

Copyrighted Material

*Using EViews for Principles of*

# ECONOMETRICS

*Third Edition*

William E. Griffiths, Ph.D. & R. Carter Hill

Guay C. Lim

Copyrighted Material

# PREFACE

This book and the EViews Student Version 6 econometric software program that is attached are supplements to *Principles of Econometrics, 3<sup>rd</sup> Edition* by R. Carter Hill, William E. Griffiths and Guay C. Lim (Wiley, 2008), hereinafter *POE*. This book is not a substitute for the textbook, nor is it a stand alone computer manual. It is a companion to the textbook, showing how to perform the examples in the textbook using EViews Student Version 6. It will be useful to students taking econometrics, as well as their instructors, and others who wish to use EViews for econometric analysis.

EViews is a very powerful and user-friendly program that is ideally suited for classroom use. You can find further details at the website <http://www.eviews.com>. The disk included with this book contains not only EViews Student Version 6, but also EViews workfiles for all the examples in *POE*, and corresponding text definition files (that can be opened with Notepad or Wordpad) of the form \*.def. These files and various other forms of support for *POE* are available at <http://www.wiley.com/college/hill> and <http://www.bus.lsu.edu/hill/poe>.

The chapters in this book parallel the chapters in *POE*. Thus, if you seek help for the examples in Chapter 11 of the textbook, check Chapter 11 in this book. However, within a Chapter the sections numbers in *POE* do not necessarily correspond to the sections in this EViews supplement.

We welcome comments on this book, and suggestions for improvement. We would like to acknowledge the valuable assistance of David Lilien from Quantitative Micro Software, the company that develops and distributed EViews. Of course, David (and EViews) are not responsible for any blunders that we may have committed..

William E. Griffiths  
Department of Economics  
University of Melbourne Vic 3010  
Australia  
[wegrif@unimelb.edu.au](mailto:wegrif@unimelb.edu.au)

R. Carter Hill  
Economics Department  
Louisiana State University  
Baton Rouge, LA 70803  
[eohill@lsu.edu](mailto:eohill@lsu.edu)

Guay C. Lim  
Melbourne Institute for Applied Economic and Social Research  
University of Melbourne Vic 3010  
Australia  
[g.lim@unimelb.edu.au](mailto:g.lim@unimelb.edu.au)

# **BRIEF CONTENTS**

1. Introduction to EViews	1
2. The Simple Linear Regression Model	36
3. Interval Estimation and Hypothesis Testing	60
4. Prediction, Goodness of Fit and Modeling Issues	70
5. The Multiple Linear Regression Model	90
6. Further Inference in the Multiple Regression Model	110
7. Nonlinear Relationships	130
8. Heteroskedasticity	147
9. Dynamic Models, Autocorrelation, and Forecasting	170
10. Random Regressors and Moment Based Estimation	194
11. Simultaneous Equations Models	211
12. Nonstationary Time Series Data and Cointegration	219
13. VEC and VAR Models: An Introduction to Macroeconometrics	227
14. Time-Varying Volatility and ARCH Models: An Introduction to Financial Econometrics	234
15. Panel Data Models	247
16. Qualitative and Limited Dependent Variables	269
17. Importing and Exporting	305
A. Review of Math Essentials	319
B. Statistical Distribution Functions	326
C. Review of Statistical Inference	338
Index	351

# CONTENTS

<b>CHAPTER 1 Introduction to EViews</b>	<b>1</b>	<b>CHAPTER 2 The Simple Linear Regression Model</b>	<b>36</b>
1.1 Using EViews for Principles of Econometrics	1	2.1 Open the Workfile	36
1.1.1 Installing EViews 6 student version	2	2.1.1 Examine the data	37
1.1.2 Checking for updates	2	2.1.2 Checking summary statistics	38
1.1.3 Obtaining data workfiles	2	2.1.3 Saving a group	39
1.2 Starting EViews	3	2.2 Plotting the Food Expenditure Data	40
1.3 The Help System	4	2.2.1 Enhancing the graph	42
1.3.1 EViews help topics	4	2.2.2 Saving the graph in the workfile	44
1.3.2 The READ ME file	5	2.2.3 Copying the graph to a document	44
1.3.3 Quick help reference	5	2.2.4 Saving a workfile	45
1.3.4 EViews Illustrated	6	2.3 Estimating a Simple Regression	45
1.3.5 Users guides and command reference	6	2.3.1 Viewing equation representations	47
1.4 Using a Workfile	7	2.3.2 Computing the income elasticity	48
1.4.1 Setting the default path	7	2.4 Plotting a Simple Regression	49
1.4.2 Opening a workfile	8	2.5 Plotting the Least Squares Residuals	51
1.4.3 Examining a single series	9	2.5.1 Using View options	51
1.4.4 Changing the sample	12	2.5.2 Using Resids	52
1.4.5 Copying a graph into a document	13	2.5.3 Using Quick/Graph	52
1.5 Examining Several Series	14	2.5.4 Saving the residuals	53
1.5.1 Summary statistics for several series	15	2.6 Estimating the Variance of the Error Term	54
1.5.2 Freezing a result	16	2.7 Coefficient Standard Errors	54
1.5.3 Copying and pasting a table	17	2.8 Prediction Using EViews	55
1.5.4 Plotting two series	17	2.8.1 Using direct calculation	55
1.5.5 A scatter diagram	18	2.8.2 Forecasting	56
1.6 Using the Quick Menu	19	<b>KEYWORDS</b>	<b>59</b>
1.6.1 Changing the sample	20	<b>CHAPTER 3 Interval Estimation and Hypothesis Testing</b>	<b>60</b>
1.6.2 Generating a new series	20	3.1 Interval Estimation	61
1.6.3 Plotting using Quick/Graph	21	3.1.1 Constructing the interval estimate	62
1.6.4 Saving your workfile	22	3.1.2 Using a coefficient vector	62
1.6.5 Opening an empty group	23	3.2 Right-tail Tests	64
1.6.6 Quick/Series statistics	25	3.2.1 Test of significance	64
1.6.7 Quick/Group statistics	26	3.2.2 Test of an economic hypothesis	65
1.7 Using EViews Functions	27	3.3 Left-tail Tests	65
1.7.1 Descriptive statistics functions	27	3.3.1 Test of significance	65
1.7.2 Using a storage vector	30	3.3.2 Test of an economic hypothesis	66
1.7.3 Basic arithmetic operations	33	3.4 Two-tail Tests	67
1.7.4 Basic math functions	34	3.4.1 Test of significance	67
<b>KEYWORDS</b>	<b>35</b>	3.4.2 Test of an economic hypothesis	69
		<b>KEYWORDS</b>	<b>69</b>
		<b>CHAPTER 4 Prediction, Goodness-of-Fit and Modeling Issues</b>	<b>70</b>
		4.1 Prediction in the Food Expenditure Model	70
		4.1.1 A simple prediction procedure	71

4.1.2 Prediction using EViews	72		
4.2 Measuring Goodness-of-Fit	75		
4.2.1 Calculating $R^2$	75		
4.2.2 Correlation analysis	76		
4.3 Modeling Issues	76		
4.3.1 The effects of scaling the data	76		
4.3.2 The log-linear model	78		
4.3.3 The linear-log model	79		
4.3.4 The log-log model	79		
4.3.5 Are the regression errors normally distributed?	80		
4.3.6 Another example	81		
4.4 The Log-Linear Model	85		
4.4.1 Prediction in the log-linear model	86		
4.4.2 Alternative commands in the log-linear model	87		
4.4.3 Generalized $R^2$	89		
KEYWORDS	89		
<b>CHAPTER 5 The Multiple Regression Model</b>	<b>90</b>		
5.1 The Workfile: Some Preliminaries	91		
5.1.1 Naming the page	91		
5.1.2 Creating objects: a group	92		
5.2 Estimating a Multiple Regression Model	94		
5.2.1 Using the Quick menu	94		
5.2.2 Using the Object menu	96		
5.3 Forecasting from a Multiple Regression Model	97		
5.3.1 A simple forecasting procedure	97		
5.3.2 Using the forecast option	99		
5.4 Interval Estimation	102		
5.4.1 The least squares covariance matrix	102		
5.4.2 Computing interval estimates	103		
5.5 Hypothesis Testing	104		
5.5.1 Two-tail tests of significance	104		
5.5.2 A one-tail test of significance	105		
5.5.3 Testing nonzero values	106		
5.6 Saving Commands	108		
KEYWORDS	109		
<b>CHAPTER 6 Further Inference in the Multiple Regression Model</b>	<b>110</b>		
6.1 $F$ and Chi-Square Tests	110		
6.1.1 Testing significance: a coefficient	111		
6.1.2 Testing significance: the model	115		
6.2 Testing in an Extended Model	116		
6.2.1 Estimating the model	116		
6.2.2 Testing: a joint $H_0$ , 2 coefficients	117		
6.2.3 Testing: a single $H_0$ , 2 coefficients	118		
6.2.4 Testing: a joint $H_0$ , 4 coefficients	121		
6.3 Including Nonsample Information	122		
6.4 The RESET Test	124		
6.4.1 The short way	124		
6.4.2 The long way	126		
6.5 Viewing the Correlation Matrix	127		
6.5.1 Collinearity: An exercise	128		
KEYWORDS	129		
<b>CHAPTER 7 Nonlinear Relationships</b>	<b>130</b>		
7.1 Polynomials	130		
7.2 Dummy Variables	134		
7.2.1 Creating dummy variables	135		
7.3 Interacting Dummy Variables	136		
7.4 Dummy Variables with Several Categories	138		
7.5 Testing the Equivalence of Two Regressions	140		
7.6 Interactions Between Continuous Variables	143		
7.7 Log-Linear Models	144		
KEYWORDS	146		
<b>CHAPTER 8 Heteroskedasticity</b>	<b>147</b>		
8.1 Examining Residuals	147		
8.1.1 Plot against observation number	148		
8.1.2 Plot against an explanatory variable	149		
8.1.3 Plot of least squares line	152		
8.2 Heteroskedasticity-Consistent Standard Errors	154		
8.3 Weighted Least Squares	155		
8.3.1 A short way	155		
8.3.2 A long way	156		
8.4 Estimating a Variance Function	157		
8.4.1 Variance function estimates	157		
8.4.2 Generalized least-squares	159		
8.5 A Heteroskedastic Partition	159		
8.5.1 Least-squares estimates: one equation	160		
8.5.2 Least-squares estimates: two equations	160		

8.5.3 Generalized least-squares estimates	162		
8.6 The Goldfeld-Quandt Test	163		
8.6.1 The wage equation	164		
8.6.2 The food expenditure equation	164		
8.7 Testing the Variance Function	166		
8.7.1 The Breusch-Pagan test	167		
8.7.2 The White test	168		
KEYWORDS	169		
<b>CHAPTER 9 Dynamic Models, Autocorrelation, and Forecasting</b>	<b>170</b>		
9.1 Least-Squares Residuals: Sugarcane Example	170		
9.1.1 Correlation between $\hat{e}_t$ and $\hat{e}_{t-1}$	172		
9.2 Newey-West Standard Errors	174		
9.3 Estimating an AR(1) Error Model	175		
9.3.1 A short way	175		
9.3.2 A long way	176		
9.3.3 A more general model	178		
9.3.4 Testing the AR(1) error restriction	178		
9.4 Testing for Autocorrelation	179		
9.4.1 Residual correlogram	179		
9.4.2 Lagrange multiplier ( <i>LM</i> ) test	182		
9.4.3 Durbin-Watson test	183		
9.5 Autoregressive Models	184		
9.5.1 Workfile structure for time series data	184		
9.5.2 Estimating an AR model	185		
9.5.3 Forecasting with an AR model	187		
9.6 Finite Distributed Lags	189		
9.7 Autoregressive Distributed Lag Models	190		
9.7.1 Graphing the lag weights	191		
KEYWORDS	193		
<b>CHAPTER 10 Random Regressors and Moment Based Estimation</b>	<b>194</b>		
10.1 The Inconsistency of the Least Squares Estimator	194		
10.2 Instrumental Variables Estimation	199		
10.3 The Hausman Test	201		
10.4 Test for Weak Instruments	202		
10.5 Test Instrument Validity	203		
10.6 A Wage Equation	204		
KEYWORDS	210		
<b>CHAPTER 11 Simultaneous Equations Models</b>	<b>211</b>		
11.1 Examining the Data	211		
11.2 Estimating the Reduced Form	212		
11.3 TSLS Estimation of an Equation	213		
11.4 TSLS Estimation of a System of Equations	214		
11.5 Supply and Demand at Fulton Fish Market	216		
KEYWORDS	218		
<b>CHAPTER 12 Nonstationary Time Series Data and Cointegration</b>	<b>219</b>		
12.1 Stationary and Nonstationary Variables	219		
12.2 Spurious Regressions	221		
12.3 Unit Root Tests for Stationarity	222		
12.4 Cointegration	224		
KEYWORDS	226		
<b>CHAPTER 13 VEC and VAR Models: An Introduction to Macroeconometrics</b>	<b>227</b>		
13.1 VEC and VAR Models	227		
13.2 Estimating a VEC Model	227		
13.3 Estimating a VAR Model	230		
13.4 Impulse Responses and Variance Decompositions	233		
KEYWORDS	233		
<b>CHAPTER 14 Time-Varying Volatility and ARCH Models: An Introduction to Financial Econometrics</b>	<b>234</b>		
14.1 Time-Varying Volatility	234		
14.2 Testing for ARCH Effects	236		
14.3 Estimating an ARCH Model	239		
14.4 Generalized ARCH	242		
14.5 Asymmetric ARCH	243		
14.6 GARCH in Mean Model	245		
KEYWORDS	246		

<b>CHAPTER 15 Panel Data Models</b>	<b>247</b>
15.1 Grunfeld Data: Two Equations	247
15.1.1 Separate least squares estimation	248
15.1.2 Stacking the data	249
15.1.3 Least squares estimation with dummy variables	251
15.1.4 Introducing the pool object	252
15.1.5 Seemingly unrelated regressions	254
15.1.6 Testing contemporaneous correlation	255
15.1.7 Testing cross-equation restrictions	256
15.2 Grunfeld Data: Ten Firms	257
15.2.1 Structuring the workfile	258
15.2.2 Fixed effects using dummy variables	258
15.2.3 Testing the effects	260
15.2.4 Pooled least squares	260
15.2.5 The fixed effects estimator	261
15.3 NLS Panel Data	263
15.3.1 Fixed effects estimation	264
15.3.2 Random effects estimation	265
15.3.3 The Hausman test	266
<b>KEYWORDS</b>	<b>268</b>
<b>CHAPTER 16 Qualitative and Limited Dependent Variables</b>	<b>269</b>
16.1 Models with Binary Dependent Variables	269
16.1.1 Examine the data	270
16.1.2 The linear probability model	271
16.1.3 The probit model	273
16.1.4 Predicting probabilities	275
16.1.5 Marginal effects in the probit model	277
16.2 Ordered Choice Models	279
16.2.1 Ordered probit predictions	281
16.2.2 Ordered probit marginal effects	284
16.3 Models for Count Data	286
16.3.1 Examine the data	288
16.3.2 Estimating a Poisson model	290
16.3.3 Prediction with a Poisson model	290
16.3.4 Poisson model marginal effects	292
16.4 Limited Dependent Variables	293
16.4.1 Least squares estimation	294

16.4.2 Tobit estimation and interpretation	296
16.4.3 The Heckit selection bias model	298
<b>KEYWORDS</b>	<b>304</b>

<b>CHAPTER 17 Importing and Exporting Data</b>	<b>305</b>
17.1 Obtaining Data from the Internet	305
17.2 Importing An Excel Data File	310
17.3 Date Conventions	313
17.4 Importing a Text (Ascii) Data File	314
17.5 Entering Data Manually	316
17.6 Exporting Data from EViews	318
<b>KEYWORDS</b>	<b>318</b>

<b>APPENDIX A Review of Math Essentials</b>	<b>319</b>
A.1 Mathematical Operations	319
A.2 Logarithms and Exponentials	320
A.3 Graphing Functions	322
<b>KEYWORDS</b>	<b>325</b>

<b>APPENDIX B Statistical Distribution Functions</b>	<b>326</b>
B.1 Cumulative Normal Probabilities	327
B.2 Using Vectors	329
B.3 Computing Normal Distribution Percentiles	331
B.4 Plotting Some Normal Distributions	332
B.5 Plotting the <i>t</i> -Distribution	335
B.6 Plotting the Chi-square Distribution	335
B.7 Plotting the <i>F</i> Distribution	336
B.8 Probability Calculations for the <i>t</i> , <i>F</i> and Chi-square	337
<b>KEYWORDS</b>	<b>337</b>

<b>Appendix C Review of Statistical Inference</b>	<b>338</b>
C.1 A Histogram	338
C.2 Summary Statistics	340
C.2.1 The sample mean	340
C.2.2 Estimating higher moments	341
C.2.3 Create a table	342
C.2.4 Using the estimates	345
C.3 Interval Estimation	346

C.4 Hypothesis Tests About the Population Mean	348
C.4.1 One-tail test using the hip data	348
C.4.2 Two-tail test using the hip data	348
C.4.3 Testing the normality of the population	349
KEYWORDS	350
<b>INDEX</b>	<b>351</b>

# CHAPTER 1

## Introduction to EViews

### CHAPTER OUTLINE

- 1.1 Using EViews for Principles of Econometrics
  - 1.1.1 Installing EViews 6 student version
  - 1.1.2 Checking for updates
  - 1.1.3 Obtaining data workfiles
- 1.2 Starting EViews
- 1.3 The Help System
  - 1.3.1 EViews help topics
  - 1.3.2 The READ ME file
  - 1.3.3 Quick help reference
  - 1.3.4 EViews Illustrated
  - 1.3.5 Users guides and command reference
- 1.4 Using a Workfile
  - 1.4.1 Setting the default path
  - 1.4.2 Opening a workfile
  - 1.4.3 Examining a single series
  - 1.4.4 Changing the sample
  - 1.4.5 Copying a graph into a document
- 1.5 Examining Several Series
  - 1.5.1 Summary statistics for several series
  - 1.5.2 Freezing a result
  - 1.5.3 Copying and pasting a table
  - 1.5.4 Plotting two series
  - 1.5.5 A scatter diagram
- 1.6 Using the Quick Menu
  - 1.6.1 Changing the sample
  - 1.6.2 Generating a new series
  - 1.6.3 Plotting using Quick/Graph
  - 1.6.4 Saving your workfile
  - 1.6.5 Opening an empty group
  - 1.6.6 Quick/Series statistics
  - 1.6.7 Quick/Group statistics
- 1.7 Using EViews Functions
  - 1.7.1 Descriptive statistics functions
  - 1.7.2 Using a storage vector
  - 1.7.3 Basic arithmetic operations
  - 1.7.4 Basic math functions

### KEYWORDS

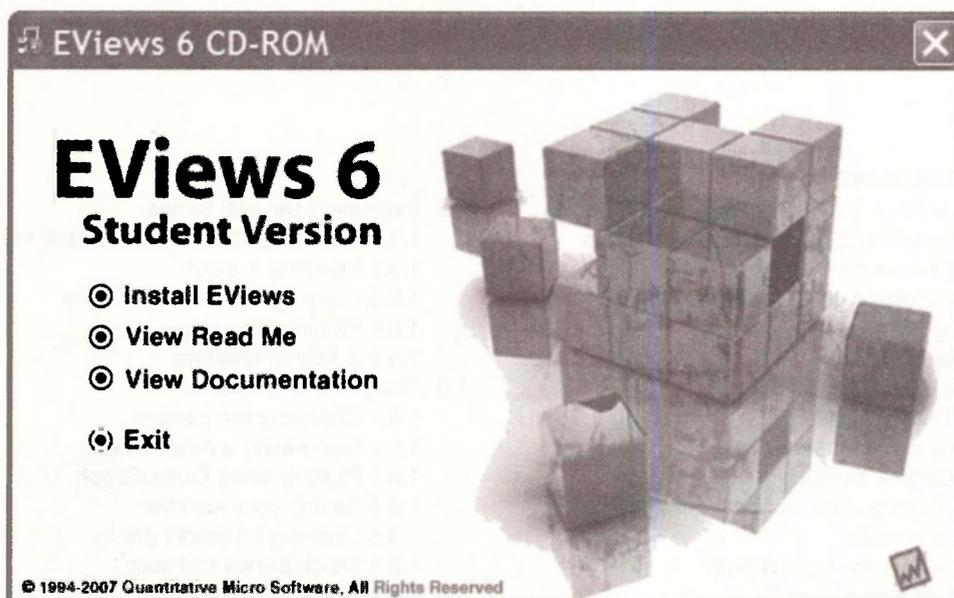
### 1.1 USING EIEWS FOR PRINCIPLES OF ECONOMETRICS

This manual is a supplement to the textbook *Principles of Econometrics, 3<sup>rd</sup> Edition*, by Hill, Griffiths and Lim (John Wiley & Sons, Inc., 2008). It is not in itself an econometrics book, nor is it a complete computer manual. Rather it is a step-by-step guide to using EViews for the empirical examples in *Principles of Econometrics*, which we will abbreviate as *POE*. This book contains a CD with **EViews Student Version 6**. We imagine you sitting at a computer with your *POE* text and *Using EViews for Principles of Econometrics, 3<sup>rd</sup> Edition* open, following along with the manual to replicate the examples in *POE*. Before you can do this you must install EViews and obtain the EViews “workfiles,” which are documents that contain the actual data.

### 1.1.1 Installing EViews 6 student version

Your copy of EViews is distributed on a single CD-ROM. EViews is a Windows-based program. First close all other applications, then insert the CD into your computer's drive and wait until the setup program launches. If the CD does not spin-up on its own, navigate to the CD drive using Windows Explorer, and click on the Setup icon (AUTORUN.EXE).

The initial screen is shown below. You should first click on **View Read Me** to see any last minute changes or instructions. Click **View Documentation** to open *EViews 6 Getting Started for the Student Version*. It describes the installation and registration process. This document can also be accessed from the **Help/Student Version Getting Started (pdf)** from the main EViews menu. You must have Adobe Acrobat to read \*.pdf files. If you do not have this software, go to [www.adobe.com](http://www.adobe.com) and download the free Adobe Reader.



When you are ready, click **Install EViews** and follow the on-screen instructions.

### 1.1.2 Checking for updates

Once installed you should visit [www.eviews.com](http://www.eviews.com) and check the “**download**” link. There you will find any updates for your software.

### 1.1.3 Obtaining data workfiles

The EViews data workfiles (with extension \*.wfl) and other resources for *POE* can be found at [www.wiley.com/college/hill](http://www.wiley.com/college/hill). Find the link “Online resources for students.” The *POE* workfiles can be downloaded in a compressed format, saved to a subdirectory (we use c:\data\evIEWS), and then expanded. In addition to the EViews workfiles, there are “*data definition*” files (\*.def) that describe the variables and show some summary statistics. The definition files are simple text files that can be opened with utilities like Notepad or Wordpad, or using a word processor. These files

\* There are a number of books listed by authors named Hill. *POE* will be one of them.

