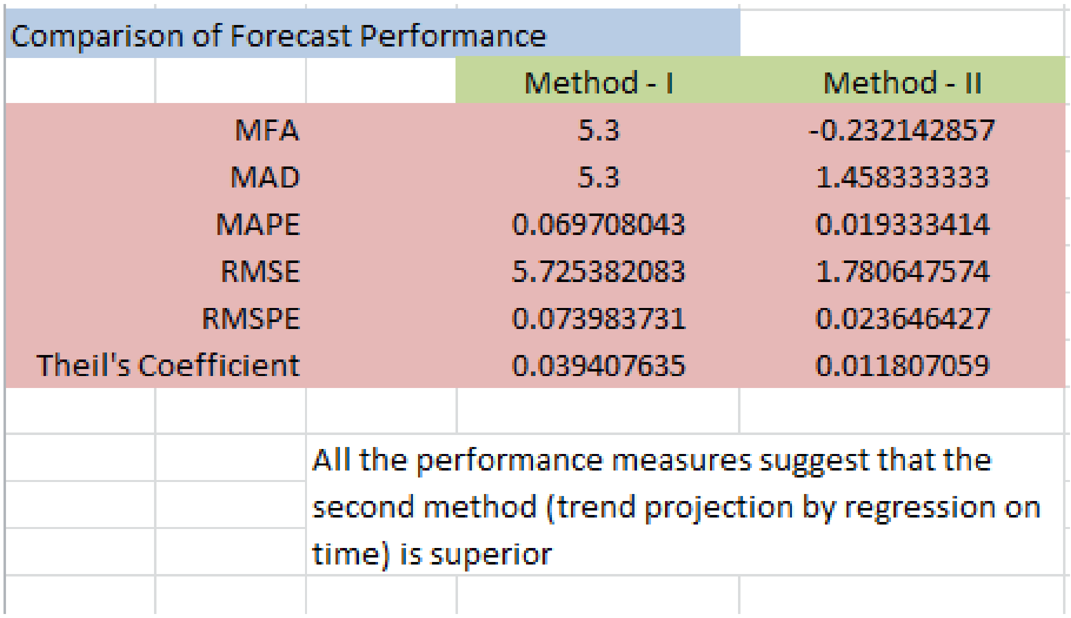


[(Y-F)/Y]^2	Y^2	F^2
1.05414E-05	4356	4384.331633
0.000717474	4624	4875.031888
5.96294E-05	5476	5391.755102
0.0016829	5476	5934.501276
0.000806521	6889	6503.270408
7.78547E-05	7225	7098.0625
0.003354921	34046	34186.95281

0.023646427	Theil's Coefficient	0.011807059

$$RMSPE = \sqrt{\frac{\Sigma(\frac{Y_t - F_t}{Y_t})^2}{T}}$$

$$U = \frac{\sqrt{\frac{1}{T}\Sigma(Y_t - F_t)^2}}{\sqrt{\frac{1}{T}\Sigma Y_t{}^2} + \sqrt{\frac{1}{T}\Sigma F_t{}^2}} = \frac{RMSE}{\sqrt{\frac{1}{T}\Sigma Y_t{}^2} + \sqrt{\frac{1}{T}\Sigma F_t{}^2}}$$



Comparison of Forecast Performance

	Method - I	Method - II
MFA	5.3	-0.232142857
MAD	5.3	1.458333333
MAPE	0.069708043	0.019333414
RMSE	5.725382083	1.780647574
RMSPE	0.073983731	0.023646427
Theil's Coefficient	0.039407635	0.011807059

All the performance measures suggest that the second method (trend projection by regression on time) is superior

Tracking Signals in Forecasting

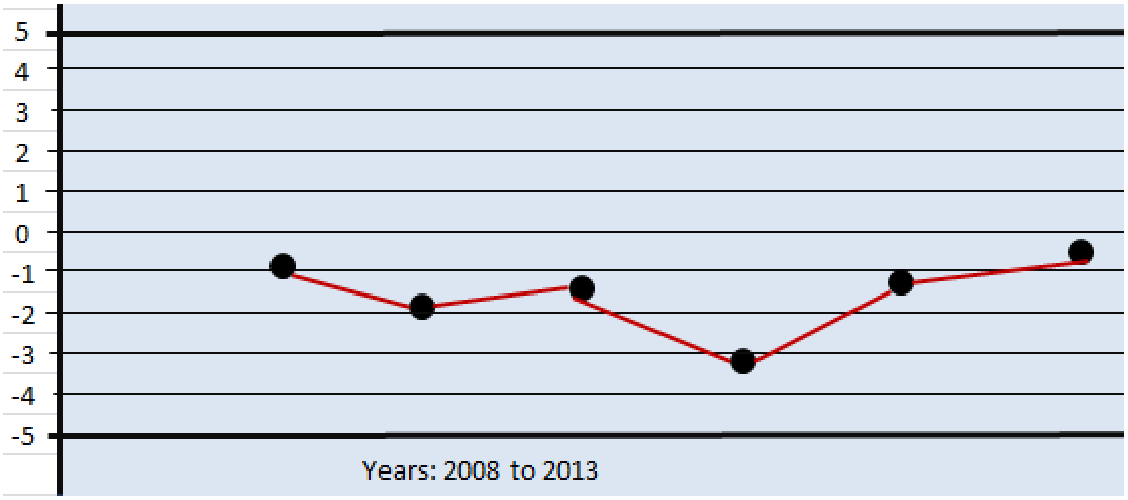
Tracking signals are used to monitor the accuracy of the forecast over time. An acceptable range according to situation is decided and we see if the tracking signal fall within this 'channel'

$$TS \;=\; cumulative\ error/MAD$$

	Y	F	Y - F	cumulative error
2008	66	66.21428571	-0.214285714	-0.214285714
2009	68	69.82142857	-1.821428571	-2.035714286
2010	74	73.42857143	0.571428571	-1.464285714
2011	74	77.03571429	-3.035714286	-4.5
2012	83	80.64285714	2.357142857	-2.142857143
2013	85	84.25	0.75	-1.392857143

|Y-F|	cumulative |Y-F|	MAD (till point)	TS=cum. Error/MAD
0.214285714	0.214285714	0.214285714	-1
1.821428571	2.035714286	1.017857143	-2
0.571428571	2.607142857	0.869047619	-1.684931507
3.035714286	5.642857143	1.410714286	-3.189873418
2.357142857	8	1.6	-1.339285714
0.75	8.75	1.458333333	-0.955102041

Let us select a safe range of -5 to 5 and plot the TSs. In practice each point will be plotted each year when data is available.



Variance of Forecast Error

Consider the following multiple regression model.

$$Y_{t+h} = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ \dots\dots.. + \beta_K X_{Kt} + e_t$$

Its OLS estimate is

$$\boldsymbol{Y_{t+h}} = \boldsymbol{\widehat{\beta}_0} + \boldsymbol{\widehat{\beta}_1 X_{1t}} + \boldsymbol{\widehat{\beta}_2 X_{2t}} + \ \dots\dots.. + \boldsymbol{\widehat{\beta}_K X_{Kt}} + \boldsymbol{\hat{e}_t}$$

And

$$Var(\boldsymbol{\widehat{Y}_{t+h}}) = Var(\boldsymbol{\widehat{\beta}_0} + \boldsymbol{\widehat{\beta}_1 X_{1t}} + \boldsymbol{\widehat{\beta}_2 X_{2t}} \dots.. + \boldsymbol{\widehat{\beta}_K X_{Kt}})$$

Predicted Standard Error is

$$se(\hat{Y}_{t+h}) = \sqrt{Var(\boldsymbol{\widehat{Y}_{t+h}})}$$

Computed in Stata by as: $predict\ s, stdp$

- Forecast Error is

$$\hat{e}_{t+h} = (Y_{t+h} - \hat{Y}_{t+h})$$

Variance of Forecast Error is

$$Var(\hat{e}_{t+h}) = (Var(Y_{t+h}) + Var(\hat{Y}_{t+h}))$$
$$= \sigma^2 + Var(\hat{Y}_{t+h})$$

Which has two components:

- Model variance (σ^2) that is larger than estimated variance
- Estimated Variance, $Var(\hat{Y}_{t+h})$ that tends to decrease with increase in T.

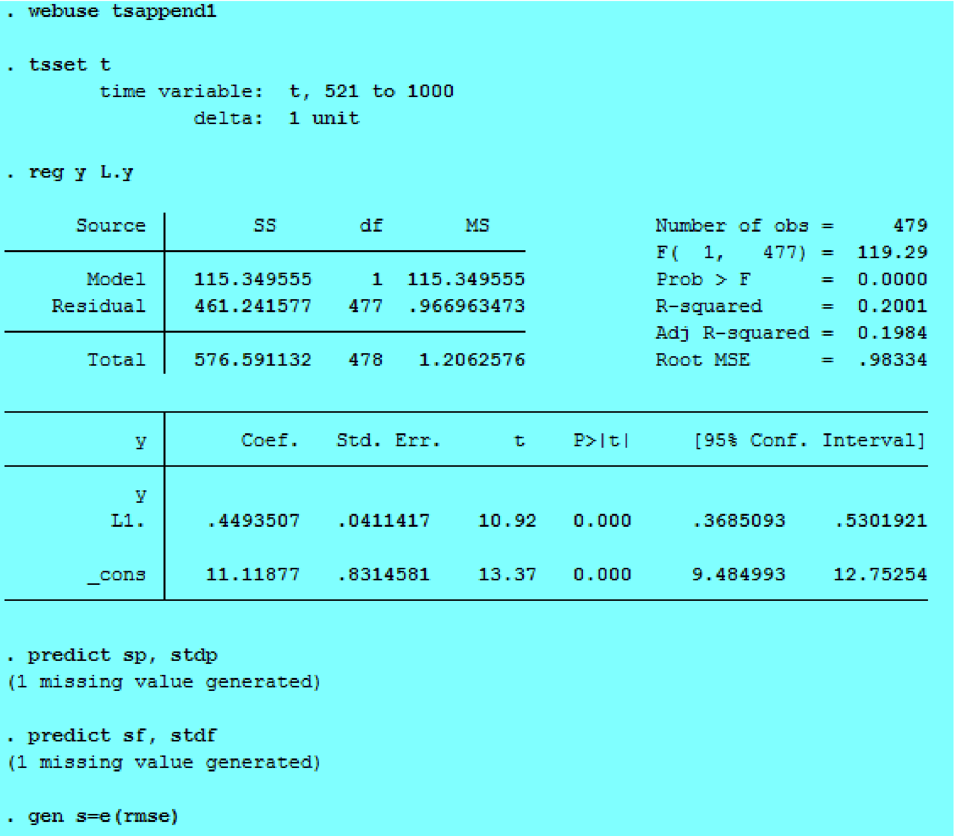
As

$$Var(\hat{e}_{t+h}) = \sigma^2 + Var(\hat{Y}_{t+h})$$

$$se(\hat{e}_{t+h}) = \sqrt{\sigma^2 + se(\hat{Y}_{t+h})^2}$$

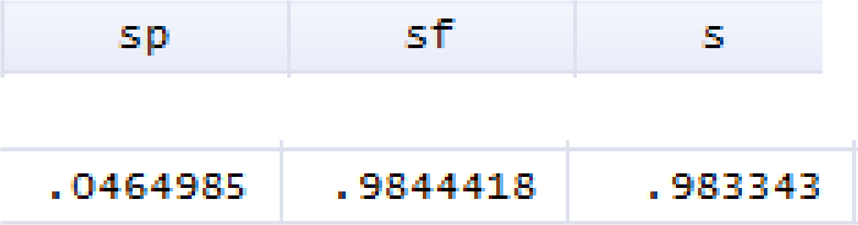
Which can be computed in Stata as: $predict\ s, stdf$

Consider the following example for explaining the concept discussed above:



```
. webuse tsappend1

. tsset t
        time variable:  t, 521 to 1000
                delta:  1 unit

. reg y L.y

      Source |       SS       df       MS              Number of obs =     479
-------------+------------------------------           F(  1,   477) =  119.29
       Model |  115.349555     1  115.349555           Prob > F      =  0.0000
    Residual |  461.241577   477  .966963473           R-squared     =  0.2001
-------------+------------------------------           Adj R-squared =  0.1984
       Total |  576.591132   478   1.2062576           Root MSE      =  .98334

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           y |
         L1. |   .4493507   .0411417    10.92   0.000     .3685093    .5301921
             |
       _cons |   11.11877   .8314581    13.37   0.000     9.484993    12.75254
------------------------------------------------------------------------------

. predict sp, stdp
(1 missing value generated)

. predict sf, stdf
(1 missing value generated)

. gen s=e(rmse)
```

From Data Editor



sp	sf	s
.0464985	.9844418	.983343

We can verify that $sf^2 = \sqrt{sp^2 + s^2}$

Lecture 43

Time Series, Cointegration and Error Correction-I

What is a time series?

An analysis of a single sequence of data is called univariate time-series analysis. An analysis of several sets of data for the same sequence of time periods is called multivariate time-series analysis or, more simply, multiple time-series analysis. For a long time there has been very little communication between econometricians and time-series analysts. Econometricians have emphasized economic theory and a study of contemporaneous relationships. Lagged variables were introduced but not in a systematic way, and no serious attempts were made to study the temporal structure of the data

Theories were imposed on the data even when the temporal structure of the data was not in conformity with the theories. The time-series analysts, on the other hand, did not believe in economic theories and thought that they were better off allowing the data to determine the model. Since the mid-1970s these two approaches—the time-series approach and the econometric approach—have been converging

Econometricians now use some of the basic elements of time-series analysis in checking the specification of their econometric models, and some economic theories have influenced the direction of time-series work.

Stationary and Nonstationary time series

Moments: a **moment** is a specific quantitative measure of the shape of a set of points. Moments are of different types e.g. raw, about mean etc.

The first raw moment ($\frac{\sum X}{n}$; do not confuse with the moments about mean) is the mean. First moment about mean is zero. Second moment about mean is the variance.

Joint distribution of $X(t_1), X(t_2), X(t_3) \ldots X(t_n)$ is complicated so we usually define the

$$mean = \mu(X) = E(X_t)$$

$$Variance = \sigma^2(t) = E(X_t^2)$$

$$AutoCovariance = \gamma(t1, t2) = Cov(X_{t1}, X_{t2})$$

The autocovariance is the covariance of the variable against a time-shifted version of itself. (is the variance if t1=t2=t)

Strict or Strong Stationary: a series is called strict stationary if the joint distribution of any set of n observations $X(t_1), X(t_2), X(t_3) \ldots X(t_n)$ is the same as joint distribution of $X(t_1 + h), X(t_2 + h), X(t_3 + h) \ldots X(t_n + h)$ for all n and h.

