BLOCK 4
ECONOMETRIC SOFTWARE PACKAGES

INTRODUCTION TO BLOCK 4

Block 4 is on Introduction to Econometric Software Packages. It has three units (Units 10 to 12). **Unit 10** is on 'Introduction to R'. Beginning with an exposure to its Graphical User Interface, the unit exposes you to the command for operations like (i) generation of matrix and data frame, (ii) importing data file, (iii) obtaining results of descriptive statistics and tests of significance, (iv) regression diagnostics and results of multiple regression and (v) panel data regression. The last section on 'panel data models' describes the commands used in R for obtaining results for (i) fixed effects model, (ii) random effects model and (iii) performing Hausman test for determining which of the two, viz. the fixed or the random effect, should be chosen.

Unit 11 is on Introduction to EViews. This unit too exposes you to perform basic functions like (i) data transformation, (ii) statistical functions and (iii) generation of graphs and scatter plots. Other areas covered for commands required to obtain results of statistical tests include (i) simple and multiple regression and (ii) panel data regression.

Unit 12 is on Introduction to Stata. The unit begins with an exposure to its Graphical User Interface in terms of its (i) main menu, (ii) command and dialogue box, (iii) the data editor, (iv) the data viewer and (v) the 'do file' editor. It then gives you an exposure on how to perform basic operations for data input. The statistical operations exposed for obtaining results include: (i) descriptive statistics, (ii) generation of new variables and (iv) changing of variable values. Other topics of statistical importance covered for commands under this software include: (i) analysis of variance and regression analysis, (ii) regression diagnostics and (iii) testing of hypothesis.

UNIT 10 INTRODUCTION TO R*

Structure

10.0 OBJECTIVES

After reading this unit, you will be able to:

- outline the features of the 'graphical user interface' of R with a brief note on installation of additional 'packages' in R;
- write the codes used in R for performing the basic mathematical operations and for generating 'matrices and data frame';
- state the code or command in R for 'importing a data file';
- describe the codes in R for obtaining the results for 'descriptive statistics' and for performing select 'tests of significance';
- illustrate the procedure for obtaining the results for 'one way ANOVA' in R;

* Dr. Sudip Mukherjee, Asst. Prof. of Economics, Dinabandhu Mahavidyalaya, Bongaon, West Bengal.

- explain the procedure for obtaining the results of 'regression' in R;
- present the details of testing for 'regression diagnostics' in R; and
- discuss the details for the testing of 'panel data regression' in R.

10.1 INTRODUCTION

R is an open source software. It is freely downloadable from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org. R Studio allows the user to run R in a user-friendly environment. R enables us to fulfil the following type of computations: (i) algebraic operations, (ii) statistical analysis and (iii) advanced visualisation of data using different graphical techniques. In this unit, we shall mainly deal with statistical analysis part of R. The unit mainly presents 'commands' or 'codes' for getting the results of 'statistical data analysis' in R. For illustrating the outputs presented by R, upon executing a command indicated, the unit uses the databases contained in R's library as well as other relevant data sets. The databases used relate to: (i) 'cars' data of R on 'distance and speed' for the results of simple regression model and test for multicollinearity, (ii) mreg data (dataset given in appendix) for results of multiple linear regression, normality test, heteroscedasticity and autocorrelation and (iii) EmplUK data of plm library (for results of panel regression).

10.2 GRAPHICAL USER INTERFACE OF R

Take a look at Fig. 10.1. The top left corner is the source window. It is used to edit the script and run it. The **R script** is where you keep a record of your work. We should write the commands on R script. On the top of the script there is a tab named Run. Run is the tab which is used to execute the command after writing in script. The window in the bottom left corner is '**console**'. The 'console' is where we see the 'output' presented by R. The 'console' can also be used to write the commands. The upper right window is the 'workplace window'. This shows all the active objects and stores all the variables used during the execution of commands. The '**history**' tab shows a list of commands used thus far. Next to it, way below, is the '**files**' tab. It shows all the 'files and folders' in the default workspace as if we are on a PC window. The '**plots**' tab shows the graphs. The above tabs, highlighted, are some of the important tabs useful in the 'user interface page' (start page) of R'.

In the R-studio (a term used to indicate the working space in R), 'packages' refer to collections of 'R functions, data and codes'. The directory is the place where packages are stored. It is called the 'library' in the 'R' environment. R comes with a standard set of packages i.e. by default R installs a set of packages (called R base) during its installation. More packages can be added-on for specific purposes.

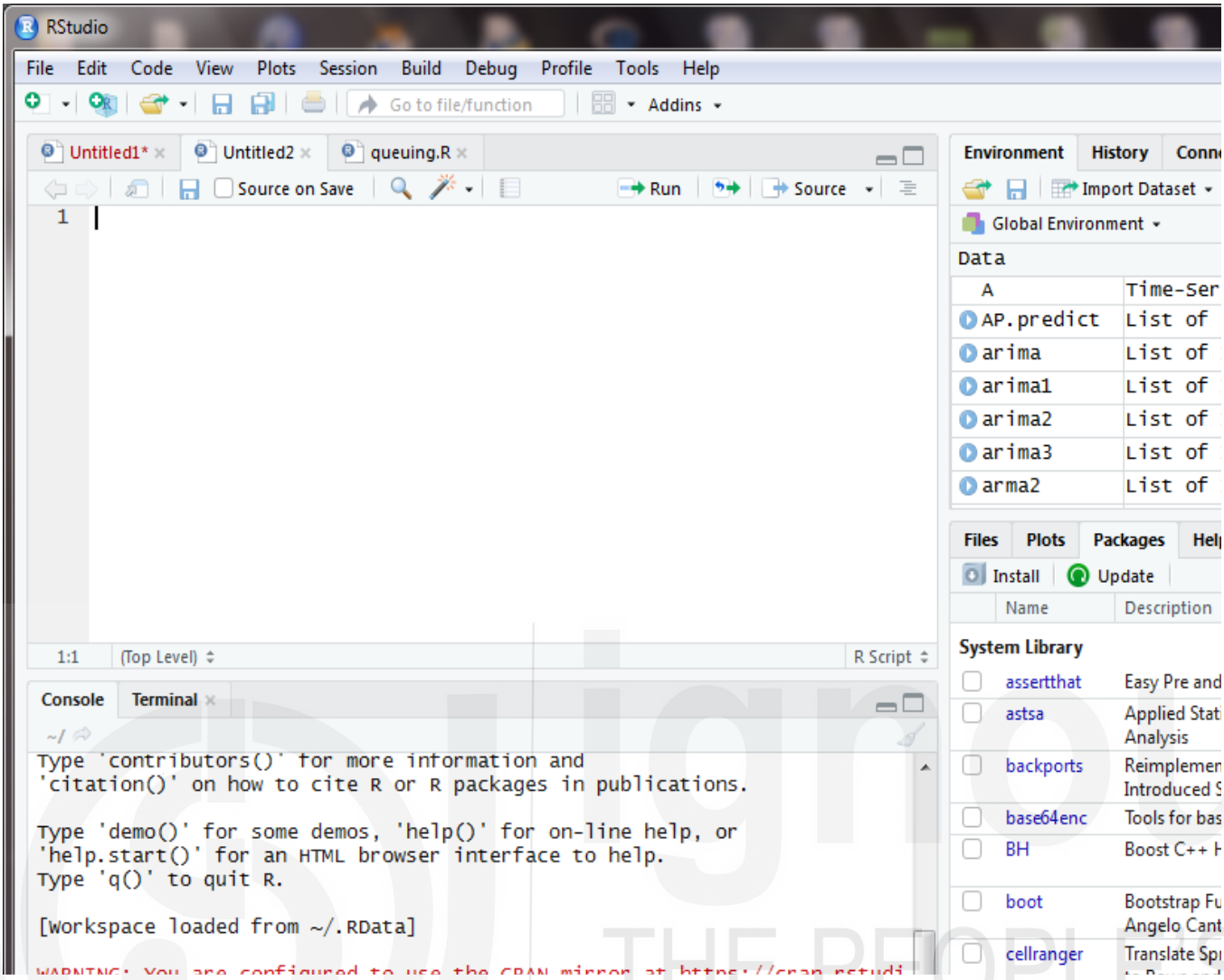


Fig. 10.1: Graphical User Interface of R Studio

When we start the R console, only the base packages are available by default. Other packages have to be explicitly 'loaded' for use by the R program. R can be extended up to 9,200+ packages available on CRAN (Comprehensive R Archive Network). Many useful 'R' functions come in packages. They are free 'libraries of codes' written by R's service community. To install an 'R' package, one needs to open an 'R' session and type in the command line:

install.packages ("<the package's name>") (10.1)

R will download the package from CRAN. Hence, one needs to be connected to the internet. Once the user has a package installed, he can use its contents in the current R session by typing/running the command:

library("<the package's name>") (10.2)

The above command (or code) will be executed by R to run the concerned package. For instance, we can install the MS EXCEL package in R by typing the command:

```
install.packages("readxl")
library("readxl")
```
(10.3)

Note that there are two lines in (10.3). Both the lines are needed to be typed in the 'R Script'. The first line 'installs' the excel package in the 'R's library'. The second line 'fetches' it from the library making it available for 'use' in the current session.

10.3 BASIC OPERATIONS AND RESULTS

R script is a plain text file in which we can write and save in R code. After opening the R studio, we first open the R script. For this, we first click on the 'file menu' in the top left corner of the opening page. We next click on 'New File' and then on 'R script'. A new R script will open. Therefore, the steps to open a new script are: File →New File→ R Script. We can use R studio as a calculator to obtain the result of any arithmetical operation. For instance, in R script we can write 4+7 and then click on run. In console window, R-studio returns the result as 11. Table 10.1 illustrates the arithmetic operator in R along with the command (or code) and result.

Table 10.1: Illustrative Codes and Result for Basic Operations in R

Description	Operator	Code	Result
Addition	+	**4+7**	**11**
Subtraction	-	**5-3**	**2**
Multiplication	*	**5*3**	**15**
Division	/	**15/3**	**5**
Exponent	^	**2^3**	**8**
Integer Division	%/%	**13%/%4**	**3**

We can create important objects in R like vector, matrix, data frame and list with simple commands. For instance, we use the c( ) function to write a vector. For example, we can create a vector named Income, say with three values 100, 120, 150, by writing in script and then clicking on run as follows: Income<-c(100,120,150). Thus, in general, for creating a vector R, the command is:

"name"<-c(x_1, x_2, …..x_n) (10.4)

Now, suppose we write similar three value entries for another vector, consumption as: consumption <–(70,90,110). Now, to subtract the 'consumption' vector from the 'income' vector to get the 'savings' vector, we use the command:

Savings<-Income-Consumption (10.5)

R will display the savings vector as: # [1] 30 30 40.

10.3.1 Matrix and Data Frame

A matrix is a rectangular array with *p rows and n columns.* An element in the *i-th row and j-th column is denoted by X[i, j], i* = 1,2,...,*n, j* = 1,2,...,*p.* In R, a 4 × 2 matrix 'X' can be created with the command*:*

>X<-matrix(nrow=4, ncol=2, data=c(1,2,3,4,5,6,7,8)) (10.6a)

The result appears in the console as follows:

```
X
[,1] [,2]
[1,] 1 5
[2,] 2 6
[3,] 3 7
[4,] 4 8
```

We can create the data frame using the following command:

data<- data.frame (Income, Consumption, Savings) (10.6_b)

data (10.6_c)

The output in console appears as:

Income	Consumption	Savings
1 100	70	30
2 120	90	30
3 150	110	40

10.3.2 Importing a Data File

The R command for importing data file needs it to be first read as:

read.<format>("file name") (10.7)

Then, to import a .csv file the command is:

read.csv(D:/mydata/test.csv, header=TRUE) (10.7.a)

Likewise, to import a .text file the command is

read.table (D:/mydata/test.txt, header=TRUE) (10.7.b)

The steps for importing an excel data file are:

Step 1: install.packages("readxl")

Step 2: library("readxl")

Step 3: File →Import→ From Excel

Step 4: click on excel tab, then click on 'browse' to select your data.

Note that giving blank space while giving command is allowed in R.

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) Given, price<-(10,15,18) and quantity<-(40,35,30), compute 'expenditure'.

……

……

……

……

……

2) Create a 'data frame' using the vectors in 1 above.

……

……

……

……

……

3) State the steps to open a new script. What role a 'script' plays in R?

……

……

……

……

……

10.4 DESCRIPTIVE STATISTICS AND TESTS OF SIGNIFICANCE

We can calculate the measures of central tendency and dispersion in R using the commands given here. Let us create a vector X using the following code:

X<-c (23, 45, 29, 56, 78, 45). We can now calculate the mean of the vector X by writing the code 'mean(X)' in the script as follows:

<-mean(X) (10.8a)

After clicking the run tab the mean value appears in the console. Likewise, the codes for calculating the median and variance for the vector X are as follows:

<-median(X) (10.8_b)

<-var(X) (10.8_c)

After calculating the variance, we can calculate the standard deviation by giving the command:

B<-var(X) (10.8d)

sd<-sqrt(B) (10.8e)

The results for the sample data X on mean, median, variance and 'sd' that R produces on console along with sample data X are as below:

```
> X<-c (23, 45, 29, 56, 78, 45)
>mean(X)
[1] 46
>median(X)
[1] 45
>var(X)
[1] 388.8
> B<-var(X)
>sd<-sqrt(B)
>sd
[1] 19.71801
```

10.4.1 Independent t-Test

We use the t-test procedure to test the hypothesis that the means of two groups are not significantly different. In other words, it is used to test the equality of means among two groups. When the groups are same (e.g. before-after/pre-post) we use paired t-test. When the groups are different, we use independent t-test. Let us consider the average weight of half kg cake packets baked by two bakers X and Y. Let the weights measured for 8 independent packets be as follows:

X: 512 530 498 540 521 528 505 523

Y: 499 500 510 495 515 503 490 511

Our objective is to test that the average weight of packets is same for the two bakers. In this example, the two bakers are different and hence the test is independent t-test. So we use the command with a two-tailed alternative as follows:

X<- c(512,530,498,540,521,528,505,523)

Y<- c(499,500,510,495,515,503,490,511)

t.test (X, Y, alternative = 'two.sided',var.equal=TRUE) (10.9)

After executing the R code for t-test, the following test results are given by R (Table 10.2).

Table 10.2: Results of Independent t-test

Two Sample t-test
data: X and Y
t = 2.9058, df = 14, p-value = 0.01151
alternative hypothesis: true difference in means is ≠ 0
95 percent confidence interval:
4.386709 29.113291
sample estimates:
mean of X mean of Y
519.625 502.875

Since $p < 0.05$, we cannot say that the evidence supports the conclusion that the null is true. We therefore accept the alternative hypothesis and conclude that 'the difference in the two means is statistically significant for the difference'.

10.4.2 Paired t-Test

In a similar way, the R command for a 'paired t-test' [with the values of data on the 'test scores' of 10 students as entered] for the difference in scores for 'before and after' the consumption of a health booster beverage is as follows:

```
Y <– c(4,3,5,6,7,6,4,7,6,2)
X <– c(5,4,6,7,8,7,5,8,5,5)
t.test (xp, yp, paired=TRUE)                    (10.10)
```

The results of the paired t-test's generated by R is as in Table 10.3.