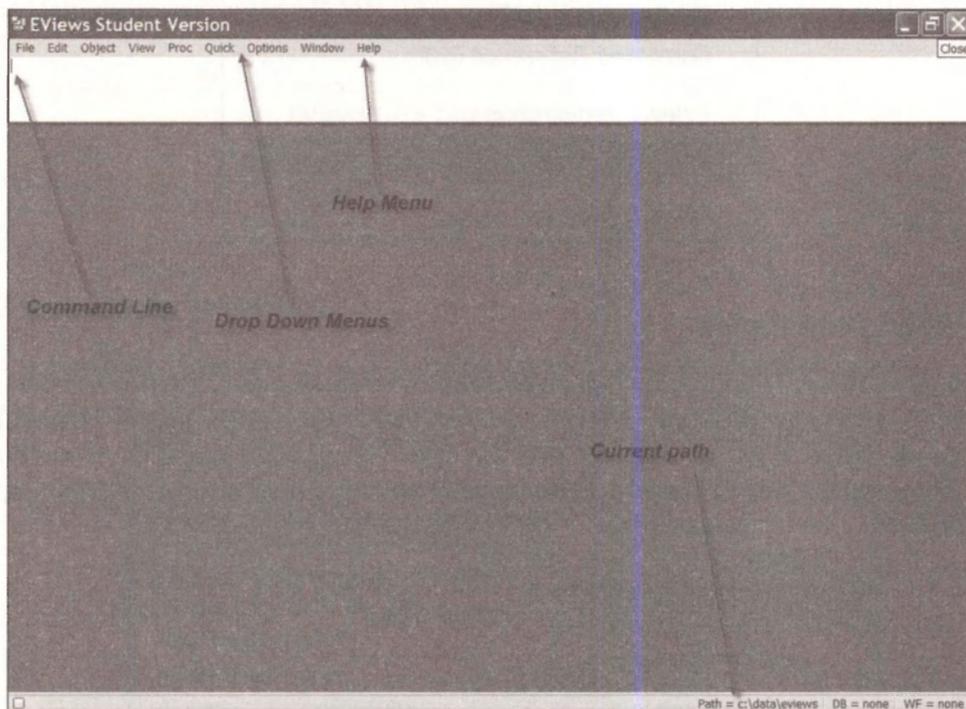
should be downloaded as well. Individual EViews workfiles, definition files, and other resources can be obtained at [www.bus.lsu.edu/hill/poe](http://www.bus.lsu.edu/hill/poe). The data files in EViews, Excel, and ASCII format, along with the definition files, are also on the EViews 6 SV CD-ROM that came with this manual.

## 1.2 STARTING EViews

To launch EViews double click the EViews 6 SV icon on the desktop, if one is present. It should resemble



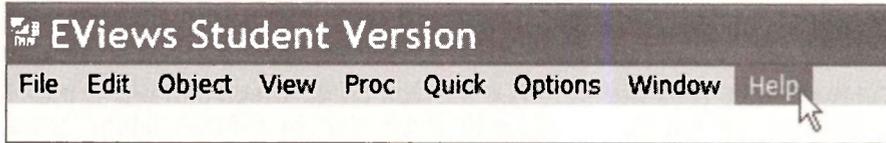
Alternatively, select EViews from the Windows Start Menu. When EViews opens you are presented with the following screen



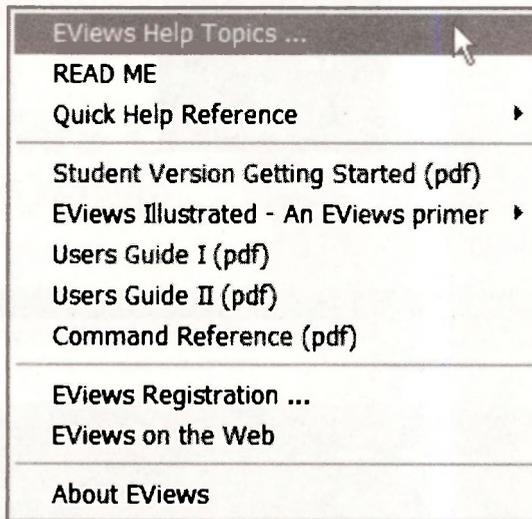
Across the top are **Drop Down Menus** that make implementing EViews procedures quite simple. Below the menu items is the **Command Line**. It can be used as an alternative to the menus, once you become familiar with basic commands and syntax. Across the bottom is the **Current Path** for reading data and saving files. The EViews **Help Menu** is going to become a close friend.

### 1.3 THE HELP SYSTEM

Click **Help** on the EViews menu

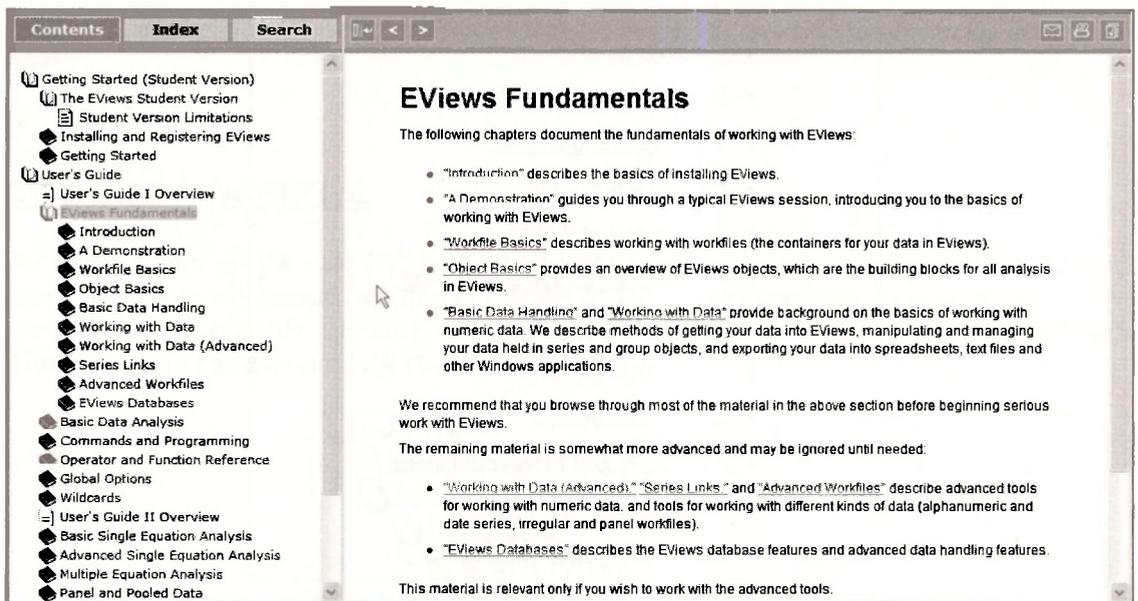


The resulting menu is



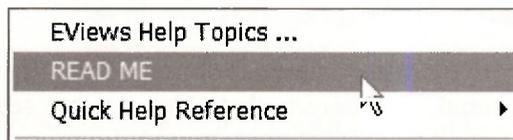
#### 1.3.1 EViews help topics

First, click on **EViews Help Topics**. Select **Getting Started (Student Version)** to obtain basic information about the Student Version of EViews. Selecting **User's Guide/EViews Fundamentals** opens a list of chapters that can take you through specifics of working with EViews. These guides will be a useful reference after you have progressed further through *Using EViews for POE*.



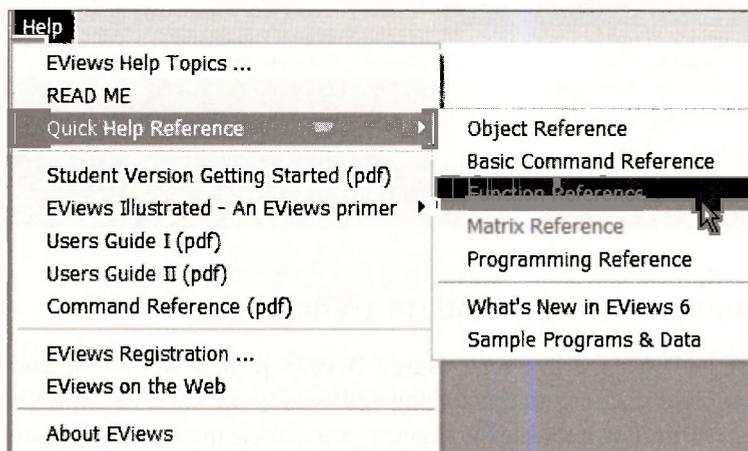
### 1.3.2 The read me file

On the **Help** menu, select **READ ME**. This opens a PDF file with the latest installation notes and errata.



### 1.3.3 Quick help reference

Select **Quick Help Reference**. You find another menu. Select **Function Reference**.



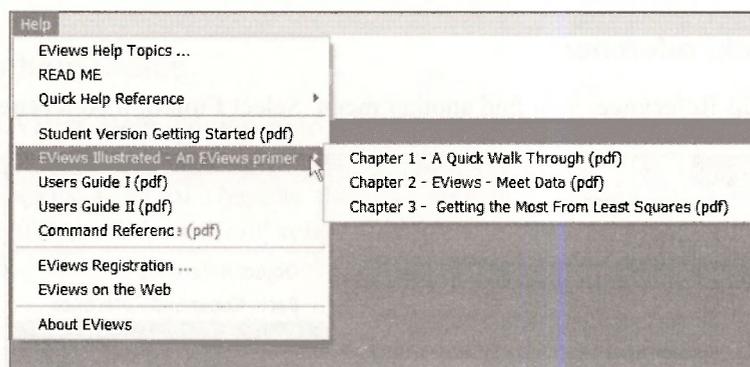
EViews has many, many functions available for easy use.

- [Operators.](#)
- [Basic mathematical functions.](#)
- [Time series functions.](#)
- [Financial functions.](#)
- [Descriptive statistics.](#)
- [Cumulative statistics functions.](#)
- [Moving statistics functions.](#)
- [Group row functions.](#)
- [By-group statistics.](#)
- [Additional and special functions.](#)
- [Trigonometric functions.](#)
- [Statistical distribution functions.](#)

You should just take a moment to examine the **Operators** (basic addition, multiplication, etc.) and the **Basic mathematical functions** (square roots, logarithms, absolute value, etc.). This **Function Reference** help is one that you will use very frequently, and to which we will refer a great deal.

### 1.3.4 EViews Illustrated

The next **Help menu** item is for sample chapters from *EViews Illustrated* by Richard Startz, which is good humored tutorial, with screen shots like you are seeing here, covering many aspects of using EViews. The first three chapters are provided.



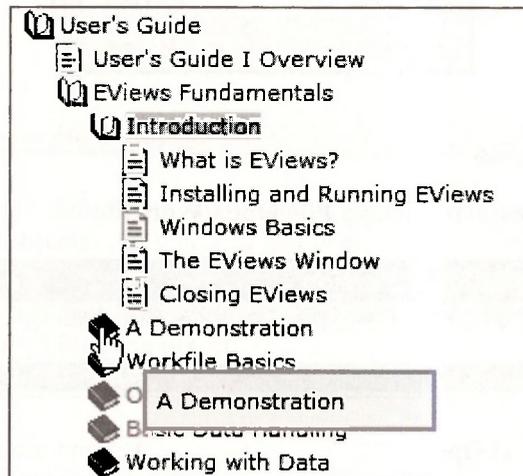
### 1.3.5 Users Guides and Command Reference

The User's Guide I (794 pages), User's Guide II (688 pages) and Command Reference (926 pages) are the complete documentation for the full version of EViews 6. While these are good rainy day reading we do not necessarily suggest you search them for information until you are more familiar with the workings of EViews. This book, *Using EViews for POE*, is an effort to

guide you through the essentials of EViews that are needed to replicate the examples in the book *POE*.

## 1.4 USING A WORKFILE

As noted earlier all the data for the book *Principles of Econometrics* is provided as EViews *workfiles*. These will be used starting in Chapter 2. To illustrate the basic functioning of EViews we will use an example provided with EViews. Click on **Help/EViews Help Topics/User's Guide/EViews Fundamentals/A Demonstration**.



The demonstration starts by importing data into EViews from an Excel spreadsheet. We will skip that step here, but if you like you can gain further practice by following along their demonstration.

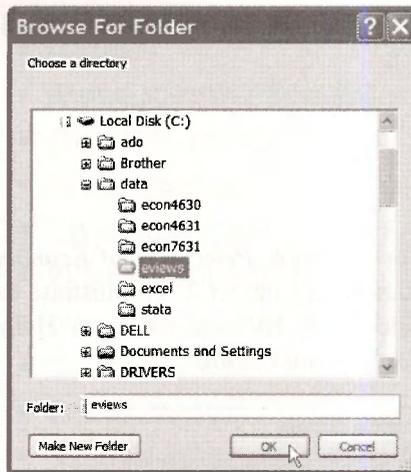
**Remark:** You will want to be able to “import” data into EViews, or enter data manually. We cover the various methods for entering data into EViews in Chapter 17 of this manual.

### 1.4.1 Setting the default path

At the bottom of the EViews screen you will see **Path = .** This determines where EViews will first look for data and save workfiles.

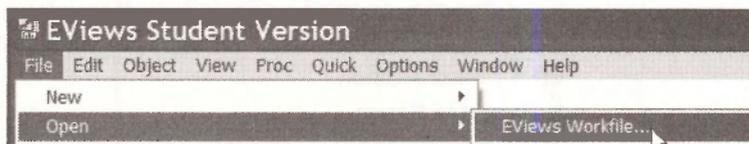
Path = c:\data\evIEWS

Double-click on **Path** and a window will open in which you can locate what you desire for your default directory. We will use the path **c:\data\evIEWS**

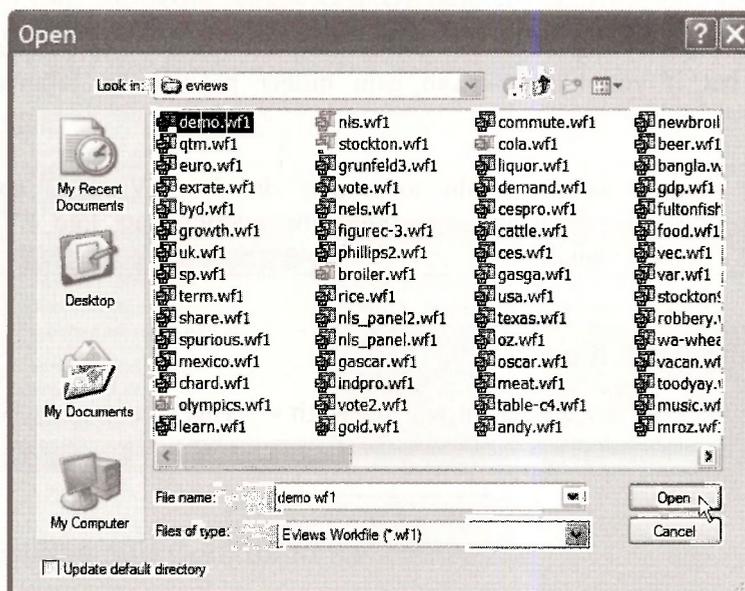


### 1.4.2 Opening a workfile

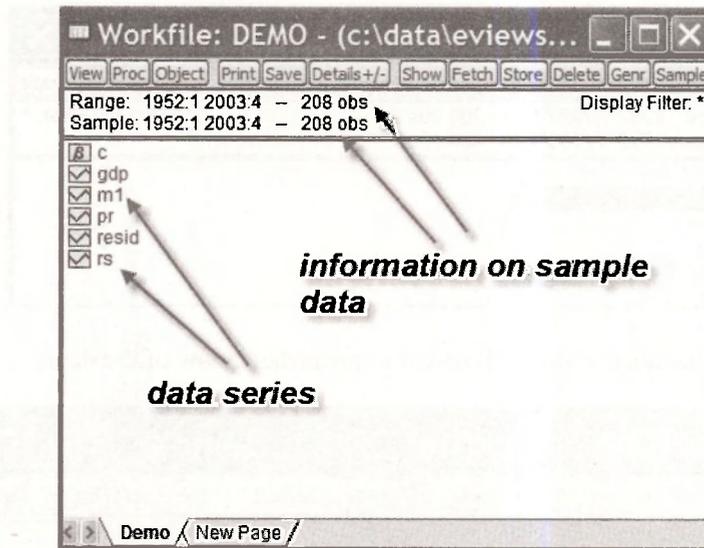
Open the workfile called *demo* by clicking **File/Open/Workfile**



Select *demo.wf1* and click on **Open**.



The workfile opens to show



Located on the left side are data series that are indicated by the icon . EViews calls the elements of the workfile **objects**. As you will discover, there are many types of objects that EViews can save into the workfile—not only series but tables, graphs, equations, and so on. As Richard Startz says, an object is a little “thingie” that computer programmers talk about. Each little icon “thingie” in the workfile is an object.

In this workfile the data series, or variables, are:

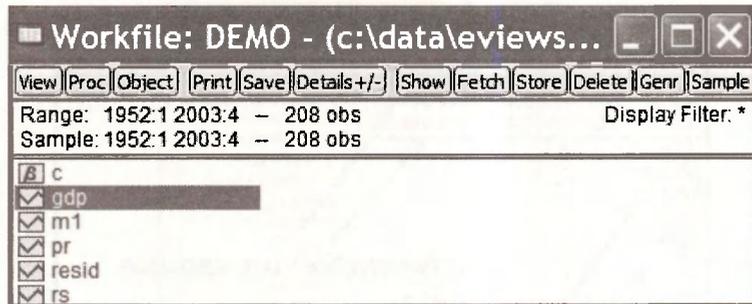
- *GDP*—gross domestic product
- *MI*—money supply
- *PR*—price level (index)
- *RS*—short term interest rate

The series **resid** and the icon labeled  $\beta$  are always present in EViews workfiles (even new ones with no data) and their use will be explained later. Across the top of the workfile are various buttons that initiate tasks in EViews, and these too will be explained later.

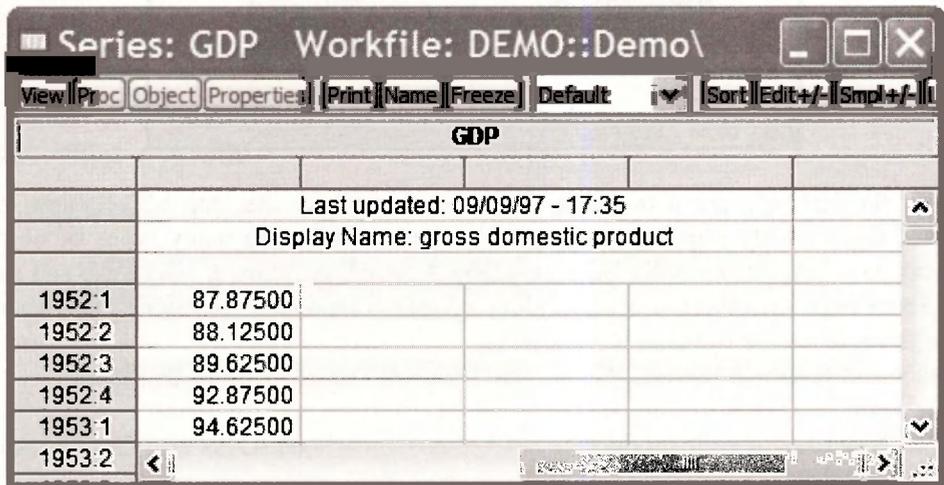
Below the buttons are **Range: 1952:1 2003:4**, which indicates that the 208 observations on the variables included run from 1952, Quarter 1, to 2003, Quarter 4. **Sample: 1952:1 2003:4** denotes the data observations EViews will use in calculations. Many times we will choose for analysis less than the full range of observations that are available, so **Sample** will differ from **Range**.

### 1.4.3 Examining a single series

It is a good idea each time you open a workfile to look at one or more series just to verify that the data are what you expect. First, select one series



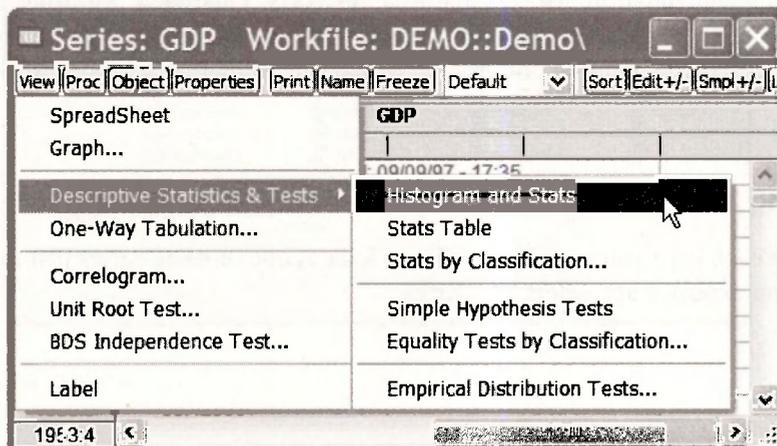
Double-click in the blue area, which will reveal a spreadsheet view of the data.



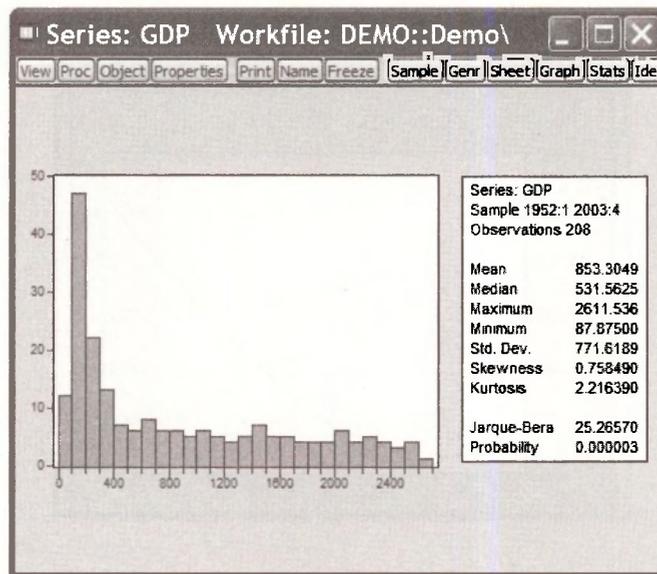
In the upper left hand corner is a button labeled **View**



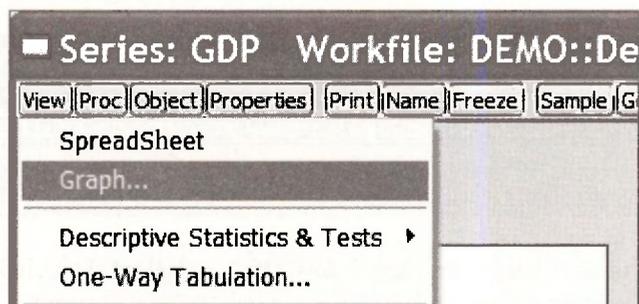
This opens a drop-down menu with a number of choices. Select **Descriptive Statistics & Tests/Histogram and Stats**.



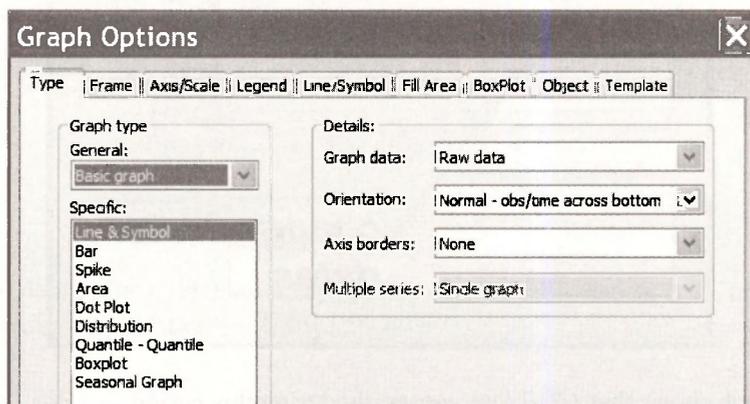
The result is



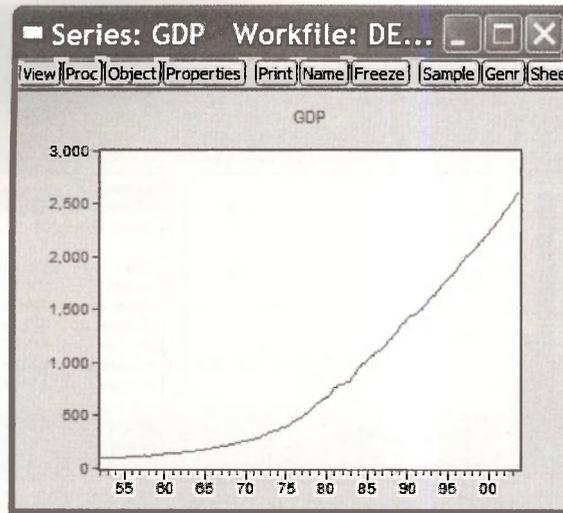
This histogram is shown with various summary statistics on the side.  
Click on **View** again. Select **Graph**.