All moments are independent of t. The mean, variance, and all higher order moments of the joint distribution of any combination of variables $X(t_1), X(t_2), X(t_3) \ldots X(t_n)$ are constant & independent of t. (Very strong assumption)

Statistical Properties do not change over time.

Now define $h = t_2 - t_1$and call it a **lag**.

As joint distribution of $X(t_1), X(t_2)$ and that of $X(t_1 + h), X(t_2 + h)$ are the same and not dependent on t_1 or t_2 ($t_2 = \; t_1 + h$) but on the difference$h$.

So autocovariance function $\gamma(t1, t2) = \gamma(h) = Cov\ (X_t, X_{t+h})$ is the autocovariance function ($acvf$) at lag h.

$Var\ (X_t) = Var(X_{t+h}) = \sigma^2 = \gamma(0)$

So the autocorrelation function at lag (h), is defined as $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

A plot of $\rho(h)$ is called correlogram.

Second Order or Weak Stationary (covariance stationarity): a series is called weakly stationary if its mean and variance do not depend on t (are constant) and the $acvf$ depends only on the lag h.

Stationary series revert in the long run to their mean.

Non-Stationarity: In real life most of the data is non-stationary. Variables that increase over time are non-stationary.

Mean and variance are not constant over time.

We need differencing and/or detrending.

We can define a non-stationary model as $X_t = \mu_t + e_t$ where mean μ_t is a function of time (linear or non-linear) and e_t is second order stationary series.

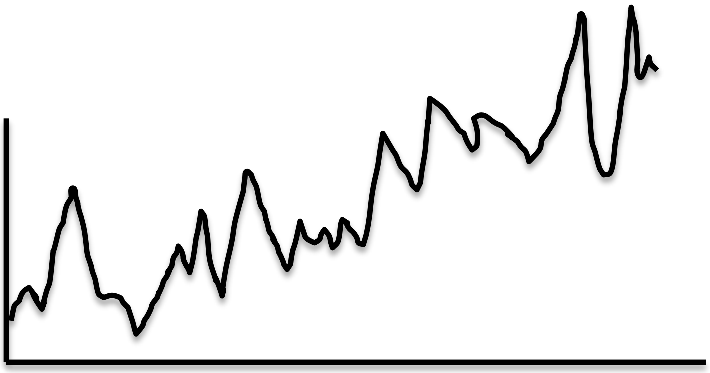
Autoregressive Model specifies that the output variable depends linearly on its own previous values.

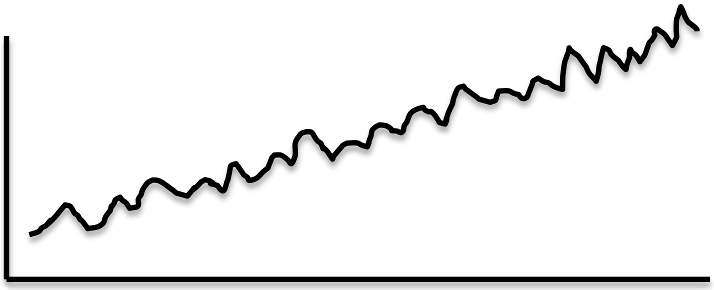
Time series models include ARMA, ARIMA (auto regressive integrated moving average) etc. (we are not going to discuss them). OLS should not be used on non-stationary data. (e.g. problem of spurious regression). We usually can transform variables by taking differences of using lags if the data becomes stationary at difference or lag.

Types of Non-Stationarity

- Random Walk with Drift: $Y_t = \; \mu + Y_{t-1} + e_t$
- Deterministic Trend Process: $Y_t = \beta_0 + \; \beta_1 t + e_t$

- To induce Stationarity both will require different treatment
    - Differencing
    - Detrending





If a non-stationary series $Y_t$is difference d times in order to become stationary, it is said to be ***integrated*** of order d.

If $Y_t \sim I(d) \quad then \ \Delta^d Y_t \sim I(0)$

$I(0)$ is a stationary series. It will cross the mean frequently.

$I(1)$ is a series containing one unit root

Most of the economic and financial series contain a single unit root (Some may be stationary)

However prices have been seen in various researches to have 2 unit roots.

UNIT ROOT TEST

A unit root test tests is used to know if a time series variable is non-stationary using an autoregressive model. The famous tests include augmented Dickey–Fuller test and the Phillipe-Perron test. The null hypothesis for these tests is the existence of a unit root.

Why do we need to test for Non-Stationarity?

- The stationarity or non-stationarity of a series can strongly influence its behavior and properties.
- Spurious regressions: If two variables are trending over time, a regression of one on the other could have a high R-square even if the two are totally unrelated.
- The parameters are misleading
- If variables in the model are not stationary, the usual "t-ratios" will not follow a t-distribution, so various tests will not be valid

Augmented Dickey-Fuller unit-root test ($dfuller$)

$dfuller$ performs the augmented Dickey-Fuller test that a variable follows a unit-root process. It adds lagged differences to the model. The null hypothesis is that the variable contains a unit root, and the alternative is that the variable was generated by a stationary process. You may optionally exclude the constant, include a trend term, and include lagged values of the difference of the variable in the regression.

Syntax: dfuller varname [if] [in] [, options]

Menu: Statistics > Time series > Tests > Augmented Dickey-Fuller unit-root test

Characteristics of Non-Stationary Series

Characteristics of the series	Dickey-Fuller Regression Model
No Constant , No time Trend	$\Delta Y_t = \gamma Y_{t-1} + u_t$
Constant, without time trend	$\Delta Y_t = \alpha + \gamma Y_{t-1} + u_t$
Constant and time trend	$\Delta Y_t = \alpha + \gamma Y_{t-1} + \gamma t + u_t$

Use constant (drift) when series fluctuates against a non-zero mean.

Phillips-Perron unit-root test ($pperron$)

pperron performs the Phillips-Perron test that a variable has a unit root. The null hypothesis is that the variable contains a unit root, and the alternative is that the variable was generated by a stationary process. pperron uses Newey-West standard errors to account for serial correlation, (so it allows for autocorrelated residuals) whereas the augmented Dickey-Fuller test implemented in dfuller uses additional lags of the first-difference variable.

Syntax: pperron varname [if] [in] [, options]

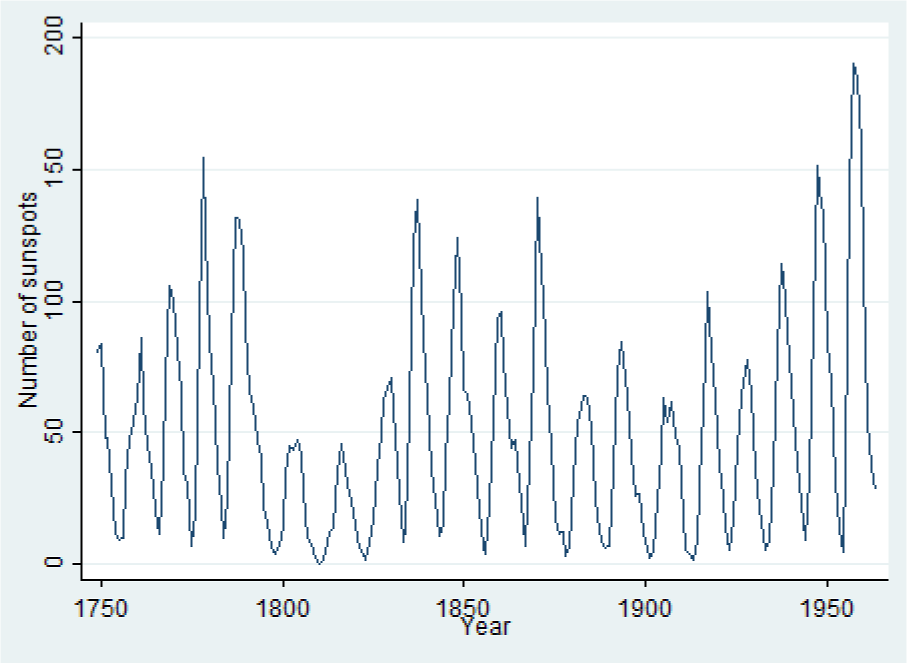
Menu: Statistics > Time series > Tests > Phillips-Perron unit-root test

Examples

webuse sunspot

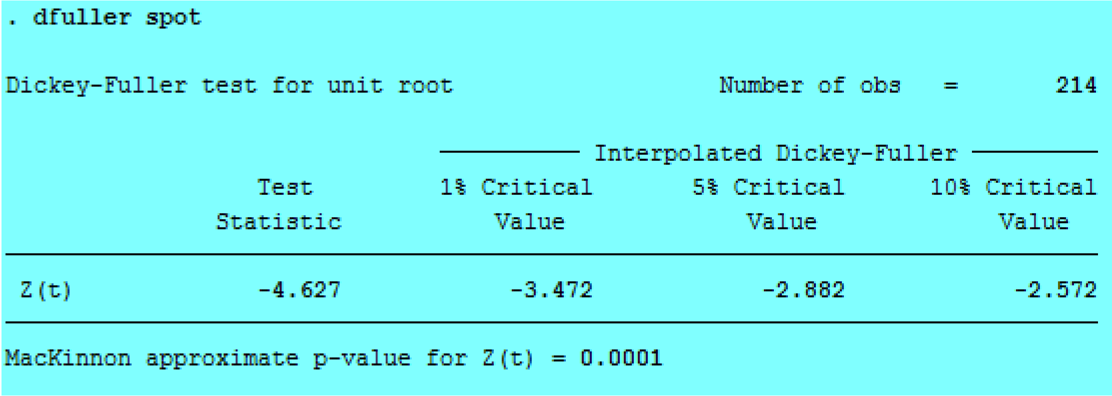
tsset time

tsline spot



Seems to be stationary

dfuller spot

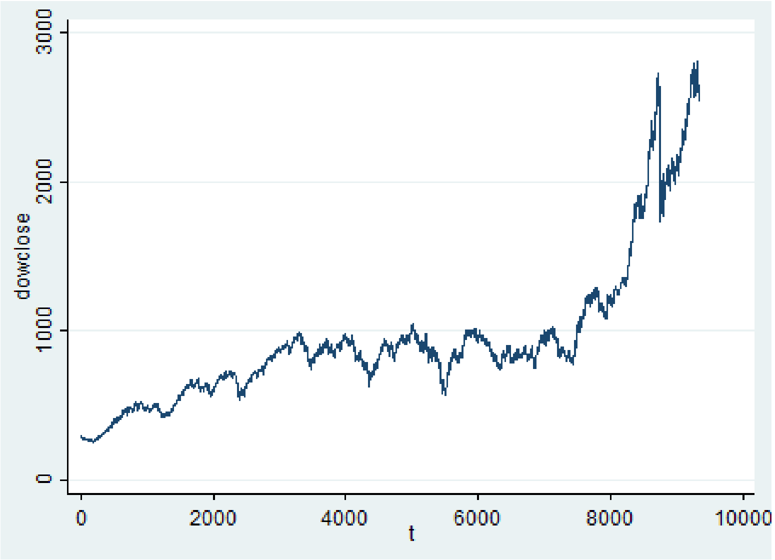


```
. dfuller spot

Dickey-Fuller test for unit root                   Number of obs   =       214

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -4.627            -3.472            -2.882            -2.572
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0001
```

webuse dow1

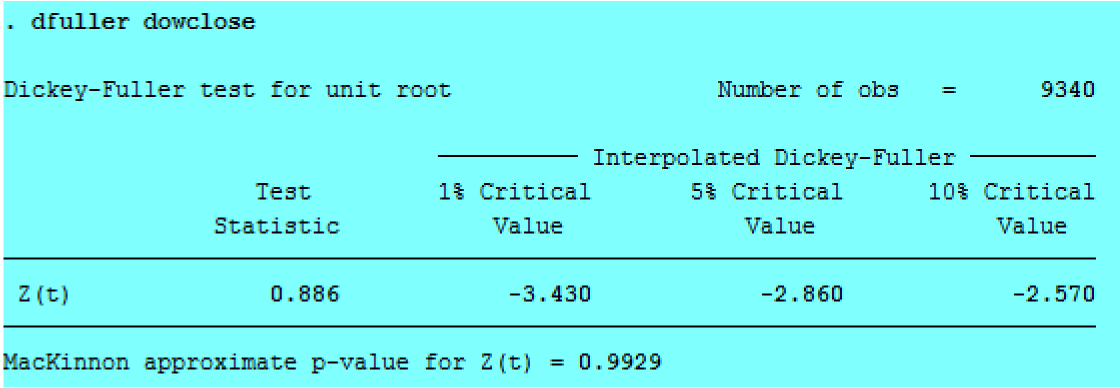
des

tsline dowclose



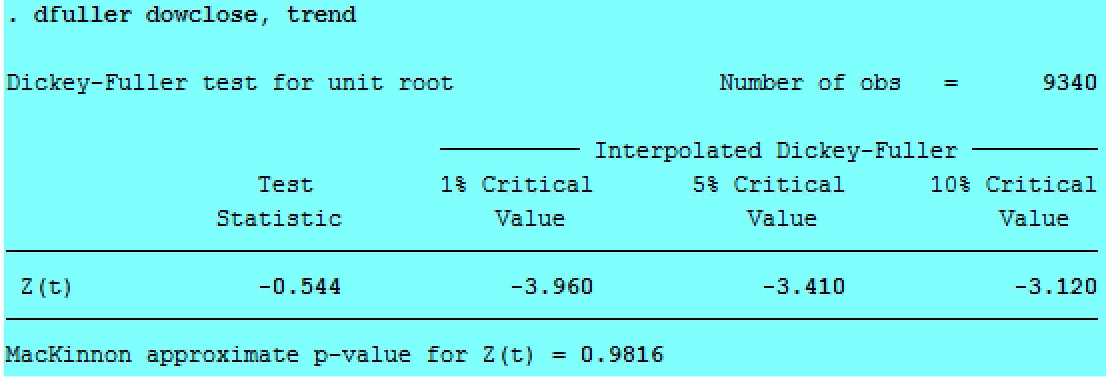
Shows random Walk with non-zero drift (use drift option)

dfuller dowclose



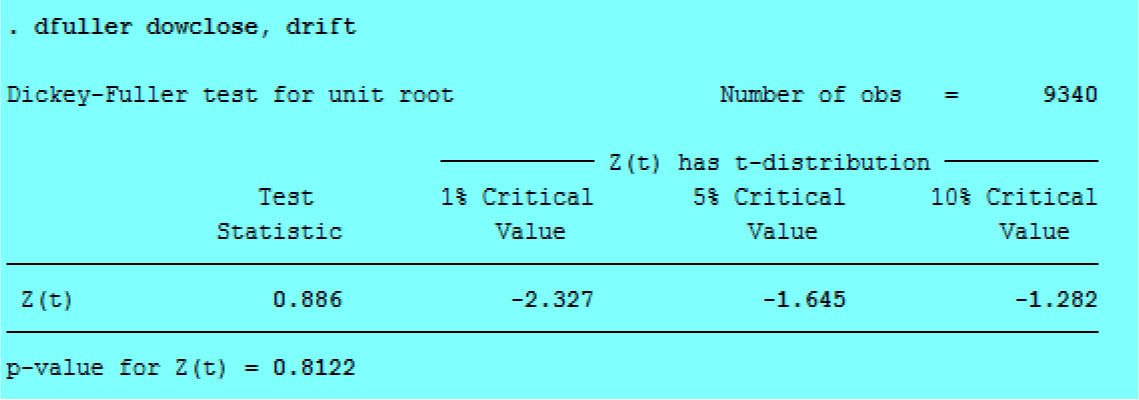
```
. dfuller dowclose

Dickey-Fuller test for unit root                   Number of obs   =      9340

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)              0.886            -3.430            -2.860            -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9929
```

trend specifies that a trend term be included in the associated regression and that the process under the null hypothesis is a random walk, perhaps with drift. This option may not be used with the noconstant or drift option.



