UNIT 12 STATA*

Structure

12.0 Objectives

12.1 Introduction

12.2 Graphical User Interface of Stata 16.0

- 12.2.1 Main Menu
- 12.2.2 Command and Dialogue Box
- 12.2.3 The Data Editor
- 12.2.4 The Data Viewer
- 12.2.5 The Do-File Editor

12.3 Basic Operations and Results

- 12.3.1 Data Input
- 12.3.2 Descriptive Statistics
- 12.3.3 Generation of New Variables
- 12.3.4 Changing the Variable Values

12.4 Analysis of Variance and Regression Analysis

- 12.4.1 One Way ANOVA
- 12.4.2 Two Way ANOVA

12.4.3 Regression Analysis

12.5 Regression Diagnostics

- 12.5.1 Test for Normality
- 12.5.2 Test for Heteroscedasticity
- 12.5.3 Test for Multicollinearity
- 12.5.4 Test for Autocorrelation
- 12.5.5 Test for Model Specification

12.6 Testing of Hypothesis

- 12.6.1 Single Hypothesis
- 12.6.2 Joint Hypothesis

12.7 Let Us Sum Up

12.8 Key Words

12.9 Suggested Books for Further Reading

12.10 Answers/Hints to Check Your Progress Exercises

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

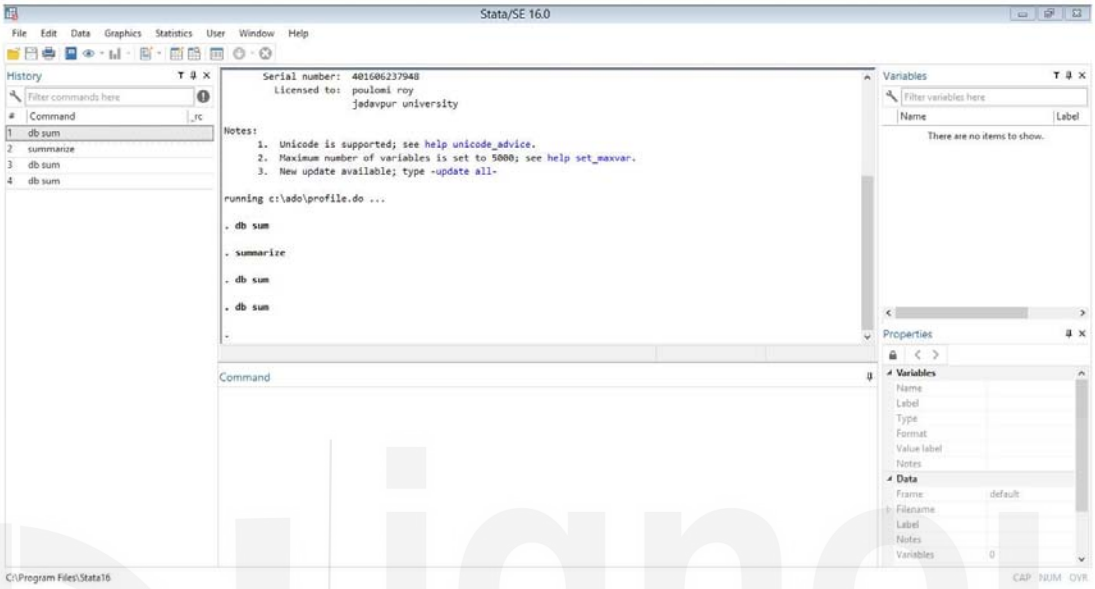

Fig. 12.1: Graphical User Interface of STATA

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.