There you will see many options. The default graph type is a **Basic Graph** with the **Line & Symbol** plotted. Select **OK**.

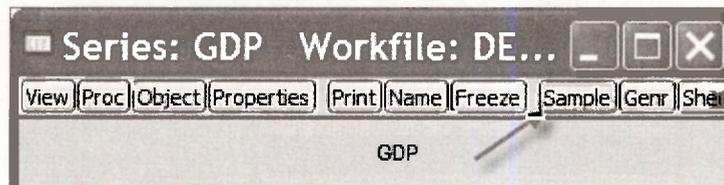


The result is a line graph. The dates are on the horizontal axis and *GDP* on the vertical axis.

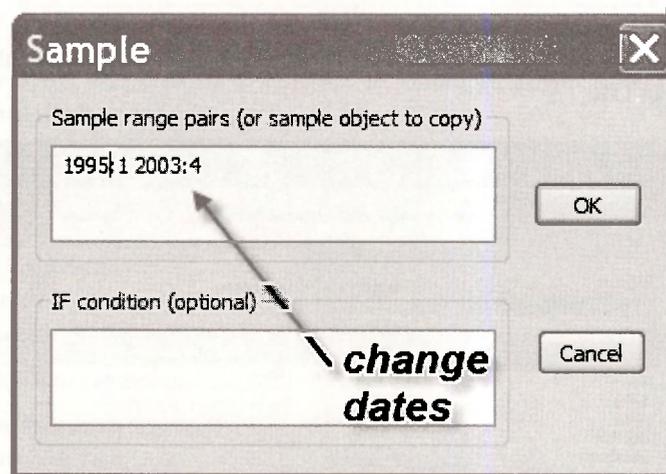


### 1.4.4 Changing the sample

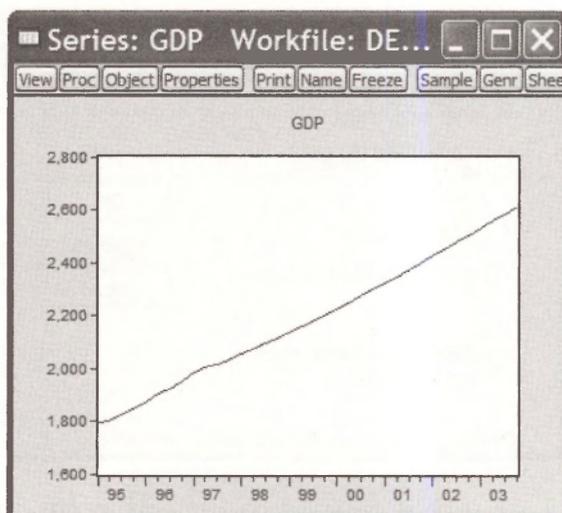
If you wish to view the graph or summary statistics for a different sample period, click on the **Sample** button. This feature works the same in all EViews windows.



In the dialog box that opens change the sample to 1995:1 to 2003:4 then click **OK**.



The resulting graph shows that *GDP* rose constantly during this period.

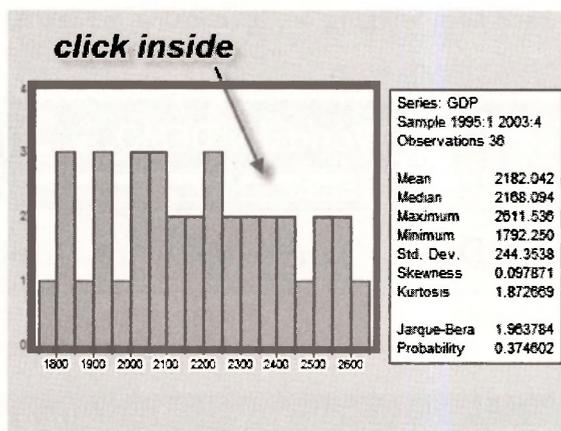


### 1.4.5 Copying graph into a document

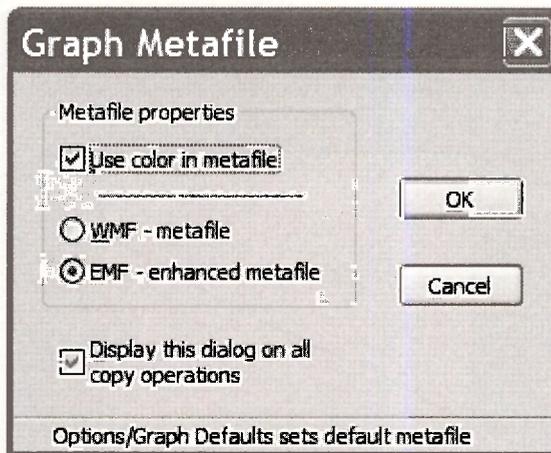
Select **View/Descriptive Statistics & Tests/Histogram and Stats**. You will find now the summary statistics and histogram of *GDP* for the period 1995:1 to 2003:4. These results can be printed by selecting the **Print** button.

**Print**

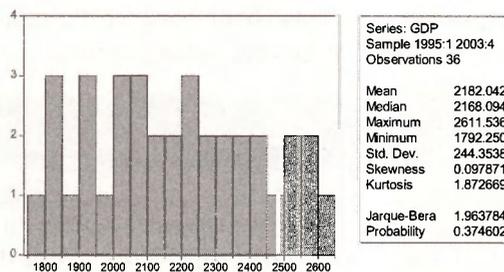
You may prefer to copy the results into a word processor for later editing and combining results. How can results be taken from EViews into a document? Click inside the histogram



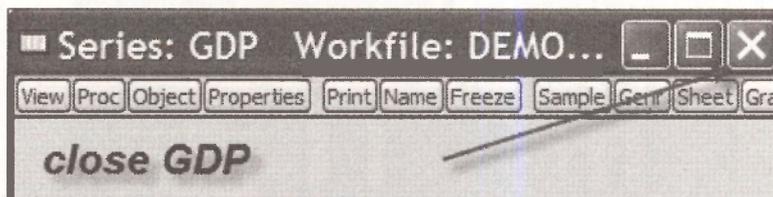
While holding down the **Ctrl** key press **C** (which we will denote as **Ctrl+C**). This is the Windows keystroke combination for **Copy**.



In the resulting dialog box you can make some choices, then click **OK**. This copies the graph into the Windows clipboard (memory). Open a document in your word processor and enter **Ctrl+V** which will **Paste** the figure into your document.

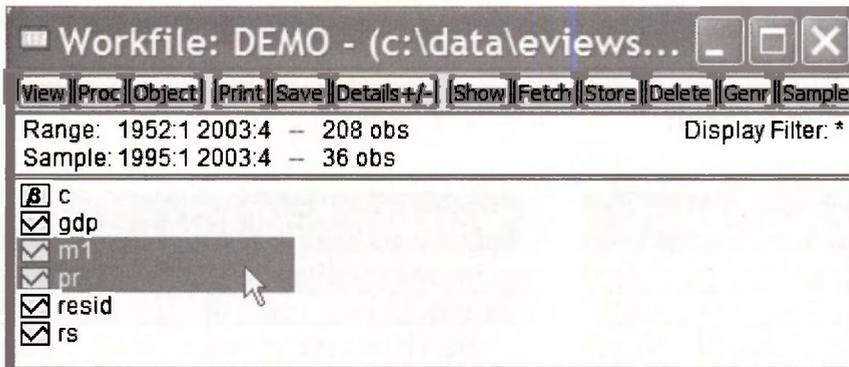


Lets close the graph we have been working on, by clicking the red **X** in the upper right hand corner of the *GDP* screen

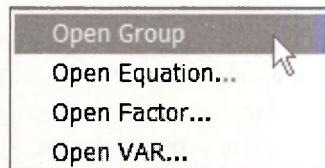


### 1.5 EXAMINING SEVERAL SERIES

Rather than examining one series at a time we can view several. In the workfile window select the series *MI* and then while holding down the **Ctrl**-key select the *PR* series. Double click inside the blue area to open what is called a **Group** of variables.



Click on **Open Group**.



A spreadsheet view of the data will open.

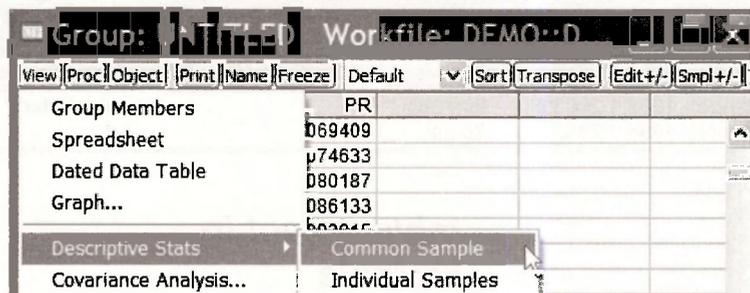
Group: UNTITLED Workfile: DEMO::D...

obs	M1	PR
1995.1	1209.235	1.069409
1995.2	1219.420	1.074633
1995.3	1204.520	1.080187
1995.4	1197.609	1.086133
1996.1		

Note that the series begins in 1995:1 because we changed the **Sample** range in Section 1.4.4.

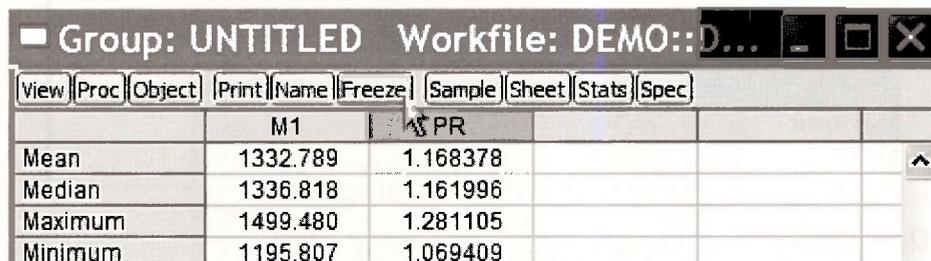
### 1.5.1 Summary statistics for several series

From the spreadsheet we can again examine the data by selecting the **View** button. Select **Descriptive Stats/Common Sample**.



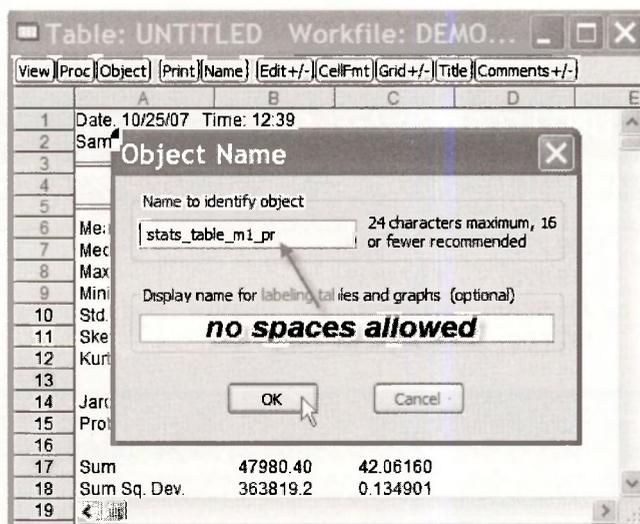
The result is a table of summary statistics is created for the two series (variables) in the group.

### 1.5.2 Freezing a result

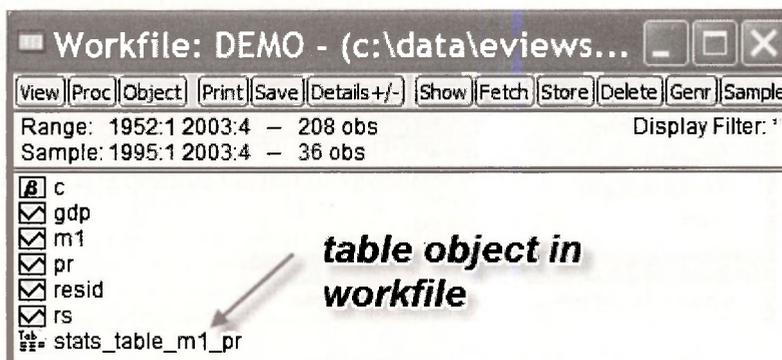


	M1	PR
Mean	1332.789	1.168378
Median	1336.818	1.161996
Maximum	1499.480	1.281105
Minimum	1195.807	1.069409

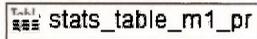
These results can be “saved” several ways. Select the **Freeze** button. This actually saves an image of the table. In the new image window, select the **Name** button. Enter a name for this image, which EViews calls an **Object**. The name should be relatively short and cannot contain any spaces. Often underscores “\_” can be used to separate words to make recognition easier.



Click **OK**, then close the **Object** by clicking on the red **X**. Check back in the workfile and you will now see a new entry, which is the table you have created.

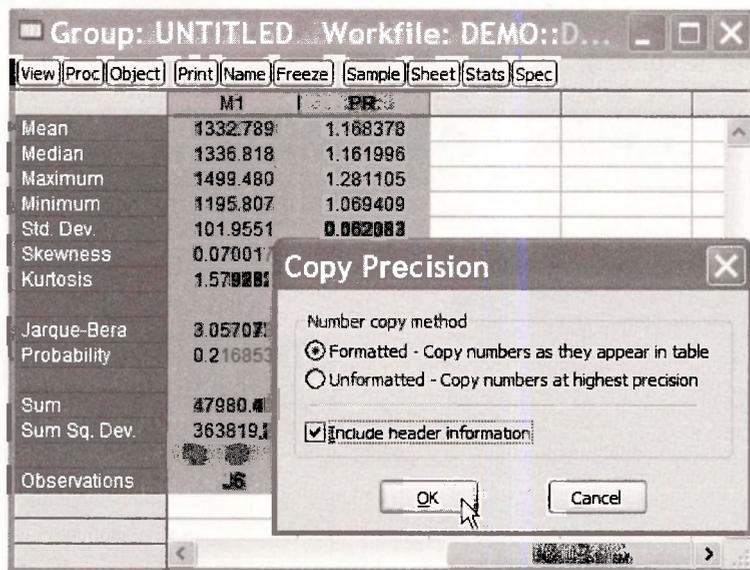


The table can be recalled at any time by double clicking the **Table icon**



### 1.5.3 Copying and pasting a table

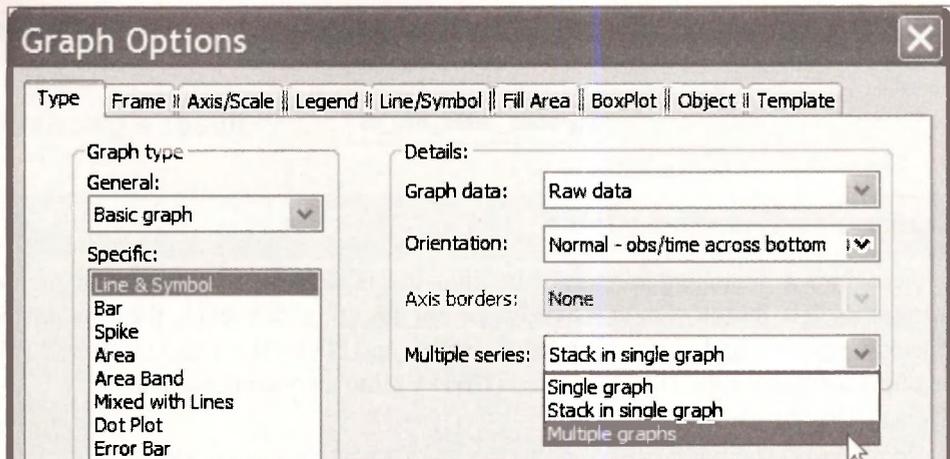
To copy these into a document directly, highlight the table of results (drag the mouse while holding down its left button), enter **Ctrl+C**. In the resulting box click the **Formatted** radio button, check the box to **Include header information**, and click **OK**. This copies the table to the Windows clipboard, which then can be pasted (**Ctrl+V**) into an open document.



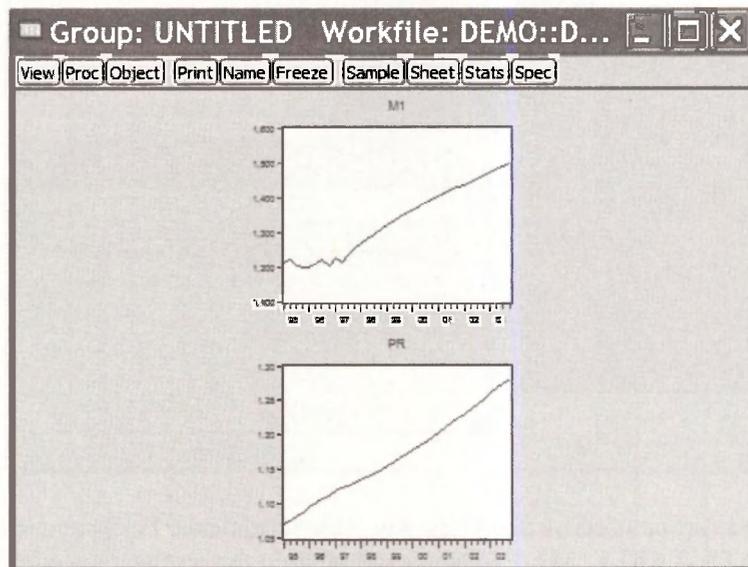
This same method can be used for basically any table in EViews. For example, if you open the saved table “**STATS\_TABLE\_M1\_PR**” you can highlight the results, then copy and paste as we have done here.

### 1.5.4 Plotting two series

Return to the spreadsheet view of the two series *M1* and *PR*. Select **View/Graph**. In the resulting dialog box, select **Multiple graphs** in the **Multiple series** menu.



Click **OK** to obtain two plots of the series

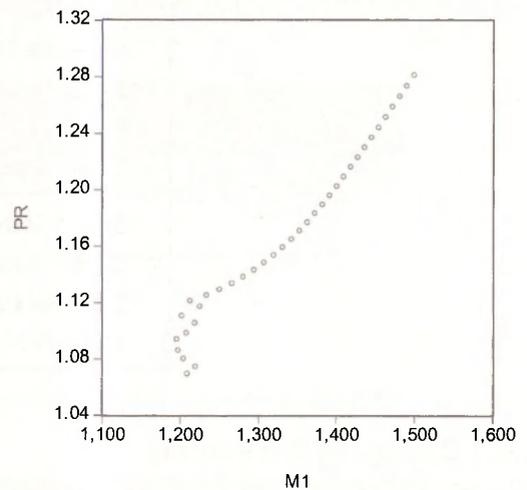
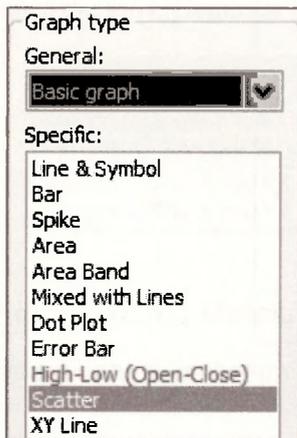


We can **Freeze** this picture, then assign it a **Name** for future reference.

### 1.5.5 A scatter diagram

A scatter diagram is a plot of the data points with one variable on one axis and the other variable on the other axis. In the **Group** screen click **View/Graph**. Select **Scatter** as the specific type of graph.

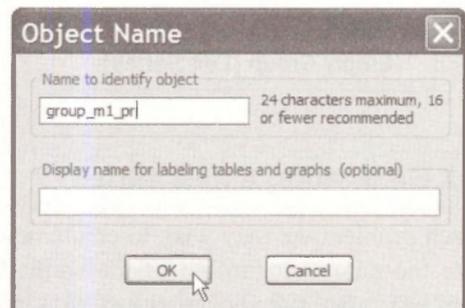
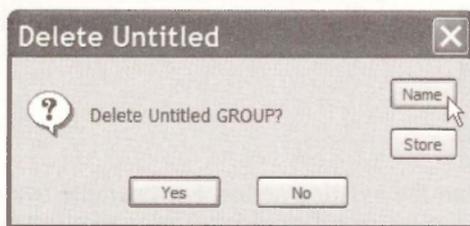
Click **OK**. Copy the graph by clicking inside the graph area, entering **Ctrl+C** to copy, then paste into a document using **Ctrl+V**. The resulting graph is on the next page. Recall that we are still operating with the sample from 1995:1 to 2003:4, which is only 36 data points.



The variable *M1* is on the horizontal axis because it is the first series in the spreadsheet view.

Get table

Click the red X in the Group window. When you do this you will be faced with some choices. The Group consists of the two series *M1* and *PR*. This Group can be saved by selecting **Name** and assigning a name.



In the workfile window you will find a new object for this group.

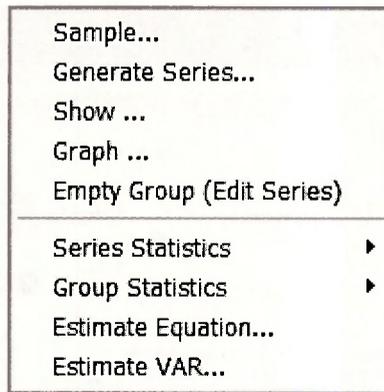
 group\_m1\_pr

## 1.6 USING THE QUICK MENU

The spreadsheet view of the data is very powerful. Another key tool is the **Quick** menu on the EViews workfile menu.

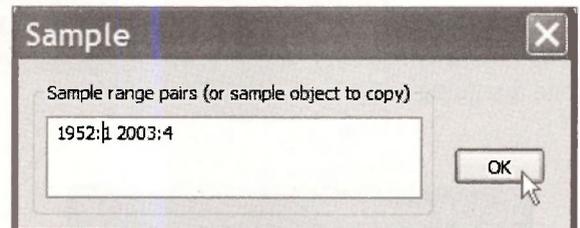
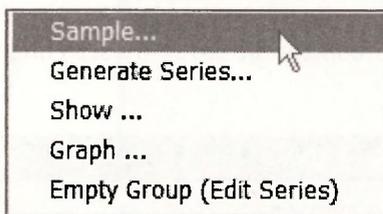


The options shown are



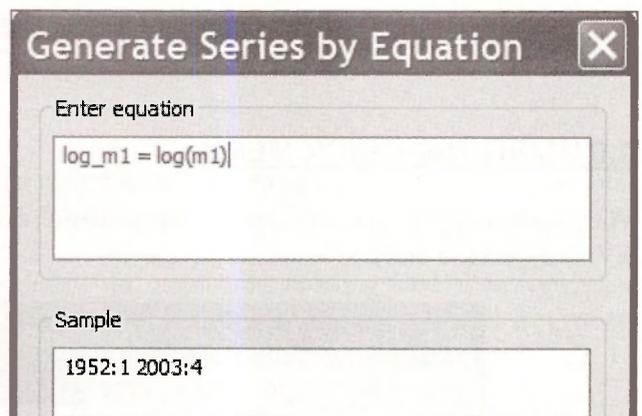
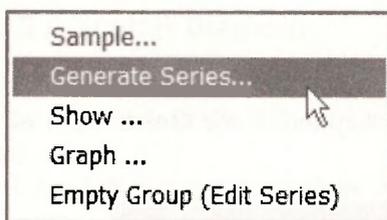
### 1.6.1 Changing the sample

By selecting **Sample** from this menu we can change the range of sample observations. Change the sample to 1952:1 to 2003:4 and click **OK**.