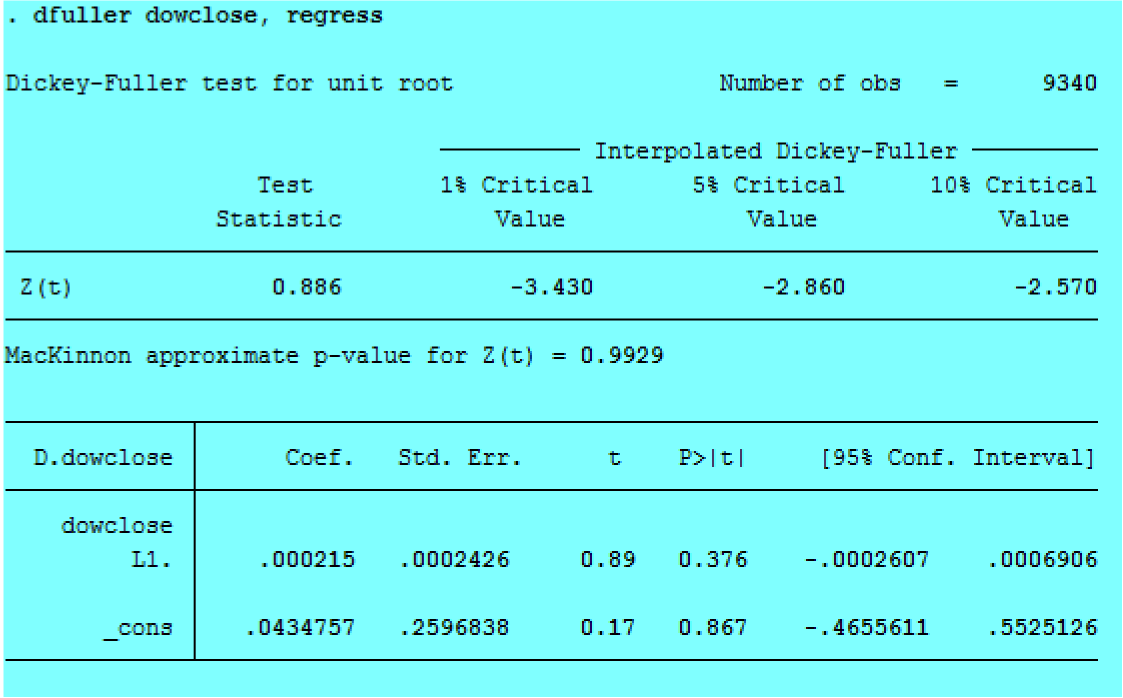
```
. dfuller dowclose, trend

Dickey-Fuller test for unit root                   Number of obs   =      9340

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -0.544            -3.960            -3.410            -3.120
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9816
```

drift indicates that the process under the null hypothesis is a random walk with nonzero drift. This option may not be used with the noconstant or trend option.



```
. dfuller dowclose, drift

Dickey-Fuller test for unit root                   Number of obs   =      9340

                                   ──────── Z(t) has t-distribution ───────
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
──────────────────────────────────────────────────────────────────────────────
 Z(t)              0.886            -2.327            -1.645            -1.282
──────────────────────────────────────────────────────────────────────────────
p-value for Z(t) = 0.8122
```

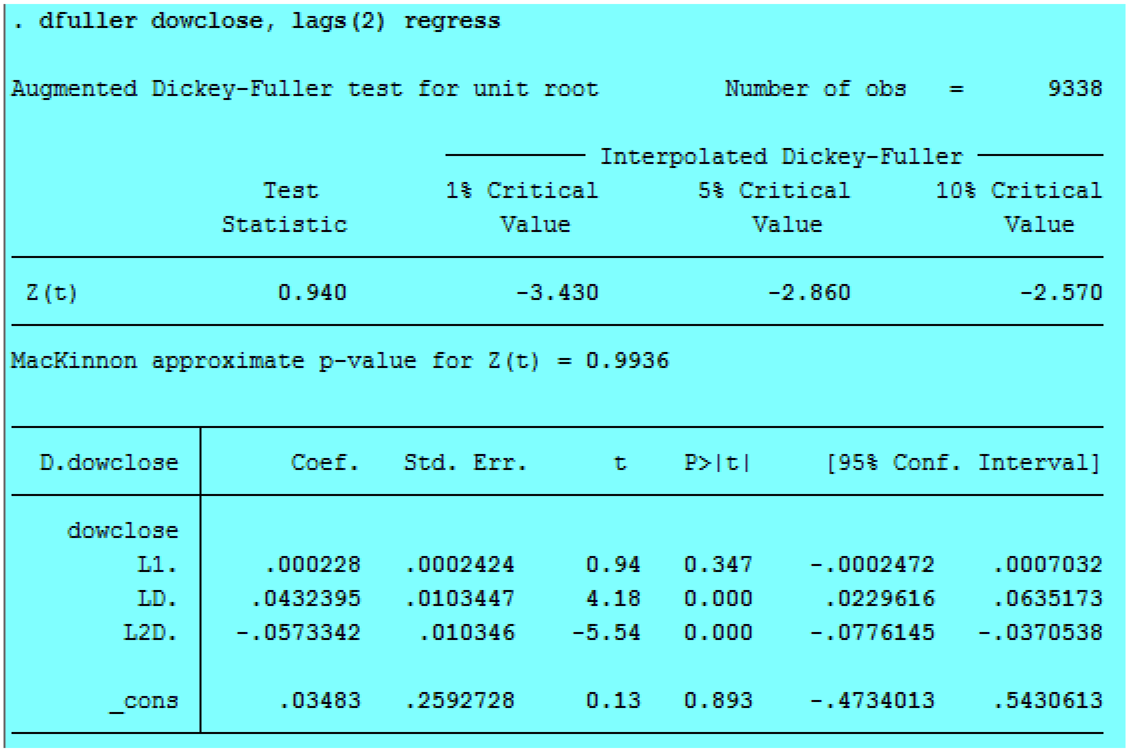
regress specifies that the associated regression table appear in the output. By default, the regression table is not produced.



```
. dfuller dowclose, regress

Dickey-Fuller test for unit root                   Number of obs   =      9340

                                   ─────── Interpolated Dickey-Fuller ──────
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
──────────────────────────────────────────────────────────────────────────────
 Z(t)              0.886            -3.430            -2.860            -2.570
──────────────────────────────────────────────────────────────────────────────
MacKinnon approximate p-value for Z(t) = 0.9929

──────────────────────────────────────────────────────────────────────────────
  D.dowclose │      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
─────────────┼────────────────────────────────────────────────────────────────
    dowclose │
         L1. │    .000215   .0002426     0.89   0.376    -.0002607    .0006906
             │
       _cons │   .0434757   .2596838     0.17   0.867    -.4655611    .5525126
──────────────────────────────────────────────────────────────────────────────
```

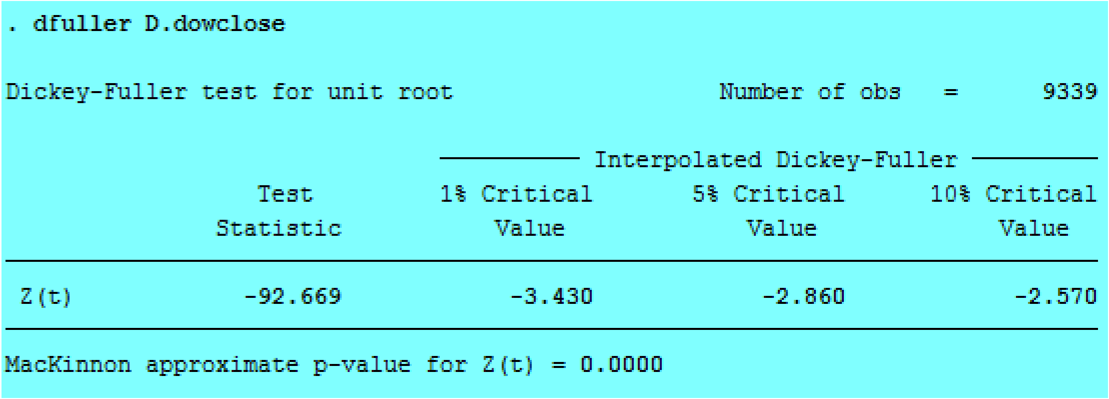
NOTE: see that coefficient of L1 is not significant

lags(#) specifies the number of lagged difference terms to include in the covariate list.



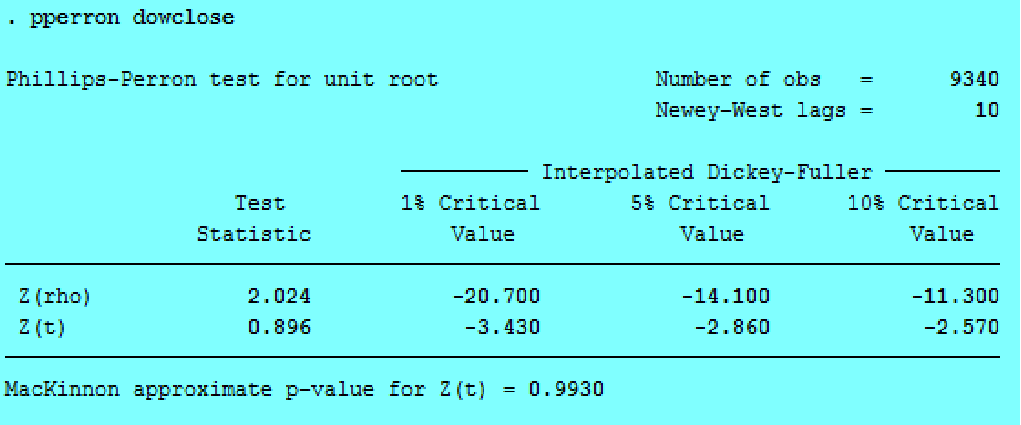
```
. dfuller dowclose, lags(2) regress

Augmented Dickey-Fuller test for unit root         Number of obs   =      9338

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             0.940           -3.430            -2.860            -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9936

------------------------------------------------------------------------------
  D.dowclose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dowclose |
         L1. |    .000228   .0002424     0.94   0.347    -.0002472    .0007032
         LD. |   .0432395   .0103447     4.18   0.000     .0229616    .0635173
        L2D. |  -.0573342    .010346    -5.54   0.000    -.0776145   -.0370538
             |
       _cons |     .03483   .2592728     0.13   0.893    -.4734013    .5430613
------------------------------------------------------------------------------
```

dfuller D.dowclose

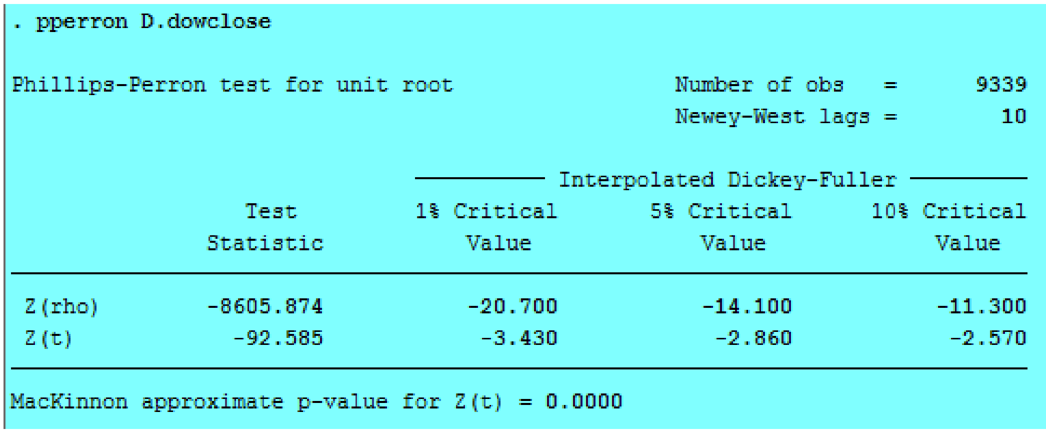


```
. dfuller D.dowclose

Dickey-Fuller test for unit root                   Number of obs   =      9339

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)           -92.669           -3.430            -2.860            -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

Philip Perron Test Example



```
. pperron dowclose

Phillips-Perron test for unit root                 Number of obs   =      9340
                                                   Newey-West lags =        10

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)            2.024         -20.700           -14.100           -11.300
 Z(t)              0.896          -3.430            -2.860            -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9930
```



```
. pperron D.dowclose

Phillips-Perron test for unit root                 Number of obs   =      9339
                                                   Newey-West lags =        10

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)        -8605.874         -20.700           -14.100           -11.300
 Z(t)            -92.585          -3.430            -2.860            -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

Lecture 44
Time Series, Cointegration and Error Correction-II

What to do in case of non-stationarity?

Differencing: throws out long term properties of the series and is good for short term modeling

Cointegration: Granger introduces this.