Table 10.3: Results of Paired t-test

data: X and Y
t = 3.3541, df = 9, p-value = 0.008468
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.325555 1.674445

The results of the paired t-test has the p-value (0.008) less than .05. Hence, we accept the alternative hypothesis. We conclude that there is a significant

difference in the points scored by the students between 'after and before'of drinking a cup of beverage.

10.4.3 One Way ANOVA

We can use the one-way ANOVA procedure to test the hypothesis that the means of two or more groups are not statistically different. To get the result in R on one-way ANOVA, we use the data on test scores in an examination. The test scores are taken as the 'dependent variable' and the type of examination administered (3 types) as the 'independent variable'. Our objective is to find out that the score obtained by the students do NOT vary due to the examination type administered. We key-in (i.e. enter) the test scores with the ANOVA command as follows:

```
type1.scores <- c(95,91,89,90,99,88,96,98,95)

type2.scores <-c(83,89,85,89,81,89,90,82,84)

type3.scores <- c(68,75,79,74,75,81,73,77,80)

Score <- c(type1.scores, type2.scores, type3.scores)

Type <- rep(c("type1", "type2", "type3"), Times=c( length(type1.scores),
length(type2.scores), length(type3.scores)))

data<- data.frame(Type, Score)

result<-aov(Score ~ Type, data = data)                    (10.11)
```

R produces the ANOVA output as in Table 10.4.

Table 10.4: Results of One Way ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Type	2	352.7	176.33	13.01	0.000149 ***
Residuals	2	4 325.3	13.56		

Significance codes: '***' 0.001 '**' 0.01 '*' 0.05

The value of F statistic is 13.01 with the 'p' value 0.0001 which is less than .05. Hence, we reject the null hypothesis of 'no difference in test scores' and accept the alternative hypothesis. We conclude that the variable 'examination type' does make a significant difference to the scores of the students.

10.5 REGRESSION

In data analysis, we run regressions to examine the causality between the dependent and independent variables. After performing the regression analysis, we aim at concluding how much would be the response in terms of change in the value of dependent variable corresponding to a 1 unit change in

the value of the independent variable. In order to see how R can be used in regression, we first take up a two variable simple linear regression model. For data, we consider the dataset named 'cars' available in the R library. The two variables of this dataset are: speed and distance (dist). We consider 'dist' as the dependent variable and 'speed' as the independent variable. To find out the causal relationship between speed and dist, we consider the equation: dist = $\beta_0 + \beta_1$ speed + u. Before attempting the regression, we have to check for the linearity of the relationship between the dependent and the independent variable. For this, we use the 'scatter diagram'. The R command for the scatter diagram is:

scatter.smooth(x=cars$speed,y=cars$dist, main="Dist ~ Speed") (10.12)

Execution of the above command makes R generate the scatter diagram as in Fig. 10.2.

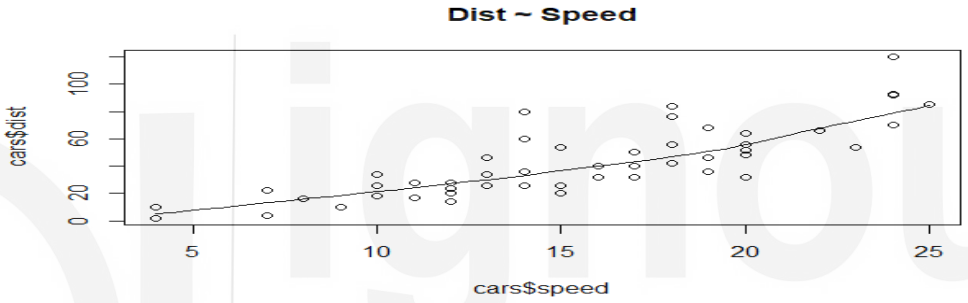


Fig. 10.2: Scatter Diagram Between Distance and Speed

It is clear from Fig. 10.2 that there is a linear relationship between distance and speed. We next calculate the correlation between the two variables using the following R command.

cor(cars$speed, cars$dist) (10.13)

The value of the correlation coefficient displayed by R is 0.80. Since the correlation coefficient of up to 0.80 is acceptable to proceed without treating for multicollinearity, we can proceed for doing the regression. In R, the command or code used for regression is 'lm()'. The following command is the R command to be used for regression.

linearMod<- lm(dist ~ speed, data=cars) (10.14_a)

summary(linearMod) (10.14_b)

anova(linearMod) (10.14_c)

The results of the above regression are presented in Table 10.5.The value of F-statistics is 89.57 with p-value zero. The model is therefore overall statistically significant. R^2=0.65 implies that 65% of the variation in distance is explained by the independent variable 'speed'. The coefficient of speed is 3.93 with p-value zero. Since the p value < 0.05, the variable 'speed' is a significant contributor to distance. Hence, if speed increases by one unit the

distance increases by 3.93 units. The regression model may therefore be written as:

'dist.' = –17.57 + 3.93 * speed.

Table 10.5: Result of Regression Analysis

```
Call:lm(formula = dist. ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
–29.069  –9.525  –2.272   9.215  43.201


Coefficients: Estimate    Std. Error  t value  Pr(>|t|)
(Intercept)  –17.5791      6.7584     –2.601   0.0123 *
Speed          3.9324      0.4155      9.464   1.49e-12 ***
---
Significance codes: '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12
```

10.5.1 Multiple Linear Regression

To explain the multiple linear regression model, we consider the dataset named "mreg". We can import the data after saving in excel as indicated in sub-section 10.3.2. The Table A-1 (given in annexure) is the data on 'total fertility rate'(TFR) for the 20 states in India. We take TFR as the dependent variable and 'female literacy'(FLIT), location of residence (URBAN) and economic status (POV) as the three independent variables. To find out the causal relationship between the dependent and the independent variables, we consider the following equation:

$$TFR = \beta_0 + \beta_1 FLIT + \beta_2 URBAN + \beta_3 POV + u \qquad (10.15)$$

The following is the command in R to perform the multiple regression.

```
model<-lm(TFR~FLIT+URBAN+POV,data=mreg)        (10.16a)
summary (model)                                (10.16b)
```

The results of the multiple regression generated by R by running the above command is presented in Table 10.5.

Table 10.6: Results of Multiple Regression

Call:lm(formula = TFR ~ FLIT + URBAN + POV, data = mreg)

Residuals:

Min	1Q	Median	3Q	Max
–0.74293	–0.24584	0.02322	0.27228	0.71634

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.179605	0.604323	6.916	3.47e-06 ***
FLIT	–0.031897	0.008894	–3.586	0.00247 **
URBAN	–0.008654	0.010862	–0.797	0.43728
POV	0.013113	0.008279	1.584	0.13278

Significance codes: '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 0.413 on 16 degrees of freedom

Multiple R-squared: 0.6505, Adjusted R-squared: 0.585

F-statistic: 9.927 on 3 and 16 DF, p-value: 0.0006166

The value of F-statistics is 9.93 with the p-value 0.0006 which is less than 0.05. Hence, the model is overall statistically significant. Further, R^2=0.65 implies that 65% of the variation in dependent variable is explained by the independent variables. The coefficient of FLIT is -0.032 with p-value 0.002 (which is < 0.05). Hence, FLIT is a significant variable. The negative sign to FLIT implies that TFR decreases by about 3 percent with a percentage increase in FLIT. The p-value associated with the coefficients of URBAN & POV are greater than 0.05 implying that these variables are not statistically significant in influencing the TFR.

10.5.2 Regression Diagnostics

We know that if we want the results from OLS method to hold with its 'best' properties, we should be sure that the dataset satisfies the assumptions of the 'classical linear regression models' (CLRM). Specifically, these assumptions relate to the following four:(i) there is no multicollinearity in our data, (ii) the variance of the residuals is constant (i.e. there is no heteroscedasticity), (iii) the values of the residuals are independent (i.e. there is no autocorrelation) and (iv) the values of the residuals are normally distributed. We shall now note the commands in R for testing or detecting for the presence or absence of the indicators behind these assumptions.

Test for Multicollinearity: . To detect this, we know that the 'variance inflation factor (VIF)' can be used. To compute the VIF, the following command is used in R.

```
library(car)                    (10.17a)
vif(model)                      (10.17b)
```

Execution of the above two commands, produces the output as in Table 10.6 by R.

Table 10.7 : VIF Values for Car Dataset

FLIT	URBAN	POV
1.309584	1.204745	1.202844

As a rule of thumb, a variable with VIF value greater than 2, needs further investigation. But in this example, the VIFs for the independent variables are all below 2. Thus, we can conclude that there is no serious multicollinearity in the data set.

Test for Heteroscedasticity: To check for the homogeneity of error variance the following is the R code.

```
library(lmtest)                 (10.18a)
bptest(model)                   (10.18b)
```

Execution of the above two commands in (10.18) produces the following output (Table 10.8) by R.

Table 10.8: Breusch-Pagan Test Result

Studentized Breusch-Pagan Test
data: model
BP = 1.3637, df = 3, p-value = 0.7141

The null hypothesis of the Breusch-Pagan (BP) test is 'error variance is constant' or 'there is no heteroscedasticity'. In this example, since the p-value is greater than 0.05, it suggests that there is no statistical significance to reject the null hypothesis. We accept the null and conclude that the error variance is constant.

Test for Autocorrelation: Another assumption for efficiency of the OLS estimator is that the values of the residuals are independent. To test for this, the following R command is used to check for the presence or otherwise of autocorrelation.

```
library(lmtest)                 (10.19a)
v<-dwtest(model)                (10.19b)
```

The result of DW test displayed by R is as in Table 10.9.

Table10.9: Result of Durbin-Watson Test

Durbin-Watson test
data: model
DW = 2.0656, p-value = 0.5605
alternative hypothesis: true autocorrelation is greater than 0

The value of the Durbin-Watson statistics is 2.066 with p-value greater than 0.05. Since the p-value is greater than 0.05, we have no basis to reject the null hypothesis which is 'there is NO autocorrelation'. The test result implies that there is no autocorrelation in the dataset.

Test for Normality: One of the assumptions of the OLS regression is error terms are normally distributed. To verify for this, we give the following command in R to obtain the quantile-quantile (Q-Q) plot.

plot(model) (10.20)

Fig.5.3 represents the normal Q-Q plot generated by R. When the dots are nearer to the diagonal line, then it can be concluded that the errors are distributed normally. In Q-Q plot, all the dots are close to the diagonal line. From the diagram we can conclude that the errors are normally distributed.

Check Your Progress 2 [answer within space given in about 50-100 words]

1) Write the command in R for obtaining the multiple linear regression for three independent variables as in (10.15). For this, consider that you are using the dataset stored in R library by the name 'mreg'. Also, state the expected relationship between the variables considered.

.....

.....

.....

.....

.....

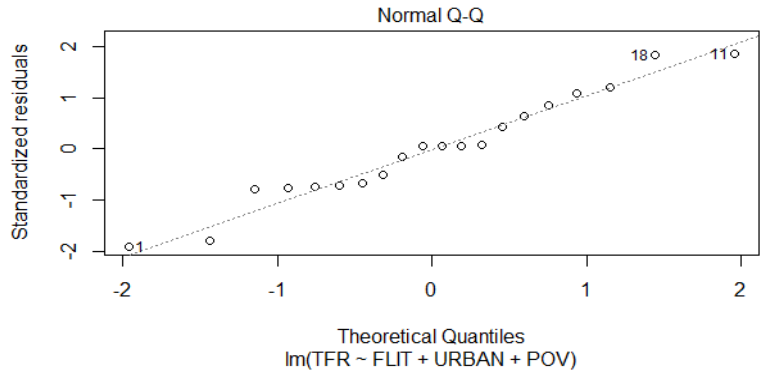


Fig. 5.3: Q-Q plot

2) In the results of the multiple linear regression obtained above, what is the significance status of FLIT? Why?

.....

.....

.....

.....

.....

10.6 PANEL DATA REGRESSION

In R, the function 'plm' is used for 'panel data regression'. We use panel data regression in data sets where both the cross section and time series dimensions are present. The elements of plm function are as follows:

Y = Dependent variable

X = independent variable

Data = data file name

Index = c("cross section id", "time series id")

Model = pooling – used for pooled regression

within- used for fixed effects model

random- used for random effects model

We write the plm function in the following form

plm(Y~X, data, index, model) (10.21)

10.6.1 Fixed Effects Model

For illustrating the results of 'panel regression', we use the 'Emp1UK' data of plm package in R. We can call for the data by using the command:

data("EmplUK") (10.22)

Suppose we are interested in finding the effect of wages on employment. Here 'employment' is the dependent variable and 'wage' is the independent variable. We 'call' the library by the command:

```
library(plm)                                                    (10.23)
```

For getting the results of panel regression for the case of the fixed effects model, the following command is to be given in R:

```
fe<-plm(emp~wage, data =EmplUK, index=c("firm","year"),
model="within")                                                 (10.24a)
summary(fe)                                                     (10.24b)
```

After clicking the run tab, the result displayed in the console by R is given in Table 10.10. The p-value of F-statistic is less than 0.05. This implies that the overall fit of the model is good. The p-value of the coefficient of wage is 0.0001 which is also less than 0.05. Hence, the variable 'wage' is a significant predictor of employment. This means, if wage increases by one unit, employment decreases by 0.13 units. The result suggests an inverse relationship between employment and wage.

Table 10.10: Results of Fixed Effects Model

```
One way (individual) effect Within Model

Call:

plm(formula = emp ~ wage, data = EmplUK, model = "within",

index = c("firm", "year"))

Unbalanced Panel: n = 140, T = 7-9, N = 1031

Residuals:

    Min.      1st Qu.       Median     3rd Qu.        Max.

-22.595206  -0.307023   0.027979   0.351118   27.454176

Coefficients:

          Estimate      Std. Error     t-value          Pr(>|t|)

wage-   0.136878      0.035533     -3.8521          0.0001255 ***

---

Significance codes: 0 '***' 0.001 '**' 0.01 '*'

Total Sum of Squares: 5030.6. Residual Sum of Squares: 4948.1

R-Squared: 0.0164.   Adj. R-Squared: -0.13832

F-statistic: 14.8391 on 1 and 890 DF. p-value: 0.00012548
```

10.6.2 Random Effects Model

Using the plm package from the R library, the codes for obtaining the random effects model results are as follows:

```
re<-plm(emp~wage, data =EmplUK, index=c("firm","year"),
model="random")                                          (10.25a)

summary(re)                                              (10.25b)
```

After clicking the run tab, the result displayed in the console is given below (Table 10.11).