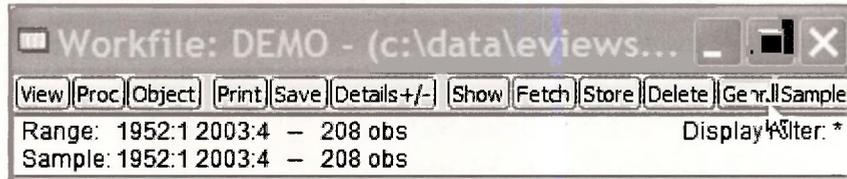


### 1.6.2 Generating a new series

In each problem we may wish to create new series from the existing series. For example, we can create the natural logarithm of the series **M1**. Select **Quick/Generate Series**. In the resulting dialog box type in the equation  $\log\_m1 = \log(m1)$ , then click **OK**. A new series will appear in the workfile. The function **log** creates the natural logarithm. All logarithms use in *Principles of Econometrics* are natural logs.



Alternatively, we can generate a new series by selecting the **Genr** button on the workfile menu. This will open the same **Generate Series** dialog box.



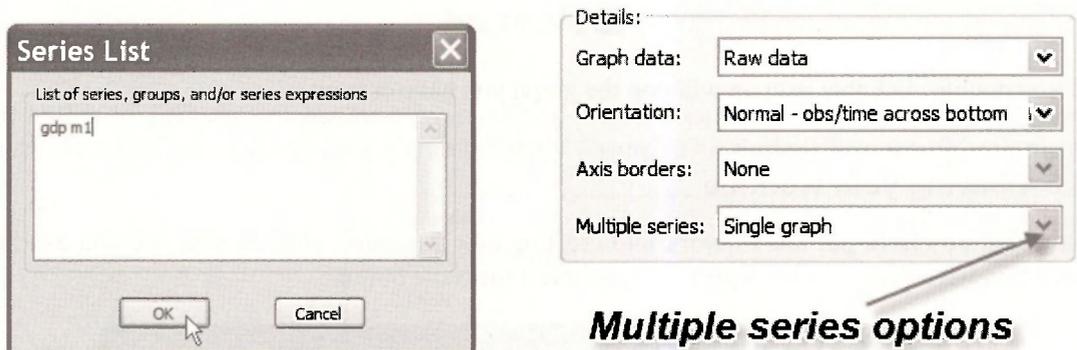
### 1.6.3 Plotting using Quick/Graph

We can create graphs from the spreadsheet view, but we can also use **Quick/Graph**.

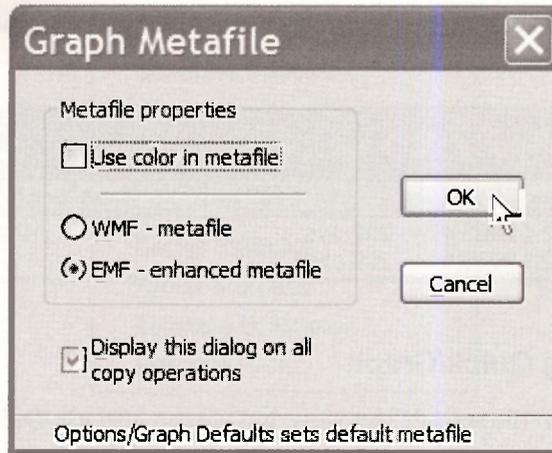


This will open the **Graph options** window. For a basic graph click **OK**.

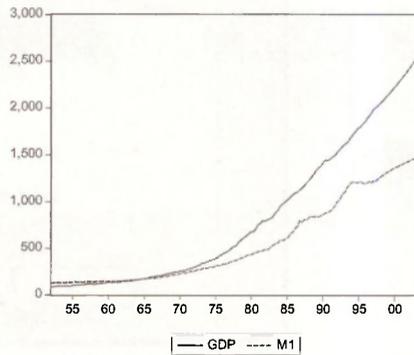
If you enter two series into the **Series List** window then the **Graph options** window will have an additional option



Click **OK**. The resulting graph shows the two series plots in a single window. In EViews the curves are in two different colors, but this will not show in a black and white document. The programmers at EViews have thought of this problem. Click inside the graph and enter **Ctrl+C** to copy. In the **Graph Metafile** box that opens uncheck the box "Use color in metafile." Click **OK**.



In your document enter **Ctrl+V** to paste the black and white graph. Now the graph lines show up as solid for *GDP* and broken for *M1* so that the difference can be viewed.



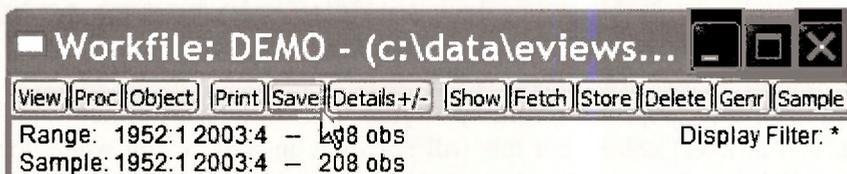
To save the graph, click **Name** and enter the name **GDP\_M1\_PLOT**. Click **OK**. Close the graph by clicking “**X**”. You will find an icon in the workfile window.



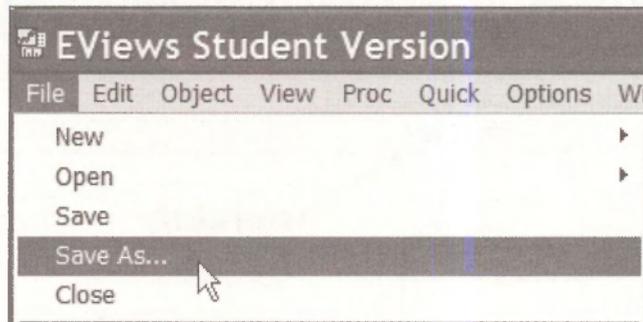
If you double click this icon up will pop the graph you have created.

### 1.6.4 Saving your workfile

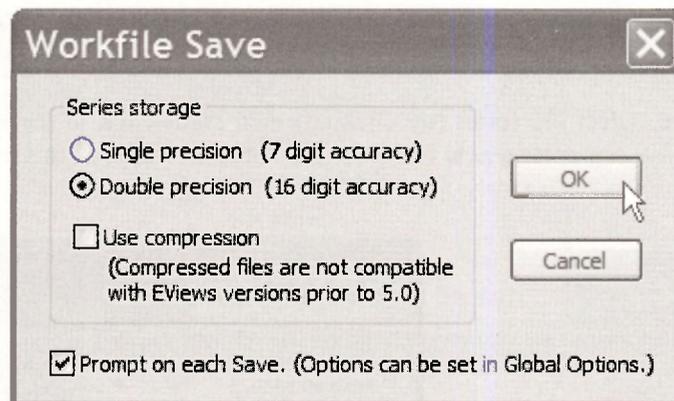
Now that you have put lots of work into creating new variables, plots, and so on, you can **Save** what you have done. On the workfile menu select the **Save** button



In the following window, IF you click **OK** then all the objects you have created will be saved into the workfile *demo.wfl*. You may wish to save these results using a different name, so that the original data workfile is not changed. To save the workfile, select **File/Save As** on the main EViews menu



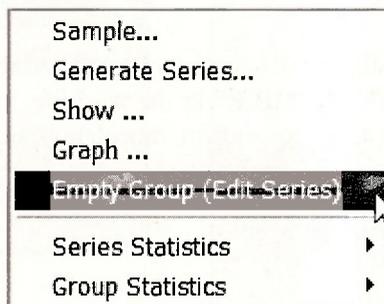
We will use the name *demo\_ch1.wfl* for this workfile. Enter this and click **OK**. You will be presented with some options. Use the default of **Double precision** and click **OK**.



You will note that the workfile name has changed.

### 1.6.5 Opening an empty group

The ability to enter data manually is an important one. In Chapter 17 we show all the ways you might enter data into EViews. Select **Quick/Empty Group (Edit Series)** from the EViews menu.



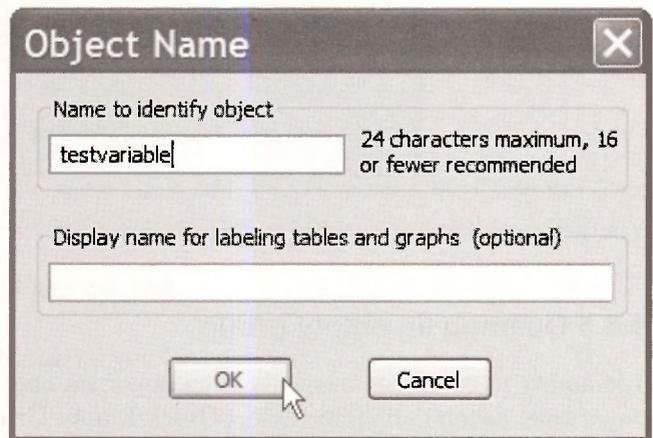
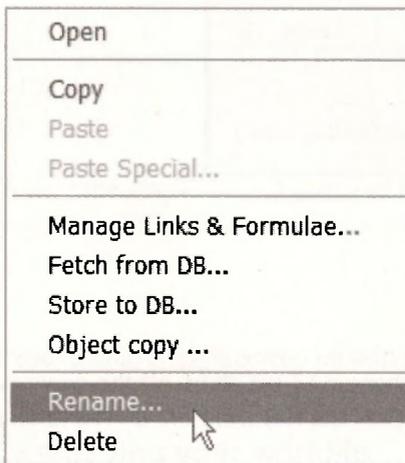
A spreadsheet opens into which you can enter new data. The default name for a new series is **SER01** that we will change. As you enter a number press **Enter** to move to the next cell. You can add new data in as many columns as you like.

obs	SER01
1952:1	1.000000
1952:2	2.000000
1952:3	2.000000
1952:4	5
1953:1	NA

When you have finished entering the data you wish, click the red X in the upper right corner of the active window. You will be asked if you want to “**Delete Untitled GROUP?**” Select **Yes**. In the workfile *demo\_ch1.wfl* you will now find the new series labeled



To change this name, select the series (by clicking) then **right-click** in the shaded area. A box will open in which you can enter a new name for the “object” which in this case is a data series. Press **OK**



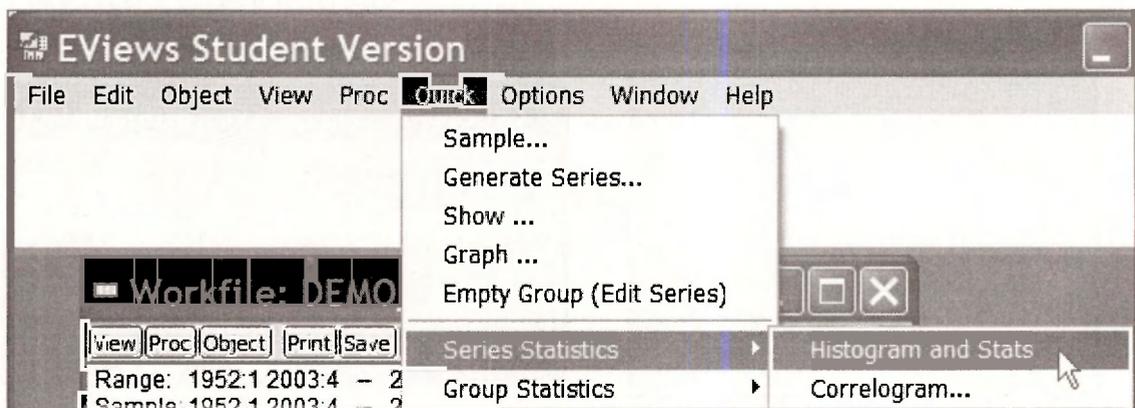
You can go through these same steps to delete an unwanted variable, such as the one we have just created. Select the series “**TESTVARIABLE**” in the workfile, and right click. Select **Delete**. In the resulting window you will be asked to confirm the deletion. Select **Yes**.



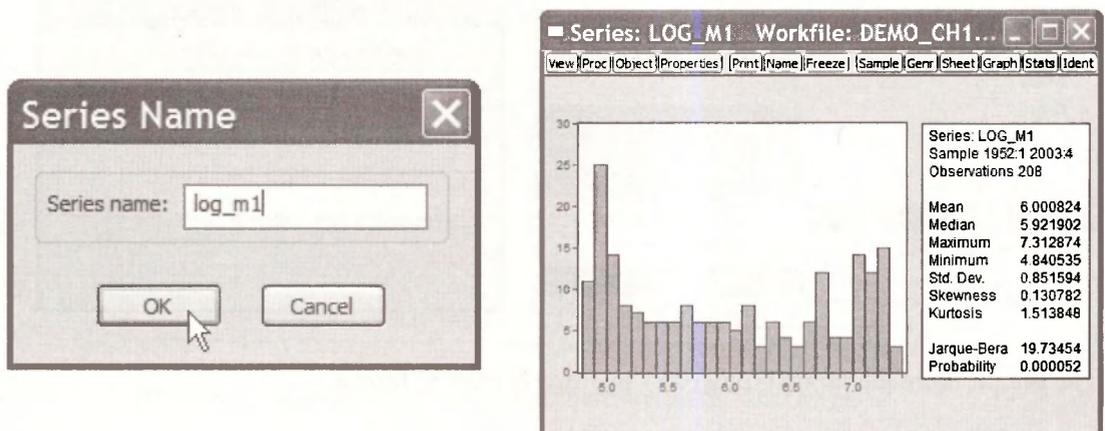
More than one series or objects can be selected for deletion by selecting one, then hold down the **Ctrl**-key while selecting others. To delete all these selected objects **right-click** in the blue area, and repeat the steps above.

### 1.6.6 Quick/Series statistics

The next item on the EViews Quick menu is **Series Statistics**. Select **Quick/Series Statistics/Histogram and Stats**

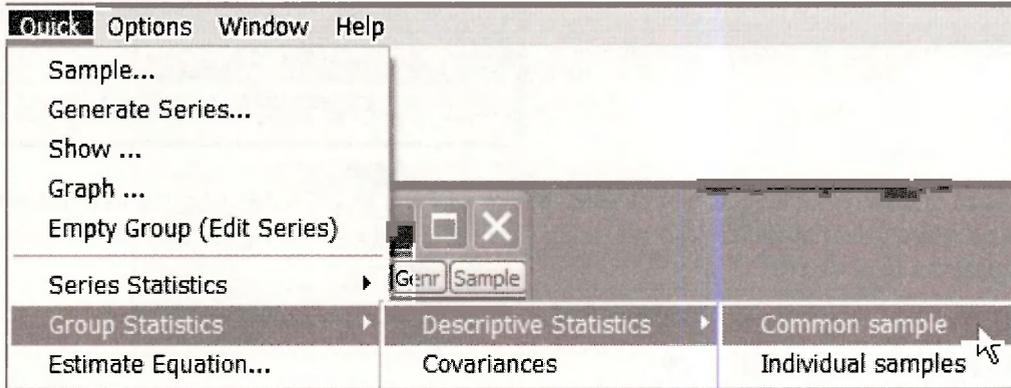


In the resulting window you can enter the name of the series (one) for a which you desire the summary statistics. Then select **OK**.

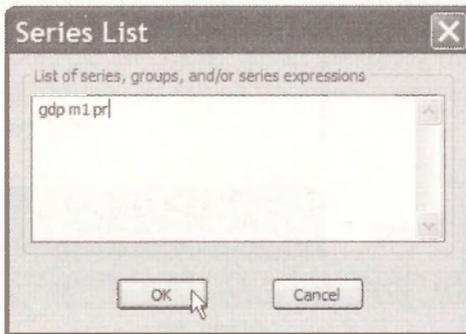


### 1.6.7 Quick/Group statistics

We can obtain summary statistics for a Group of series by choosing **Quick/Group Statistics**.

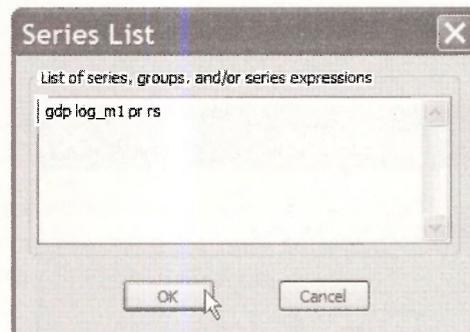
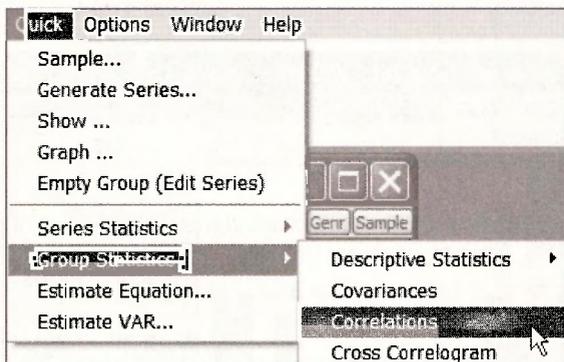


Enter the series names into the box and press **OK**. This will create the summary statistics table we have seen before. You can **Name** this group, or **Freeze** the table, or copy and paste using **Ctrl+C** and **Ctrl+V**.



View	Proc	Object	Print	Name	Freeze	Sample	Sheet	Stats	Spec
				GDP		M1		PR	
Mean				853.3049		569.3548		0.605202	
Median				531.5625		373.1375		0.490262	
Maximum				2611.536		1499.480		1.281105	
Minimum				87.87500		126.5370		0.197561	
Std. Dev.				771.6189		451.3036		0.365495	
Skewness				0.758490		0.726813		0.402946	
Kurtosis				2.216390		2.029020		1.620387	

Another option under **Quick/Group Statistics** is **Correlations**. Enter the names of series for which the sample correlations are desired and click **OK**.



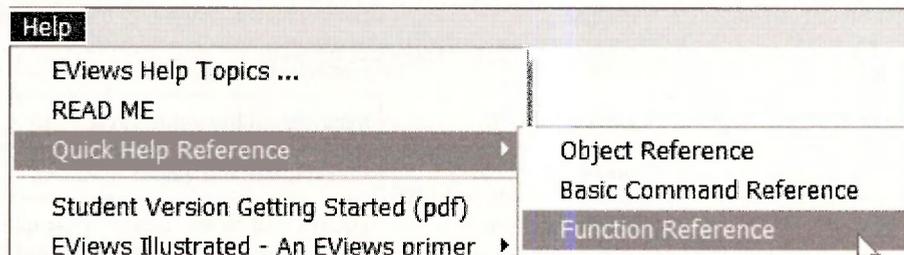
The sample correlations are arranged in an array, or matrix, format.

The screenshot shows the EViews software interface with a window titled "Group: UNTITLED Workfile: DEM...". The menu bar includes View, Proc, Object, Print, Name, Freeze, Sample, Sheet, Stats, and Spec. The main display area shows a "Correlation" matrix for the variables GDP, LOG M1, PR, and RS. The matrix is symmetric, with diagonal elements all equal to 1.000000. The off-diagonal elements represent the pairwise correlations between the variables.

	GDP	LOG M1	PR	RS
GDP	1.000000	0.959003	0.986551	0.168504
LOG M1	0.959003	1.000000	0.987383	0.346364
PR	0.986551	0.987383	1.000000	0.268010
RS	0.168504	0.346364	0.268010	1.000000

## 1.7 USING EIEWS FUNCTIONS

Now we will explore the use of some EViews functions. Select **Help/Quick Help Reference/Function Reference**.



### 1.7.1 Descriptive statistics functions

Select **Descriptive Statistics** from the list of material links.

## Operator and Function Reference

This material is divided into several topics:

- [Operators.](#)
- [Basic mathematical functions.](#)
- [Time series functions.](#)
- [Financial functions.](#)
- [Descriptive statistics.](#)
- [Cumulative statistics functions.](#)

Some of the descriptive statistics functions listed there are on the next page.

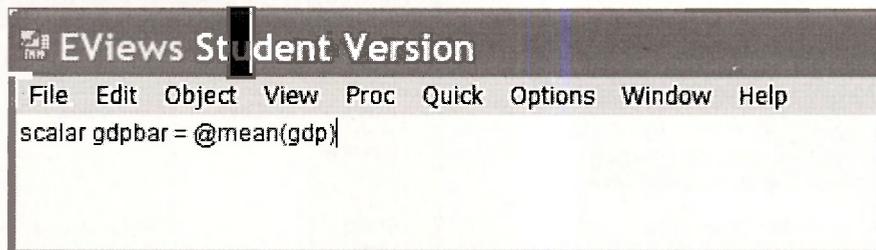
## Descriptive Statistics Functions in EViews 6 Student Version

Function	Name	Description
@cor(x, y[, s])	correlation	the correlation between X and Y.
@cov(x, y[, s])	covariance	the covariance between X and Y (division by $n$ ).
@covp(x, y[, s])	population covariance	the covariance between X and Y (division by $n$ ).
@covs(x, y[, s])	sample covariance	the covariance between X and Y (division by $n-1$ ).
@inner(x, y[, s])	inner product	the inner product of X and Y.
@obs(x[, s])	number of observations	the number of non-missing observations for X in the current sample.
@nas(x[, s])	number of NAs	the number of missing observations for X in the current sample.
@mean(x[, s])	mean	average of the values in X.
@median(x[, s])	median	computes the median of the X (uses the average of middle two observations if the number of observations is even).
@min(x[, s])	minimum	minimum of the values in X.
@max(x[, s])	maximum	maximum of the values in X.
@stdev(x[, s])	standard deviation	square root of the unbiased sample variance (sum-of-squared residuals divided by $n-1$ ).
@stdevp(x[, s])	population standard deviation	square root of the population variance (sum-of-squared residuals divided by $n$ ).
@stdevs(x[, s])	sample standard deviation	square root of the unbiased sample variance. Note this is the same calculation as @stdev.
@var(x[, s])	variance	variance of the values in X (division by $n$ ).
@varp(x[, s])	population variance	variance of the values in X. Note this is the same calculation as @var.
@vars(x[, s])	sample variance	sample variance of the values in X (division by $n-1$ ).
@skew(x[, s])	skewness	skewness of values in X.
@kurt(x[, s])	kurtosis	kurtosis of values in X.
@sum(x[, s])	sum	the sum of X.
@prod(x[, s])	product	the product of X (note this function could be subject to numerical overflows).
@sumsq(x[, s])	sum-of-squares	sum of the squares of X.

In this table of functions you will note that these functions begin with the “@” symbol. Also, these functions return a single number, which is called a **scalar**. In the commands the variables, or series, are called **x** and **y**. The bracket notation “[,s]” is optional and we will not use it. These functions are used by typing commands into the **Command Window** and pressing **Enter**. For example, to compute the sample mean of *GDP* type

**scalar gdpbar = @mean(gdp).**

The command window looks like this.



At the bottom of the EViews screen you will note the message

GDPBAR successfully created

In the workfile window the new object is denoted with “#” that indicates a scalar.

 gdpbar

We called the sample mean **GDPBAR** because sample means are often denoted by symbols like  $\bar{x}$  which is pronounced “x-bar.” In the “text messaging” world in which you live, simple but meaningful names will occur to you naturally.