Cointegration

Two series of variables may be integrated, but their difference (or any linear combination) may be stationary. This means that each variable wanders quite far from its mean, but the two series wander very near each other. The value/ location of one variable could be told by looking at the other variable. They have an equilibrium relationship (never expected to drift too far). Deviations from this equilibrium will be corrected over time.

Cointegration & Error Correction Model Using Stata

Engle-Granger Cointegration Analysis: STEPS

- Test individual variables for unit root
- Estimate the static regression
- Test for unit roots in the error of the static regression; if residuals are stationary, series are cointegrated.
- Finally we can use Error Correction Model

Granger: two or more integrated time series that are cointegrated have an error correction representation.

If x and y are cointegrated, then, by the Granger Representation Theorem, we can model y and x as being in an error correcting relationship.

Basically, it has y and x being in an equilibrium relationship, with the short run behavior of y being a function of the short run behavior of x and an equilibrating factor. It estimates the speed at which a dependent variable Y returns to equilibrium after a change in an independent variable - X.

The EC model is:

$$\Delta Y_t = \beta \Delta X_t + \rho \left(Y_{t-1} - \gamma X_{t-1}\right) + \epsilon_t$$

Where $Y_{t-1} - \gamma X_{t-1}$ is the predicted error generated from a basic regression Y on X.

where ϵ_t is stationary (test for this with Dickey-Fuller). β is the short run effect of x on y and γ is the speed of equilibration, with giving the long run relation between y and x.

NOTE: ECM is normally used on non-stationary data but can be used for stationary data.

Cointegration and Error Correction Model in Detail

Stationary Data means that data has a finite mean and variance that do not depend on time. Data reverts to mean in the long run. It crossed the mean quite often.

Usually time series data is not stationary but is integrated.

Integrated time series data:

- Does not revert to the mean
- Usually moves in a random walk
- Previous changes are reflected in the current value
- May have infinite variance and no appropriate mean
- Shocks are permanently incorporated

If a non-stationary series $Y_t$is difference d times in order to become stationary, it is said to be ***integrated*** of order d.

Two series are cointegrated if

- They are integrated of the same order
- There is a linear combination of the two series that is stationary i.e. integrated of order zero I(0)
- Cointegrated data do not drift very far from each other.
- Deviation from equilibrium will be corrected over time

Critical Values and Power				
Statistics	**Name**	**1%**	**5%**	**10%**
1	**DF**	**4.07**	**3.37**	**3.03**
2	**ADF**	**3.77**	**3.17**	**2.84**

Example:

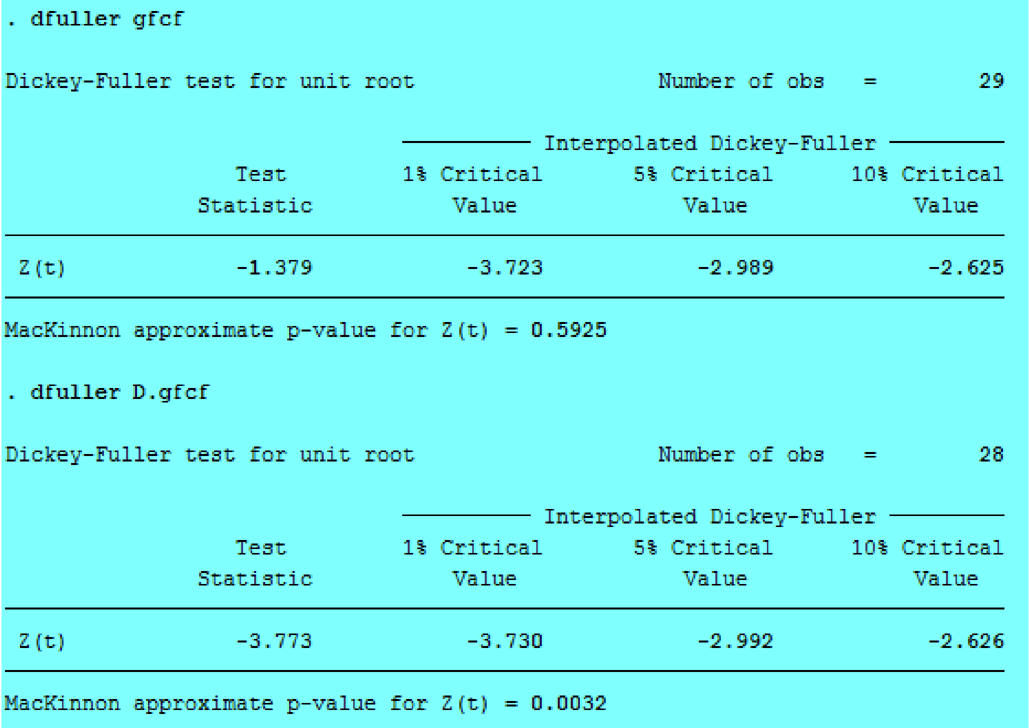
Consider the file ECM.dta provided to you.

It contains Pakistani data (GDP, GFCF, imports etc.) from WDI

Year 1983 to 2012

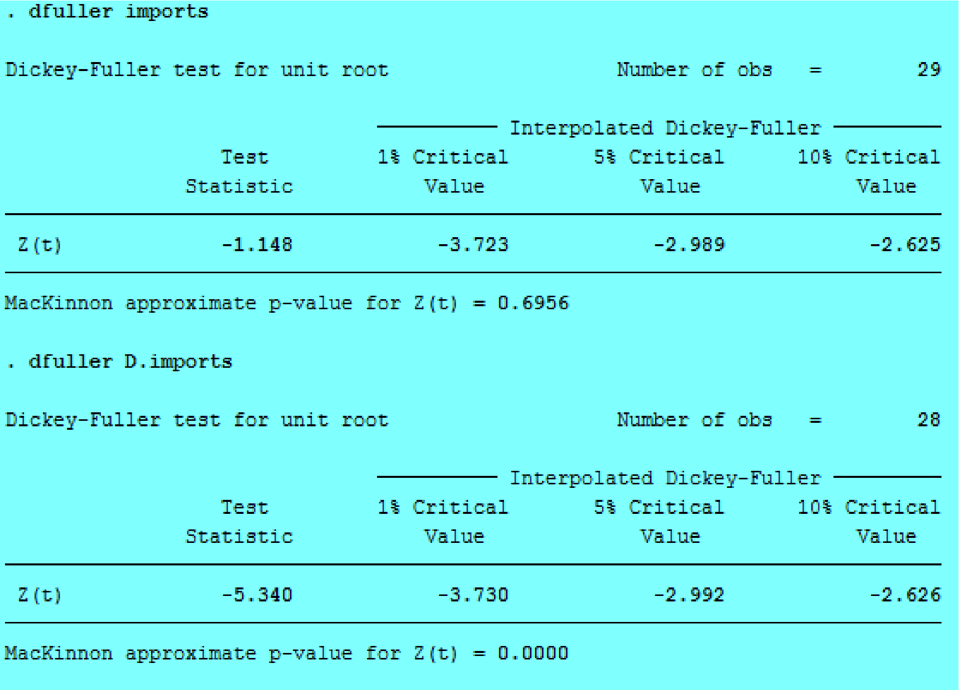
Suppose we focus on GFCF and IMPORTS

Are they 'integrated'?



```
. dfuller gfcf

Dickey-Fuller test for unit root                   Number of obs   =        29

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.379            -3.723            -2.989            -2.625
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.5925

. dfuller D.gfcf

Dickey-Fuller test for unit root                   Number of obs   =        28

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -3.773            -3.730            -2.992            -2.626
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0032
```

GFCF is integrated at first difference.



```
. dfuller imports

Dickey-Fuller test for unit root                   Number of obs   =        29

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.148            -3.723            -2.989            -2.625
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.6956

. dfuller D.imports

Dickey-Fuller test for unit root                   Number of obs   =        28

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -5.340            -3.730            -2.992            -2.626
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

Imports is also integrated at first difference. Both are integrated at the same level i.e. first difference

Are GFCF and imports cointegrated?

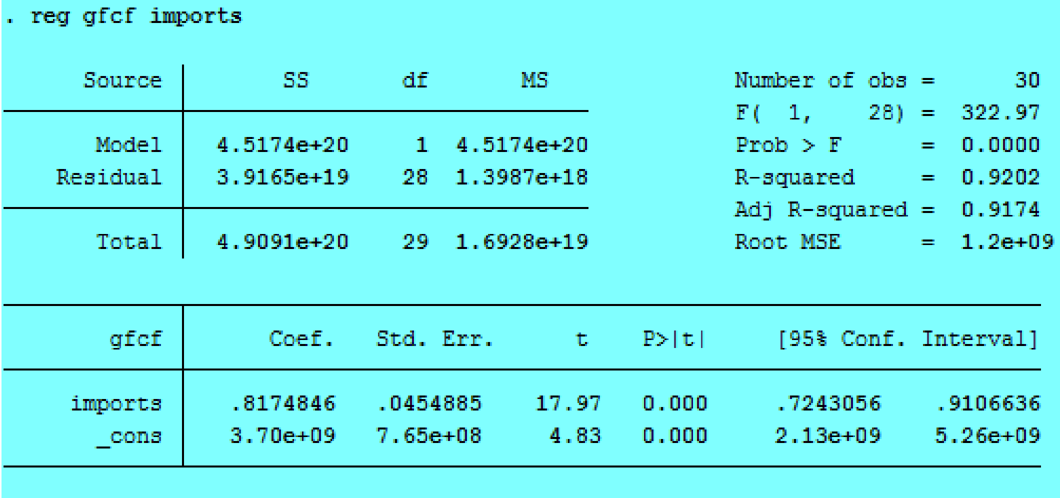
We use Engle Granger Test. It has three steps.

- Run a basic regression (long run)
- Predict the errors
- Run the regression of First difference of residuals on lag of residuals and on lag of first difference of residuals

If the coefficient of lag of residuals is significant, the series are cointegrated.

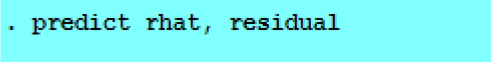
IMPORTANT: the t-values reported in simple regression are not appropriate so we use the EG critical values.

Step I



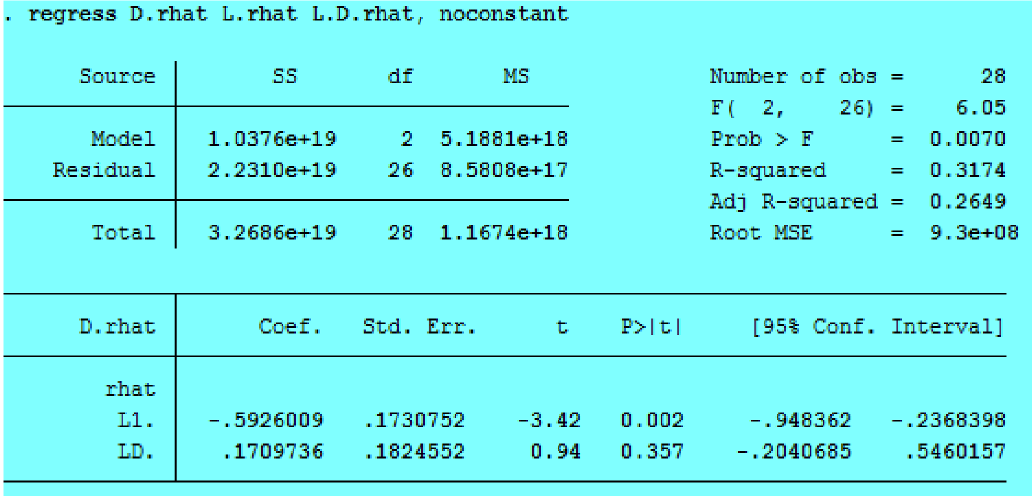
```
. reg gfcf imports

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =  322.97
       Model |  4.5174e+20     1  4.5174e+20           Prob > F      =  0.0000
    Residual |  3.9165e+19    28  1.3987e+18           R-squared     =  0.9202
-------------+------------------------------           Adj R-squared =  0.9174
       Total |  4.9091e+20    29  1.6928e+19           Root MSE      =  1.2e+09

------------------------------------------------------------------------------
        gfcf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     imports |   .8174846   .0454885    17.97   0.000     .7243056    .9106636
       _cons |   3.70e+09   7.65e+08     4.83   0.000     2.13e+09    5.26e+09
------------------------------------------------------------------------------
```

Step II



```
. predict rhat, residual
```

Step III



```
. regress D.rhat L.rhat L.D.rhat, noconstant

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  2,    26) =    6.05
       Model |  1.0376e+19     2  5.1881e+18           Prob > F      =  0.0070
    Residual |  2.2310e+19    26  8.5808e+17           R-squared     =  0.3174
-------------+------------------------------           Adj R-squared =  0.2649
       Total |  3.2686e+19    28  1.1674e+18           Root MSE      =  9.3e+08