Table 10.11: Results of Random Effects Model

```
One way (individual) effect Random Effect Model

 Call:

plm(formula = emp ~ wage, data = EmplUK, model = "random",

 index = c("firm", "year"))


Unbalanced Panel: n = 140, T = 7-9, N = 1031

Effects:

varstd.devshare

idiosyncratic      5.560        2.358          0.022

individual         248.638      15.768         0.978

 Min.      1st Qu.       Median     Mean      3rd Qu.       Max.

 0.9436    0.9436        0.9436     0.9450    0.9472        0.9502

Residuals:

      Min.      1st Qu.       Median      Mean      3rd Qu.      Max.

   -19.4797    -0.6047       -0.3073     -0.0075   0.1078       29.4516

Coefficients:

                Estimate      Std. Error     z-value          Pr(>|z|)

(Intercept)   11.557156     1.584263       7.2950           2.987e-13 ***

wage          -0.141571       0.035293       -4.0113          6.038e-05 ***

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05
```

Total Sum of Squares: 5847.5
Residual Sum of Squares: 5750.7
R-Squared: 0.016588
Adj. R-Squared: 0.015633
Chi sq: 17.3224 on 1 DF, p-value: 3.1545e-05

The p-value of Chi square statistic is less than 0.05. This implies that the overall fit of the model is good. The p-value of the coefficient of 'wage' is 0.000 which is also less than 0.05. Hence, the variable 'wage' is a significant predictor of employment. This means, if wage increases by one unit, employment decreases by 0.14 units. The result suggests an inverse relationship between employment and wage.

10.6.3 Hausman Test

We can run Hausman test to decide between fixed and random effects. The null hypothesis of the Hausman test is 'the required model is of random effect'. Therefore, if the p-value of the Hausman test is greater the 0.05, then we reject the null and use the 'fixed effects' model. The code for Hausman test in R is:

"phtest()" (10.26)

By running (10.26) on our dataset, R produces the following output (Table 10.12).

Table 10.12: Results of Hausman Test

data: emp ~ wage
chisq = 1.2968, df = 1, p-value = 0.2548
alternative hypothesis: our model is inconsistent

The p-value of Hausman test is greater than 0.05. Hence, we reject the null hypothesis of 'random effect'. We conclude that we must go for the 'fixed effects' model.

10.7 LET US SUM UP

The unit has presented the R commands for performing t-test, one-way ANOVA, regression analysis, tests for normality, multicollinearity, auto correlation, heteroscedasticity and panel regressions. The codes or command for all these tests are indicated. We have used the datasets in the library of R for some of these tests. For a few others, we have entered the sample data

into the command. A summary of datasets used from the R-library, or entered sample data directly, by 'type of test' used for illustrating the various statistical tests in the unit is given in Table 10.13.

Note: The objectives indicated in Section 10.0 are generally stated. For questions in 'assignments' and 'term end exam' on these lines, you would be required to present the command in R for getting the results followed by 'presentation of the output' and comments by way of 'interpretation of the results'. For this, it is enough if major indicators (e.g. value of the test statistic, estimated co-efficient, value of p, etc.) are given. The values so given in the 'output' could be imaginary and not actual. Using them, in the interpretation of the results, you can show your analytical awareness on the expected 'direct or inverse' relationship between the variables, the justification to 'reject or not reject' the null hypothesis based on the p-value, etc. This part would therefore be not common and would carry the scope to vary between answers of different learners.

Table 10.13: Summary of Data Sets Used in the Unit

Sl. No.	Type of Test	Sample Data	Database from R Library	Sample Size
1	Independent t-Test	Yes	-	16
2	Paired t-Test	Yes	-	20
3	One Way ANOVA	Yes	-	27
4	Regression	No	'Cars'	50
5	Multiple Regression	No	mreg*	20
6	VIF Test	No	'mreg'	20
7	BP Test	No	'mreg'	20
8	DW Test	No	'mreg'	20
9	FE Model (PR)	No	'EmplUK'	1031/140 (N/n)
10	RE Model (PR)	No	'EmplUK'	1031/140 (N/n)
11	Hausman Test	No	'EmplUK'	1031/140 (N/n)

* Data set for 20 States is given in Annexure

10.8 KEY WORDS

library : This is the directory where the packages are stored. It is called as the "library" in the 'R' environment.

c() : This is the function to write a vector in R.

data.frame() : This is the function to write data frame.

t.test : This is the command for obtaining the results of t-test in R.

lm : This is the command for obtaining the results of regression in R.

plm : This is the command for obtaining the results of panel regression in R.

phtest() : This is the code for the application of Hausmann Test in R.

10.9 SUGGESTED BOOKS FOR FURTHER READING

1) Mailund, Thomas (2017). Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist, Apress.

2) Heumann, Christian, Shalabh, Michael Schomaker (2016). Introduction to Statistics and Data Analysis with Exercises, Solutions and Applications in R, Springer.

3) Bhaumik, Sankar (2017). Principles of Econometrics, A Modern Approach Using Eviews, Oxford University Press.

10.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Expenditure = price * quantity. Hence, Expenditure = <-(400,525,540)

2) <-data frame (price, quantity, expenditure)

3) File → New File → R Script. The 'R script' is where all records of our work is kept. The commands are written into the R script.

Check Your Progress 2

1) The commands in R for this multiple regression are as in (10.16). The dependent variable is the 'total fertility rate' TFR. It is assumed or hypothesized to have a linear relationship between three independent variables viz. female literary (FLIT), location of residence (URBAN) and economic status (POVERTY).

2) FLIT is a significant determinant of TFR because the p-value associated with the estimated coefficient of FLIT is < 0.05. In such situations, we cannot say there is evidence to say that the null is true (which in this case is: FLIT is NOT related to TFR). We therefore reject the null hypothesis

and accept the alternative hypothesis. This means FLIT is a significant determinant of TFR. This means, if FLIT increases, TFR decreases i.e. there is an inverse relationship between the two.

Annexure: TFR Dataset for States of India

(percent)

State	TFR	FLIT	URBAN	POV
Andhra Pradesh	1.8	50.4	27.3	15.8
Assam	2.4	54.6	12.9	19.7
Bihar	4	33.1	10.5	41.4
Chhattisgarh	2.6	51.9	20.1	40.9
Gujarat	2.4	57.8	37.4	16.8
Haryana	2.7	55.7	28.9	10
Himachal Pradesh	1.9	67.4	9.8	14
Jammu & Kashmir	2.4	43	24.8	5.4
Jharkhand	3.3	38.9	22.2	40.3
Karnataka	2.1	56.9	34	25
Kerala	1.9	87.7	26	15
Madhya Pradesh	3.1	50.3	26.5	38.3
Maharashtra	2.1	67	42.4	30.7
Orissa	2.4	50.5	15	46.4
Punjab	2	63.4	33.9	8.4
Rajasthan	3.2	43.9	23.4	22.1
Tamil Nadu	1.8	64.4	44	22.5
Uttar Pradesh	3.8	42.2	20.8	32.8
Uttarakhand	2.6	59.6	25.7	39.6
West Bengal	2.3	59.6	28	24.7

Source: Bhaumik(2017), Principles of Econometrics.

UNIT 11 INTRODUCTION TO EVIEWS*

Structure

* As stated in the course Introduction, although time series as a topic by itself is not a part of this course, a small bit necessary for this unit is used in Section 11.4.

11.0 OBJECTIVES

After reading this unit, you will be able to:

- outline with illustrations how the 'basic functions' can be performed in EViews;
- describe the procedure of obtaining the results of 'simple and multiple regression' in EViews; and
- explain the procedure of obtaining the results of regression for 'panel data' in EViews.

11.1 INTRODUCTION

Econometrics deals with three types of data: cross-sectional, time series, and panel (longitudinal). In a time series data, you observe the behaviour of a

* Dr. Poulomi Roy, Jadavpur University.

single entity over multiple time periods. This can range from high frequency data such as financial data (hours, days), to lower frequency data like months or quarterly or annual. Such lower frequency data could be in industrial production, inflation, unemployment rate, GDP, etc. In a cross-section, you analyse data from multiple entities at a single point of time. One big difference between time series analysis and cross-sectional analysis is that the order of the observation numbers does not matter in cross-sections. In time series data, you would lose some of the most interesting features of the data if you shuffled the observations. Panel data is a combination of time series and cross-sectional data since multiple entities are observed at multiple time periods. EViews allows you to work with all three types of data. It is the most commonly used software for time series analysis. You can work with it easily even with cross-sections and/or panel data. EViews allows you to save results within a program and to 'retrieve' these results for further calculations later.

There are two Handbooks for EViews. Of the two, the User's Guide is very useful. It explains all the functions of EViews in a step-by-step manner. In particular, the 'Command and Programming Reference' is useful if you want to write batch programmes to perform a sequence of steps. Both the manuals are available in the EViews help menu itself (as .pdf). The searchable help index in the EViews (help menu: Help > Index and Help > Find) is also very helpful. The best way for getting an idea of how EViews works is to experiment with the commands presented here in this unit and try other commands and options on your own and learn.

You can access most of the EViews functions via menus. You can just browse through the menus, and find the appropriate command. You will be guided through several windows that prompt you for the information required to perform the command. EViews organises data, graphs, output, etc. as objects. Each of these objects can be copied, saved, copy-pasted into other Windows programmes to be used for further analysis. A collection of objects can be saved together in a 'workfile'. Since EViews creates new objects with everything you do, it is sensible to delete unimportant intermediate results to avoid a messy workfile. Please note that you cannot mix data series of different frequencies (annual, quarterly, monthly, weekly, daily) within the same work file page. To create a new workfile, you have to click: File > New > Workfile. If you have quarterly data from the first quarter of 1950 until the last quarter of 2005, then the command is: Select: "dated - regular frequency, frequency 'quarterly', start date: '1950:1', end date '2005:4' ". Note that there are always two series in the workfile: 'c', and 'resid'. They stand for 'coefficients' and 'residuals'. Every time you estimate something, the coefficients are stored in 'c' and the residuals in 'resid'.

11.2 BASIC FUNCTIONS

Even though the copy-paste method is easy enough, there is a second more direct way to import data into EViews from, say, Excel. This involves

starting with a new workfile in EViews. For this, you have to click on 'Proc /Import /Import from File'. A dialog box will open. You will have to first specify the location where your data file resides. After you double click on the file, another dialog box opens. Once you click through a series of 'Next' options, and finally 'Finish', all the data, including the variable names, would have been read in by the EViews. Note that EViews allow you to import other types of data files, for instance, data file saved in STATA format.

11.2.1 Data Transformation

The command 'genr' creates new variables and modifies existing variables. An example is: 'genr expn = expn_stu/1000'. This command generates the expenditure variable by dividing the original data on 'expenditure of students' by 1,000. Likewise, 'genr**avginc2** = avginc^2' and 'genr**lavginc** = log(avginc)' creates the variable 'average income' on 'expenditure of students' and 'log of average income' respectively. Note that 'genr' is the term for generation and ávginc2' is the new variable generated. The numeral 2 indicates the new variable is raised to the power 2. There is no space allowed in EViews between the command and the name of the variable or operation. Note that the command 'genr testscr = testscr/100' simply modifies an existing variable. Many actions can be done more quickly in the command window instead of browsing through the menus. The command window is the white window below the menu bar. Transformations can be more easily specified in the 'command window' like: lgdp = ln(gdp). This command generates a new series 'lgdp' as the natural log of gdp. Some of the most frequently used operators are: + for addition, – for subtraction, * for multiplication, / for division and ^ for exponentiation. Log(X) calculates the natural logarithm of X and exp(X) computes the exponent of X. EViews uses several relational operators. These are: < for less than, > for greater than, <= for less than or equal to, >= for greater than or equal to and <> for not equal to.

11.2.2 Statistical Functions

EViews provides functions that provide access to the density or probability functions, cumulative distribution, quantile functions and random number generators for a number of standard statistical distributions. For instance, the following command:

genr**result** = @cchisq (7.78, 4) (11.1_a)

generates the output for the chi-square test as:

χ^2 (4)Pr(Y ≤ 7.78) = 0.90 (11.1_b)

For a complete table and descriptions of distribution functions you can use, refer to the 'statistical distribution functions' under the Help section.

11.2.3 Graphs and Scatter Plots

Suppose you want to make a plot of the four variables rgdprconsrgovrinv. Note that all the four variables are written together without allowing for space in between them. Here, 'r' stands for real and gdp, cons, gov and inv stand for GDP, consumption, government and investment respectively. Select the four variables rgdp rcons rgov rinv by keeping the 'Ctrl' key pressed and click on the four series with your mouse pointer. Then, right click with the mouse and select Open > As Group. Then click View > Graph > Line on the group window. If you like to create multiple separate graphs, use View > Multiple Graphs > Line. To preserve your work, click 'Freeze' on the group window. Then in the new graph window, click Name and enter the desired graph name. You can name your graph by clicking on 'Add Text'. To shade a certain time period, you must click Line/Shade and then select Shaded area and enter the time span that you wish to shade. If you want to shade several periods which are not connected, you have to repeat this procedure several times.

Exporting EViews graphs: EViews generates graphs, but they may not look the way you want them to. In order to use an EViews graph in another program, you can save it as a Windows metafile ("*.wmf"). Then, click 'Proc > Save graph to disk' in the graph window. You have to Freeze the graph first. Choose 'metafile' for Word, or the ".eps" format, and choose a file path. The picture can then be imported into a Word document using Insert > Picture > From File in Word.

To create a scatter diagram of X against Y, just type: scat X, Y. Or, you can make a group containing X and Y. You can then select View > Graph > Scatter in the group window.

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) State the steps for importing a file into EViews when you are not using the direct method of copy-paste.

……

……

……

……

……

……

……

……

……

2) Outline the steps for 'exporting a graph' from EViews.

……

……

……

……

……

……

11.3 SIMPLE AND MULTIPLE REGRESSION

The graph of the regression line does not allow us to make exact quantitative statements about the relationship between variables. For this, we must know the exact values of the slope and the intercept. For instance, in applications, we want to predict the effect of increase by one unit in the explanatory variable (e.g. the student-teacher ratio) on the dependent variable (e.g. the test scores). To answer the questions relating to such precise nature of the relationship between large classes and poor student performance, we need to estimate the regression intercept and slope. A regression line, as we know, fits a line through the observations in the scatter plot according to some principle. We can, for instance, draw a line from the test score for the lowest student-teacher ratio to the test score for the highest student-teacher ratio, ignoring all the observations in between. Or, we could sort the data by student-teacher ratio and split the sample in half so that the observations with the lowest ten student-teacher ratios are in one set, and the observations with the highest ten student-teacher ratios are in the other set. For each of the two sets, we could calculate the average student-teacher ratio and the corresponding average test score, and then connect the two resulting points. Or, we could just eyeball the relationship. Some of these principles have better properties than others to infer the true underlying (population) relationship from the given sample. We know that the principle of ordinary least squares (OLS) will give us very desirable properties under certain restrictive assumptions. If the dependent variable *Y*, is only determined by a single explanatory variable *X*, in a linear fashion like:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ , } i = 1,2,3,\ldots..,N \qquad (11.2_a)$$

with 'u' representing the error term, or the random disturbance not accounted for by the linear equation, then the task is to estimate the values of β_0 *and* β_1. If we have values for these coefficients, then β_1 describes the effect of a unit increase in *X* on *Y*. Often, a regression line is a linear approximation of an underlying relationship. The intercept β_0 only has a useful meaning if the observations occur in the data around *X* = 0. The command for regressing a variable *Y* on a constant (intercept) and another variable *X* in EViews is:

ls *Y* c *X* $\qquad (11.2_b)$

where 'ls' stands for least squares and *'c'* is the letter in EViews reserved for the constant (which is actually a vector of 1's). We need to type the variable names in that order since EViews takes the first variable as the dependent variable. It does not matter if we place the constant before the explanatory variable or after. Alternatively, we can start in the Main menu, click on Object, then the New Object and finally on 'Equation'. The same dialog box will open. Now, to get heteroscedasticity corrected regression result, we have to type:

ls(h) *Y* c *X* (11.2_c)

where 'h' within parentheses indicates that we are using heteroscedasticity-robust standard errors. 'c' stands for the intercept. The regression output appears as in Table 11.1 below.