To view this scalar object double click on it, or type **show gdpbar** in the Command window. At the bottom of the EViews screen you will see

Scalar GDPBAR = 853.304863221

The sample mean of *GDP* during the sample period is 853.305.

Scalars you have created can be used in further calculations. For example, enter the following commands by typing them into the command window and pressing **Enter**

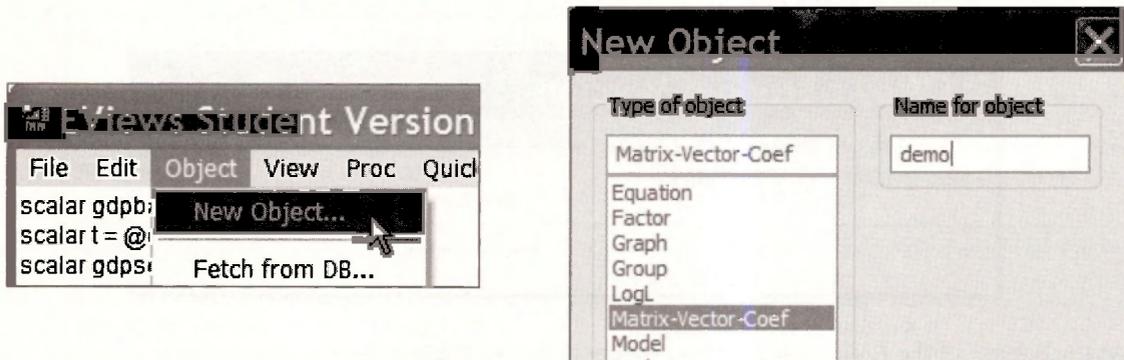
**scalar t = @obs(gdp)**  
**scalar gdpse = @stdev(gdp)**  
**scalar z = (gdpbar - 800)/(gdpse/@sqrt(t))**

The value of **z** is 0.996, and is the test statistic value for the null hypothesis that the population mean of *GDP* equals 800. In the workfile are now objects for each of the scalars created.

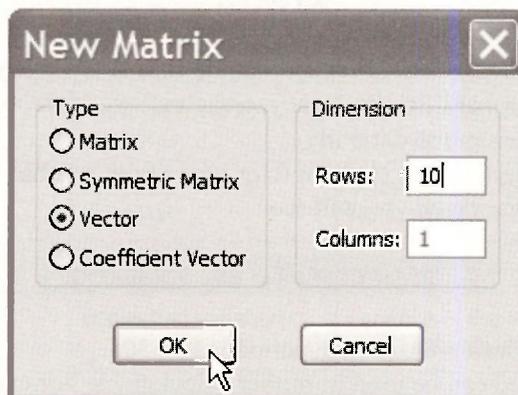
### 1.7.2 Using a storage vector

The creation of scalars leads to inclusion of additional objects into the workfile, and the scalars cannot be viewed simultaneously. One solution is to create a storage vector into which these scalars can be placed.

On the EViews menu bar select **Object/New Object**. In the resulting dialog box select **Matrix-Vector-Coeff** and enter an object name, say **DEMO**. Click **OK**.



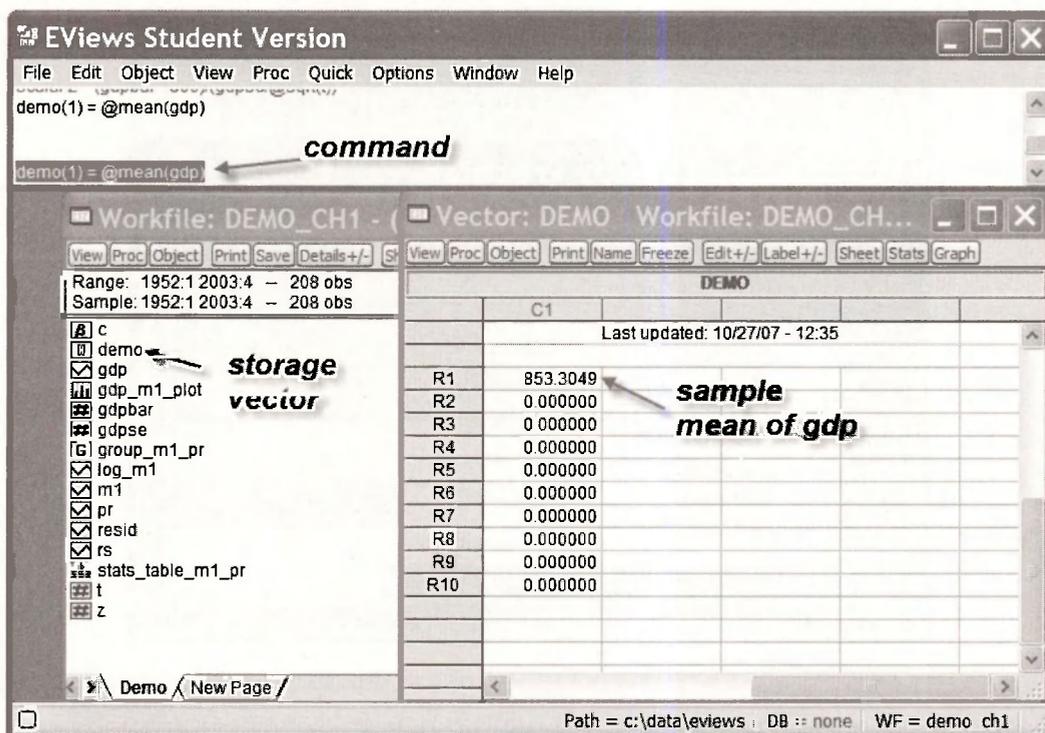
A dialog box will open asking what type of “new matrix” you want. To create a storage vector (an array) with 10 rows select the radio button **Vector**, enter 10 for Rows, and click **OK**.



A spreadsheet will open with rows labeled R1 to R10. Now enter into the Command window the command

```
demo(1) = @mean(gdp)
```

When you press **Enter** the value in row R1 will change to 853.3049, the sample mean of GDP, as shown on the next page.



Now enter the series of commands, pressing **Enter** after each.

**demo(2) = @obs(gdp)**

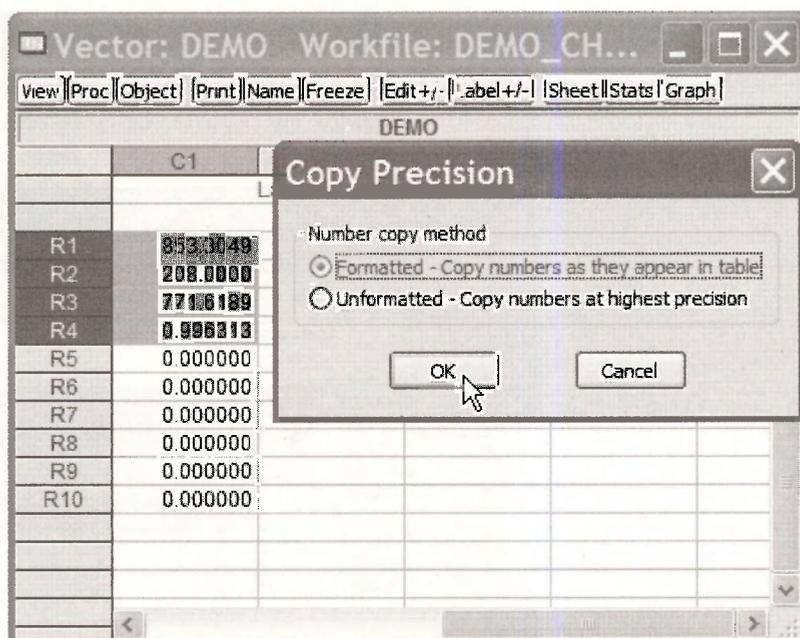
**demo(3) = @stdev(gdp)**

**demo(4) = (gdpbar - 800)/(gdpse/@sqrt(t))**

Each time a command is entered a new item shows in the vector.

DEMO	
C1	
Last updated: 10/27/07 - 12:49	
R1	853.3049
R2	208.0000
R3	771.6189
R4	0.996313

The advantage of this approach is that the contents of this table can be copied and pasted into a document for easy presentation. Highlight the contents, enter **Ctrl+C**. Choose the **Formatted** radio button and **OK**.



In an open document enter **Ctrl+V** to paste the table of results.

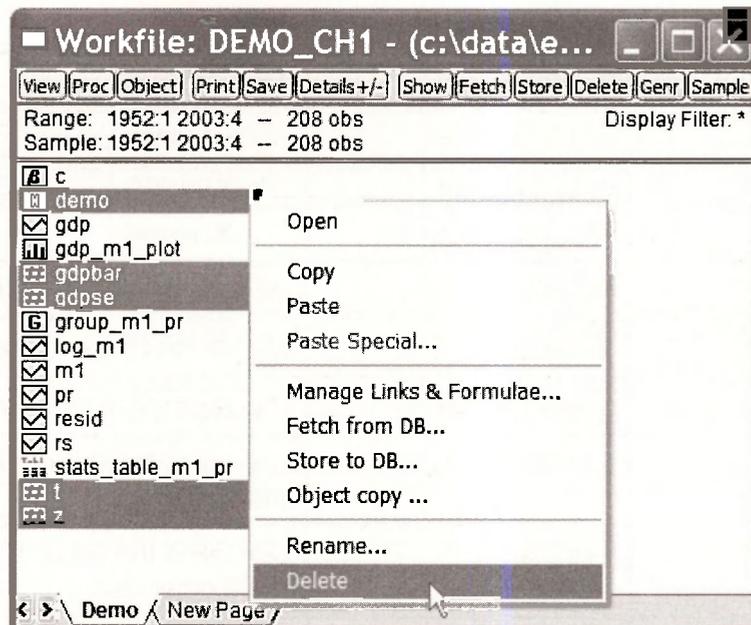
R1	853.3049
R2	208.0000
R3	771.6189
R4	0.996313

You can now edit as you would any table.

Demo vector	
GDP mean	853.3049
T sample size	208.0000
GDP Std Dev	771.6189
Z statistic	0.996313

We created many tables in the book *Principles of Econometrics* using this method.

To keep our workfile tidy, delete the scalar and vector objects that have no further use. Click the vector object **DEMO** and then while holding down the **Ctrl**-key, click on the scalars. Right-click in the blue-shaded area and select **Delete**.



If you feel confident you can choose **Yes to All**



### 1.7.3 Basic arithmetic operations

The basic arithmetic operations can be viewed at **Help/Quick Help Reference/Function Reference**

## Operator and Function Reference

This material is divided into several topics:

- Operators.
- Basic mathematical functions.

The list of operators is given on the next page. These operators can be used when working with series, such as in an operation to generate a new series, *RATIO1*, such as 3 times the ratio of *GDP* to *M1*:

$$\text{series ratio1} = 3^*(\text{gdp}/\text{m1})$$

**Basic Arithmetic Operations**

Expression	Operator	Description
+	add	$x+y$ adds the contents of X and Y.
-	subtract	$x-y$ subtracts the contents of Y from X.
*	multiply	$x*y$ multiplies the contents of X by Y.
/	divide	$x/y$ divides the contents of X by Y.
^	raise to the power	$x^y$ raises X to the power of Y.
>	greater than	$x>y$ takes the value 1 if X exceeds Y, and 0 otherwise.
<	less than	$x<y$ takes the value 1 if Y exceeds X, and 0 otherwise.
=	equal to	$x=y$ takes the value 1 if X and Y are equal, and 0 otherwise.
<>	not equal to	$x<>y$ takes the value 1 if X and Y are not equal, and 0 if they are equal.
<=	less than or equal to	$x<=y$ takes the value 1 if X does not exceed Y, and 0 otherwise.
>=	greater than or equal to	$x>=y$ takes the value 1 if Y does not exceed X, and 0 otherwise.

**1.7.4 Basic math functions**

The basic math functions can be viewed at **Help/Quick Help Reference/Function Reference**

**Operator and Function Reference**

This material is divided into several topics:

- [Operators.](#)
- [Basic mathematical functions.](#)
- [Time series functions.](#)

Some of these functions are listed below. Note that common ones like the absolute value (**abs**), the exponential function (**exp**), the natural logarithm (**log**) and the square root (**sqr**) can be used with or without the @ sign.

## Selected Basic Math Functions

Name	Function	Examples/Description
@abs(x), abs(x)	absolute value	@abs(-3)=3.
@exp(x), exp(x)	exponential, $e^x$	@exp(1)=2.71813.
@fact(x)	factorial, $x!$	@fact(3)=6, @fact(0)=1.
@inv(x)	reciprocal, $1/x$	inv(2)=0.5
@mod(x,y)	floating point remainder	returns the remainder of x/y with the same sign as x. If y=0 the result is 0.
@log(x), log(x)	natural logarithm, $\log_e(x)$	@log(2)=0.693..., log(@exp(1))=1.
@round(x)	round to the nearest integer	@round(-97.5)=-98, @round(3.5)=4.
@sqrt(x), sqr(x)	square root	@sqrt(9)=3.

## Keywords

arithmetic operators  
 basic graph  
 close series  
 copy precision  
 copying a table  
 copying graph  
 correlation  
 Ctrl+C  
 Ctrl+V  
 data definition files  
 data range  
 descriptive statistics  
 EViews functions  
 freeze  
 function reference  
 generate series  
 genr

graph metafile  
 graph options  
 group: empty  
 help  
 histogram  
 math functions  
 multiple graphs  
 name  
 object name  
 open group  
 open series  
 path  
 quick help reference  
 quick/empty group  
 quick/generate series  
 quick/graph

quick/group statistics  
 quick/sample  
 quick/series statistics  
 quick/show  
 sample range  
 sample range: change  
 scalars  
 scatter diagram  
 series  
 series: delete  
 series: rename  
 spreadsheet view  
 vectors  
 workfile: open  
 workfile: save  
 workfiles

# CHAPTER 2

## The Simple Linear Regression Model

### CHAPTER OUTLINE

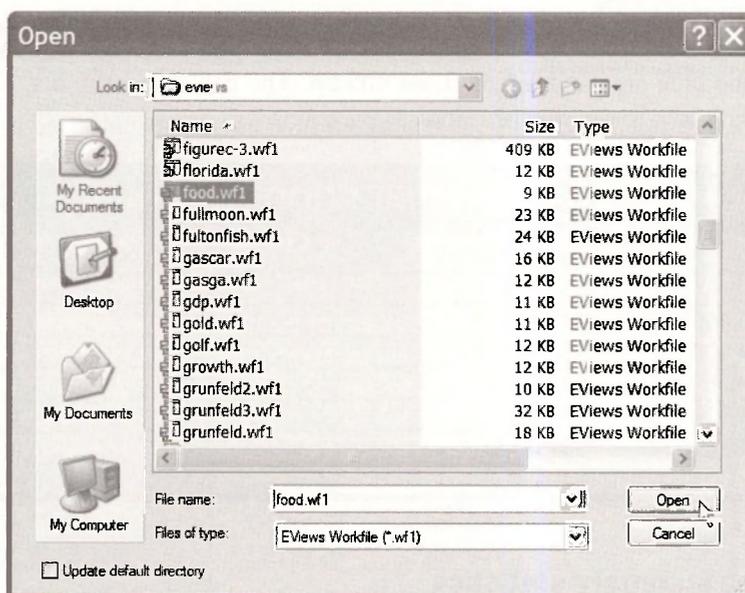
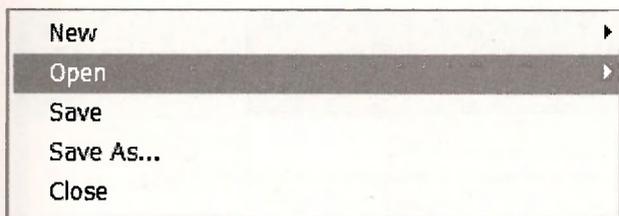
- 2.1 Open the Workfile
  - 2.1.1 Examine the data
  - 2.1.2 Checking summary statistics
  - 2.1.3 Saving a group
- 2.2 Plotting the Food Expenditure Data
  - 2.2.1 Enhancing the graph
  - 2.2.2 Saving the graph in the workfile
  - 2.2.3 Copying the graph to a document
  - 2.2.4 Saving a workfile
- 2.3 Estimating a Simple Regression
  - 2.3.1 Viewing equation representations
  - 2.3.2 Computing the income elasticity
- 2.4 Plotting a Simple Regression
- 2.5 Plotting the Least Squares Residuals
  - 2.5.1 Using View options
  - 2.5.2 Using Resids
  - 2.5.3 Using Quick/Graph
  - 2.5.4 Saving the residuals
- 2.6 Estimating the Variance of the Error Term
- 2.7 Coefficient Standard Errors
- 2.8 Prediction Using EViews
  - 2.8.1 Using direct calculation
  - 2.8.2 Forecasting

### KEYWORDS

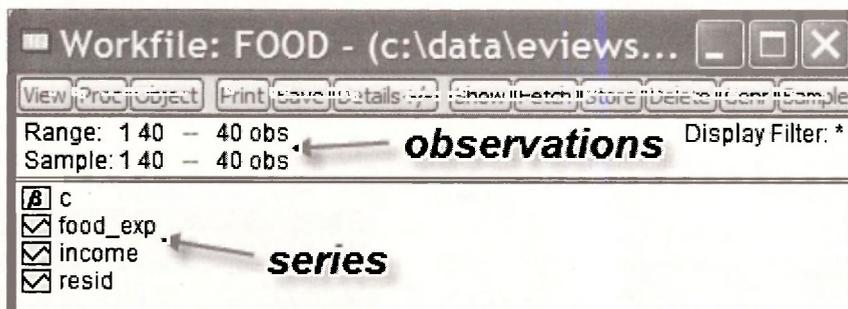
In this chapter we introduce the simple linear regression model and estimate a model of weekly food expenditure. We also demonstrate the plotting capabilities of EViews and show how to use the software to calculate the income elasticity of food expenditure, and to predict food expenditure from our regression results.

### 2.1 OPEN THE WORKFILE

The data for the food expenditure example are contained in the workfile *food.wfl*. Locate this file and open it by selecting **File/Open/EViews Workfile**

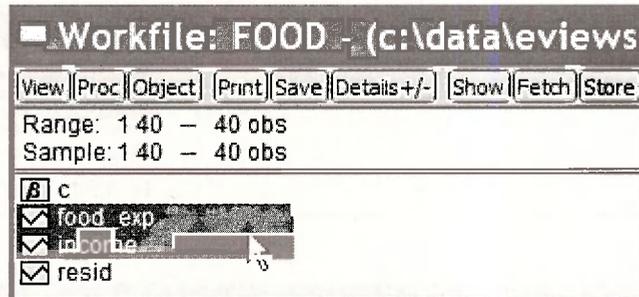


The initial workfile contains two variables *INCOME*, which is weekly household income, and *FOOD\_EXP*, which is household weekly household food expenditure. See the definition file *food.def* for the variable definitions.



### 2.1.1 Examine the data

When ever opening a new workfile it is prudent to examine the data. Select *INCOME* by clicking it, and then while holding the **Ctrl**-key select *FOOD\_EXP*.



Double-click in the blue area and select **Open Group**. The data appear in a spreadsheet format, with *INCOME* first since it was selected first.

obs	INCOME	FOOD_EXP
1	3.690000	115.2200
2	4.390000	135.9800
3	4.750000	119.3400
4	6.030000	114.9600
5	12.47000	187.0500
6		

### 2.1.2 Checking summary statistics

In the definition file *food.def* we find variable definitions and summary statistics.

```

Obs:    40

1. food_exp (y)           weekly food expenditure in $
2. income   (x)           weekly income in $100
  
```

Variable	Obs	Mean	Std. Dev.	Min	Max
food_exp	40	283.5735	112.6752	109.71	587.66
income	40	19.60475	6.847773	3.69	33.4

To verify that the workfile we are using agrees, select **View/Descriptive Stats/Common Sample**.

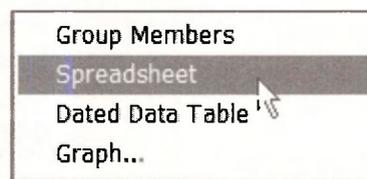


The resulting summary statistics agree with the information in the *food.def* which assures us that we have the correct data.

	INCOME	FOOD EXP
Mean	19.60475	283.5735
Median	20.03000	264.4800
Maximum	33.40000	587.6600
Minimum	3.690000	109.7100
Std. Dev.	6.847773	112.6752
Skewness	-0.626507	0.492083
Kurtosis	3.279728	2.851522
Jarque-Bera	2.747156	1.651045
Probability	0.253199	0.438006
Sum	784.1900	11342.94
Sum Sq. Dev.	1828.788	495132.2
Observations	40	40

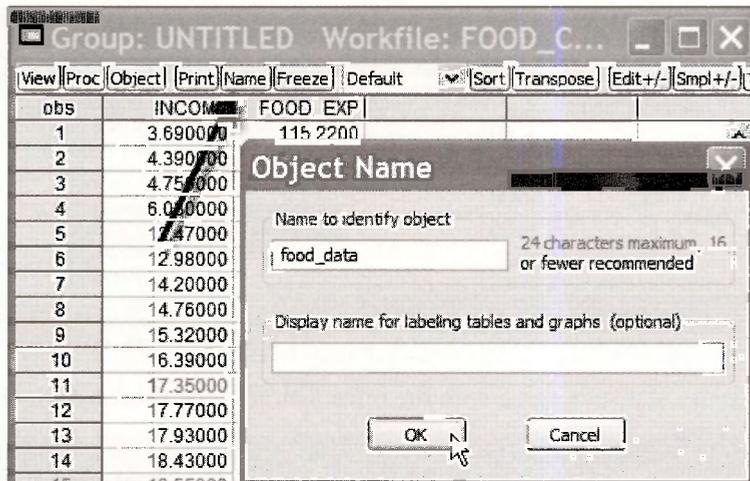
To return to the spreadsheet view, select **View/Spreadsheet**

	INCOME	FOOD_EXP
Mean	19.60475	283.5735
Median	20.03000	264.4800

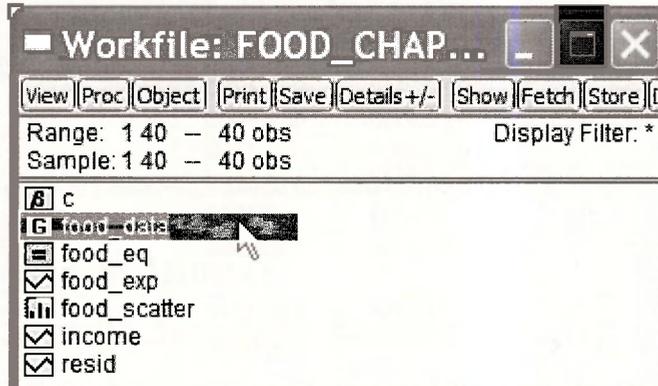


### 2.1.3 Saving a group

It is often useful to save a particular group of variables that are in a spreadsheet. From within the Group screen select **Name** and then assign an **Object Name**. Click **OK**.

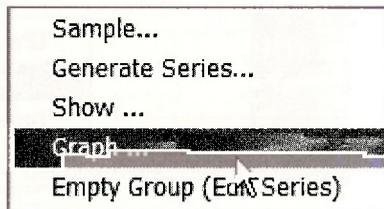


The new object in the workfile is a **Group** named **FOOD\_DATA**.