------------------------------------------------------------------------------
      D.rhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rhat |
         L1. |  -.5926009   .1730752    -3.42   0.002     -.948362   -.2368398
         LD. |   .1709736   .1824552     0.94   0.357    -.2040685    .5460157
------------------------------------------------------------------------------
```

The coefficient of first lag of residuals is -0.5926 and the t-value is -3.42. Remember that we should not use the p-values here but look at the critical values of EG. The critical value at 5% is -3.37.

Our value, -3.42 < -3.37, so hypothesis of no Cointegration is rejected. The series are Cointegrated.

If $< t_c$, we reject the null hypothesis that the least square residuals are non-stationary (no cointegration). Here the least square residuals are stationary and the series GFCF and Imports are Cointegrated.

Engle and Granger suggested a model for cointegrated series

The EC model is:

$$\Delta Y_t = \beta \Delta X_t + \rho \left(Y_{t-1} - \gamma X_{t-1}\right) + \epsilon_t$$

Where $Y_{t-1} - \gamma X_{t-1}$ is the predicted error generated from a basic regression Y on X and where ϵ_t is stationary (test for this with Dickey-Fuller). β is the short run effect of x on y and γ is the speed of reverting to equilibrium, with giving the long run relation between y and x.

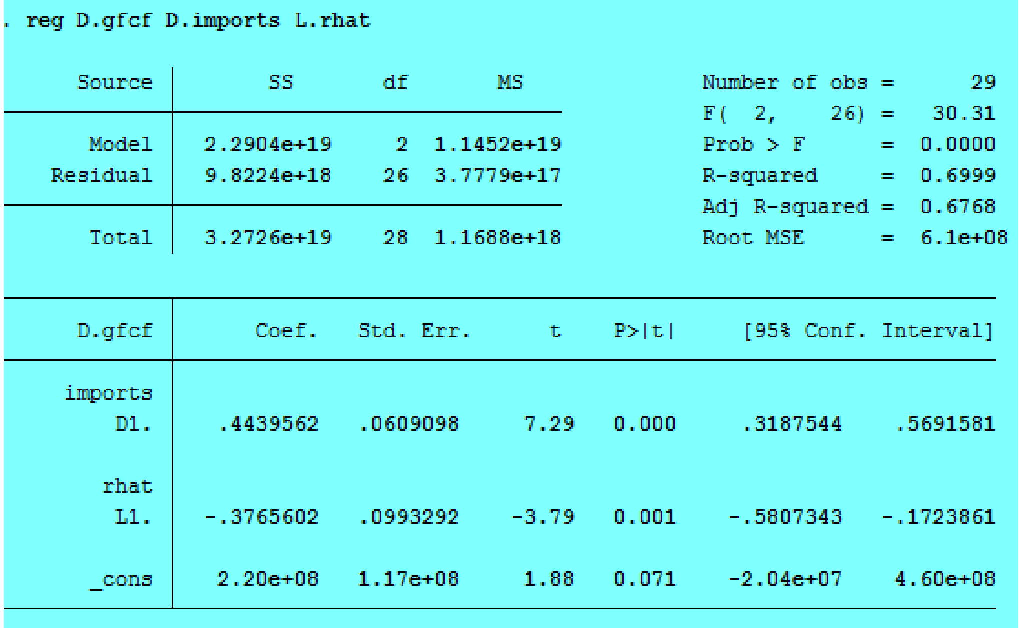
using the file ECM.dta,

$tsset\ year$

$regress\ gfcf\ imports$

$predict\ rhat, residual$

Now using the residuals etc.



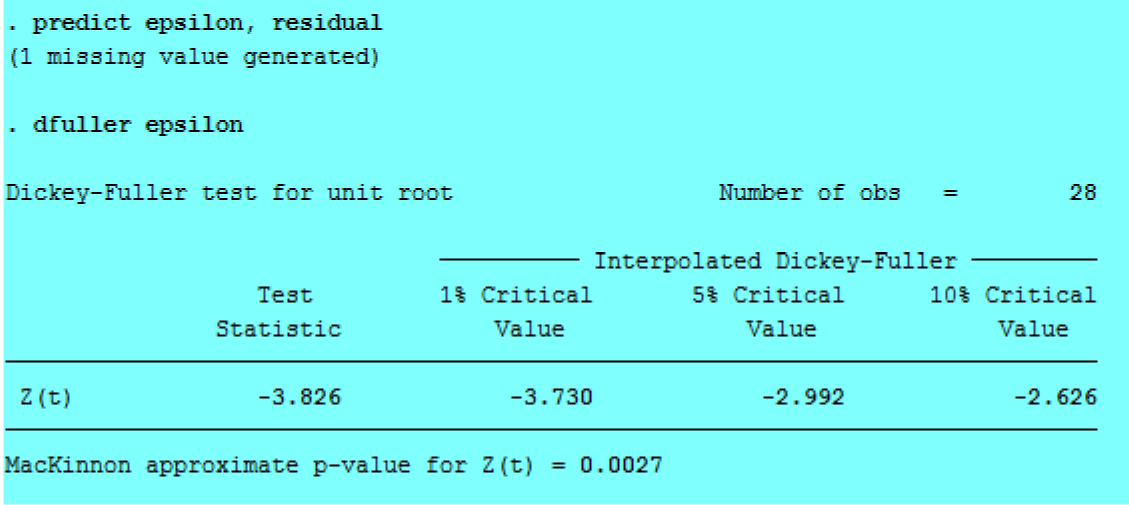
```
. reg D.gfcf D.imports L.rhat

      Source |       SS       df       MS              Number of obs =      29
-------------+------------------------------           F(  2,    26) =   30.31
       Model |  2.2904e+19     2  1.1452e+19           Prob > F      =  0.0000
    Residual |  9.8224e+18    26  3.7779e+17           R-squared     =  0.6999
-------------+------------------------------           Adj R-squared =  0.6768
       Total |  3.2726e+19    28  1.1688e+18           Root MSE      =  6.1e+08

------------------------------------------------------------------------------
      D.gfcf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     imports |
         D1. |   .4439562   .0609098     7.29   0.000     .3187544    .5691581
             |
        rhat |
         L1. |  -.3765602   .0993292    -3.79   0.001    -.5807343   -.1723861
             |
       _cons |   2.20e+08   1.17e+08     1.88   0.071    -2.04e+07    4.60e+08
------------------------------------------------------------------------------
```

The short run effect of imports on GFCF is positive and significant.

The long term relation is established by the coefficient -0.3765 which is also significant.

The speed shows that the deviations from equilibrium are corrected at 37.5% in one time period.

Check if the error generated here is stationary



```
. predict epsilon, residual
(1 missing value generated)

. dfuller epsilon

Dickey-Fuller test for unit root                   Number of obs   =        28

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -3.826            -3.730            -2.992            -2.626
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0027
```

Additional or alternative lags and deterministic terms may be added.

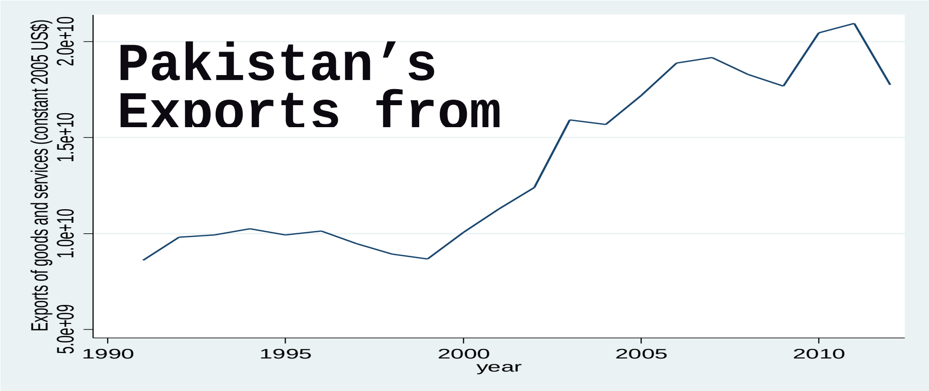
Lecture 45

Time Series, Cointegration and Error Correction-III

As you know, we define 'Time Series' as follows:

Variables recorded over time

- Annually (Macro data)
- Biannually (financial data: banks biannual reports)
- Quarterly (Macro data, agricultural data,)
- Monthly (agricultural data, market prices)
- Daily (Stock exchange, Foreign ExchangeRates, commodity)



Example Graph for Time Series Data

An Autoregressive Process is when we have variable depending linearly on its own past values (lags)

This can be as weighted sum of past values.

AR (p) process

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots\dots\ \varphi_p y_{t-p} + \epsilon_t$$

AR (1) process

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \epsilon_t$$

Stata example: AR (3) model

$reg\ gdp\ L.gdp\ L2.gdp\ L3.gdp$

VAR (Vector Autoregressive Model is a multivariate time-series regression of each dependent variable on lags of itself and on lags of all the other dependent variables

VAR (1) process for y and z

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 z_{t-1} + u_t$$

$$z_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + v_t$$

Stata example: VAR (1) model

$tsset\ time$

$var\ gdp\ exports, lags(1)$

$If\ lags\ are\ not\ specified, 2\ is\ the\ default$

For VAR both series should be stationary. Dickey Fuller test can be used to know when (what level) data is stationary. For Dickey Fuller test we need to know the appropriate lag (Schwert's rule of thumb can be used)

- If p is too small then the remaining serial correlation in the errors will bias the test.
- If p is too large then the power of the test will suffer.
- Monte Carlo experiments suggest it is better to error on the side of including too many lags.

Schwert's (1989) rule of thumb

This rule of thumb helps in having an optimum lag

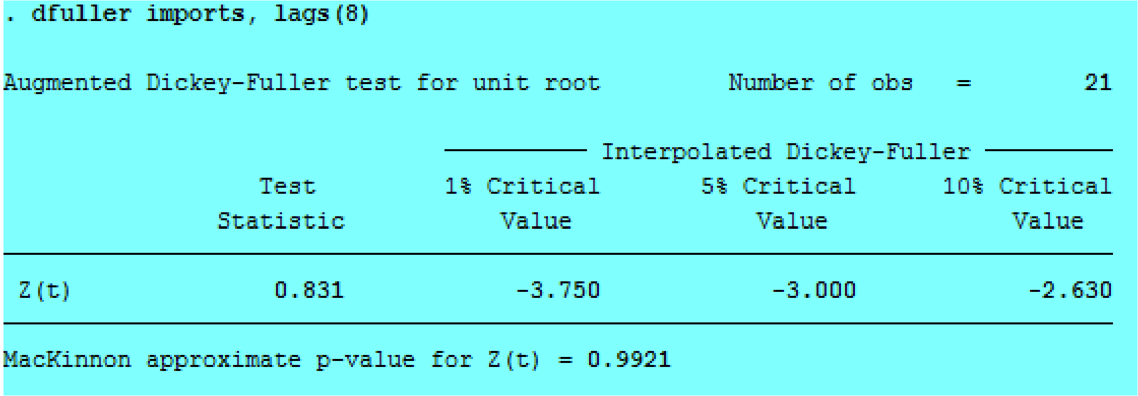
$$p_{max} = [12.(\frac{T}{100})^{\frac{1}{4}}]$$

Where $T = number\ of\ observations$ and [] $means\ integer\ part$

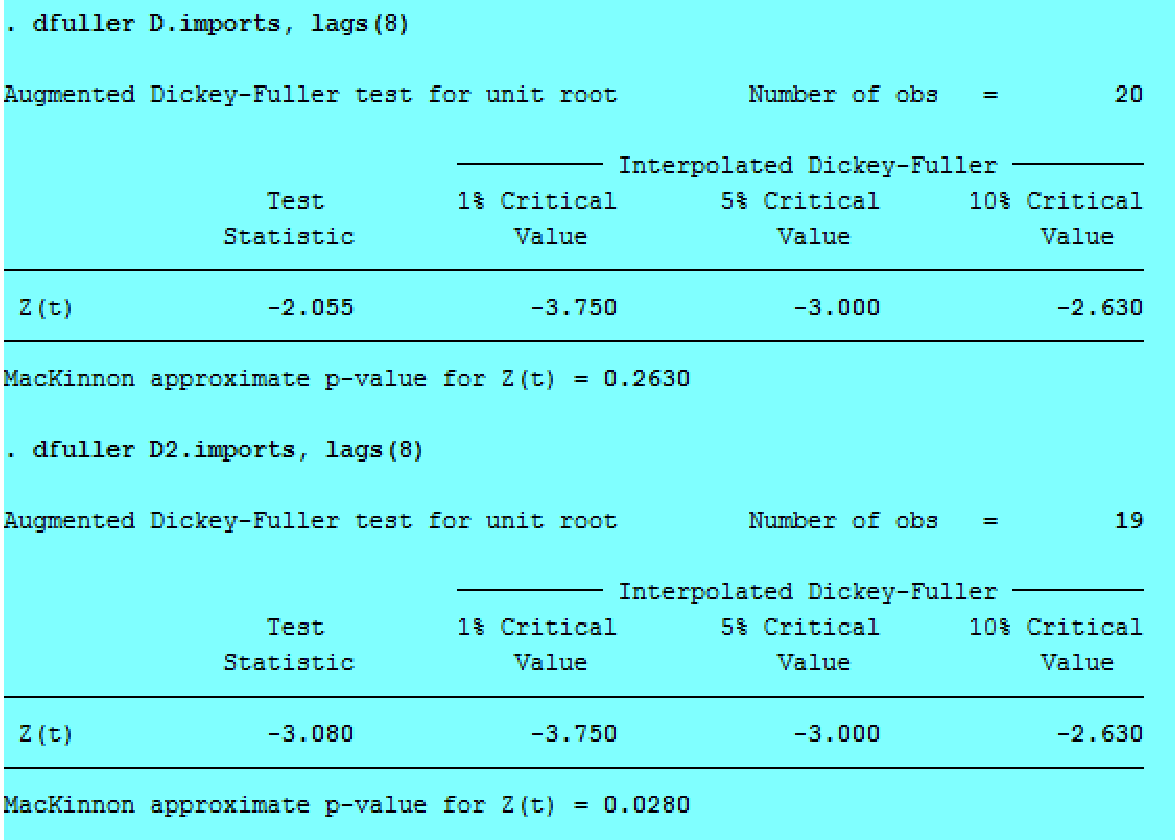
In ECM.dta with 30 observations, lags = 8

The Default lag length for the $dfuller$ command is zero.

Here are some examples using the file ECM.dta



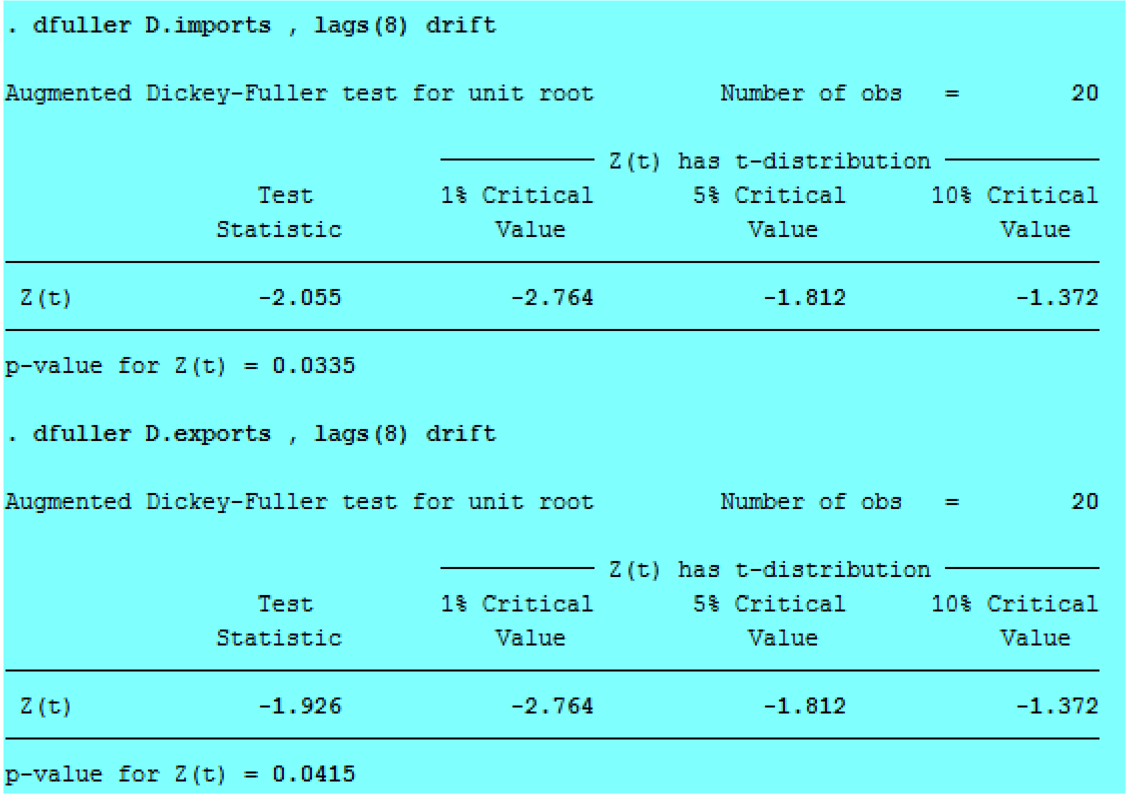
```
. dfuller imports, lags(8)