Table 11.1: Simple Regression Results in EViews

Dependent Variable: Y

Method: Least Squares

Date: 11/10/20

Time: 16:56

Sample: 1 10

Included observations: 10

Huber-White-Hinkley (HC1) heteroscedasticity consistent standard errors and covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	618.8527	51.06075	12.11993	0.0000
X	0.606961	2.333492	0.260108	0.8013

R-squared 0.006824	Mean dependent var 631.3500
Adjusted R-squared -0.117323	S.D. dependent var 9.264418
S.E. of regression 9.792813	Akaike info criterion 7.578031
Sum squared resid 767.1935	Schwarz criterion 7.638548
Log likelihood -35.89016	Hannan-Quinn criterion 7.511644
F-statistic 0.054969	Durbin-Watson stat 0.853391
Prob(F-statistic) 0.820522	Wald F-statistic 0.067656
Prob(Wald F-statistic) 0.801349	

The results of Table 11.1 pertain to the data in the source: http://fmwww.bc.edu/ec-p/data/stockwatson/caschool.dta According to these

results, an increase in X by one-unit, results in an increase in Y by 0.6 points. Using the standard reporting format of regression results, we can write the results as:

$$Y = 618.9 + 0.61X,\ R^2 = 0.007, SER = 9.8$$

$$(51.1) \qquad (2.33)$$

Note that being a small sample, the regression's R^2 value is quite low and the slope coefficient is also not statistically significant.

Multiple Regression: Economic theory often suggests that most of the variables are not influenced by a single variable but by a multiple of variables. An extension of the simple regression model is the multiple regression model. It incorporates more than one regressor like:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots . + \beta_k X_{ki} + u_i,\ i = 1,\ldots,n \qquad (11.3)$$

To estimate the coefficients of the multiple regression model, we proceed in a similar way as in the case of simple regression model. The only difference is that we need to list the additional explanatory variables. In general, the command is:

ls (options) Y c X_1 X_2 ... X_k $\qquad (11.4_a)$

The word 'options' can be omitted, in which case, the command simply becomes:

ls Y c X_1 X_2 ... X_k $\qquad (11.4_b)$

This is the default version of the regression command which gives homoscedasticity-only standard errors. To get heteroscedasticity consistent standard error, we must put 'h' option within the bracket just as in case of simple regression. That is:

ls (h) Y c X_1 X_2 ...X_k $\qquad (11.4c)$

The regression output generated by EView is as in Table 11.2 below.

Table 11.2: Regression Results of Multiple Regression

Dependent Variable: Y
Method: Least Squares
Date: 11/10/20
Time: 18:22
Sample: 1 420
Included observations: 420
Huber-White-Hinkley (HC1) heteroscedasticity consistent standard errors and covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	658.5520	8.641528	76.20782	0.0000
X_1	-0.734326	0.256780	-2.859744	0.0045
X_2	-0.175534	0.033661	-5.214827	0.0000
X_3	11.56897	1.818811	6.360731	0.0000
X_4	-0.398234	0.033174	-12.00438	0.0000

R-squared 0.796213	Mean dependent var 654.1565 Adjusted
R-squared 0.794248	S.D. dependent var 19.05335
S.E. of regression 8.642569	Akaike info criterion 7.163110
Sum squared resid 30998.01	Schwarz criterion 7.211208
Log likelihood -1499.253	Hannan-Quinn criterion 7.182121
F-statistic 405.3592	Durbin-Watson stat 1.522233
Prob(F-statistic) 0.000000	Wald F-statistic 417.1963
Prob(Wald F-statistic) 0.000000	

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) State the command for a simple regression in EViews. State also the corresponding command for a heteroscedasticity corrected regression result.

……

……

……

……

……

2) State the corresponding two commands in EViews for a multiple regression.

……

……

……

……

……

……

11.4 TIME SERIES REGRESSION

For studying time series data in EViews, the order of observations is very important. Since much information about the current behaviour of a variable is contained in its own previous values i.e. its own 'lags', we refer to '(t-1) as a one period lag and '(t+1)' as a one period lead. In general, we can consider (t–j) lags in the regression equation. For this, we must create the required number of 'time dummies'. To enter data in EViews, we first open the Excel spreadsheet and copy the data for real GDP variable Y into an EViews data file.

11.4.1 Time Dummy and Taking Differences

One important task in time series analysis is the creation of a variable that represents the time trend. EViews has a time trend variable already built-in. If we type @TREND instead of a series name, then the EViews returns a time trend that increases by one for each observation of the workfile. Likewise, for taking differences, EViews provides a shortcut to compute the first and second differences. If we type d(X) instead of a series name, it means we want to use the first difference of X. We must now take an example to understand better how EViews helps us to obtain the regression results for analysing the time series data. For this, let us consider the example available in the Stock and Watson (2018) website, on quarterly data for real GDP. You can open an Excel spreadsheet and copy the data for real GDP into an EViews data file. In order to do this, you have to fist open a new workfile ('Open a New Workfile'), chose the frequency as 'quarterly' under 'data specification', and designate 1955:1 and 2017:4 as the start and end dates. Next, you can copy-paste the Excel data for the GDPC1 column into the new group ['Quick' then 'Empty Group (Edit Series)']. Rename the variable gdpr (by closing the group and right clicking on ser 01 to rename). Next, save the workfile and name it by any name (e.g. 'ch15usmacro.wf1').

Next, suppose we want to convert real quarter-wise-GDP into an annualised rate of GDP growth. We can do this by the following transformation:

$$g_{Y,t} = \frac{GDP_t - GDP_{t-1}}{GDP_{t-1}} \times 400$$

$$= \left(\frac{GDP_t}{GDP_{t-1}} - 1\right) \times 400 \approx 400 \times [\ln(GDP_t) - \ln(GDP_{t-1})]$$

$$=400 \times \Delta \ln(GDP_t) \qquad (11.5)$$

Now, the first program file created in EViews is as shown in Table 11.3.

Table 11.3: Variables Generated for Annualised GDP Growth in EViews

genr**lgdpr** = log(gdpr)
genr**gdpgr** = 400*(lgdpr - lgdpr(-1))

The above programme sheet or file is given to help you notice how the lines inserted (i.e. commands given), generates the log of real GDP, the lag of real GDP, and the growth rate variables. Also notice how the lagged variables for past values are created. We have generated a lag by adding a '(–1)' after the name of the particular variable in the 'genr' statement or command. In a spreadsheet, this amounts to copying an entire data series and pasting it into a new column one observation down (since the first observation becomes the second observation). The procedure generalized to higher lags remains the same [e.g. $X_{(t-12)}$ is X(–12)]. To underscore this point, note that when working with time series data, lags frequently used can be incorporated by simply creating lagged variables by entering (–i) immediately after the variable name as shown below.

genr**dinf**=inf-inf(-1) (11.7_a)

genr**yeardinf**=inf- inf(-4) (11.7_b)

The first command above (11.7_a) generates the quarterly change in the inflation rate (assuming that you work with quarterly data). The second command (11.7_b) generates the variable measuring the annual change in the inflation rate. The 'd' in the command indicates 'difference' (e.g. dinf = difference in inflation).

11.4.2 Unit Root Test

To perform the 'unit root test', we have to first double click on the 'series name' to open the series window and then select 'View/Unit Root Test…'. We must specify four sets of options to carry out a unit root test. The first three settings, appearing on the left-hand side of the dialog box (Fig. 11.2), determine the basic form of the unit root test. The fourth set of option appearing on the right-hand side of the dialog box, consist of test-specific advanced settings. You only need to concern yourself with these settings to customise the calculation of your unit root test. Specifically, therefore, the following are the steps.

First, use the top most drop down menu to select the type of unit root test that you wish to perform. You may choose one of six tests: ADF, DFGLS, PP, KPSS, ERS, and NP.

Second, specify whether you wish to test for a unit root in the level, first difference, or second difference of the series.

Third, choose your exogenous regressors. You can choose to include a constant, a constant and linear trend, or neither. There are limitations on these choices for some of the tests.

Finally, by clicking on **OK**, EViews computes the test using the specified settings. You can further customise your test using the advanced settings portion in the dialog box. The advanced settings for both the ADF and DFGLS tests allow you to specify how the lagged difference terms are to be

included in the ADF test equation. You may choose to let the EViews automatically select, or you may specify a fixed positive integer value. If you choose automatic selection, you are given the additional option of selecting both the information criterion and the maximum number of lags to be used in the selection procedure. Now, let us suppose we have chosen to estimate an ADF test by including a constant in the regression and by employing an automatic lag length selection using a Schwarz Information Criterion (SIC) with a maximum lag length of 14.

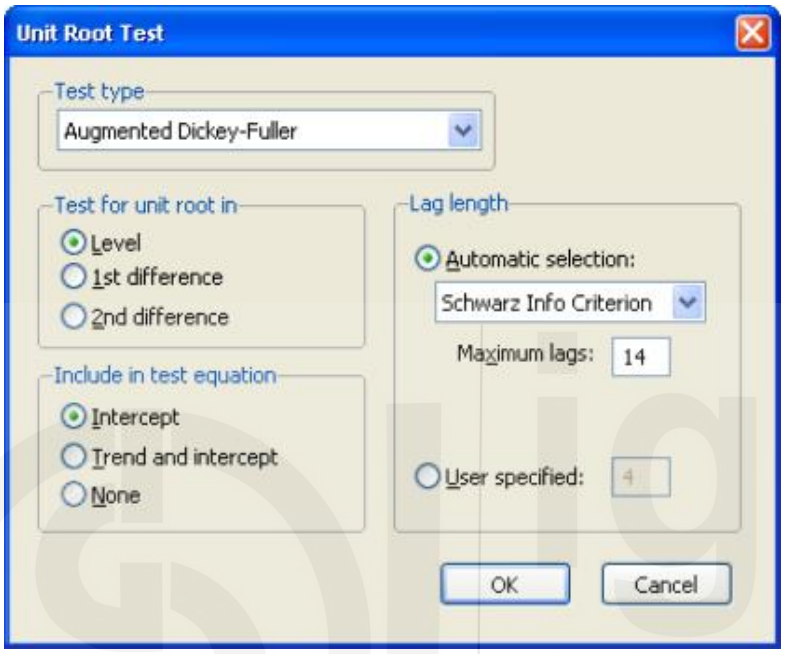


Fig. 11.2: Dialog Box of Unit Root Test

If we apply these settings to the data on the U.S. one-month Treasury bill (TBILL) data (for the period from March 1953 to July 1971 (database drawn from 'Hayashi_92.WF1'), we get the results presented in Table 11.4. The first part of the unit root output provides information about the form of the test in terms of: (i) the type of test, (ii) the exogenous variables used and (iii) the lag length used. The output contains the test results, associated critical values, and the *p*-value as shown in the Table 11.4 below.

Table 11.4: ADF Test Results for Dataset TBILL

Null Hypothesis: TBILL has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic based on SIC, MA

Augmented Dickey-Fuller test statistic		
Test critical values:	1% level	
	5% level	
	10% level	

*MacKinnon (1996) one-sided p-values.

Source: tp://www.EViews.com/help/helpintro.html#page/content%2Fadvtimeser-Unit_Root_Testing.html%23ww184599.

The ADF statistic value is –1.417. The associated one-side *p*-value (for a test with 221 observations) is 0.573. You can see that EViews reports the critical

values at the 1%, 5% and 10% levels of significance. Notice here that the statistic value is greater than the critical values. Hence, we do not reject the null at conventional test sizes. The second part of the output shows the intermediate test equation that EViews used to calculate the ADF statistic. This output is as shown in Table 11.5.

Table 11.5: Equation Results of ADF Test

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(TBILL)
Method: Least Squares
Date: 08/08/06 Time: 13:55
Sample: 1953M03 1971M07
Included observations: 221

	Coefficient	Std. E
TBILL(-1)	-0.022951	0.016
D(TBILL(-1))	-0.203330	0.067
C	0.088398	0.056
R-squared	0.053856	Mean d
Adjusted R-squared	0.045175	S.D. de
S.E. of regression	0.371081	Akaike
Sum squared resid	30.01882	Schwar
Log likelihood	-92.99005	Hannan
F-statistic	6.204410	Durbin-
Prob(F-statistic)	0.002395	

Check Your Progress 3 [answer within the space given in about 50-100 words]

1) How is a 'lagged variable' created in EViews?

.....

.....

.....

.....

.....

2) Specify the steps involved in giving the command to perform a 'unit root test'.

.....

.....

.....

.....

.....

.....

11.5 PANEL DATA REGRESSION

The panel data approach pools time series data with cross-sectional data. Depending on the application, it can comprise of individuals, firms, countries, or regions over a specific time period. The general structure of such a model is:

$$Y_{it} = a_0 + bX_{it} + u_{it} \qquad (11.8)$$

where $u_{it} \sim IID(0, \sigma^2)$ and i = 1, 2,….,N are individual level observations, and t = 1,2,….,T are time series observations indicated in the sub-script 't'. It is assumed that Y_{it} is a continuous variable. The observations of each cross-section unit are stacked over time on top of one another. This is the standard pooled model where intercepts and slope coefficients are homogeneous across all N cross-sections and through all T-time periods. The pooled OLS estimator captures both the 'between' and 'within' dimensions of the data. But it does not do so efficiently since each observation is given equal weight in estimation. We know that: (i) the unbiasedness and consistency properties of the estimator requires that the explanatory variables are uncorrelated with any omitted factors and (ii) there are two types of panel estimators viz. Fixed Effect (FE) estimator and Random Effect (RE) estimator. Let us now take an illustration.

11.5.1 Panel Regression Results in EViews: An Illustration

To illustrate the estimation of panel equations in EViews, we first consider an example involving unbalanced panel data. For this, we use the data from Harrison and Rubinfeld (1978) for the study of 'hedonic pricing' (Harrison_panel.WF1). The data are well known and used as a example dataset in many sources (e.g. Baltagi, 2005, p. 171). The data consists of 506 census tract observations on 92 towns (in the Boston area) with the group sizes ranging from 1 to 30. The dependent variable of interest is the logarithm of the 'median value' (MV) of owner occupied houses (MV). The regressors include various measures of housing desirability.