## 2.2 PLOTTING THE FOOD EXPENDITURE DATA

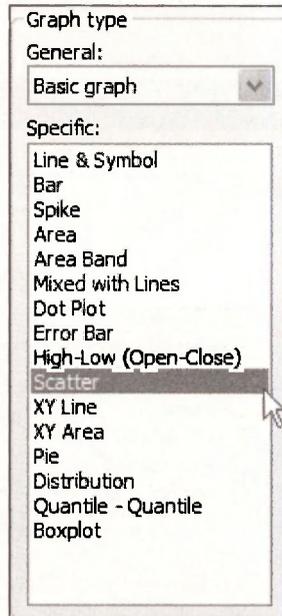
With any software there are several ways to accomplish the same task. We will make use of EViews “drop-down menus” until the basic commands become familiar. Click on **Quick/Graph**



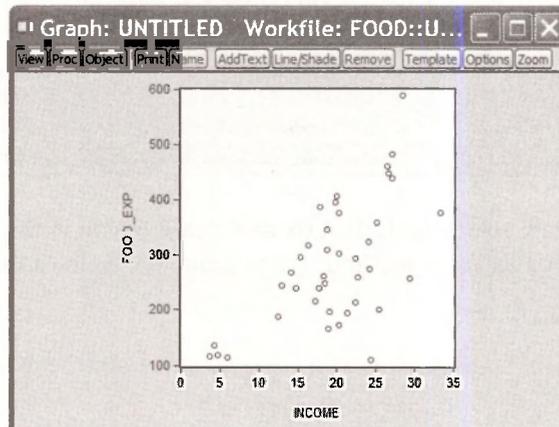
In the dialog box type the names of the variables with the x-axis variable coming first!



In the **Graph Options** box select **Scatter** from among the **Basic graphs**.



A plot appears, to which we can add labels and a title.

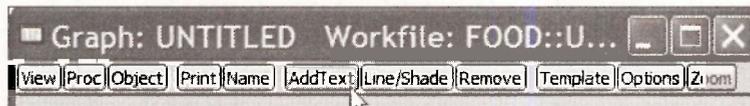


## 2.2.1 Enhancing the graph

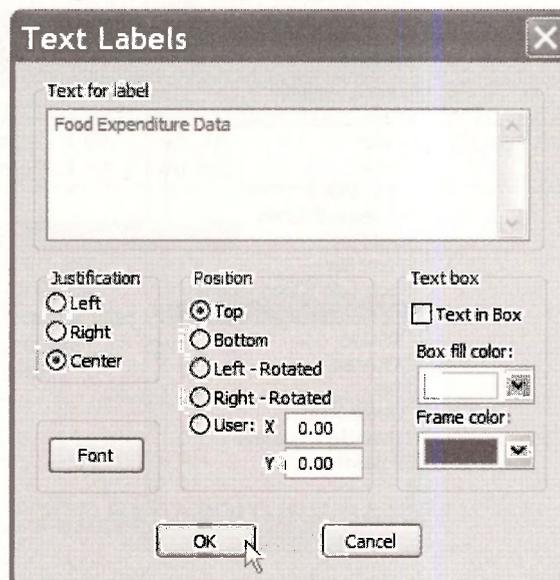
While the basic graph is fine, for a written paper or report you it can be improved by

- adding a title
- changing vertical axis scale

These tasks are easily accomplished. To add a title, click on **AddText** on the Graph menu.

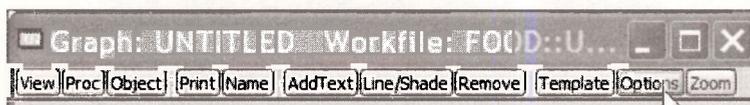


In the resulting dialog box you will be able to add a title, specify the location of the title, and use some stylistic features.

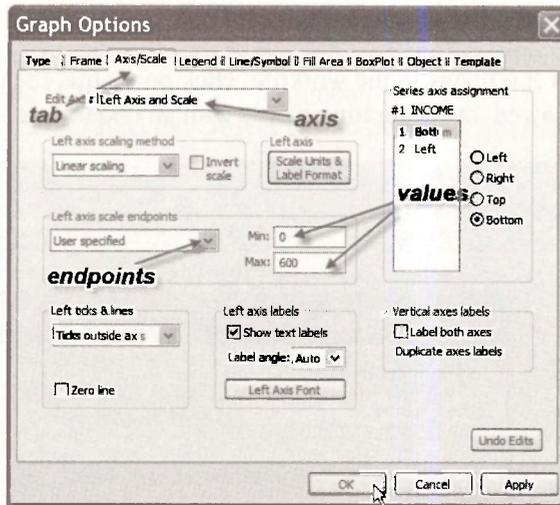


To have the title centered at the top click the appropriate options and type in the title. Click **OK**.

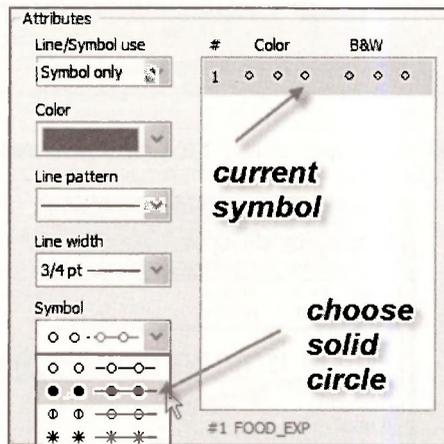
To alter the vertical axis so that it begins at zero, click on **Options** on the Graph Menu



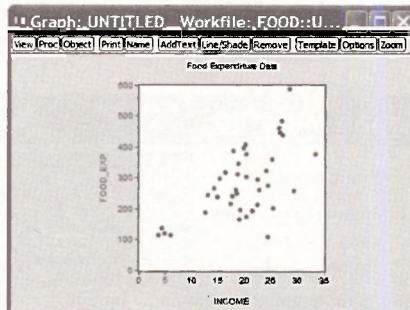
Click on the **Axes/Scale** tab, select the **Left Axis and Scale** option in the drop down box. Choose **User specified** in the **Left axis scale method**. Enter 0 and 600 as the **Min** and **Max** values. Click **OK**.



To change the “empty circles” used in the graph to “filled circles”, again choose **Options**, but select the **Line/Symbol** tab.



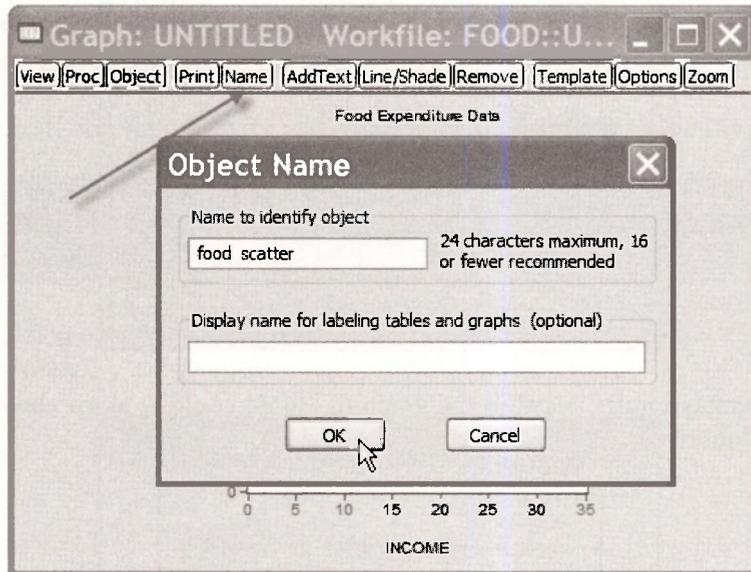
Click on **Symbol pattern** and choose the style you want. Note that other options are available. Click **OK**. The resulting graph is now



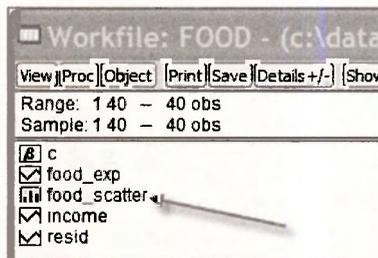
Explore the other tabs on the **Options** menu to see all the features.

### 2.2.2 Saving the graph in the workfile

To save the graph, so that it remains in the workfile, click on **Name**, then enter a name. Note that separate words are not allowed, but separating words with an underscore is an alternative.

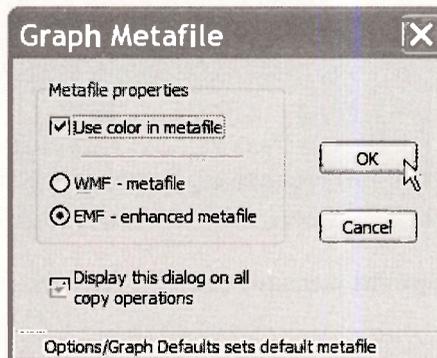


In the workfile, you will find an icon representing the graph just created



### 2.2.3 Copying the graph to a document

As is usual with Windows based applications, we can copy by clicking somewhere inside the graph, to select it, then **Ctrl+C**. Or in the main window click on **Edit/Copy**



The dialog box that shows up allows you to choose the file format. Switch to your word processor and simply paste the graph (**Ctrl+V**) into the document.

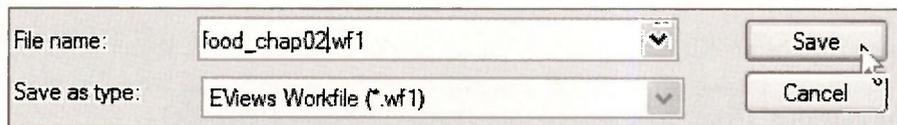
To save the graph to disk, select the **Object** button on the Graph menu.



Select **View Options/Save graph to disk**. In the resulting dialog box you have several file types to choose from, and you can select a name for the graph image.

### 2.2.4 Saving a workfile

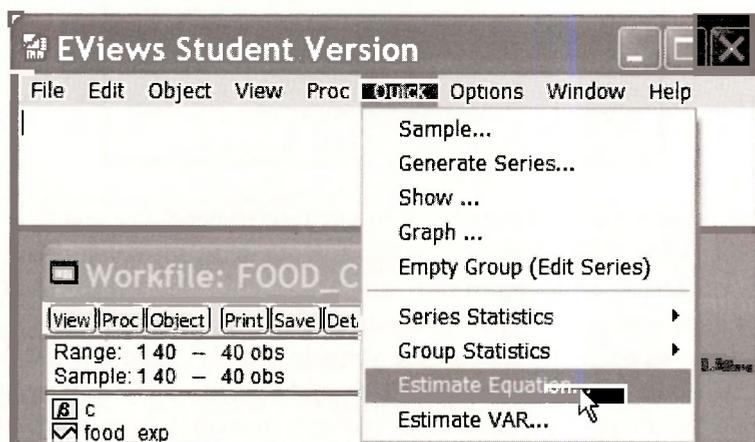
You may wish to save your workfile at this point. If you select the **Save** button on the workfile menu the workfile will be saved under its current name *food.wf1*. It might be better to save this file under a new name, so that the original workfile remains untouched. Select **File/Save As** on the EViews menu and select a simple but informative name.



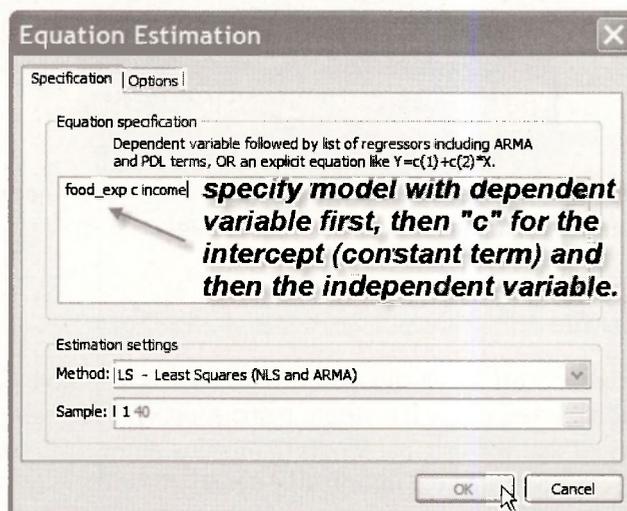
We will name it *food\_chap02.wf1*

## 2.3 ESTIMATING A SIMPLE REGRESSION

To estimate the parameters  $b_1$  and  $b_2$  of the food expenditure equation, we select **Quick/Estimate Equation** from the EViews menu.



In the **Equation Specification** dialog box, type the dependent variable *FOOD\_EXP* (the  $y$  variable) first, C (which is EViews notation for the intercept term, or constant), and then the independent variable *INCOME* (the  $x$  variable). Note in the **Estimation settings** window, the **Method** is **Least Squares** and the **Sample** is **1 40**. Click **OK**.



The estimated regression output appears. EViews produces an equation object in its default Stats view. We can name the equation object to save it permanently in our workfile by clicking on **Name** in the equation's toolbar. We have named this equation **FOOD\_EQ**.

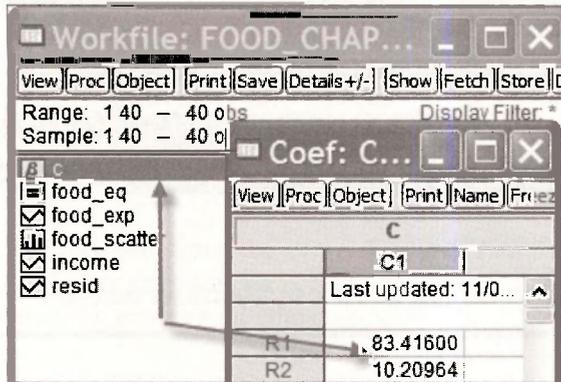
	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

R-squared	0.385002	Mean dependent var	283.5735
Adjusted R-squared	0.368818	S.D. dependent var	112.6752
S.E. of regression	89.51700	Akaike info criterion	11.87544
Sum squared resid	304505.2	Schwarz criterion	11.95988
Log likelihood	-235.5088	Hannan-Quinn criter.	11.90597
F-statistic	23.78884	Durbin-Watson stat	1.893880
Prob(F-statistic)	0.000019		

Note the estimated coefficient  $b_1$ , the intercept in our food expenditure model is recorded as the coefficient on the variable C in EViews. C is the EViews term for the constant in a regression model. Note that we cannot name any of our variables C since this term is reserved exclusively for the constant or “intercept” in a regression model. Our EViews output shows  $b_1 = 83.4160$ . The estimated value of the slope coefficient on the variable weekly income ( $X$ ) is  $b_2 = 10.2096$ , as reported in *POE*, Chapter 2.3.2. The interpretation of  $b_2$  is: for every \$100 increase in weekly income we estimate that there is about a \$10.21 increase in weekly food expenditure, holding all other factors constant.

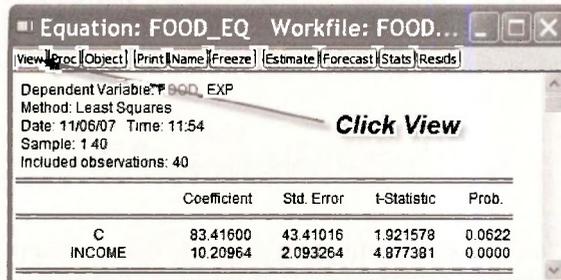
In the workfile window, double click on the vector object **C**. It always contains the estimated coefficients from the most recent regression.



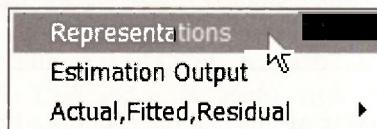
The vector **RESID** always contains the least squares residuals from the most recent regression. We will return to this shortly.

### 2.3.1 Viewing equation representations

One EViews button that we will use often is the **View** button in a regression window

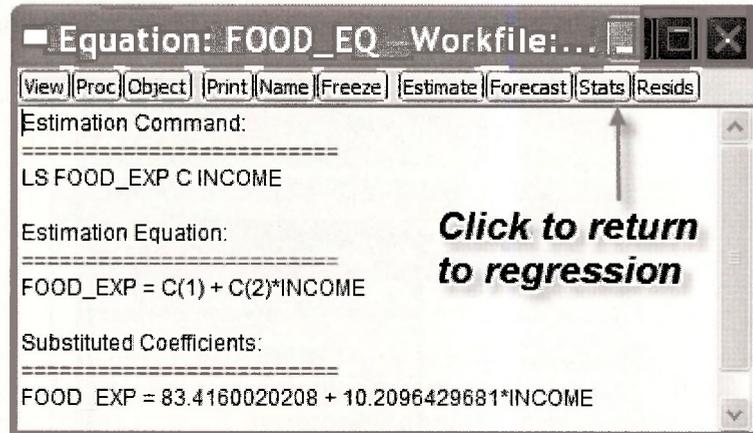


On the drop down menu list click **Representations**



The resulting display shows three things:

- The **Estimation Command** is what can be typed into the command line to obtain the equation results.
- The **Estimation Equation** that shows the coefficients and how they are linked to the variables on the equation's right side: C(1) is the intercept and C(2) is the slope
- The **Substituted Coefficients** displays the fitted regression line.



To return to the regression window click **Stats**.

### 2.3.2 Computing the income elasticity

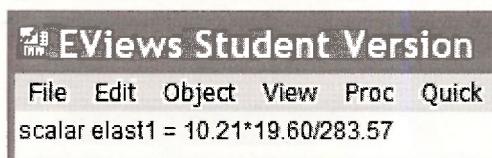
As shown in equation (2.9) of *POE* the income elasticity is defined to be

$$\varepsilon = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x} \cdot \frac{x}{E(y)} = \hat{\beta}_2 \cdot \frac{x}{E(y)}$$

which is then implemented by replacing unknowns by estimated quantities,

$$\hat{\varepsilon} = b_2 \cdot \frac{\bar{x}}{\bar{y}} = 10.21 \times \frac{19.60}{283.57} = 0.71$$

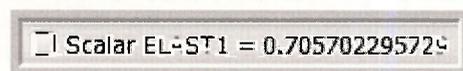
We can use EViews as a “calculator” by simply typing into the command line



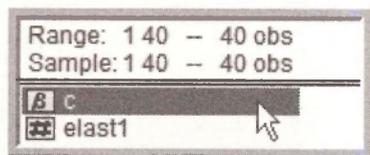
then pressing **Enter**. The word **scalar** means that the result is a single number. An icon appears in the workfile,



Double-click in the shaded area, and in the lower left corner of the EViews screen you will find the result



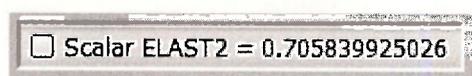
While this gives the answer, there is something to be said for using the power of EViews to simplify the calculations. EViews saves the estimates from the most recent regression in the workfile. They are obtained by double clicking the “ $\beta$ ” icon



These coefficients can be accessed from the array `@coefs`. Also, EViews has functions to compute many quantities. The arithmetic mean is computed using the function `@mean`. Thus the elasticity can also be obtained by entering into the command line

```
scalar elast2 = @coefs(2)*@mean(income)/@mean(food_exp)
```

The result is slightly different than the first computation because in the first we used “rounded off” values of the sample means.



Because the array `@coefs` is not permanent, you may want to save the slope estimate as a separate quantity by entering the commands

```
scalar b2 = @coefs(2)
scalar elast3 = b2*@mean(income)/@mean(food_exp)
```

However, the coefficient array can always be retrieved if the food equation has been saved and named. Recall that we did save it with the name `FOOD_EQ`. By saving the equation we also save the coefficients, which can be retrieved from the array `FOOD_EQ.@coefs`.

```
scalar elas = food_eq.@coefs(2)*@mean(income)/@mean(food_exp)
```

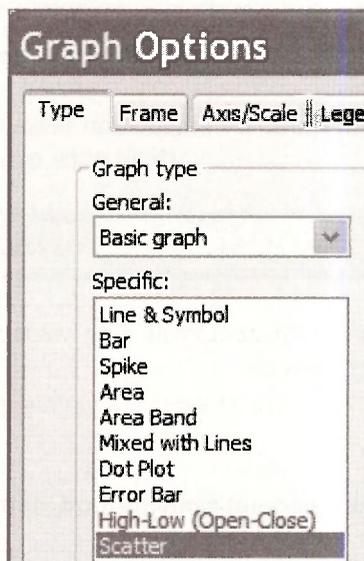
We have some surplus icons in our workfile now. Keep `B2` and `ELAS`. To clean out the other elasticities, highlight (hold down `Ctrl` and click each), right-click in the blue area, and select `Delete`. `Save` the workfile.

## 2.4 PLOTTING A SIMPLE REGRESSION

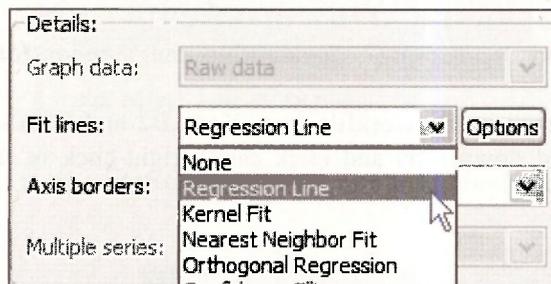
Select `Quick/Graph` from the EViews menu. In the `Series List` dialog box enter `INCOME` and `FOOD_EXP`.



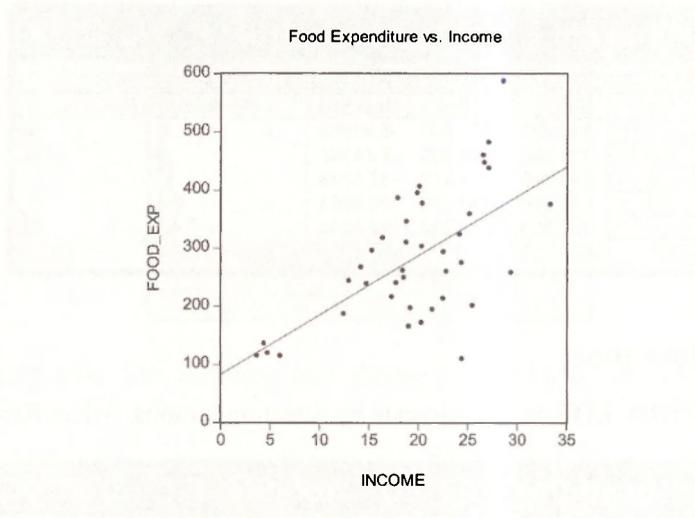
On the **Type** tab select **Scatter**



In the **Details** section, using the **Fit lines** drop down menu, select **Regression Line**.



Edit the scale of the vertical axis, choose solid circles for data points, and add a title as shown in Section 2.2.1. Click inside the graph, enter **Ctrl+C**, **OK**, and then paste into a document using **Ctrl+V**. The graph should look like this.



Return to EViews and in the Graph window select the **Name** button and assign a name to this object, such as **FITTED\_LINE**.

## 2.5 PLOTTING THE LEAST SQUARES RESIDUALS

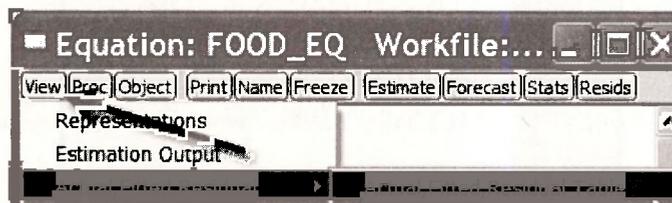
The least squares residuals are defined as

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2x_i$$

As you will discover these residuals are important for many purposes. To view the residuals open the saved regression results in **FOOD\_EQ** by double clicking the icon.

### 2.5.1 Using View options

Within the equation **FOOD\_EQ** window, click on **View** then **Actual, Fitted, Residual**. There you can select to view a table or several graphs.