Augmented Dickey-Fuller test for unit root         Number of obs   =        21

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             0.831            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9921
```



```
. dfuller D.imports, lags(8)

Augmented Dickey-Fuller test for unit root         Number of obs   =        20

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -2.055            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2630

. dfuller D2.imports, lags(8)

Augmented Dickey-Fuller test for unit root         Number of obs   =        19

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -3.080            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0280
```

Imports are, in fact, stationary at second level

The command $dfuller$ for random walk with drift

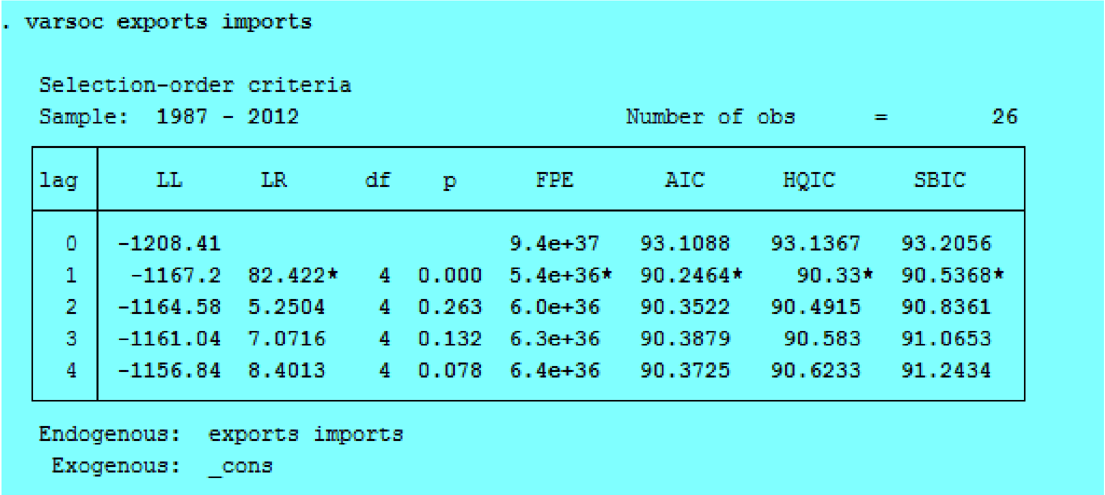


```
. dfuller D.imports , lags(8) drift

Augmented Dickey-Fuller test for unit root         Number of obs   =        20

                               ---------- Z(t) has t-distribution ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.055            -2.764            -1.812            -1.372
------------------------------------------------------------------------------
p-value for Z(t) = 0.0335

. dfuller D.exports , lags(8) drift

Augmented Dickey-Fuller test for unit root         Number of obs   =        20

                               ---------- Z(t) has t-distribution ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.926            -2.764            -1.812            -1.372
------------------------------------------------------------------------------
p-value for Z(t) = 0.0415
```

If they follow random walk with drift, exports and imports are integrated of order 1

Applying VAR

First determine the optimal lag length for VAR. The command is *varsoc* (number of max lags can be given)

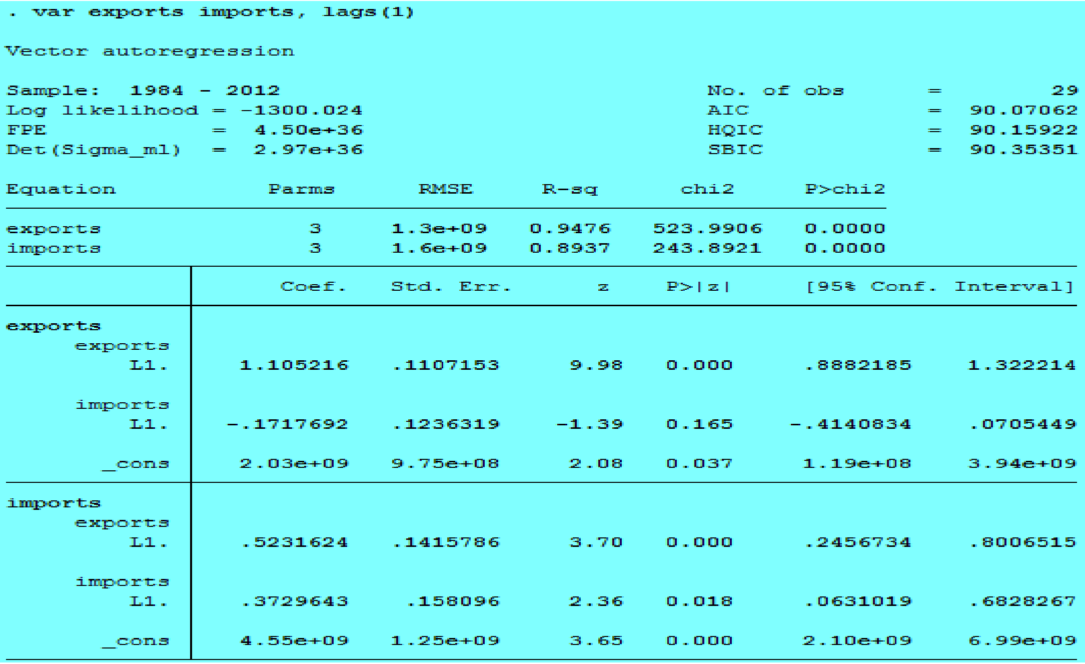


```
. varsoc exports imports

   Selection-order criteria
   Sample:  1987 - 2012                         Number of obs      =        26
  +---------------------------------------------------------------------------+
  |lag |    LL      LR      df    p      FPE       AIC      HQIC      SBIC    |
  |----+----------------------------------------------------------------------|
  |  0 | -1208.41                      9.4e+37   93.1088   93.1367   93.2056  |
  |  1 |  -1167.2  82.422*   4  0.000  5.4e+36*  90.2464*    90.33*  90.5368* |
  |  2 | -1164.58  5.2504    4  0.263  6.0e+36   90.3522   90.4915   90.8361  |
  |  3 | -1161.04  7.0716    4  0.132  6.3e+36   90.3879    90.583   91.0653  |
  |  4 | -1156.84  8.4013    4  0.078  6.4e+36   90.3725   90.6233   91.2434  |
  +---------------------------------------------------------------------------+
   Endogenous:  exports imports
    Exogenous:  _cons
```

AIC: Akaike Information Criterion, should be minimized; suggested lag is ONE

The model can be estimated as follows:

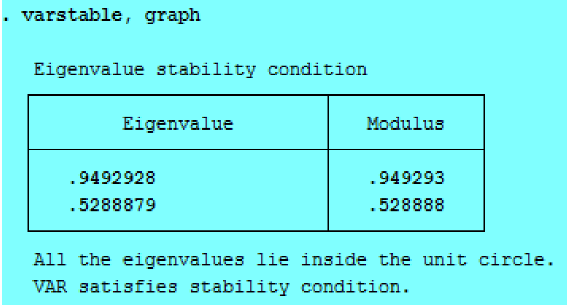
Default lag=2 so we need to specify ONE (suggested by AIC). Here the Sign and significance, not the magnitude, are important.



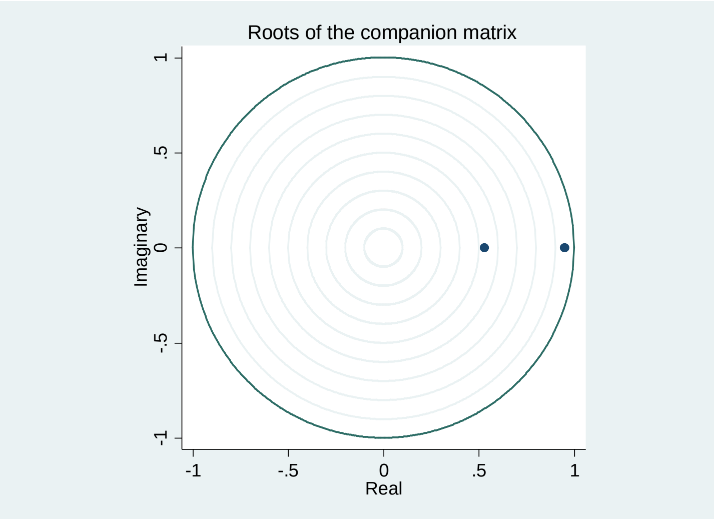
```
. var exports imports, lags(1)

Vector autoregression

Sample:  1984 - 2012                               No. of obs      =        29
Log likelihood = -1300.024                         AIC             =  90.07062
FPE            =  4.50e+36                         HQIC            =  90.15922
Det(Sigma_ml)  =  2.97e+36                         SBIC            =  90.35351

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
exports               3     1.3e+09   0.9476   523.9906   0.0000
imports               3     1.6e+09   0.8937   243.8921   0.0000
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
exports      |
     exports |
         L1. |   1.105216   .1107153     9.98   0.000     .8882185    1.322214
             |
     imports |
         L1. |  -.1717692   .1236319    -1.39   0.165    -.4140834    .0705449
             |
       _cons |   2.03e+09   9.75e+08     2.08   0.037     1.19e+08    3.94e+09
-------------+----------------------------------------------------------------
imports      |
     exports |
         L1. |   .5231624   .1415786     3.70   0.000     .2456734    .8006515
             |
     imports |
         L1. |   .3729643    .158096     2.36   0.018     .0631019    .6828267
             |
       _cons |   4.55e+09   1.25e+09     3.65   0.000     2.10e+09    6.99e+09
------------------------------------------------------------------------------
```

Post Estimation Test of stability of the model: eigen values should be less than one

After running the model, use the *Varstable, gra*



```
. varstable, graph

  Eigenvalue stability condition

  +----------------------------------+
  |      Eigenvalue     |   Modulus  |
  |---------------------+------------|
  |   .9492928          |   .949293  |
  |   .5288879          |   .528888  |
  +----------------------------------+

  All the eigenvalues lie inside the unit circle.
  VAR satisfies stability condition.
```



ARMA (Autoregressive Moving Average)

ARIMA (Autoregressive Integrated Moving Average) model

ARIMA is a A combination of AR and MA processes where AR is autoregressive, I is Integrated and MA is an abbreviation for a Moving Average process (a process containing past realizations of the variables own residual)

ARMA (p,q) process

$$y_t = \varphi_0 + \epsilon_t + \varphi_1 e_{t-1} + \varphi_2 e_{t-2} + \dots\dots \varphi_p e_{t-q}$$

ARMA (1,0) is as AR(1)

ARMA (0,1) is as MA(1)

ARMA (1,1) process can be given by the equation

$$y_t = c + \alpha_1 y_{t-1} + \beta_1 \epsilon_{t-1} + \epsilon_t$$

Series where I=0 is stationary.

arima is a maximum likelihood estimation (not OLS). We are discussing Univariate ARIMA

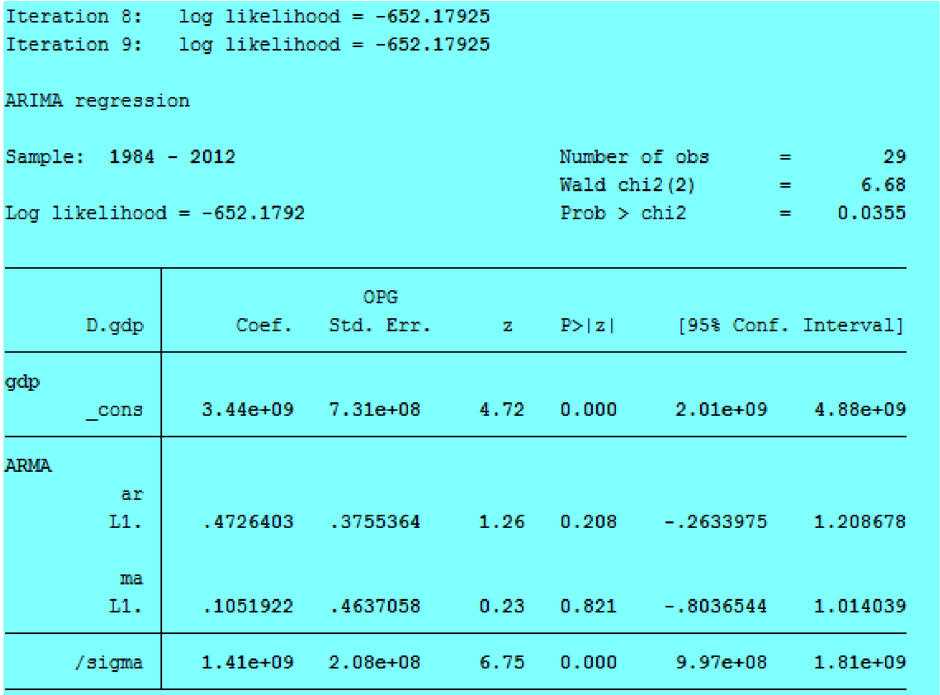
Stata Example:

We must run arima on a stationary series

Stationary level must be determined. Then the Stata command will be

$arima\ gdp, arima\ (1,1,1)$ which is the same as: $arima\ gdp, ar(1)\ ma(1)$

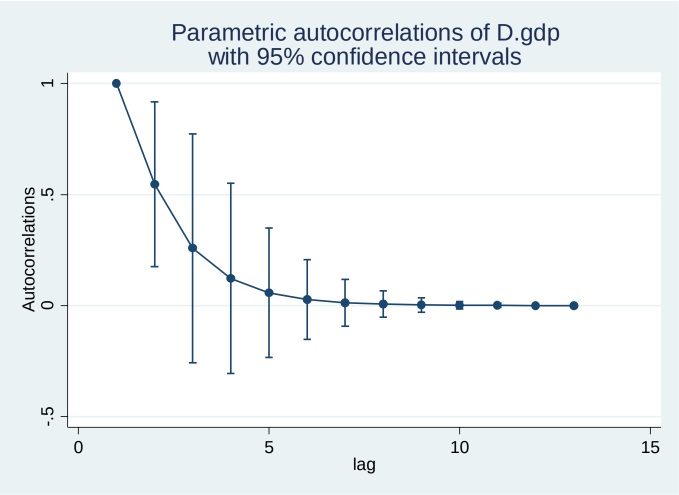
$arima\ gdp, arima\ (1,1,1)$ will give the result



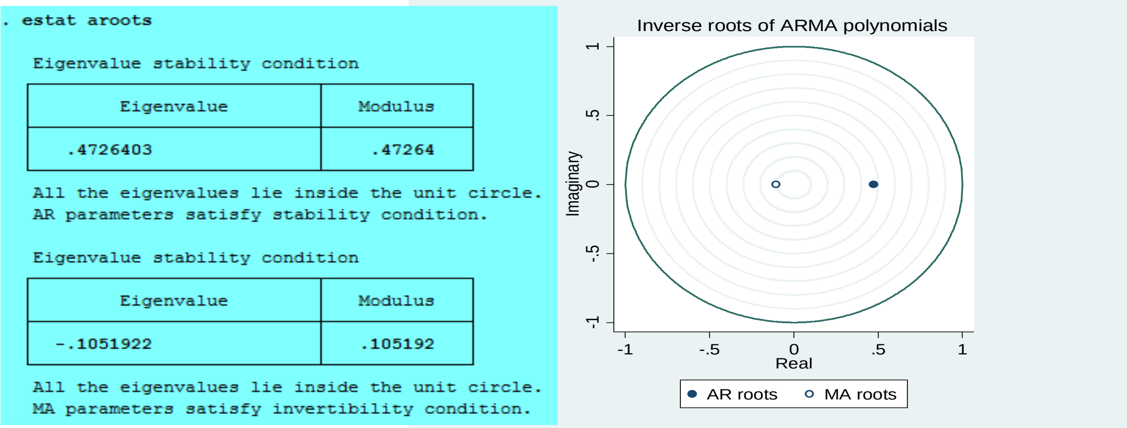
```
Iteration 8:   log likelihood = -652.17925
Iteration 9:   log likelihood = -652.17925