We must begin by structuring our workfile as an undated panel. For this, click on the 'Range' description in the workfile window. Then, select 'Undated Panel' and enter 'TOWNID' as the 'identifier series'. EViews will prompt you twice to create a CELLID series to uniquely identify observations. Click on OK to both questions to accept your settings. The workfile so selected are like Tables 11.6_a and 11.6_b. EViews restructures your workfile so that it is an unbalanced panel workfile. The top portion of the workfile window will change to show the undated structure which has 92 cross-sections and a maximum of 30 observations in a cross-section. To estimate the panel regression equation, we have to open the equation specification dialog by selecting 'Quick/Estimate Equation' from the Main EViews menu (Table 11.7). First, we shall estimate a fixed effects specification of our hedonic housing equation. The dependent variable in our

specification is the 'median value' MV and the regressors are: (i) the crime rate (CRIM), (ii) a dummy variable for the property along Charles River (CHAS), (iii) air pollution (NOX), (iv) average number of rooms (RM),

Table 11.6$_a$: Workfile Facsimile I

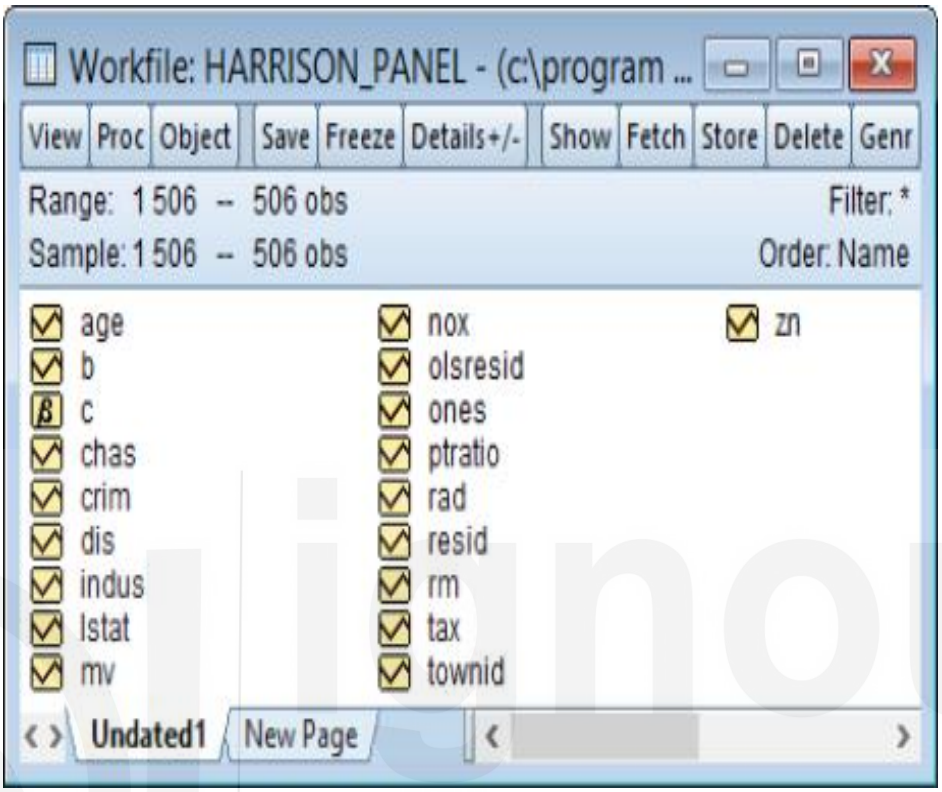
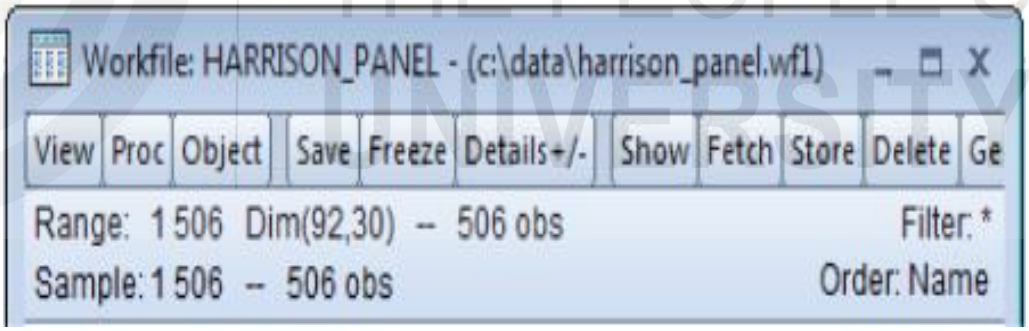


Table 11.6$_b$: Workfile Facsimile II



(v) proportion of older units (AGE), (vi) distance from employment centres (DIS), (vii) proportion of African-Americans in the population (B) and (viii) the proportion of lower status individuals (LSTAT). Note that you may include a constant term C in the specification. Since we are estimating a fixed effect specification, EViews will add one if it is not present so that the fixed effect estimates are relative to the constant term and add up to zero. Next, click on the 'Panel Options' tab and select 'Fixed' for the cross-section effects. We will leave the remaining settings at their defaults. Now, click on OK to accept the specification. EViews returns the estimated results as in Table 11.8.

Table 11.7: Equation Estimation Facsimile

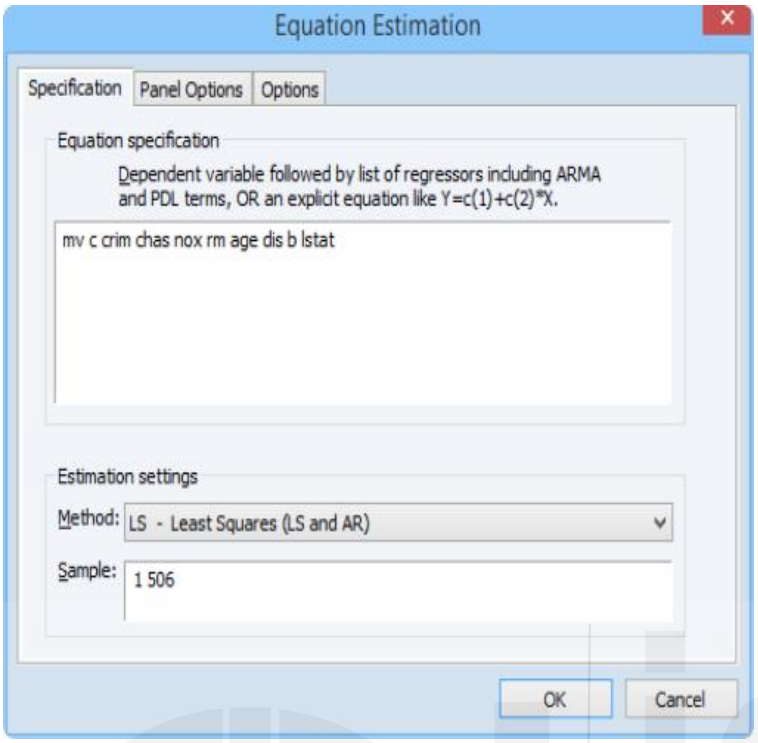


Table 11.8: Results of Panel Data Estimation by EViews

Dependent Variable: MV
Method: Panel Least Squares
Date: 08/23/06 Time: 14:29
Sample: 1 506
Periods included: 30
Cross-sections included: 92
Total panel (unbalanced) observations: 506

	Coefficient	Std.
C	8.993272	0.13
CRIM	-0.625400	0.10
CHAS	-0.452414	0.29
NOX	-0.558938	0.13
RM	0.927201	0.12
AGE	-1.406955	0.48
DIS	0.801437	0.71
B	0.663405	0.10
LSTAT	-2.453027	0.25

Effects Specificati

Cross-section fixed (dummy variables)

R-squared	0.918370	Mean
Adjusted R-squared	0.898465	S.D. d
S.E. of regression	0.130249	Akaike
Sum squared resid	6.887683	Schwa
Log likelihood	369.1080	Hanna
F-statistic	46.13805	Durbir
Prob(F-statistic)	0.000000	

Source: http://www.EViews.com/help/helpintro.html#page/content%2Fpanel-Panel_Estimation_Examples.html%23

We may click on the 'Estimate' button to modify the specifications. For instance, we can modify the specification to include the additional regressors (ZN, INDUS, RAD, TAX, PTRATIO) used in estimation, change the cross-section effects to be estimated as a random effect, and use the 'Options' settings to set the random effects computation method. Then, the top portion of the resulting output as generated by EViews is as shown in Table 11.9.

Table 11.9: Panel Data Estimation Results for RE Model

Dependent Variable: MV
Method: Panel EGLS (Cross-section random eff
Date: 08/23/06 Time: 14:34
Sample: 1 506
Periods included: 30
Cross-sections included: 92
Total panel (unbalanced) observations: 506
Wallace and Hussain estimator of component va

	Coefficient	Std. Er
C	9.684427	0.2076
CRIM	-0.737616	0.1089
ZN	0.072190	0.6846
INDUS	0.164948	0.4263
CHAS	-0.056459	0.3040
NOX	-0.584667	0.1298
RM	0.908064	0.1237
AGE	-0.871415	0.4871
DIS	-1.423611	0.4627
RAD	0.961362	0.2806
TAX	-0.376874	0.1866
PTRATIO	-2.951420	0.9583
B	0.565195	0.1061
LSTAT	-2.899084	0.2493

Effects Specification

Cross-section random
Idiosyncratic random

Source: http://www.EViews.com/help/helpintro.html#page/content%2Fpanel-Panel_Estimation_Examples.html%23

11.5.2 Hausman test

The Hausman test compares the random effects estimator with the fixed effect estimator. If the null is rejected, it favours the fixed effect model relative to the random effect model. The null and alternative hypotheses are expressed as:

H_0: Random effects are independent of explanatory variables

H_1: H_0 is not true

For the Hausman Test in EViews, we have to first load the file and create group data. Then, to perform fixed effect estimation, we have to sequentially follow the path indicated below.

Quick>>Estimate equation>>Panel option>>Fixed>>ok (11.9)

Next, for performing the random effects estimation, we again have to sequentially click the path indicated below.

Quick>>Estimate equation>>Panel option>>Random>>ok (11.10)

We interpret the results by rejecting the null hypothesis if the probability value is statistically significant at 5% level. It implies that the individual effect 'a_i' is correlated with the explanatory variables. Therefore, we use the fixed effect estimator to run the analysis. Otherwise, we use the random effects estimator.

Check Your Progress 4 [answer within the space given in about 50-100 words]

1) State the steps for estimating a 'panel regression equation' in EViews.

.....

.....

.....

.....

.....

2) Specify the command for Hausman test in EViews.

.....

.....

.....

.....

.....

.....

11.6 LET US SUM UP

The unit has outlined the basic functions in EViews and then introduced you to obtaining the results of: (i) simple and multiple regression and (ii) panel data regression. For a complete list of commands, you must consult the EViews Command and Programming Reference or the User's Guide. Alternatively, you can use the 'Help' command inside EViews. Under the Find tab in EViews, there are Help Topics. Here, you can search for whatever you are looking for. The best way to learn how to use the program is to spend some time exploring and playing with it.

11.7 KEY WORDS

ls Y c X : Is the regression command in EViews for simple regression

ls (h) Y c X : Generates heteroscedasticity-robust standard errors for the intercept.

ls Y c X_1 X_2 … X_k : Is the command in EViews for multiple regression.

ls (h) Y c X_1 X_2 … X_k : Is the command in EViews for multiple regression for getting heteroscedasticity-robust standard errors.

11.8 SUGGESTED BOOKS FOR FURTHER READING

1) Agung, I. G. N. (2011). *Time series data analysis using EViews*. John Wiley & Sons.

2) Agung, I. G. N. (2013). *Panel data analysis using EViews*. John Wiley & Sons.

3) EViews 11 Help Topics, http://www.EViews.com/help/helpintro.html

4) McCullough, B. D. (1999). Econometric software reliability: EViews, LIMDEP, SHAZAM and TSP.

5) Vogelvang, B. (2005). *Econometrics: theory and applications with EViews*. Pearson Education.

11.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) (i) start with a new workfile. (ii) click on 'Proc /Import /Import from File'. (iii) specify the location where your data file resides. (iv) double click on that file and click successively a series of 'next' options. (v) finally click on 'finish'.

2) (i) save it as a Windows metafile ("*.wmf"). (ii) click Proc > Save graph to disk in the graph window. (iii) Freeze the graph. (iv) Choose 'metafile' for Word, or the ".eps" format, and choose a file path. (v) Finally use: Insert > Picture > From File in Word.

Check Your Progress 2

1) ls Y c X and ls(h) Y c X respectively.

2) 'ls Y c X_1 X_2 ... X_k' and ls (h) Y c X_1 X_2 ...X_k respectively.

Check Your Progress 3

1) It is generated by adding a '(–1)' after the name of the particular variable in the 'genr' statement or command. The generalised command for higher lags is $X_{(t-r)}$ is X(–r)].

2) (i) select the type of unit root test that you wish to perform (e.g. ADF). (ii) specify the level at which you want to perform the unit root test i.e. at first difference level or second difference level, etc. (iii) choose your exogenous variables with 'c' or without 'c' and (iv) click OK.

Check Your Progress 4

1) We have to first open the equation specification dialog by selecting 'Quick/Estimate Equation' from the Main EViews menu. (ii) We then have to specify the dependent variable and the regressors. (iii) We must specify the constant C for estimating a FE panel. (iv) We now click on the 'Panel Options' tab and select 'Fixed' for the cross-section effects. (v) We finally click on OK to generate the results. To change to the RE computations, we click on 'Estimate' button and modify the settings in the 'options' section to set for the random effect method.

2) For the FE and the RE tests, we sequentially have to click the path as in: Quick>>Estimate equation>>Panel option>>Fixed>>ok and Quick>>Estimate equation>>Panel option>>Random>>ok respectively.

UNIT 12 STATA*

Structure

12.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the 'Graphical User Interface of Stata 16.0' in terms of its basic features;

* Dr. Poulomi Roy, Jadavpur University.

- state the three approaches of executing the Stata command;
- describe the 'basic operations' of Stata 16.0 in terms of 'data input' and obtaining 'results of descriptive statistics';
- specify the uses of the six options that appear below a dialogue box in a Stata's 'summary statistical window';
- illustrate how generation/transformation of variables can be made on Stata 16.0;
- discuss the procedure of obtaining the results of 'analysis of variance' and 'regression analysis' in Stata 16.0;
- describe how diagnostic tests can be conducted on regression results obtained; and
- outline the procedure for testing single and joint hypothesis.

12.1 INTRODUCTION

STATA is one of the widely used statistical packages. It was created in 1985 by Stata Corp. It is a well accepted statistical package used in many fields like economics, political science, sociology, etc. It is very useful for data management, statistical analysis and graphical representation of data. Over 35 years, its popularity has increased due to its effectiveness in transforming and processing large data set.

12.2 GRAPHICAL USER INTERFACE OF STATA 16.0

The main windows of Stata 16 can be divided into four parts: command window, review or history window, variable window and result window. Brief outline of functions performed by these windows are as below.

a) **Command Window:** In the command window one can write different commands to be executed. Each time only one command line can be executed. Some of the useful keys of this window are: page-up key, page-down key and tab key.

b) **History Window:** The top left panel of the Stata window is the history window. This window keeps history of command lines that were executed in the command window.

c) **Variables Window:** This is the right panel of Stata window. It lists the variables, the label names and their format.

d) **Results Window:** This is the middle panel which displays the results or the Stata output.