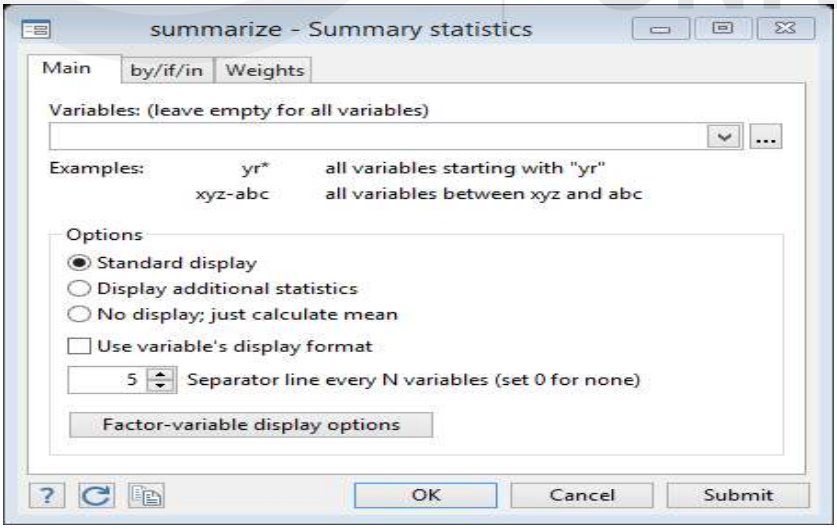

Fig. 12.3: Dialogue Box for Summary Statistics

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).

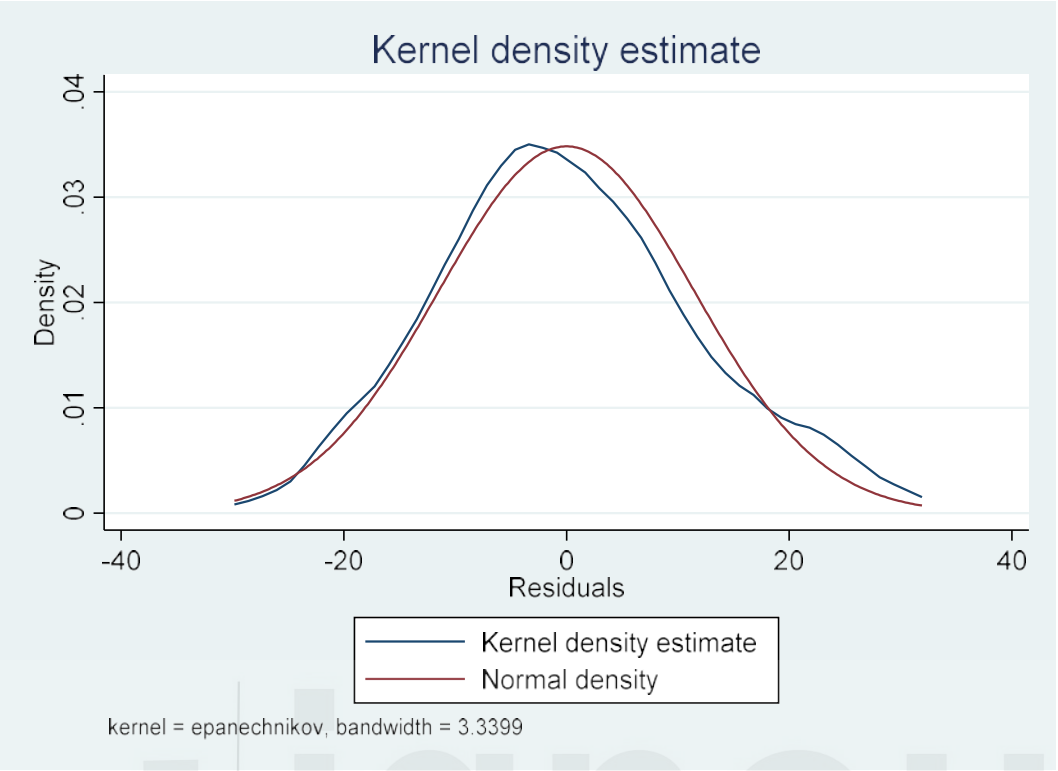

Fig. 12.4: Kernel Density Graph of Error Terms for Normality Test

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.

ignou THE PEOPLE'S UNIVERSITY

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.

ignou
THE PEOPLE'S
UNIVERSITY

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