ARIMA regression

Sample:  1984 - 2012                            Number of obs      =        29
                                                Wald chi2(2)       =      6.68
Log likelihood = -652.1792                      Prob > chi2        =    0.0355

------------------------------------------------------------------------------
             |                 OPG
       D.gdp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
gdp          |
       _cons |   3.44e+09   7.31e+08     4.72   0.000     2.01e+09    4.88e+09
-------------+----------------------------------------------------------------
ARMA         |
          ar |
         L1. |   .4726403   .3755364     1.26   0.208    -.2633975    1.208678
             |
          ma |
         L1. |   .1051922   .4637058     0.23   0.821    -.8036544    1.014039
-------------+----------------------------------------------------------------
      /sigma |   1.41e+09   2.08e+08     6.75   0.000     9.97e+08    1.81e+09
------------------------------------------------------------------------------
```

ARIMA Post Estimation Commands in Stata

estat acplot estimates autocorrelation and covariances

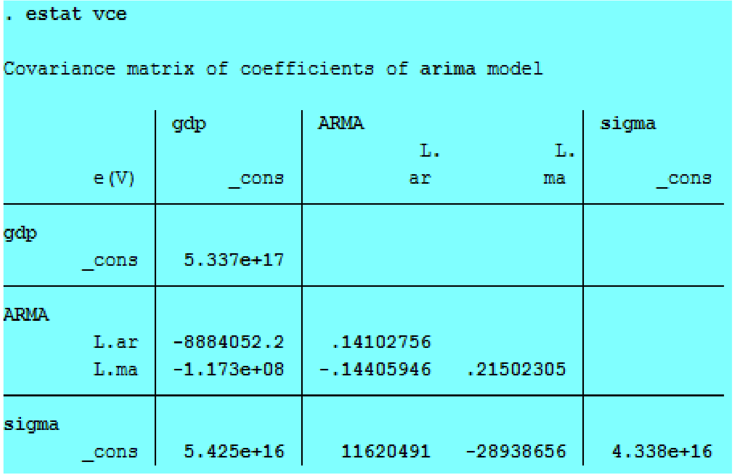


estat aroots Checks the stability conditions



```
. estat aroots

  Eigenvalue stability condition

  +---------------------------+
  |   Eigenvalue  |  Modulus  |
  |---------------+-----------|
  |    .4726403   |   .47264  |
  +---------------------------+
   All the eigenvalues lie inside the unit circle.
   AR parameters satisfy stability condition.

  Eigenvalue stability condition

  +---------------------------+
  |   Eigenvalue  |  Modulus  |
  |---------------+-----------|
  |   -.1051922   |  .105192  |
  +---------------------------+
   All the eigenvalues lie inside the unit circle.
   MA parameters satisfy invertibility condition.
```

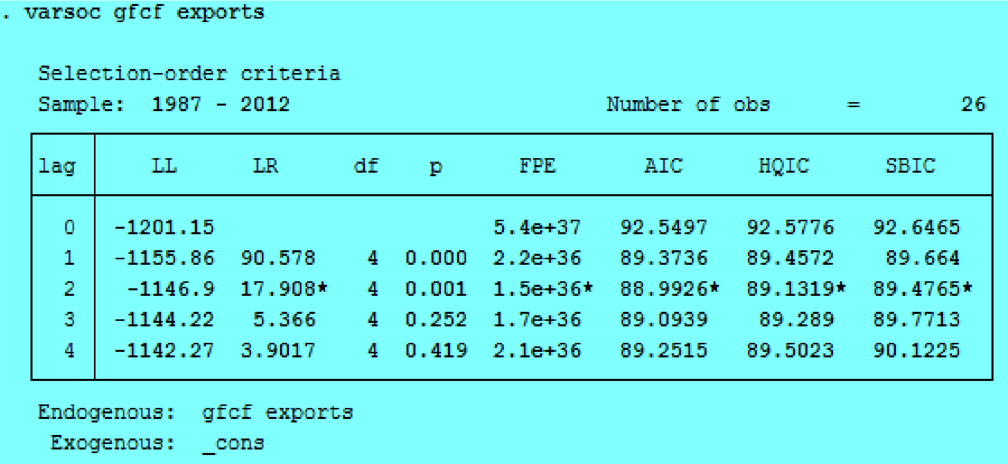
estat vce gives the Variance covariance matrix of the estimates



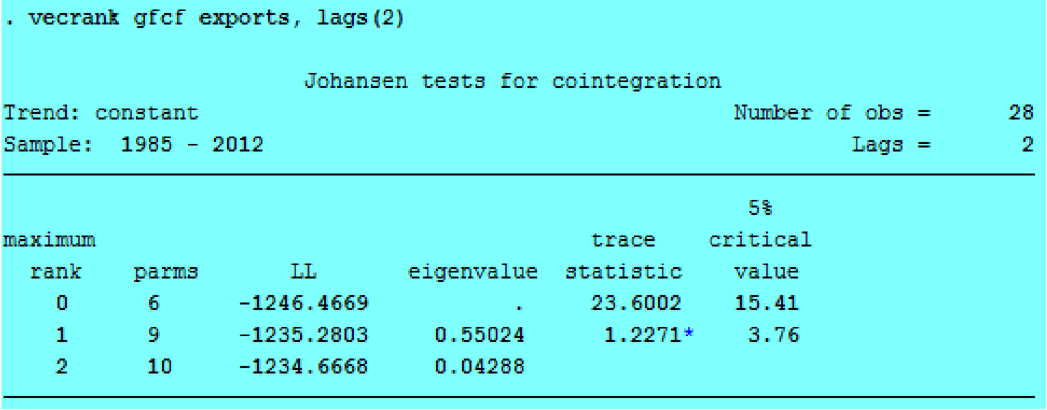
```
. estat vce

Covariance matrix of coefficients of arima model

             |      gdp   |          ARMA            |    sigma
             |            |         L.          L.   |
        e(V) |      _cons |         ar          ma   |    _cons
-------------+------------+--------------------------+-----------
gdp          |            |                          |
       _cons |  5.337e+17 |                          |
-------------+------------+--------------------------+-----------
ARMA         |            |                          |
        L.ar | -8884052.2 |  .14102756               |
        L.ma | -1.173e+08 | -.14405946    .21502305  |
-------------+------------+--------------------------+-----------
sigma        |            |                          |
       _cons |  5.425e+16 |   11620491    -28938656  |  4.338e+16
```

Johansen's Test for Cointegration

This test is based on maximum likelihood estimation and two statistics: maximum eigenvalues and a trace-statistics. $vecrank$ is the command in Stata. We use one lag as suggested by $varsoc$ for VAR model; then use the command $vecrank\ exports\ imports, lags(1)$



```
. varsoc gfcf exports

   Selection-order criteria
   Sample:  1987 - 2012                             Number of obs      =        26
  +---------------------------------------------------------------------------+
  |lag |    LL      LR      df    p      FPE       AIC       HQIC      SBIC    |
  |----+----------------------------------------------------------------------|
  |  0 | -1201.15                      5.4e+37   92.5497   92.5776   92.6465   |
  |  1 | -1155.86  90.578    4  0.000  2.2e+36   89.3736   89.4572    89.664   |
  |  2 |  -1146.9  17.908*   4  0.001  1.5e+36*  88.9926*  89.1319*  89.4765*  |
  |  3 | -1144.22   5.366    4  0.252  1.7e+36   89.0939    89.289   89.7713   |
  |  4 | -1142.27  3.9017    4  0.419  2.1e+36   89.2515   89.5023   90.1225   |
  +---------------------------------------------------------------------------+
   Endogenous:  gfcf exports
    Exogenous:  _cons
```



```
. vecrank gfcf exports, lags(2)