The user can select one among two possible formats for Stata's windows settings viz. compact window setting and wide screen layout. The selection path is as follows:

Main Menu: Edit → Preferences→Load preference set→Compact window settings/layout

or

Main Menu: Edit → Preferences→Load preference set→Wide screen Layout

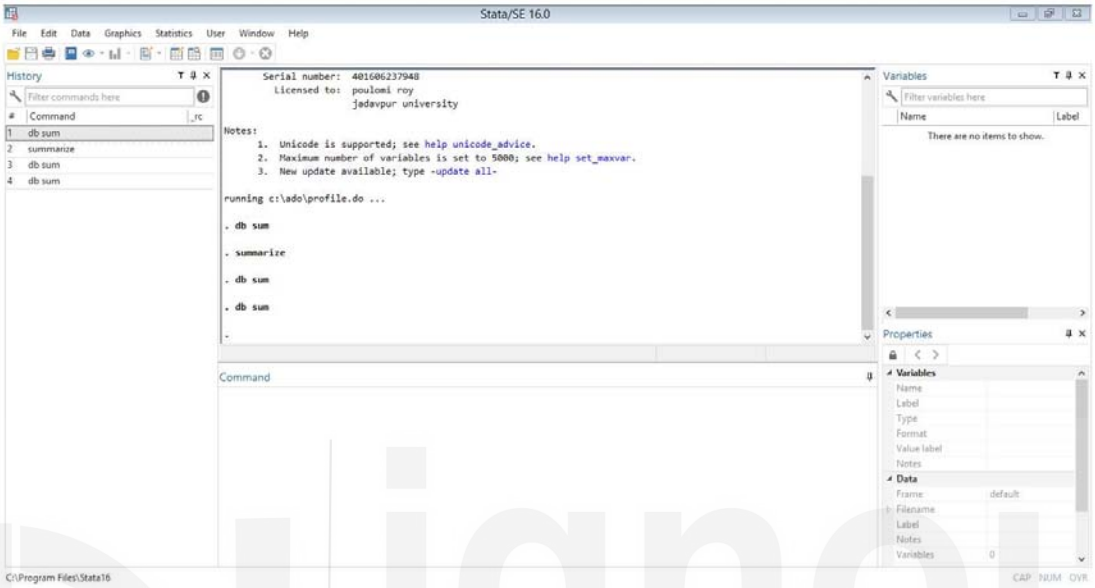


Fig. 12.1: Graphical User Interface of STATA

The user can save any one of its preferred widow layout.

12.2.1 Main Menu

STATA main menu contains 'file, edit, data, graphics and statistics' functions. The 'tool bar' in Statacontains a set of icons to allow accessing different tools such as the viewer, the do file editor or the data editor. There are 12 specific functions which can be performed by the 'main menu'. The 12 functions are listed with a relevant icon appearing from left to right as shown in Fig. 12.2.

1 2 3 4 5 6 7 8 9 10 11 12

Fig. 12.2: Main Menu Options

An outline of the functions performed by each of the above 12 icons is given below:

1)  Opens the Stata data file (as extension *.dta)

2)  Saves the active data file.

3) prints results, contents of the Stata viewer or graphics.

4) adds, suspends or resumes the .log file.

5)  opens the Stata viewer.

6)  shows the graphical window at the front.

7)  opens the .do file with the Stata do editor.

8)  edits the opened data file.

9)  provides a visual view of the data.

10)  is the variables manager

11)  continues the execution of the Stata program in break.

12)  stops the execution of the Stata program.

12.2.2 Command and Dialogue Box

The method of executing the Stata commands is through the 'dialogue box'. To use the dialog box for a desired command, two options are available. The first is to select the command item from Stata's main menu like: Main menu: Statistics→Summary, Tables and Tests→summary and descriptive statistics→Summary statistics. The second is to type the command db followed by the command of interest and then to click **Enter.** For instance, if we type '**db summarize'** in the command window, the following dialogue box will open.

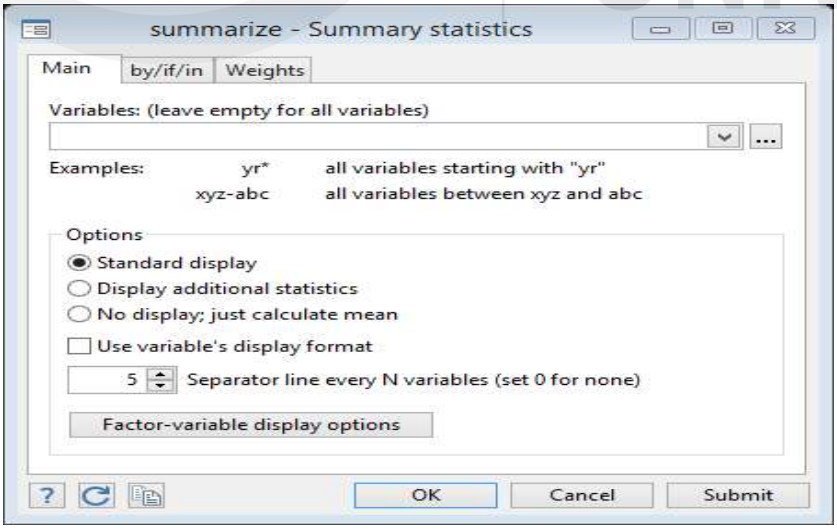


Fig. 12.3: Dialogue Box for Summary Statistics

There are six buttons appearing in the bottom of the dialogue box in the Fig. 12.3 above. The use of these buttons are as follows:

Displays the Stata help file.

Reset: Initialises the dialogue box fields to their default values.

Copies in the clipboard the syntax that will be generated after clicking on **OK.**

OK Executes the command and closes the dialogue box.

Cancel Closes the dialogue box without executing the command.

Submit Executes the command without closing the dialogue box.

After selecting the options, if we click on 'submit' or 'ok' then the command will be executed. The command will be displayed in the result window.

12.2.3 The Data Editor

For entering new data, or changing an already entered data, the 'edit icon' is to be used. One can also type 'edit' in the command window and 'open' a data file in the edit mode. In the data editor, columns represent variables and rows represent observations. We can use the copy/paste command from a sheet of other software such as Excel to paste data into the data editor.

There are two types of data: numeric (Ex. 100/100.11) and string. The latter are alphanumeric (e.g. Sector). String variables appear in red. An example of string variable is '**sysuse auto'.** Let us say 'mpg' (miles per gallon) and 'price' (price of petrol per litre) are two variables for which data is entered. Then, a command like:

edit mpg price in 1/20 if price>=10000 (12.1)

is executed as follows. A data window will open up to show variables 'mpg' and 'price' of first 20 rows if price is greater than or equal to 10000.

12.2.4 The Data Viewer

The format of the Data Viewer is similar to that of the Data Editor. However, this tool can be used only to visualise the data set but not for changing it. To display the Data Viewer, we first click on the icon **Data Viewer.** Alternatively, we can type the command **browse.** If we merely want to explore the dataset without changing it, it is safer to use the **Browse** to avoid making undesired changes. Using the Data Editor, we can edit a subset of variables or observations. For instance, let us say we type as follows:

browse mpg price in 1/20 if price>=10000 (12.2)

By default, the values of labelled variables (e.g. 1 for rural area and 2 for urban area) are displayed in the Data Editor or the Data Viewer. If we wish to edit the values, we must add the option 'nolabel' (e.g. edit area, nolabel). While the alpha-numeric contents (string) are displayed in red, labels are displayed in blue. There are two options for preserving and restoring. Using

'preserve', we canensure that the data will be preserved after the program is terminated. Then, using 'restore', we can access the pre-served data in a later sitting.

12.2.5 The Do-File Editor

The Do-File Editor allows us to perform the writing and editing functions. It also allows the editing or executing of a part or all of the Stata command lines in the current do file. The icons on the Do-File editor performs the following 12 functions.

 **New:** Opens a new do-file in a new Do-File Editor.

 **Open:** Opens a new do-file from a disk in a new Do-File Editor.

 **Save:** Saves the current do-file to disk.

 **Print:** Prints the contents of the Do-File editor window.

 **Find:** Locates and opens the Find/Replace dialog.

 **Cut:** Cuts the selected text to the clip board.

 **Copy:** Copies the selected text to the clip board.

 **Paste:** Pastes the selected text from the clip board to the current do file.

 **Undo:** Cancels the last change.

 **Preview:** Opens a viewer window to display the contents of the Do-File Editor window.

 **Run:** Runs the do file command lines, showing all commands and their output. If a text is highlighted, it is interpreted by Stata as **Run the Selected Text.**

 **Do:** Runs the commands in the do files without showing any output. If text is highlighted, it is interpreted by Stata as **Do the Selected Text.**

Check Your Progress 1 [answer within the space given in about 50-100 words]

1) State the names of four main windows of Stata 16. What are its two window layouts?

……

……

……

……

……

2) List the first five functions contained in the main Stata menu.

……

……

……

……

……

3) Mention the two types of data with examples.

……

……

……

……

……

4) State the important functions of the 'Do-File'.

……

……

……

……

……

12.3 BASIC OPERATIONS AND RESULTS

Some of the pre-defined functions already installed in Stata 16.0 are stated below:

- abs(x) to generate the absolute value of the variable x
- exp(x) to get the exponential function of x.
- int(x) reports the integer by truncating x e.g. int(5.2) = 5 and int(–5.2) = –5.
- The function trunc(x) produces the same result.
- Ln(x) generates the natural logarithm of x.
- max(x_1, x_2, , x_n) gives the maximum value of x_1, x_2, ..., x_n.
- min(x_1, x_2, , x_n) gives the minimum value of x_1, x_2, , x_n.
- sqrt(x) generates the square root of x.

12.3.1 Data Input

Stata can open only one dataset at any time. It holds the entire dataset in random-access (or virtual) memory. Before opening a new data set, one has to close a opened data set by using the command **clear**. The command **input** allows inserting the data directly with the command window. This command is useful with a small number of observations. For instance, we may need to input the data for three variables: size, weight and income. We write: '**input** size weight income'. Here size indicates household size. The Stata allows entering the data in three fields as follows:

5 120 2000

7 180 1500

3 100 3000

2 200 4000

4 140 2500

end

We write 'end' as above to indicate data entry is over.

12.3.2 Descriptive Statistics

The command **summarize** provides descriptive statistics for numerical variables. If no option is specified, Stata produces for each variable of the 'varlist' the number of observations (Obs), the average (Mean), the standard deviation (Std. Dev.), the minimum value (Min.) and the maximum value (Max.). The option 'detail' generates more detailed statistics, such as kurtosis and skewness. The command:

bysorteducation **: summarize** income (12.3)

produces the descriptive statistics of the variable 'income' for different education categories. The command **tabstat** makes it possible to produce almost the same results as **summarize**, but allows greater flexibility in the choice of the descriptive statistics. For instance, '**tabstat** income size, stats (mean, median, variance, sd, skewness)'produces the mean, the median, the variance, the standard deviation, and the skewness of the variables income and size.

12.3.3 Generation of New Variables

The command generate (or gen) can be used to generate new variables. The command 'gen' is used for simple arithmetic computations. The command 'egen'is used when computation is based on whole or part of the observations. For instance, 'genln_inc=ln(income)' generates the value of the natural log of 'income'. Likewise, **egen**inc_m = mean(income) yields 'mean of income values'.

12.3.4 Changing the Variable Values

The 'replace' command can be used to modify the contents of the existing variable. For instance,

replacesize = 6 if age > 46 (12.4)

replaces the contents of the variable size by 6 if age is higher than 46. Likewise,

genx = "poor" in 1/10 (12.5)

creates a variable string labelled as' poor' in the first 10 observations. Similarly, we can use:

replacex = non-poor if x > 1000 (12.6)

replaces all values higher than 1000 by the string 'non-poor'.

12.4 ANALYSIS OF VARIANCE AND REGRESSION ANALYSIS

As we know, analysis of variance is a statistical test for comparing the means of continuous variables in two or more independent groups. Suppose the group size is three. One may be interested in executing three separate group to group comparisons. Each of these comparisons, if done independently of each other, fails to consider the total data comprehensively thereby increasing the probability of committing type I error. The 'analysis of variance' (ANOVA) addresses the overall question of 'whether there are significant differences among the groups' without addressing the difference between any two groups in particular. The technique therefore mainly examines the 'within and between' groups variations.

12.4.1 One Way ANOVA

We run a one way ANOVA when there is only one treatment group. Syntax to be used is 'anova or just one way'. For instance, let us say we run an experiment by varying the amount of fertilizer used in growing apple trees to test four concentrations with the aim of measuring the 'average weight of the fruit'. Consider the data set as depicted in Table 12.1.

Table 12.1: One Way ANOVA Data

input treatment	weight
1	117.5
1	113.8
1	104.4
2	48.9

```
2 50.4
2 58.9
3 70.4
3 86.9
4 87.7
4 67.3
end
```

To obtain one-way ANOVA results, we type:

```
anova weight treatment (12.7)
```

Stata gives one way ANOVA result as in Table 12.2. Notice that for the number of observations in the sample of 10, $R^2 = 0.9147$ which is the ratio of 'Model SS' (5295.5443) to 'Total SS' (5789.136). It indicates that 91.47% variation in the weight of the apple is explained by the fertilizer used. The F statistic is 21.46 and has a significance level of .0013 i.e. the overall model is significant at 0.13% level of significance. This leaves the confidence level low at 87 percent (0.87), far from the usual level of 95 percent. Since there is only one treatment group in the model, the first and second line are identical. The third line indicates the residual part. The sum of 'model (between) and residual (within) sum of squares' is equal to the total sum of squares.

Table 12.2: One Way ANOVA Result

```
      Number of obs =        10    R-squared     =  0.9147
      Root MSE      =   9.07002    Adj R-squared =  0.8721

      Source | Partial SS        df        MS        F     Prob>F
-------------+----------------------------------------------------
       Model |  5295.5443         3   1765.1814     21.46  0.0013
             |
   treatment |  5295.5443         3   1765.1814     21.46  0.0013
             |
    Residual |  493.59167         6   82.265278
-------------+----------------------------------------------------
       Total |   5789.136         9   643.23733
```

12.4.2 Two Way ANOVA

Two way ANOVA technique is used when there are two or more treatment groups. We use the data file of records for each of the 58 patients for their 'drug, disease and change in systolic blood pressure levels'. Consolidated data for these 58 patients is (Table 12.3.) are grouped for three types of diseases (1, 2, 3). Four types of drugs (1, 2, 3, 4) are administered on these patients. The data is classified and arranged as 'distribution of patients by disease and drugs. The second panel of Table 12.3 presents the two way

ANOVA results generated by Stata. The command that we type for obtaining this result is:

anova drug systolic # disease (12.8)

Note that the data in the first panel of above table is not balanced since there are unequal number of patients corresponding to the different drug-disease cell. But we don't need balanced data to perform a two-way factorial ANOVA in STATA. The STATA output reveals that that $R^2 = 0.4560$ which indicates that 45.6% of variation in systolic blood pressure is explained by the model. The first line of the ANOVA table reports model sum of squares i.e. 4259.3385 with d.f. 11. Value of F statistic is 3.51 with its corresponding p value of 0.0013. This indicates that overall model can predict the variation in systolic blood pressure at less than one percent level of significance (0.13%). The 2nd line of the ANOVA table reports that the partial sum of square corresponding to drug is 2997.4719 with an F statistic value of 9.05 with its probability value 0.0001. This means that the variation explained on systolic blood pressure by the effectiveness of drug is at a level of less than one percent significance. The conclusion is that the disease group and the interaction between drug and disease do not affect variation in the systolic blood pressure even at 10% level of significance. There are other factors which must be considered.