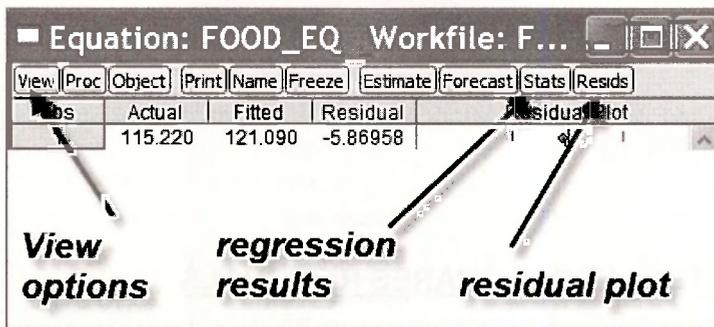


If you select **Actual, Fitted, Residual Table** you will see the values of the dependent variable  $y$ , the predicted (fitted) value of  $y$ , given by  $\hat{y} = b_1 + b_2x$  and the least squares residuals, along with a plot.

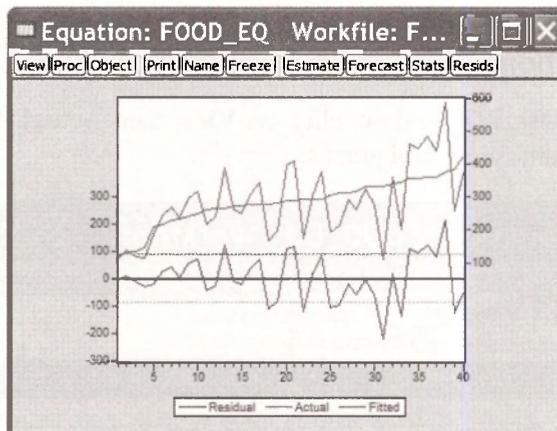
obs	Actual	Fitted	Residual	Residual Plot
1	115.220	121.090	-5.86958	
2	135.980	128.236	7.74367	
3	119.340	131.912	-12.5718	
4	114.960	144.980	-30.0201	
5	187.050	210.730	-23.6802	
6				

### 2.5.2 Using Resids plot

Within the object **FOOD\_EQ** you can navigate by selecting buttons. Select **Resids**.



The result is a plot showing the least squares residuals (lower graph) along with the actual data (*FOOD\_EXP*) and the fitted values. When using this plot note that the horizontal axis is the **observation number** and not *INCOME*. In this workfile the data happen to be sorted by income, but note that the fitted values are not a straight line. When examining residual plots, a **lack of pattern** is consistent with the assumptions of the simple regression model.

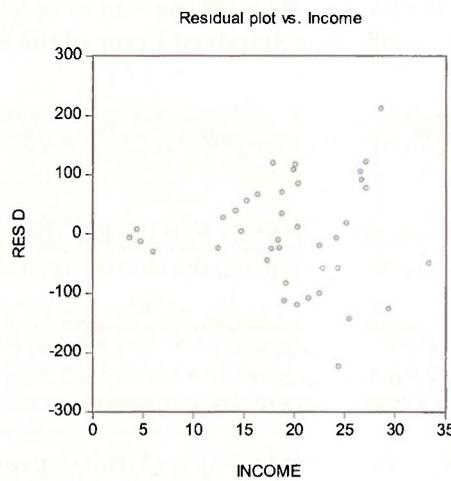


### 2.5.3 Using Quick/Graph

To create a graph of the residuals against income we can use the fact the EViews saves the residuals from the most recent regression in the series labeled **RESID**. Click on **Quick/Graph**. In the dialog box enter *INCOME* (x-axis comes first) and **RESID**.



Choose the **Scatter** plot. The resulting plot shows how the residuals relate to the values of income.



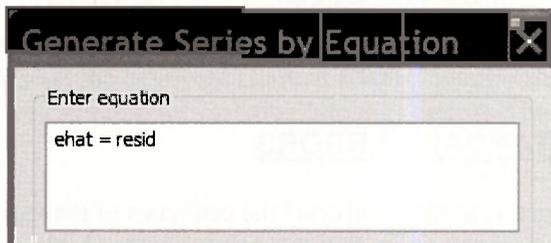
Save this plot by selecting **Name** and assigning **RESIDUAL\_PLOT**.

### 2.5.4 Saving the residuals

To save these residuals for later use, we must **Generate** a new variable (series). In the workfile screen click **Genr** on the menu.



In the resulting dialog box create a new variable called **EHAT** that contains the residuals



Click **OK**. Alternatively, simply type into the command line

**series ehat = resid**

## 2.6 ESTIMATING THE VARIANCE OF THE ERROR TERM

The estimator for  $\sigma^2$ , the variance of the error term is

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{N-2} = \frac{\text{Sum squared resid}}{N-2}$$

where **Sum squared resid** is the EViews name for the sum of squared residuals. The square root of the estimated error variance is called the **Standard Error of the Regression** by EViews,

$$\text{S.E. of regression} = \hat{\sigma} = \sqrt{\frac{\sum e_i^2}{N-2}} = \sqrt{\hat{\sigma}^2}$$

Open the regression equation we have saved as **FOOD\_EQ**. Below the estimation results you will find the Standard Error of the Regression and the sum of squared least squares residuals.

S.E. of regression	89.51700
Sum squared resid	304505.2

Also reported are the sample mean of the  $y$  values (**Mean dependent variable**)

$$\text{Mean dependent var} = \bar{y} = \frac{\sum y}{N}$$

The sample standard deviation of the  $y$  values (**S.D. dependent var**) is

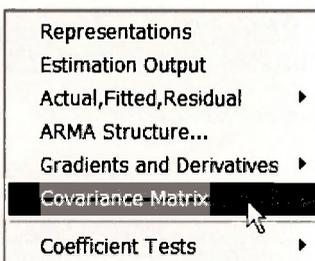
$$\text{S.D. dependent var} = s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{N-1}}$$

These are

Mean dependent var	283.5735
S.D. dependent var	112.6752

## 2.7 COEFFICIENT STANDARD ERRORS

The estimated error variance is used to construct the estimates of the variances and covariances of the least squares estimators as shown in *POE* equations (2.20)-(2.22). These estimated variances can be viewed from the **FOOD\_EQ** regression by clicking on **View/Covariance Matrix**



The elements are arrayed as

$$\begin{bmatrix} \widehat{\text{var}}(b_1) & \widehat{\text{cov}}(b_1, b_2) \\ \widehat{\text{cov}}(b_1, b_2) & \widehat{\text{var}}(b_2) \end{bmatrix}$$

In EViews they appear as

Coefficient Covariance Matrix			
	C	INCOME	
C	1884.442	-85.90316	
INCOME	-85.90316	4.381752	

The highlighted value is the estimated variance of  $b_2$ . If we take the square roots of the estimated variances, we obtain the standard errors of the estimates. In the regression output these standard errors are denoted **Std. Error** and are found right next to the estimated coefficients.

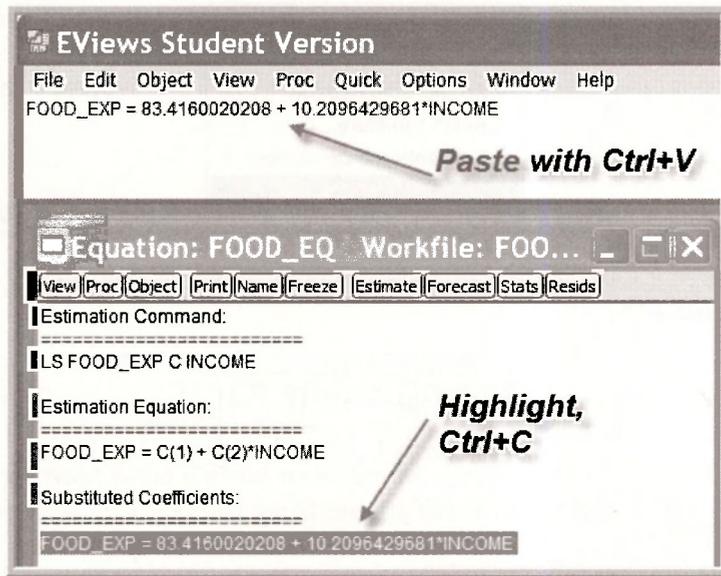
Variable	Coefficient	Std. Error
C	83.41600	43.41016
INCOME	10.20964	2.093264

## 2.8 PREDICTION USING EIEWS

There are several ways to create forecasts in EViews, and we will illustrate two of them.

### 2.8.1 Using direct calculation

Open the food equation **FOOD\_EQ**. Click on **View/Representations**. Select the text of the equation listed under **Substituted Coefficients**. We can choose **Edit/Copy** from the EViews menubar, or we can simply use the keyboard shortcut **Ctrl+C** to copy the equation representation to the clipboard. Finally, we can paste the equation into the command line.



To obtain the predicted food expenditure for a household with weekly income of \$2000, edit the command line to read

$$\text{scalar FOOD\_EXP\_HAT} = 83.4160020208 + 10.2096429681*20$$

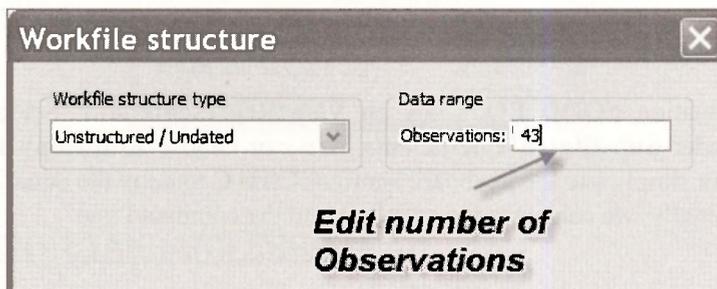
Press **Enter**. The resulting scalar value is



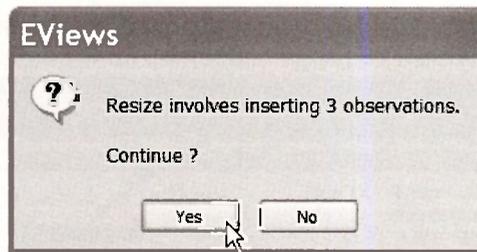
which is correct to more decimals than the value 287.61 we report in Chapter 2.3.3b.

### 2.8.2 Forecasting

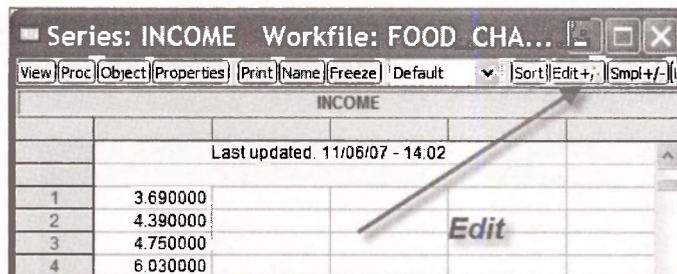
A more general, and flexible, procedure uses the power of EViews. In order to predict we must enter additional  $x$  observations at which we want predictions. In the main workfile window, double-click **Range**. This workfile has an **Unstructured/Undated** structure. Change the number of observations to 43.



Click **OK**. EViews will check with you to confirm your action.



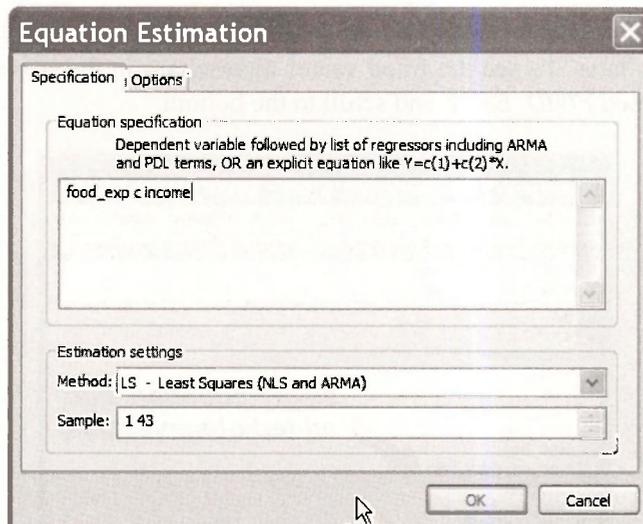
Next, double-click on *INCOME* in the main workfile to open the series, and click the **Edit+/-** button in the series window, which puts EViews in edit mode.



Scroll to the bottom and you see NA in the cells for observations 41-43. Click the cell for observation 41 and enter 20. Enter 25 and 30 in cells 42 and 43, respectively. When you are done, click the **Edit+/-** button again to turn off the edit mode.

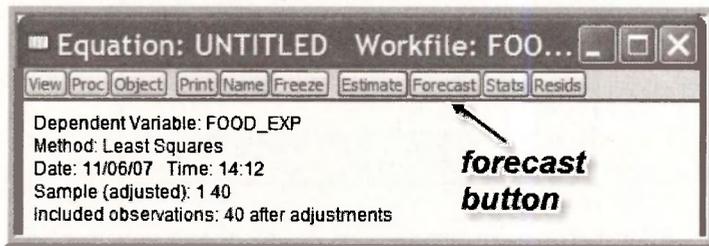
Now we have 3 extra *INCOME* observations that do not have *FOOD\_EXP* observations. When we do a regression EViews will toss out the missing observations, but it will use the extra *INCOME* values when creating a forecast.

To forecast, first re-estimate the model with the original data. This step is not actually necessary, but we want to illustrate a point. Click on **Quick/Estimate Equation**. Enter the equation. Note in the dialog box that the **Sample** is 1 to 43.

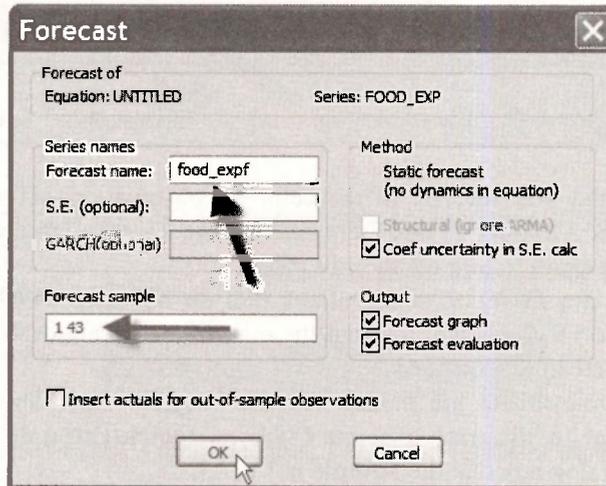


The estimation results are the same, and EViews tells us that the **Included observations** are 40 **after adjustments**. The 3 observations with not values for *FOOD\_EXP* were discarded.

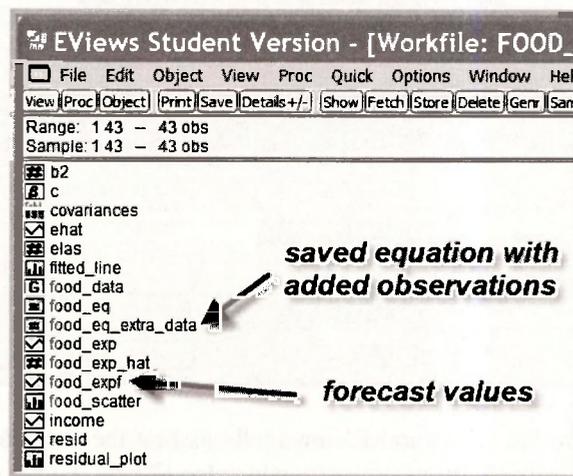
To forecast with the estimated model, click on the **Forecast** button in the equation window.



The **Forecast** dialog box appears. EViews automatically assigns the name **FOOD\_EXPF** to the forecast series, so if you want a different name enter it. The **Forecast sample** is 1 to 43. Predictions will be constructed for the 40 samples values and for the 3 new values of *INCOME*. For now, ignore the other options. Click **OK**.



A graph appears showing the fitted line for observations 41-43 along with lines labeled  $\pm 2$  S.E. We will discuss these later. To see the fitted values themselves, in the workfile window, double click on the series named *FOOD\_EXPF* and scroll to the bottom.



There you see the three forecast values corresponding to incomes 20, 25 and 30. The value in cell 41 is 287.6089, which is the same predicted value we obtained earlier in Chapter 2.3.3b.

While this approach is somewhat more laborious, by using it we can generate forecasts for many observations at once. More importantly, using EViews to forecast will make other options available to us that simple calculations will not.

## Keywords

coefficient vector	graph copy to document	residuals
covariance matrix	graph options	S.D. dependent variable
descriptive statistics	graph regression line	S.E. of regression
edit +/-	graph save	sample range
elasticity	graph symbol pattern	scalar
equation representations	graph title	scatter diagram
equation save	group: open	spreadsheet
error variance	mean dependent variable	standard errors
estimate equation	object: name	Std. Error
forecast	quick/estimate equation	sum of squared resid
generate series	quick/graph	workfile: open
genr	resid	workfile: save
graph axes/scale	residual table	

# CHAPTER 3

## Interval Estimation and Hypothesis Testing

### CHAPTER OUTLINE

- 3.1 Interval Estimation
  - 3.1.1 Constructing the interval estimate
  - 3.1.2 Using a coefficient vector
- 3.2 Right-tail Tests
  - 3.2.1 Test of significance
  - 3.2.2 Test of an economic hypothesis
- 3.3 Left-tail Tests
  - 3.3.1 Test of significance
  - 3.3.2 Test of an economic hypothesis
- 3.4 Two-tail Tests
  - 3.4.1 Test of significance
  - 3.4.2 Test of an economic hypothesis

### KEYWORDS

In this chapter we continue to work with the simple linear regression model and our model of weekly food expenditure. To begin, open the food expenditure workfile *food.wf1*. On the EViews menu choose **File/Open** and then open the file. So that the original file is not altered save this under a new name. Select **File/Save As** then name the file *food\_chap03.wf1*. Estimate the simple regression

$$FOOD\_EXP = \beta_1 + \beta_2 INCOME + e$$

The estimation can be carried out by entering into the command line

**ls food\_exp c income**

Alternatively, on the EViews menu select **Quick/Estimate Equation**, then fill in the dialog box with the equation specification and click **OK**.

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

food\_exp c income

Name the resulting regression results **FOOD\_EQ** by selecting the **Name** button and filling in the **Object name**.

Equation: FOOD\_EQ Workfile: FO...

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: FOOD\_EXP  
Method: Least Squares  
Date: 11/08/07 Time: 09:04  
Sample: 1 40  
Included observations: 40

	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

R-squared	0.385002	Mean dependent var	283.5735
Adjusted R-squared	0.368818	S.D. dependent var	112.6752
S.E. of regression	89.51700	Akaike info criterion	11.87544
Sum squared resid	304505.2	Schwarz criterion	11.95988
Log likelihood	-235.5088	Hannan-Quinn criter.	11.90597
F-statistic	23.78884	Durbin-Watson stat	1.893880
Prob(F-statistic)	0.000019		

### 3.1 INTERVAL ESTIMATION

For the regression model  $y = \beta_1 + \beta_2 x + e$ , and under assumptions SR1-SR6, the important result that we use in this chapter is given in equation (3.3) of *POE*.

$$t = \frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(N-2)} \text{ for } k = 1, 2$$

Using this result we can show that the interval  $b_k \pm t_c \text{se}(b_k)$  has probability  $1 - \alpha$  of containing the true but unknown parameter  $\beta_k$ , where the “critical value”  $t_c$  from a  $t$ -distribution such that  $P(t > t_c) = P(t < -t_c) = \alpha/2$

To construct interval estimates we will use EViews’ stored regression results. We will also make use of EViews built in statistical functions. For each distribution (see **Function reference** in EViews Help) four statistical functions are provided. The two we will make use of are the **cumulative distribution** (CDF) and the **quantile** (Inverse CDF) functions.

The for the  $t$ -distribution the CDF is given by the function **@ctdist(x,v)**. This function returns the **probability** that a  $t$ -random variable with  $v$  degrees of freedom falls to the left of  $x$ . That is,

$$@ctdist(x, v) = P[t_{(v)} < x]$$

The quantile function **@qtdist(p,v)** computes the **critical value** of a *t*-random variable with **v** degrees of freedom such that probability **p** falls to the left of it. For example, if we specify **tc=@qtdist(.975,38)**, then

$$P[t_{(38)} < tcrit] = .975.$$

To construct the interval estimates we require the least squares estimates  $b_k$  and their standard errors  $se(b_k)$ . After each regression model is estimated the coefficients and standard errors are saved in the arrays **@coefs** and **@stderrs**. However they are saved only until the next regression is run, at which time they are replaced. If you have **named** the regression results, as we have (**FOOD\_EQ**) then the coefficients are saved as well, with the names **food\_eq.@coefs** and **food\_eq.@stderrs**, respectively.

### 3.1.1 Constructing the interval estimate

Since we have estimated only one regression we can use the simple form for the saved results. Thus **@coefs(2) = b2** and **@stderrs(2) = seb2**. To generate the 95% confidence interval  $[b_2 - t_{.975}se(b_2), b_2 + t_{.975}se(b_2)]$  enter the following commands in the EViews command window, pressing the <Enter> key after each:

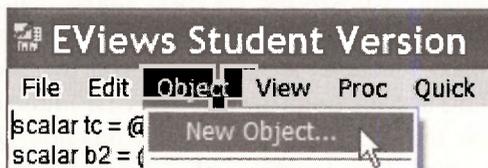
```
scalar tc = @qtdist(.975,38)
scalar b2 = @coefs(2)
scalar seb2 = @stderrs(2)
scalar b2_lb = b2 - tc*seb2
scalar b2_ub = b2 + tc*seb2
```

These scalar values show up in the workfile with the symbol #. For example, the value of the lower bound of the interval estimate is

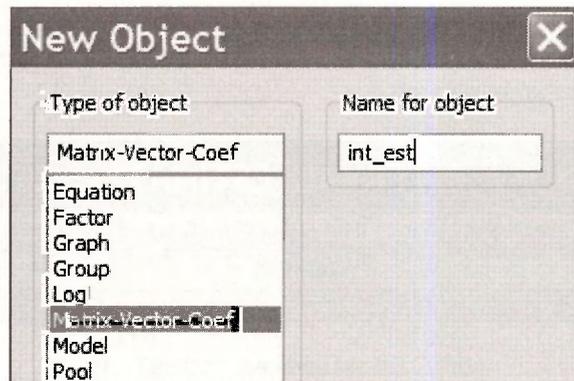
Scalar B2\_LB = 5.97205249155

### 3.1.2 Using a coefficient vector

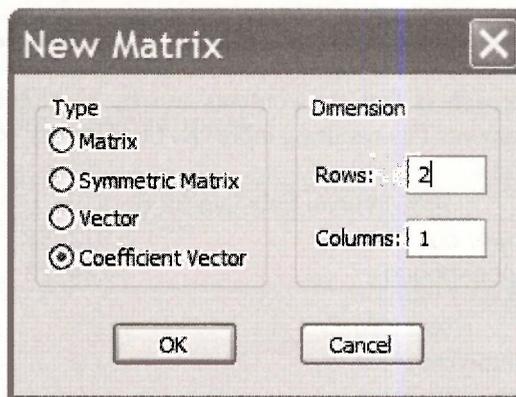
While the above approach works perfectly fine, it may be nicer for report writing to store the interval estimates in an array and construct a table. On the main EViews Menu select **Objects/New Object**



We will create a **Matrix-Vector-Coeff** named **INT\_EST**



It will be a coefficient vector that has 2 rows and 1 column



Click **OK**, and the empty array appears. Instead of all that pointing and clicking, you can simply enter on the command line

```
coef(2) int_est
```

Now, enter the commands

```
int_est(1) = @coefs(2) - @qtdist(.975,38)*@stderrs(2)
int_est(2) = @coefs(2) + @qtdist(.975,38)*@stderrs(2)
```

Here we have used the EViews saved results directly rather than create scalars for each elements. The vector we created is

INT_EST	
	Last updated: 11/08/07 - 09:28
R1	5.972052
R2	1.111723

Click on **Freeze** and then **Name**. We chose the name **B2\_INTERVAL\_ESTIMATE** and it looks like this:

INT_EST					
	A	B	C	D	E
1		C1			
2		Last updated: 11/08/07 - 09:28			
3					
4	R1	5.972052			
5	R2	14.44723			
6					
7					

The advantage of this approach is that the contents can be highlighted, copied (**Ctrl+C**) and pasted (**Ctrl+V**) into a document. The resulting table can be edited as you like

95% Interval Estimate for Beta 2	
lower bound	5.972052
upper bound	14.44723

## 3.2 RIGHT-TAIL TESTS

### 3.2.1 Test of significance

To test the null hypothesis that  $\beta_2 = 0$  against the alternative that it is positive ( $> 0$ ), as described in Chapter 3.4.1a of *POE*, requires us to find the critical value, construct the  $t$ -statistic, and determine the  $p$ -value.