                        Johansen tests for cointegration
Trend: constant                                         Number of obs =      28
Sample:  1985 - 2012                                             Lags =       2
-------------------------------------------------------------------------------
                                                                5%
maximum                                              trace   critical
  rank    parms       LL        eigenvalue   statistic    value
    0       6     -1246.4669          .        23.6002    15.41
    1       9     -1235.2803     0.55024        1.2271*    3.76
    2      10     -1234.6668     0.04288
-------------------------------------------------------------------------------
```

Variables are cointegrated. If rank would have been zero there would have been no Cointegration.

Vector Error Correction Model

ECM cannot be used in complex situations like more number of nonstationary variables

A vector error correction model (VECM) adds error correction features to a multi-factor model such as a vector autoregressive model. It is nothing but multivariate specification of ECM.

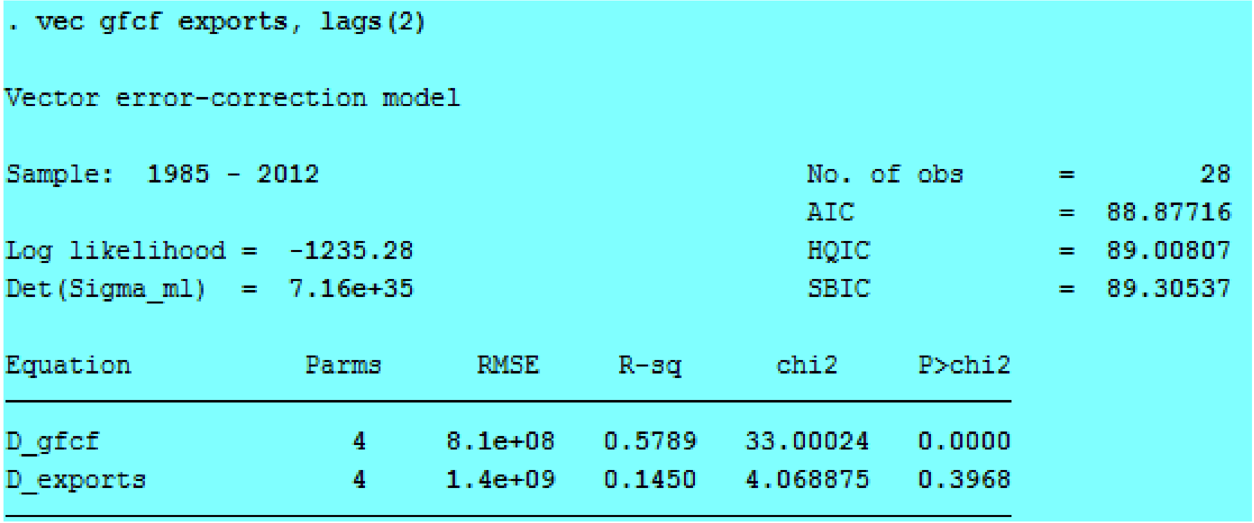
The command in stata is vec. Here there is one lag less than that of VAR but Stata will automatically subtract the lag and you do not need to do that.

The syntax of vec is $vec\ varlist\ [if]\ [in]\ [, options]$.

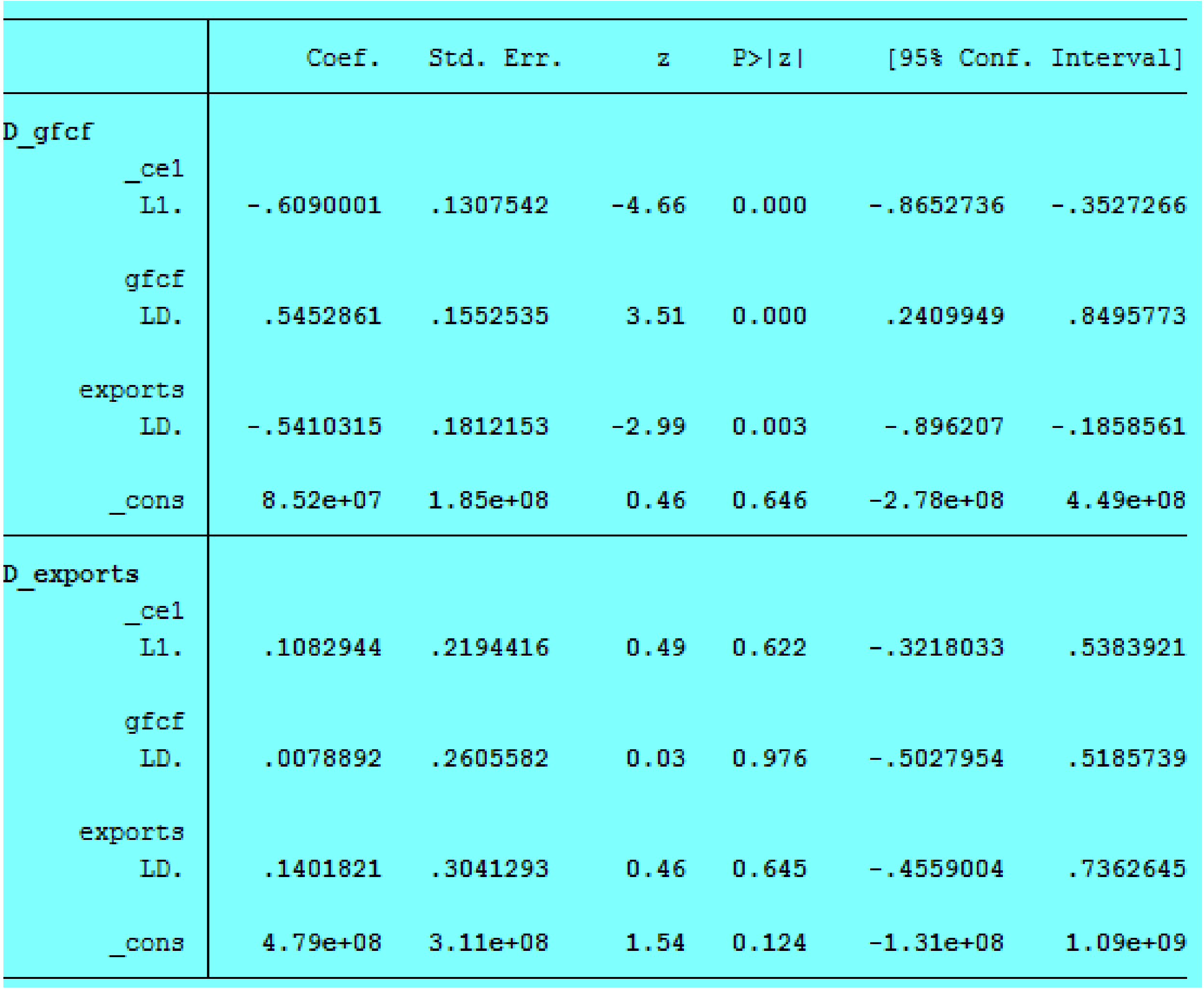
For the menu we can click on

Statistics > Multivariate time series > Vector error-correction model (VECM)

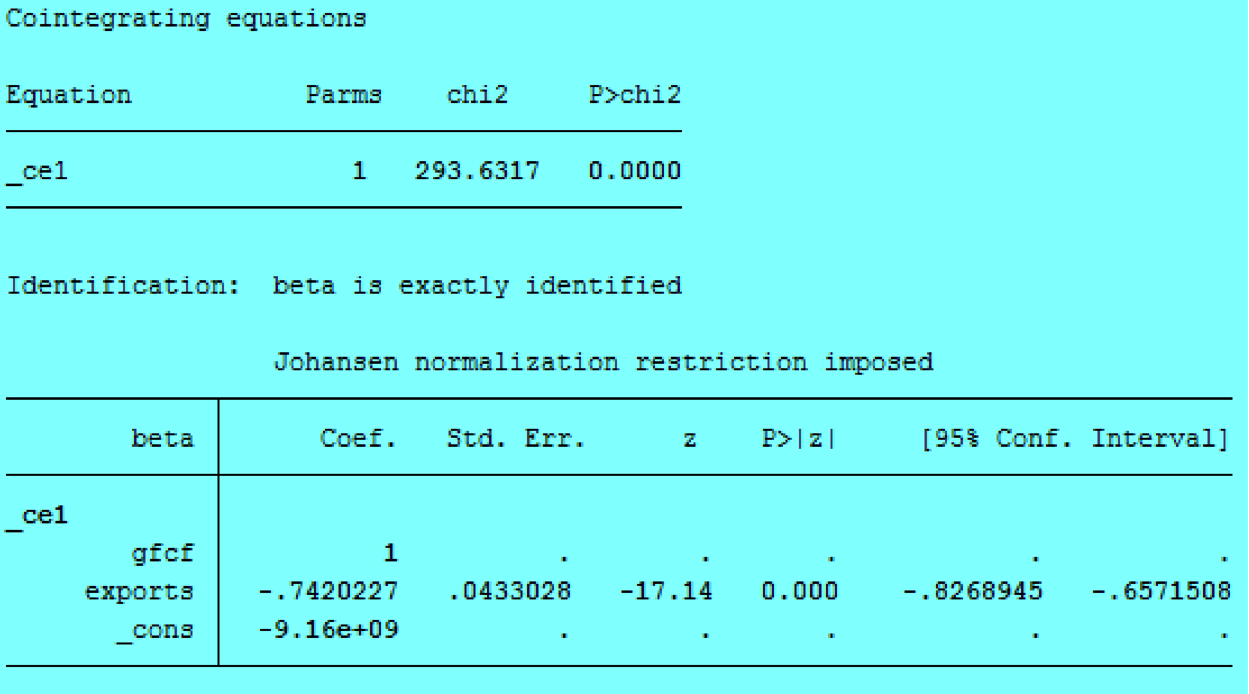
Here is an example:



```
. vec gfcf exports, lags(2)

Vector error-correction model

Sample:  1985 - 2012                            No. of obs      =        28
                                                AIC             =  88.87716
Log likelihood =  -1235.28                      HQIC            =  89.00807
Det(Sigma_ml)  =  7.16e+35                      SBIC            =  89.30537

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
D_gfcf               4     8.1e+08   0.5789   33.00024   0.0000
D_exports            4     1.4e+09   0.1450   4.068875   0.3968
----------------------------------------------------------------
```

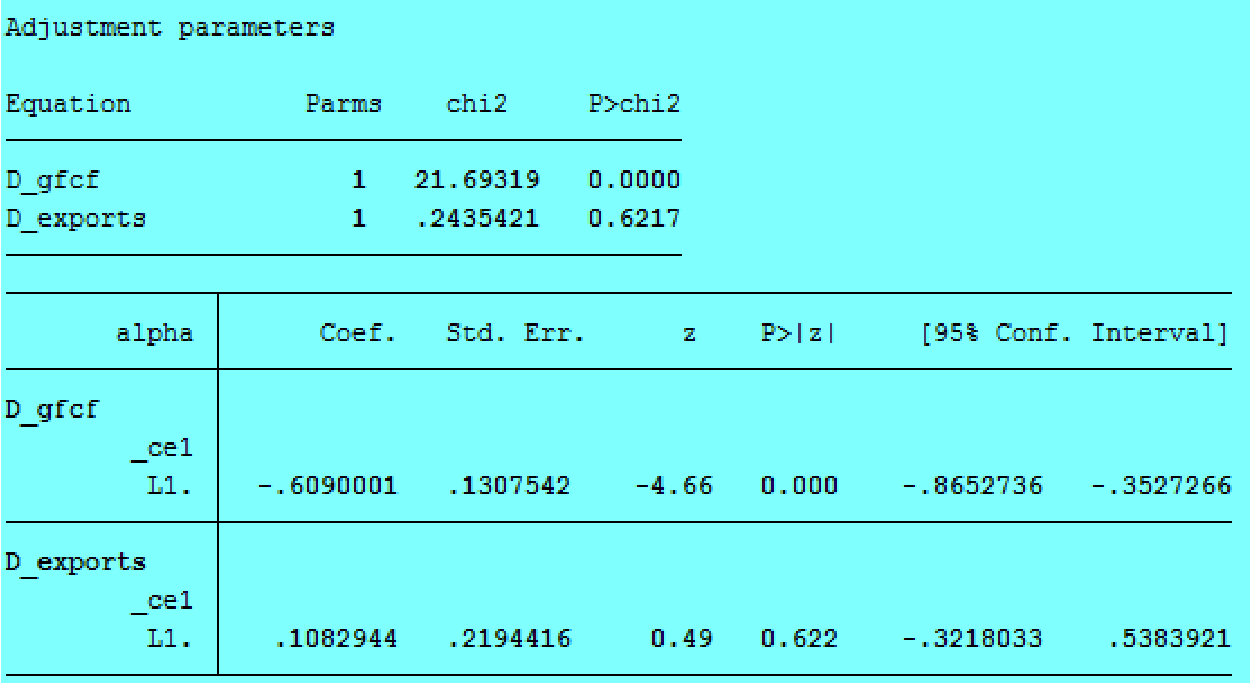


```
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
D_gfcf       |
        _ce1 |
         L1. |  -.6090001   .1307542    -4.66   0.000    -.8652736   -.3527266
             |
        gfcf |
         LD. |   .5452861   .1552535     3.51   0.000     .2409949    .8495773
             |
     exports |
         LD. |  -.5410315   .1812153    -2.99   0.003     -.896207   -.1858561
             |
       _cons |   8.52e+07   1.85e+08     0.46   0.646    -2.78e+08    4.49e+08
-------------+----------------------------------------------------------------
D_exports    |
        _ce1 |
         L1. |   .1082944   .2194416     0.49   0.622    -.3218033    .5383921
             |
        gfcf |
         LD. |   .0078892   .2605582     0.03   0.976    -.5027954    .5185739
             |
     exports |
         LD. |   .1401821   .3041293     0.46   0.645    -.4559004    .7362645
             |
       _cons |   4.79e+08   3.11e+08     1.54   0.124    -1.31e+08    1.09e+09
------------------------------------------------------------------------------
```



```
Cointegrating equations

Equation           Parms    chi2     P>chi2
-------------------------------------------
_ce1                   1   293.6317   0.0000
-------------------------------------------

Identification:  beta is exactly identified

                 Johansen normalization restriction imposed
------------------------------------------------------------------------------
        beta |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_ce1         |
        gfcf |          1          .        .       .            .           .
     exports |  -.7420227   .0433028   -17.14   0.000    -.8268945   -.6571508
       _cons |  -9.16e+09          .        .       .            .           .
------------------------------------------------------------------------------
```

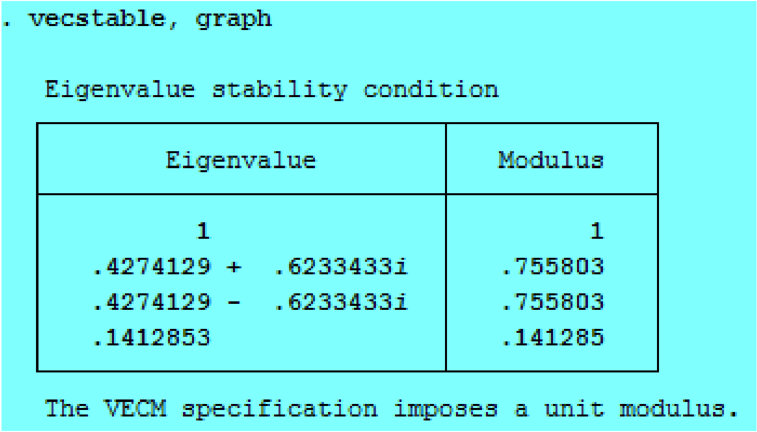
If we use the option *alpha*, we get the short run adjustment parameters as well. This would be in addition to the previous results as



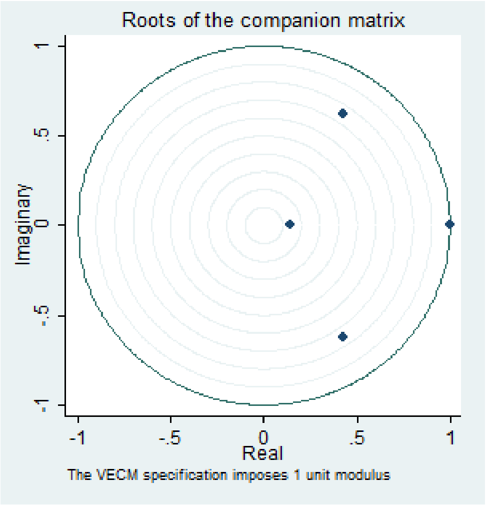
```
Adjustment parameters

Equation           Parms    chi2     P>chi2
-------------------------------------------
D_gfcf                 1   21.69319   0.0000
D_exports              1   .2435421   0.6217
-------------------------------------------

------------------------------------------------------------------------------
       alpha |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
D_gfcf       |
        _ce1 |
         L1. |  -.6090001   .1307542    -4.66   0.000    -.8652736   -.3527266
-------------+----------------------------------------------------------------
D_exports    |
        _ce1 |
         L1. |   .1082944   .2194416     0.49   0.622    -.3218033    .5383921
------------------------------------------------------------------------------
```

It looks like gfcf responds faster than exports in case of changes or shocks.

Post Estimation commands for vec



```
. vecstable, graph

  Eigenvalue stability condition

  ┌───────────────────────────┬──────────────┐
  │        Eigenvalue         │   Modulus    │
  ├───────────────────────────┼──────────────┤
  │            1              │            1 │
  │    .4274129 +   .6233433i │      .755803 │
  │    .4274129 -   .6233433i │      .755803 │
  │    .1412853               │      .141285 │
  └───────────────────────────┴──────────────┘
   The VECM specification imposes a unit modulus.
```

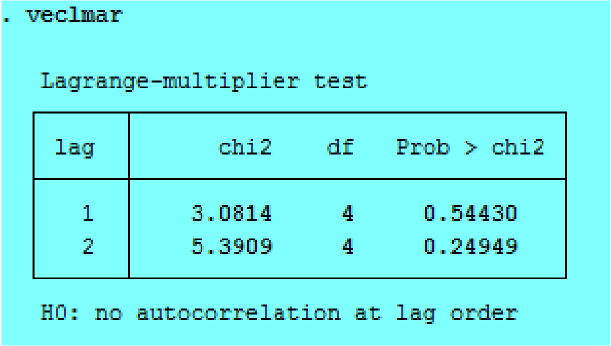


Roots of the companion matrix

The VECM specification imposes 1 unit modulus

Usually eigenvalues should be within the circle

Post Estimation: *veclmar* test the autocorrelation of residuals



```
. veclmar

   Lagrange-multiplier test

  ┌──────┬──────────────────────────────────────┐
  │ lag  │        chi2     df    Prob > chi2    │
  ├──────┼──────────────────────────────────────┤
  │   1  │      3.0814      4      0.54430      │
  │   2  │      5.3909      4      0.24949      │
  └──────┴──────────────────────────────────────┘
   H0: no autocorrelation at lag order
```

Here we do not detect autocorrelation.