Table 12.3: Distribution of Patients by Disease and Drugs and 2-Way ANOVA

```
. tabulate drug disease
```

	Patient's Disease			
Drug Used	1	2	3	Total
1	6	4	5	15
2	5	4	6	15
3	3	5	4	12
4	5	6	5	16
Total	19	19	20	58

```
                    Number of obs =        58    R-squared     =  0.4560
                    Root MSE      =   10.5096    Adj R-squared =  0.3259
```

Source	Partial SS	df	MS	F	Prob>F
Model	4259.3385	11	387.21259	3.51	0.0013
drug	2997.4719	3	999.15729	9.05	0.0001
disease	415.87305	2	207.93652	1.88	0.1637
drug#disease	707.26626	6	117.87771	1.07	0.3958
Residual	5080.8167	46	110.45254		
Total	9340.1552	57	163.86237		

12.4.3 Regression Analysis

Let us now consider how to get results of Ordinary Least Squares (OLS) estimation in regression analysis. We consider data on the 'mileage rating' (measured by 'miles per gallon' i.e. 'mpg') and the weight of 74 automobiles. The variables in our data are mpg, weight, and make of automobile classified into domestic and foreign. The variable on make of car is thus a dummy assuming the value 1 for foreign and 0 for domestic automobiles. We wish to fit the model: mpg = β_0 + $\beta_1$weight + $\beta_2$foreign + u. For this, we give the command as follows:

```
regress mpg weight foreign                                    (12.9)
```

The command yields the following result of regression by Stata.

Table 12.4: Results of Regression by OLS Estimation

Source	SS	df	MS	Number of obs	=	74
				F(2, 71)	=	69.75
Model	1619.2877	2	809.643849	Prob > F	=	0.0000
Residual	824.171761	71	11.608053	R-squared	=	0.6627
				Adj R-squared	=	0.6532
Total	2443.45946	73	33.4720474	Root MSE	=	3.4071

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	-.0065879	.0006371	-10.34	0.000	-.0078583	-.0053175
foreign	-1.650029	1.075994	-1.53	0.130	-3.7955	.4954422
_cons	41.6797	2.165547	19.25	0.000	37.36172	45.99768

The first panel of the above Stata output reports ANOVA results. The observed value of F-statistic is 69.75 with corresponding probability value zero. This indicates that the hypothesis of 'joint coefficients of weight and make of car as foreign is zero' is rejected. R^2 = 0.6627 implies that the 66.27% of variation in 'mpg'is explained by the two independent variables (weight and foreign made car) considered here.

In the second panel of regression result, the first line reports coefficient of weight (–0.007) and its corresponding probability value (0.00). The negative value of the coefficient indicates that as the weight of automobile increases, it significantly reduces the 'mileage per gallon'. The variable foreign does not have a significant influence on the variation in mpg as its probability value is more than 10% (0.13 or 13%).

Check Your Progress 2 [answer within the space given in about 50-100 words]

1) State some of the pre-defined functions installed in Stata 16.0.

……

……

……

……

……

……

2) When is the command ‘replace’ useful in Stata. Give examples.

……

……

……

……

……

……

3) You wish to examine the impact of capital and output on employment levels. Which command would you give in Stata for this?

……

……

……

……

……

4) In (3) above, you wish to deflate data on output and capital to bring them to a constant base. Which command of Stata would you use for this?

……

……

……

……

……

12.5 REGRESSION DIAGNOSTICS

Let us now learn how we can use Stata to check ‘how well our data meets the assumptions of OLS regression’. In particular, we will consider the following assumptions.

- **Normality** i.e. the errors should be normally distributed.
- **Homogeneity of Variance** (homoscedasticity) i.e. the error variance are constant.
- **Multicollinearity** i.e. two or more independent variables are highly correlated.

- **Independence/Autocorrelation** i.e. the errors associated with one observation are not correlated with the errors of any other observation i.e. the error terms are independent.
- **Model Specification** – the model is properly specified (i.e. all relevant variables are included and irrelevant variables are not included).

12.5.1 Test for Normality

Normality of residuals is required for the validity of hypothesis testing. This is because the normality assumption is required for the p-values of t and *F*-tests to be valid. Recall that normality is not required to obtain unbiased estimates of the regression coefficients [as the OLS regression merely requires that the residuals (errors) are IID (i.e. identically and independently distributed)]. To test for this, after we run a regression analysis, we use the **predict** command to estimate the residuals and then use the command **kdensity** to perform the test for the normality of the residuals. We now type the following command to open the data file: 'sysusebplong.dta'. The data file contains blood pressure (bp) measures of male and female patients over the three age groups 30-45, 46-59, 60+. Here, age-grp is used to consider different dummies for different age groups.To check the normality of the residuals, let us consider the regression results of 'regbpsexage-grp'. We can now generate the 'predicted values of bp' by Stata by giving the command:

predict r, residual (12.10)

Once the residuals are generated, we can use the **kdensity** (**kdensity** stands for kernel density estimate) command of Stata to produce a 'kernel density plot'. We can further use the **normal** option for obtaining an overlaid normal density curve on the kdensity plot. The command for this is:

kdensity r, normal (12.11)

The Stata produces a plot as in Fig. 12.4 for the illustrative data considered here. The Figure shows that Kernel density curve and the normal density curves are very similar indicating that residuals are normally distributed.

12.5.2 Test for Heteroscedasticity

The homogeneity of variance of the residuals is needed for OLS estimates to be efficient. This means that if the model is well-fitted, there should be no pattern to the residuals plotted against the fitted values. In other words, if the variance of the residuals is non-constant, then the residual variance is characterised as 'heteroscedastic'. We are aware that there are graphical and non-graphical methods for detecting heteroscedasticity. Stata command provides the option for testing for heteroscedasticity between two tests viz. 'hettest' (performs Cook and Weisberg test for heteroscedasticity) and imtest (computes the White's general test for Heteroscedasticity).

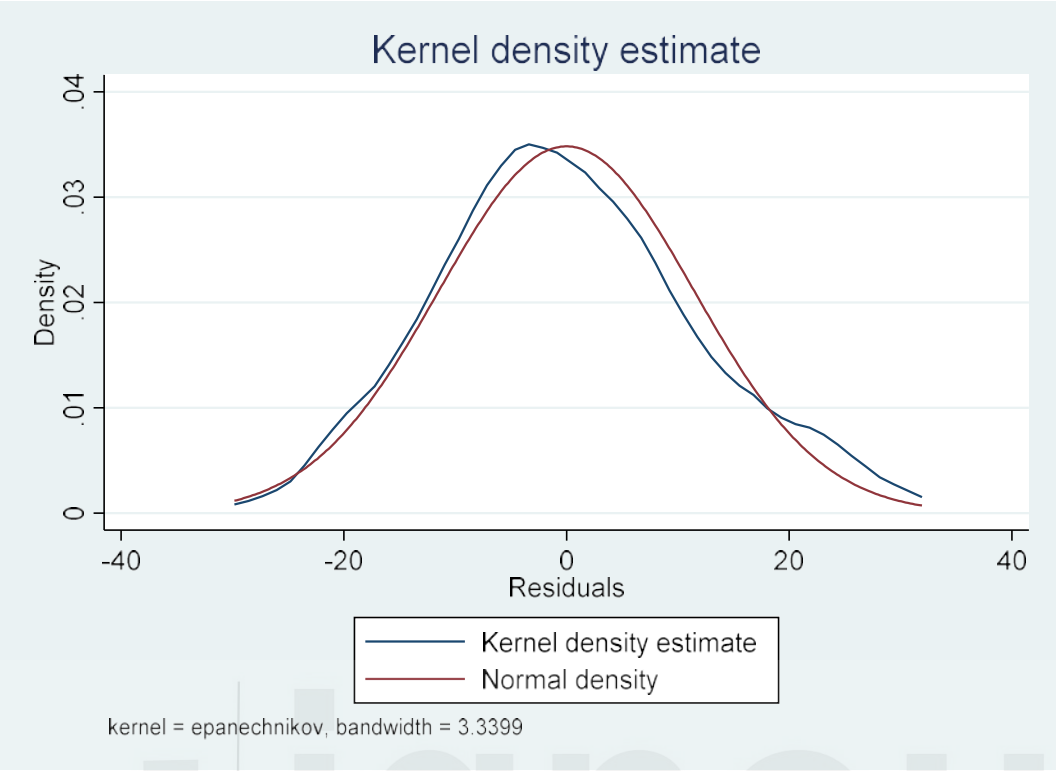


Fig. 12.4: Kernel Density Graph of Error Terms for Normality Test

Let us now type the following command in the command window of STATA to open the auto file from the system:

sysuseauto.dta (12.12a)

The file will open in Stata. Now, first type the following regression command:

reg price mpg foreign (12.12b)

and then type the post-regression command to test for heteroscedasticity as:

estathettest (12.13)

For the data considered on the 'blood pressure' values of sample patients, the results of 'hettest' as in (12.13) gives the result as follows (Table 12.5a).

Table 12.5a: Results of BP Test for Heteroscedasticy

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of bp

         chi2(1)      =     0.60
         Prob > chi2  =   0.4393
```

Based on the above values, since p > 0.05, we do not reject the null hypothesis of homoscedasticity. Alternatively, we may use the following command:

estatimtest …….. (12.14)

The Stata gives the following output (Table 12.5b).

Table 12.5b: Results for Heteroscedasticity by IM Test

```
Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	10.19	5	0.0701
Skewness	8.53	3	0.0362
Kurtosis	1.49	1	0.2216
Total	20.21	9	0.0166

Here also, since the probability value (0.07) is greater than 0.05, we do not reject the null hypothesis at 5% level of significance. This means we accept the null hypothesis that there is 'no' heteroscedasticity.

12.5.3 Test for Multicollinearity

When there is a perfect linear relationship among the predictors, the estimates for a regression model cannot be uniquely computed. We are aware that the term collinearity implies that two variables are near perfect linear combinations of one another and that when more than two variables are involved, the situation is characterised as multicollinearity. The primary concern is that as the degree of multicollinearity increases, the estimated coefficients of the regression model become unstable with the standard errors for the estimated coefficients getting wildly inflated. As a result, the observed t-statistic becomes low. To detect this, we know that 'variance inflation factor (VIF)' can be used to test for multicollinearity. The **vif** command in Stata, after the regression results are obtained, to check for multicollinearity, is given as follows:

estatvif (12.15)

The Stata gives the results for multicollinearity in terms of VIF as follows:

Table 12.6: Values of VIF for Data Considered in 12.5.1 (Normality Test for 'regbp gender age-group)

Variable	VIF	1/VIF
1.sex	1.00	1.000000
agegrp		
2	1.33	0.750000
3	1.33	0.750000
Mean VIF	1.22	

As a thumb rule, a variable whose VIF values are greater than 2, merits further investigation. The values of VIFs for the X_i's of the above example are well below 2.0. Thus, we conclude that there is no sign of multicollinearity.

12.5.4 Test for Autocorrelation

For OLS estimates to be efficient, we know that the errors associated with one observation should not be correlated with the errors of any other observation. When we have data where we have reason to believe there is time-series pattern, we must use the **dwstat** command to perform a Durbin-Watson test for correlated residuals. The command for this test in Stata is:

dwstat (12.16)

First use the data set running the following command: sysuse nlsw88.dta. Autocorrelation test can be checked in Stata in time series data only. The above data is a cross section data. So, after opening the above data, we force Stata to assume idcode as a time series variable by typing 'tsetidcode'. Now, let us consider the results of a regression performed as 'reg wage grade' on a sample of 2244 observations i.e. persons for whom data on 'wage' is regressed on their 'grade' data. The results of this regression are given in Table 12.7.

Table 12.7: Results of Regression on Wage and Grade

Source	SS	df	MS	Number of obs	=	2,244
				F(1, 2242)	=	265.57
Model	7874.79847	1	7874.79847	Prob > F	=	0.0000
Residual	66479.532	2,242	29.6518876	R-squared	=	0.1059
				Adj R-squared	=	0.1055
Total	74354.3305	2,243	33.1495009	Root MSE	=	5.4454

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grade	.7431729	.0456033	16.30	0.000	.6537438	.832602
_cons	-1.965886	.6083143	-3.23	0.001	-3.158804	-.7729677

Now, type: dwstat to get the Durbin Watson test statistic value as:

```
. dwstat

Number of gaps in sample:  1232

Durbin-Watson d-statistic(  2,  2244) =  .8291537
```

The Durbin-Watson statistic has a range from 0 to 4 with a midpoint of 2. The observed value in our example is small (0.83). This is not surprising since our data is not truly time-series.

12.5.5 Tests for Model Specification

A model specification error can occur when one or more relevant variables are omitted from the model or one or more irrelevant variables are included in the model. Model specification errors affect the estimate of regression coefficients. Let us consider here the results of the regression 'regbp gender age-group' on 240 patients. Now to check whether the model has any omitted variable or not, we use the following command in Stata.

estatovtest (12.17)

Stata presents the results of the Ramsey RESET test as follows:

Table 12.8: Results of Model Specification by Ramsey RESET Test

```
Ramsey RESET test using powers of the fitted values of bp
       Ho:  model has no omitted variables
                 F(2, 234) =      1.63
                  Prob > F =      0.1988
```

The **ovtest** command indicates that we fail to reject the null hypothesis of 'no omitted relevant variables even at 10% Level of significance. This is because the probability value of computed F is more than 0.10. Ovtest therefore tells us that we have no specification error. We therefore do not have to reconsider our model.

12.6 TESTING OF HYPOTHESIS

Ordinary least square (OLS) estimates are best linear unbiased estimates i.e. the estimators have the smallest variance. Large value of a coefficient of independent variable does not by itself indicate that that there is a significant impact on the dependent variable. Hypothesis testing allows us to carry out the right inferences about population parameters using data from a sample. In order to test a hypothesis, we may recall that we perform the following steps:

i) formulate a null hypothesis and an alternative hypothesis on population parameters.

ii) build a statistic from the sample data to test the hypothesis made.

iii) follow a decision rule defined by theory to reject (or not reject) the null hypothesis specified.

We now show here how the hypothesis testing is carried out in Stataby using the Stata example data available in 'http://www.Stata-press.com/data/r13/census'. The data corresponds to the 1980 census data for the 50 States of US. There are three variables in the data file:'brate' for the records on birth rate in each state, 'medage' for their median age and 'region' to indicate the region where the state is located. The region variablesare denoted by: 1 if the state is in the Northeast, 2 if the state is in the North Central, 3 if the state is in the South and 4 if the state is in the West. We now

type following command to open the data file in STATA: sysusecensus.dta and then use the following Stata command to get the regression results.

regressbratemedage region (12.18)

The results of the above regression are presented in Table 12.9.

Table 12.9: Regression Results for the US Census Data

Source	SS	df	MS	Number of obs	=	50
				F(5, 44)	=	100.63
Model	38803.4208	5	7760.68416	Prob > F	=	0.0000
Residual	3393.39921	44	77.1227094	R-squared	=	0.9196
				Adj R-squared	=	0.9104
Total	42196.82	49	861.159592	Root MSE	=	8.782

brate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
medage	-109.0958	13.52452	-8.07	0.000	-136.3527	-81.83892
c.medage#c.medage	1.635209	.2290536	7.14	0.000	1.173582	2.096836
region						
NCentral	15.00283	4.252067	3.53	0.001	6.433353	23.57231
South	7.366445	3.953335	1.86	0.069	-.6009775	15.33387
West	21.39679	4.650601	4.60	0.000	12.02412	30.76946
_cons	1947.611	199.8405	9.75	0.000	1544.859	2350.363

The above results indicate that the observed F-statistic is 100.63 which is very high with the corresponding probability value at 0.000. This indicates the independent variables, all taken together, affects the variation in birth rate significantly. The R-square value and the adjusted R-square values are found to be very high. This indicates that more than 91% variations in the dependent variable is explained by the explanatory variables included in the model.

12.6.1 Single Hypothesis

We can test the hypothesis that the coefficient on region 3 (Southern region) is zero by the command: 'test 3.region=0'. Likewise, to test the null hypothesis that the South (3) and NE region (1) have equal impact on the dependent variable, we give the command:

(1) 3. region = 0 (12.19)

For 12.19, Stata gives the following result:

```
 ( 1)  3.region = 0

       F(  1,    44) =    3.47
            Prob > F =    0.0691
```

We observe that F-statistic is 3.47 with a probability value of 0.0691. This indicates that the null hypothesis that the south and NE region have equal impact on the dependent variable is rejected at 10% level but not at 5% level

of significance. Now suppose we want to test that the coefficient of North Central (2) is 21 then we type (the command):

test 2.region=21 (12.20)

The result that Stata produces is as below:

```
 ( 1)  2.region = 21

       F(  1,    44) =    1.99
            Prob > F =    0.1654
```

We find that we cannot reject the hypothesis at any significance level below 16.5%. Let us now test something more difficult: whether the coefficient on 2.region is the same as the coefficient on 4.region? We type:

test 2.region=4.region (12.21)

Stata produces the following result:

2.region - 4.region = 0

F(1, 44) = 2.84

Prob> F = 0.0989

We reject the null hypothesis at 10% level but not at 5% level of significance.

12.6.2 Joint Hypothesis

Suppose we want to test whether the regional variables taken as a whole are significant or not. This is an example of joint hypothesis testing where we test whether the coefficients on 2.region, 3.region and 4.region are simultaneously zero. To perform this test we type the following command:

test (2.region=0) (3.region=0) (4.region=0) (12.22)

Stata produces the following result:

(1) 2.region = 0

(2) 3.region = 0

(3) 4.region = 0

F(3, 44) = 8.85

Prob> F = 0.0001

We reject the null hypothesis as the probability value is close to zero and the computed value of F is high (8.9).

12.7 LET US SUM UP

The unit specifies the command to be given in Stata for testing various statistical contexts. In particular, the unit has indicated the Stata commands for performing tests for one-way ANOVA, two-way ANOVA, performing a regression analysis, using the results of regression for performing tests for normality, multicollinearity, auto correlation and heteroscedasticity. Commands for performing test for model specification by Ramsey RESET test is also illustrated. At various places, the required background that you would have studied in your course on Introductory Econometrics are recapitulated here.

12.8 KEY WORDS

Do-File Editor : Allows to perform the writing and editing function in Stata. Specifically, it helps in performing various functions like open, save, print, find, copy, cut, paste, undo, preview, etc.

Predict : Is the command used in Stata for estimating the value of residuals after we run a regression analysis. This is needed for obtaining the kdensity plot/graph for verifying the normality of the residuals.

'dwstat' : This is the command in Stata for obtaining the value of Durbin Watson statistic to perform the test for autocorrelation.

'estatvif' : This is the command for obtaining the results required in Stata for testing for multicollinearity.

'estate ovtest' : Is the command in Stata for obtaining the results required for performing the model specification test by the Ramsey RESET test.

12.9 SUGGESTED BOOKS FOR FURTHER READING

1) Adkins L C (2011). *Using Stata for Principles of Econometrics*, Wiley Global Education.

2) Kohler U & Kreuter F (2005). *Data Analysis Using Stata*, Stata Press.

12.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Command window, review/history window, variable window and result window. The two window Layouts are compact window setting and wide screen layout.

2) File, edit, data, graphics and statistics.

3) Numeric (e.g. 200, 411). String – these are alpha numeric data (e.g. sector, gender).

4) Open, save, print, find, cut, copy, paste, preview, run, do, etc.

Check Your Progress 2

1) Absolute value of (X), exponential value of (X), natural logarithmic value of (X), etc.

2) It is useful in modifying variables. e.g. replace by 'poor' if income is < 1000, replace by 'non-poor' if income is > 1000.

3) (regress employment capital output)

4) replace and generate.

GLOSSARY

2SLS Method

Two-Stage least squares (**2SLS**) regression analysis is a statistical technique that is used in the analysis of structural equations. This technique is the extension of the OLS method. It is used when the dependent variable's error terms are correlated with the independent variables.

Adaptive Expectations Model

Adaptive expectations is a hypothesised process by which people form their expectations about what will happen in the future based on what has happened in the past. For instance, if inflation has been higher than expected in the past, people would revise expectations for the future.

Autocorrelation

A situation in which *cov* (U_i, U_j) $\neq 0$.

Auto-Regressive Model of Regression

An autoregressive model is when a value from a time series is regressed on previous values from that same time series. In this type of models, the dependent variable in the previous time period becomes the predictor.

c()

This is the function to write a vector in R.

Correct Specification

Correct specification of the model is one which represents the true relationship between the regressors and the regressand.

data.frame()

This is the function to write data frame in R.

Distributed Lag Models

This is a model for time series data in which a regression equation is used to predict the current values of a dependent variable based on both the current values of an explanatory variable and the lagged values of the explanatory variables.

Do-File Editor

Allows to perform the writing and editing function in Stata. Specifically, it helps in performing various functions like open, save, print, find, copy, cut, paste, undo, preview, etc.

'dwstat'

This is the command in Stata for obtaining the value of Durbin Watson statistic to perform the test for autocorrelation.

Empirical Research

It deals with quantitative data collection and analysis. It seeks to explore relationships

between a dependent variable and a set of independent variables.

Endogeneity

In two variable regression model, we assume one variable as endogenous or dependent and the other one is independent or explanatory. The explanatory variable is assumed to be strictly exogenous or pre-determined in nature. If this property does not hold, we cannot easily estimate the unknown parameters (viz. regression coefficients) by applying the method of least squares. This causes the problem of endogeneity because of simultaneous dependence of the two variables.

'estatvif'

This is the command for obtaining the results required in Stata for testing for multicollinearity.

Fixed Effects Estimator

Refers to a 'pooled OLS estimator' based on the time-demeaned variables' used in the panel regression models.

Hausman Test

This is a test procedure proposed by Hausman which helps us to make a choice between a 'random effects' model and a 'fixed effects' model. The test statistic [as in equation (9.12)in Unit 9] is distributed as χ^2 with the d.f. equal to 'the rank of the difference in the two covariance matrices' i.e. the covariance matrix of the 'consistent estimator' and the 'efficient estimator'. It is based on the difference between the random effects and fixed effects estimates.

Heteroscedasticity

Refers to a situation where the variance of error term is not common to all the U_is. This means $V(U_i) \neq V(U_j)$.

Inclusion of Irrelevant Variable

This results in a situation which, in comparative terms, is far less serious than the omission of a relevant variable. This is because of the sustenance of the three basic properties viz. unbiasedness, consistency and validation of test results.

Indirect Least Squares Method

Indirect least squares is an approach in econometrics where the coefficients in a simultaneous equations model are estimated from the reduced form model using ordinary least squares. Once the coefficients are estimated the model is put back into the structural form.

In-sample forecasting

An in-sample forecast employs a subset of the dataset to forecast the values within the estimation period and compare them to the actual outcomes. In other words, in-sample forecasting is done to assess how well the chosen model fits the data in a given sample.

Instrumental Variable Method

Instrumental variables (IVs) are used to control for confounding and measurement error in observational studies. They allow for the possibility of making causal inferences with observational data. The IVs can adjust for both observed and unobserved confounding effects.

Irrelevant Variable

This refers to a variable without the theoretical backing but of which the researcher is unsure and hence prefers to have it included in the regression model. Such inclusion, called over-specification, does not harm the basic properties of the model. This means that the estimates of regression co-efficients are biased but consistent so that the results of hypothesis testing remain valid.

library

This is the directory where the packages are stored in R. It is called as the "library" in the 'R' environment.

Linear Probability Model

It is stated as in Equations (5.4) and (5.5) of Unit 5 where dependent variable takes two values: 0 and 1. In this, we have to deal with situations of non-normality of disturbance terms, heteroscedastic variances of disturbance terms and non-fulfilment of conditional probability values being within the limits of '0' and '1'.

Linear Static Panel Data Model

It is a model like in Equation (9.1) of Unit 9. In this, a separate term for 'unobserved effects', c_i, assumed to be random variables, is introduced. Further, in order to be able to assume the effect of c_i to be 'zero' on an average, we introduce a separate intercept term as in Equation (9.2) of Unit 9.

lm

This is the command for obtaining the results of regression in R.

Logit Model

This is a model as stated in equation (5.10). This is non linear both in x's and β. Hence, the OLS method of estimation does not work here. We can estimate the parameters by using the ML method as outlined in equation

	(5.16). The test results of this model are based on standard normal *Z* values.
ls Y c X_1 X_2 … X_k	Is the command in EViews for multiple regression.
ls (h) Y c X	Generates heteroscedasticity-robust standard errors for the intercept in Eviews.
ls (h) Y c X_1 X_2 … X_k	Is the command in EViews for multiple regression for getting heteroscedasticity-robust standard errors.
ls Y c X	Is the regression command in EViews for simple regression
Measurement Error in Dependent Variable	A situation in which both the estimates of coefficients and the variance estimates remains unbiased. The latter will be larger leading to loss in precision. Hence, it is viewed as relatively less serious.
Measurement Error in Independent Variables	A situation where estimators would be biased and inconsistent with the extent of bias increasing with increase in sample size. It is hence relatively viewed more serious.
Model Specification	Model specification refers to the description of the process by which the dependent variable is estimated by a set of independent variables considered.
Multicollinearity	A situation in which regressors are excessively correlated rendering the OLS estimates to lose its best properties.
'estatovtest'	This the command in Stata for obtaining the results required for performing the model specification test by the Ramsey RESET test.
Omission of a Relevant Variable	The consequences of this are: (i) estimated slope coefficients will be biased, (ii) OLS estimates are inconsistent so much so that even large samples would not eliminate this and (iii) both the confidence interval and the result of hypothesis testing are seriously compromised.
Order Condition	This condition is based on a counting rule of the variables included and excluded from the particular equation. It is a necessary but not sufficient condition for the identification of an equation.

Out-of-Sample Forecasting An out-of-sample forecasting uses all the values in the available data in the sample to predict the future value of the regressand.

Panel Data Refers to data collected for two time points on same sample units. In other words, unlike in 'pooled cross section', no new random samples are used in the two surveys. Data is collected on same variables. This is particularly useful for assessing the effect of 'lags' which is usually there in govt. policies introduced.

phtest() This is the code for the application of Hausmann Test in R.

plm This is the command for obtaining the results of panel regression in R.

Pooled Cross Section Data Refers to data collected on same cross sections at two different points of time but pooled for the purpose of analysis. By combining the two samples, we get increased degrees of freedom or higher 'n'. Data is pooled to assess the impact of a new government policy. In other words, pooled cross section data helps us in assessing the before/after effects.

Predict Is the command used in Stata for estimating the value of residuals after we run a regression analysis. This is needed for obtaining the kdensity plot/graph for verifying the normality of the residuals.

Probit Model This is defined as: the econometric model based on 'dummy dependent variable'. The form of the Probit model depends on the 'cumulative distribution function' (CDF) where we usually work with the 'normal variate'. As in the Logit model, the test results of this model are based on the Z values.

Problem of Identification By identification problem, we basically mean whether the numerical estimates of the parameters of a structural equation can be obtained from the estimated 'reduced form coefficients'. If this cannot be done, then we say that the particular equation is unidentified or under-identified.

Quasi-Demeaned Data Refers to the procedure adopted for transforming the variables to remove the effect of their means. It is a partial (quasi) cleaning process since the 'random effects transformation subtracts a fraction of the time average effect'.

Rank Condition The Rank condition states that in a system of G equations any particular equation is identified 'if and only if' it is possible to construct at least one non-zero determinant of order (G-1) from the coefficients of the variables excluded from that particular equation but contained in the other equations of the model. This condition is called Rank condition because it refers to the rank of the matrix of parameters of excluded variables.

Recursive Model A recursive model is a special case of an equation system where the endogenous variables are determined one at a time in sequence. Thus, the right-hand side of the equation for the first endogenous variable includes no endogenous variables but includes only exogenous variables.

Reduced Form of Equations The reduced form method is a single equation method. It is applied to one equation of a system at a time.

Relevant Variable This is a variable with the theoretical justification for inclusion in the model.

Research Design It is a blue print for the whole research study. It covers the sampling, statistical and operational aspects/issues. Both the areas of logic and logistics are covered under research design.

Research Methodology This is a broad term commonly used. It includes both the research design and the research methods. The philosophy behind research (i.e. the perspective with which data is collected), the characteristics by which the variables are governed (without a knowledge of which interpretation of data collected is not possible), etc. are all covered under Research Methodology.

Research Methods Refers only to the 'tools and techniques' of survey. Tools help us in conducting the survey. Techniques help in analysing the data.

Restricted Model This is the model which imposes some restrictions on the values of one or more of the coefficients of the model.

Simple Linear Regression Model

This is a model with only one independent variable. For such a model, $R^2 \geq \bar{R}^2$. In simple linear regression, it is not required to conduct 'text for individual' significance as well as joint significance. This is because for such models, square of '*t*' value is equal to F.

Simultaneous Equation Bias

This refers to the bias arising from the application of classical least squares to an equation belonging to a system of simultaneous relations. It arises from the dependence of the explanatory variables with the error term (u_i).

t.test

This is the command for obtaining the results of t-test in R.

SELECTED READING

1) Dougherty C (2011). *Introduction to Econometrics*, Oxford, Oxford University Press.

2) G S Maddala and Kajal Lahiri (2009). Introduction to Econometrics, 4th Edition, Wiley.

3) Greene W H (2016). *Econometrics,* Prentice Hall.

4) Gujarati D, Porter D C and Gunasekhar S (2018). Basic Econometrics, 5th Edition, McGraw Hill Education (India) Pvt. Ltd.

5) Gujarati D N and Porter D C (2009). *Basic Econometrics* (Fifth Edition), McGraw Hill.

6) Gujarati D N and Porter D C (2010). *Essentials of Econometrics,* Fourth Edition), McGraw Hill.

7) Johnston J, Econometric Methods, McGraw-Hill.

8) Kmenta J, Elements of Econometrics, Macmillan, New York.

9) Koutsoyiannis A (2001). Theory of Econometrics, Palgrave Macmillan Limited.

10) Wooldridge J M (2009). *Econometrics*, Cengage Learning.

ignou
THE PEOPLE'S
UNIVERSITY