- If we choose the  $\alpha = .05$  level of significance, then the critical value is the 95<sup>th</sup> percentile of the  $t_{(38)}$  distribution.
- The  $t$ -statistic is the ratio of the estimate  $b_2$  over its standard error,  $se(b_2)$ .
- The  $p$ -value is the area to the right of the calculated  $t$ -statistic (since it is a right-tail test). This value is one minus the cumulative probability to the left of the  $t$ -statistic.

The simplest set of commands is (do not type the comments in *italic font*)

```
scalar tc95 = @qtdist(.95,38)           t-critical right tail
scalar tstat = b2/seb2                 t-statistic
scalar pval = 1 - @ctdist(tstat,38)    right-tail p-value
```

Alternatively, use the vector approach outlined in the previous section

```

coef(10) t1
t1(1) = b2
t1(2) = seb2
t1(3) = b2/seb2
t1(4) = @qtdist(.95,38)
t1(5) = 1-@ctdist( t1(3),38 )

```

*storage vector*  
*estimate*  
*standard error*  
*t-statistic*  
*t-critical right tail*  
*right-tail p-value*

Use the results of this vector to construct a table, such as

Right Tail Test of Significance	
b <sub>2</sub>	10.20964
se(b <sub>2</sub> )	2.093264
t-stat	4.877381
critical value	1.685954
p-value	9.73E-06

### 3.2.2 Test of an economic hypothesis

To test the null hypothesis that  $\beta_2 < 5$  against the alternative  $\beta_2 > 5$  the same steps are executed, except for the construction of the  $t$ -statistic.

```

coef(10) t2
t2(1) = b2
t2(2) = seb2
t2(3) = (b2-5)/seb2
t2(4) = @qtdist(.95,38)
t2(5) = 1-@ctdist( t2(3),38 )

```

*storage vector*  
*estimate*  
*standard error*  
*t-statistic*  
*t-critical right tail*  
*right-tail p-value*

Yielding

Right Tail Test Beta 2 = 5	
b <sub>2</sub>	10.20964
se(b <sub>2</sub> )	2.093264
t-stat	2.488766
critical value	1.685954
p-value	0.008658

## 3.3 LEFT-TAIL TESTS

### 3.3.1 Test of significance

To test the null hypothesis that  $\beta_2 > 0$  against the alternative that it is negative ( $< 0$ ) requires us to find the critical value, construct the  $t$ -statistic, and determine the  $p$ -value.

- If we choose the  $\alpha = .05$  level of significance, then the critical value is the 5<sup>th</sup> percentile of the  $t_{(38)}$  distribution.
- The  $t$ -statistic is the ratio of the estimate  $b_2$  over its standard error,  $se(b_2)$ .
- The  $p$ -value is the area to the left of the calculated  $t$ -statistic (since it is a left-tail test). This value is given by the cumulative probability to the left of the  $t$ -statistic.

The simplest set of commands is (do not type the comments in italic font)

```

scalar tc05 = @qtdist(.05,38)           t-critical left tail
scalar tstat = b2/seb2                 t-statistic
scalar pval = @ctdist(tstat,38)       left-tail p-value

```

Alternatively

```

coef(10) t3                            storage vector
t3(1) = b2                               estimate
t3(2) = seb2                             standard error
t3(3) = b2/seb2                          t-statisitic
t3(4) = @qtdist(.05,38)                 t-critical left tail
t3(5) = @ctdist( t3(3),38 )             left-tail p-value

```

Yielding

Left Tail Test of Significance	
$b_2$	10.20964
$se(b_2)$	2.093264
t-stat	4.877381
critical value	-1.685954
p-value	0.999990

Note that we fail to reject the null hypothesis in this case, as expected.

### 3.3.2 Test of an economic hypothesis

To test the null hypothesis that  $\beta_2 > 12$  against the alternative that  $\beta_2 < 12$ , we use the same steps as above, except for the construction of the  $t$ -statistic.

```

coef(10) t4                            storage vector
t4(1) = b2                               estimate
t4(2) = seb2                             standard error
t4(3) = (b2-12)/seb2                    t-statisitic
t4(4) = @qtdist(.05,38)                 t-critical left tail
t4(5) = @ctdist( t4(3),38 )             left-tail p-value

```

Yielding,

Left Tail Test that Beta 2 > 12	
$b_2$	10.20964
$se(b_2)$	2.093264
t-stat	-0.855295
critical value	-1.685954
p-value	0.198874

The  $t$ -statistic value  $-0.85$  does not fall in the rejection region, and the  $p$ -value is about  $.20$ , thus we fail to reject this null hypothesis.

### 3.4 TWO-TAIL TESTS

#### 3.4.1 Test of significance

The two tail test of the null hypothesis that  $\beta_2 = 0$  against the alternative that  $\beta_2 \neq 0$  we require the same test elements

- If we choose the  $\alpha = .05$  level of significance, then the right-tail critical value is the 97.5-percentile of the  $t_{(38)}$  distribution and the left tail critical value is the 2.5-percentile.
- The  $t$ -statistic is the ratio of the estimate  $b_2$  over its standard error,  $se(b_2)$ .
- The  $p$ -value is the area to the left of minus the absolute value of the calculated  $t$ -statistic plus the area to the right of the absolute value of the calculated test statistic (since it is a two-tail test). This value is given by the cumulative probability to the left of the  $-|t\text{-statistic}|$  and  $1 -$  the cumulative probability to the right of  $|t\text{-statistic}|$ .

A simple set of commands is (do not type the comments in *italic* font)

```

scalar tc975 = @qtdist(.975,38)           t-critical right tail
scalar tc025 = @qtdist(.025,38)         t-critical left tail
scalar tstat = b2/seb2                   t-statistic
scalar leftpval = @ctdist(-abs(tstat),38) left-tail p-value
scalar rightpval = 1-@ctdist(abs(tstat),38) right-tail p-value
scalar pval2 = leftpval+rightpval        two tail p-value

```

The two tail  $p$ -value is

Scalar PVAL2 = 1.94586166181e-005

The test is carried out by EViews each time a regression model is estimated. If we examine **FOOD\_EQ**, in the column labeled **t-statistic** is the ratio of the **Coefficient** to **Std. Error**. The column labeled **Prob.** contains the two-tail  $p$ -value for the test of significance. Note that the very small  $p$ -value is rounded to zero (to 4 places). For practical purposes this is enough since levels of significance below  $.001$  are hardly ever used.

Dependent Variable: FOOD_EXP				
Method: Least Squares				
Date: 11/08/07 Time: 09:04				
Sample: 1 40				
Included observations: 40				
	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

To use the coefficient vector approach

<b>coef(10) t5</b>	<i>storage vector</i>
<b>t5(1) = b2</b>	<i>estimate</i>
<b>t5(2) = seb2</b>	<i>standard error</i>
<b>t5(3) = b2/seb2</b>	<i>t-statistic</i>
<b>t5(4) = @qtdist(.025,38)</b>	<i>t-critical left tail</i>
<b>t5(5) = @qtdist(.975,38)</b>	<i>t-critical right tail</i>
<b>t5(6) = @ctdist(-abs(t5(3)),38)</b>	<i>left tail p-value portion</i>
<b>t5(7) = 1 - @ctdist(abs(t5(3)),38)</b>	<i>right tail p-value portion</i>
<b>t5(8) = t5(6) + t5(7)</b>	<i>two tail p-value</i>

The result is as follows. Here we have copied the results from EViews at the highest precision to show that the  $p$ -value works out to be the same as reported above.

Two Tail Test that Beta 2 = 0	
b <sub>2</sub>	10.2096429681
se(b <sub>2</sub> )	2.09326353144
t-stat	4.87738061395
left critical value	-2.02439416391
right critical value	2.02439416391
left portion p-value	9.72930830907e-06
right portion p-value	9.72930830911e-06
two tail p-value	1.94586166182e-05

### 3.4.2 Test of an economic hypothesis

To test the null hypothesis that  $\beta_2 = 12.5$  against the alternative  $\beta_2 \neq 12.5$  the steps are the same as those above, except for the construction of the  $t$ -statistic.

<code>coef(10) t6</code>	<i>storage vector</i>
<code>t6(1) = b2</code>	<i>estimate</i>
<code>t6(2) = seb2</code>	<i>standard error.</i>
<code>t6(3) = (b2-12.5)/seb2</code>	<i>t-statistic</i>
<code>t6(4) = @qtdist(.025,38)</code>	<i>t-critical left tail</i>
<code>t6(5) = @qtdist(.975,38)</code>	<i>t-critical right tail</i>
<code>t6(6) = @ctdist(-abs(t6(3)),38)</code>	<i>left tail p-value portion</i>
<code>t6(7) = 1 - @ctdist(abs(t6(3)),38)</code>	<i>right tail p-value portion</i>
<code>t6(8) = t6(6) + t6(7)</code>	<i>two tail p-value</i>

Which yields

Two Tail Test that Beta 2 = 12.5	
$b_2$	10.20964
$se(b_2)$	2.093264
t-stat	-1.094156
left critical value	-2.024394
right critical value	2.024394
left portion p-value	0.140387
right portion p-value	0.140387
two tail p-value	0.280774

### Keywords

@coefs	critical value	Prob.
@ctdist	hypothesis test	p-value
@qtdist	hypothesis test: left-tail	scalar
@stderrs	hypothesis test: one-tail	significance test
abs	hypothesis test: right-tail	t-distribution CDF
absolute value	hypothesis test: two-tail	t-distribution critical value
coefficient vector	interval estimation	t-statistic

# CHAPTER 4

## Prediction, Goodness-of-Fit, and Modeling Issues

### CHAPTER OUTLINE

- 4.1 Prediction in the Food Expenditure Model
  - 4.1.1 A simple prediction procedure
  - 4.1.2 Prediction using EViews
- 4.2 Measuring Goodness-of-Fit
  - 4.2.1 Calculating  $R^2$
  - 4.2.2 Correlation analysis
- 4.3 Modeling Issues
  - 4.3.1 The effects of scaling the data
  - 4.3.2 The log-linear model
  - 4.3.3 The linear-log model
  - 4.3.4 The log-log model
  - 4.3.5 Are the regression errors normally distributed?
  - 4.3.6 Another example
- 4.4 The Log-Linear Model
  - 4.4.1 Prediction in the log-linear model
  - 4.4.2 Alternative commands in the log-linear model
  - 4.4.3 Generalized  $R^2$

KEYWORDS

### 4.1 PREDICTION IN THE FOOD EXPENDITURE MODEL

In this chapter we continue to work with the simple linear regression model and our model of weekly food expenditure. To begin, open the food expenditure workfile *food.wfl*. On the EViews menu choose **File/Open** and then open the file. So that the original file is not altered save this under a new name. Select **File/Save As** then name the file *food\_chap04.wfl*. Estimate the simple regression

$$FOOD\_EXP = \beta_1 + \beta_2 INCOME + e$$

The estimation can be carried out by entering into the command line

**ls food\_exp c income**

Alternatively, on the EViews menu select **Quick/Estimate Equation**, then fill in the dialog box with the equation specification and click **OK**.

Name the resulting regression results **FOOD\_EQ** by selecting the **Name** button and filling in the **Object name**.

#### 4.1.1 A simple prediction procedure

In Chapter 2.8 of this manual we illustrated a simple procedure for obtaining the predicted value of food expenditure for a household with income of \$2,000 per week. We also showed that EViews can be used to generate forecasts automatically, both the for sample values and for new *INCOME* observations that we append to the workfile by increasing its range. If you need to review those steps do so now.

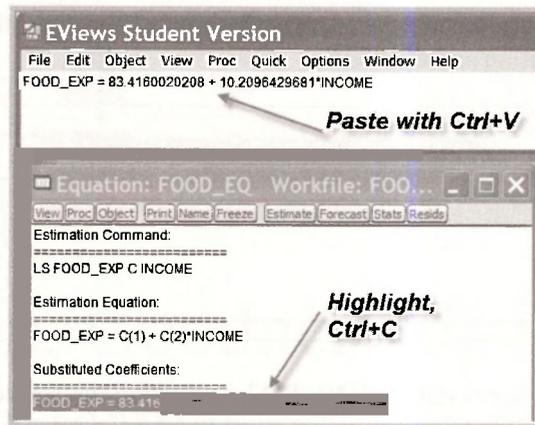
What we can add now that we did not have before is the standard error of the forecasted value. The estimated variance of the forecast error is

$$\widehat{\text{var}}(f) = \hat{\sigma}^2 \left[ 1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

A convenient form for calculation in the simple regression model is

$$\widehat{\text{var}}(f) = \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2 \widehat{\text{var}}(b_2)$$

Open the food equation **FOOD\_EQ**. Click on **View/Representations**. Select the text of the equation listed under **Substituted Coefficients**. We can choose **Edit/Copy** from the EViews menubar, or we can simply use the keyboard shortcut **Ctrl+C** to copy the equation representation to the clipboard. Finally, we can paste the equation into the command line.



To obtain the predicted food expenditure for a household with weekly income of \$2000, edit the command line to read

```
scalar food_exp_hat = 83.4160020208 + 10.2096429681*20
```

Press **Enter**. The resulting scalar value is

```
 Scalar FOOD_EXP HAT = 287.608861383
```

which is correct to more decimals than the value 287.61.

The prediction interval requires the critical value from the  $t_{(38)}$  distribution. For a 95% prediction interval the required critical value is  $t_c$  is the 97.5-percentile, which is obtained as

```
scalar tc = @qtdist(.975,38)
```

The prediction interval is obtained by entering the following commands (do not type the comments in *italic font*).

<b>scalar sig2 = (@se)^2</b>	<i>@se = std error of regression</i>
<b>scalar n = @regobs</b>	<i>@regobs = N</i>
<b>scalar varb2 = (@stderrs(2))^2</b>	<i>@stderrs = std. errors of b</i>
<b>scalar xbar = @mean(income)</b>	<i>@mean = sample mean</i>
<b>scalar varf = sig2 + sig2/n + ((20-xbar)^2)*varb2</b>	<i>^2 raises to power 2</i>
<b>scalar sef = @sqrt(varf)</b>	<i>@sqrt = square root</i>
<b>scalar yhat_lb = food_exp_hat - tc*sef</b>	<i>lower bound of interval</i>
<b>scalar yhat_ub = food_exp_hat + tc*sef</b>	<i>upper bound of interval</i>
<b>show yhat_lb</b>	<i>show lower bound of interval</i>
<b>show yhat_ub</b>	<i>show upper bound of interval</i>

The resulting prediction interval values are:

```
 Scalar YHAT_LB = 104.132276898
```

```
 Scalar YHAT_UB = 471.085445868
```

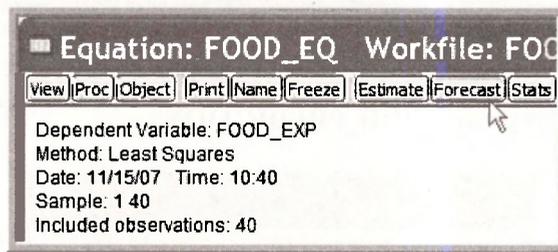
These results are correct, but obtaining a prediction interval this way each time would be tedious. Now we use the power of EViews.

#### 4.1.2 Prediction using EViews

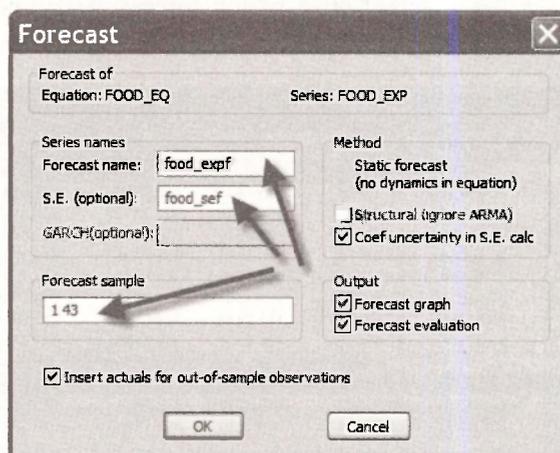
The above procedure for computing a prediction interval works for the simple regression model. EViews makes computing standard errors of forecasts simple. In Section 2.8.2 of this manual we extended the range of the workfile and entered 3 new observations for *INCOME* = 20, 25 and 30. Follow those same steps again to insert the same three observation values. The steps are:

- Double click on **Range** in the main workfile window.
- Change the number of observations to 43 and click **OK**.
- Double click on *INCOME* in the main workfile to open the series.
- Click **Edit+/-** to open edit mode.
- Enter 20, 25 and 30 in cells 41-43.
- Click **Edit+/-** to close edit mode.

In the **FOOD\_EQ** window click on **Forecast**



In the dialog box that opens enter names for the **Forecast** and the **S.E.**, which is for the standard error of the forecast. Make sure the forecast sample is set to 1-43 and click **OK**.

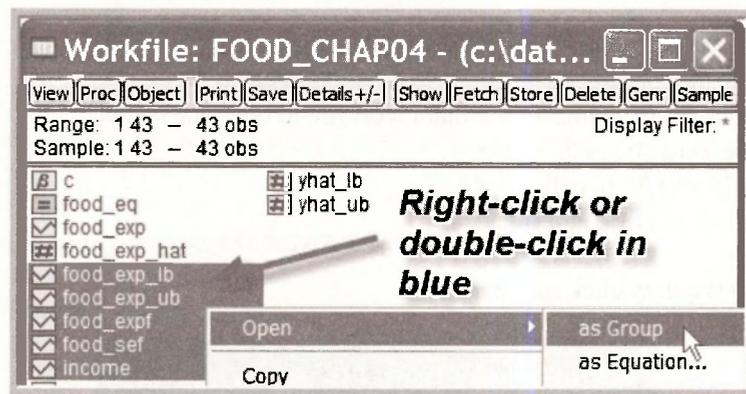


The resulting window shows the predicted values and a 95% prediction interval for the observations in their given order. For cross sectional observations this is not so useful. We will come back to it later.

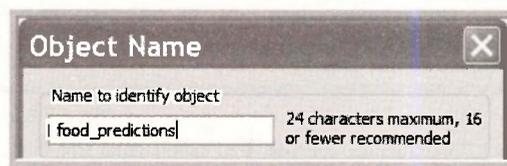
Enter the following commands into the command line, or use the **Genr** button to open a dialog box in which the new series can be defined.

```
series food_exp_ub = food_expf + tc*food_sef
series food_exp_lb = food_expf - tc*food_sef
```

Create a group by clicking on *INCOME*, *FOOD\_SEF*, *FOOD\_EXP\_LB*, *FOOD\_EXPF*, *FOOD\_EXP\_UB*. To do this, click each one while holding down **Ctrl**. Right click in the shaded area and select **Open/ as Group**.



Click on **Name** and call this group **FOOD\_PREDICTIONS**.



Scroll to the bottom to see the standard error of the forecast and prediction intervals for the specified values of income. Note that the value for = 20 are as we constructed manually.

The screenshot shows the EViews Group: FOOD\_PREDICTIONS window. The table displays data for observations 32 through 43. The columns are labeled 'obs', 'INCOME', 'FOOD\_SEF', 'FOOD\_EXP\_LB', 'FOOD\_EXPF', and 'FOOD\_EXP\_UB'. The data for observation 41 is circled in red.

obs	INCOME	FOOD_SEF	FOOD_EXP_LB	FOOD_EXPF	FOOD_EXP_UB
32	25.20000	91.38274	155.7043	340.6990	525.6937
33	25.50000	91.46535	158.6000	343.7619	528.9238
34	26.61000	91.80770	169.2396	355.0946	540.9496
35	26.70000	91.83798	170.0972	356.0135	541.9297
36	27.14000	91.99143	174.2788	360.5057	546.7326
37	27.16000	91.99861	174.4685	360.7099	546.9514
38	28.62000	92.57296	188.2118	375.6160	563.0201
39	29.40000	92.91955	195.4737	383.5795	571.6853
40	33.40000	93.11642	231.8008	424.4161	516.9759
41	20.00000	90.63284	104.1323	287.6089	471.0954
42	25.70000	93.33003	153.7591	338.6571	523.5459
43	30.00000	93.20474	201.0222	389.7053	578.3884

## 4.2 MEASURING GOODNESS-OF-FIT

### 4.2.1 Calculating $R^2$

The usual  $R^2 = 1 - SSE/SST$  is reported in the EViews regression output. In the **FOOD\_EQ** window it is reported just below the regression coefficients

	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000
R-squared	0.385002	Mean dependent var		283.5735
Adjusted R-squared	0.368818	S.D. dependent var		112.6752
S.E. of regression	89.51700	Akaike info criterion		11.87544
Sum squared resid	304505.2	Schwarz criterion		11.95988

The elements required to compute it in this window are shown as well. The sum of squared least squares residuals ( $SSE$ ) is given by

$$\text{Sum squared resid} \quad 304505.2$$

The total sum of squares ( $SST$ ) can be obtained from

$$\text{S.D. dependent var} \quad 112.6752$$

Recall that the sample standard deviation of the  $y$  values (**S.D. dependent var**) is

$$\text{S.D. dependent var} = s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N - 1}}$$

Thus if we square this value, and multiply by  $N - 1$  we will have it. That is

$$\sum (y_i - \bar{y})^2 = (N - 1)s_y^2$$

You can do this by hand, or recall that after a regression model is estimated many useful items are saved by EViews, including

**@sddep**

*standard deviation of the dependent variable*

**@ssr**

*sum of squared residuals*

Then, to calculate  $R^2$  use the commands

```
scalar sst = (N-1)*(@sddep)^2
scalar r2 = 1-@ssr/sst
```

The value  $N$  had already been calculated as

```
scalar n = @regobs
```

### 4.2.2 Correlation analysis

In the simple regression model we can compute  $R^2$  as the square of the correlation between  $X$  and  $Y$  or the square of the correlation between  $Y$  and its predicted values. The EViews function `@cor` computes the correlation between two variables.

```
scalar r2_xy=(@cor(income,food_exp))^2
scalar r2_yyhat = (@cor(food_exp, food_expf))^2
```

**NOTE:** These calculations were carried out with the **Range** and **Sample** each set to 1 to 43 from our work with prediction in Section 4.1.1 above. However, because there is no value (NA) for *FOOD\_EXP* for observations 41 to 43, EViews discarded observations 41-43 in the calculations.

## 4.3 MODELING ISSUES

### 4.3.1 The effects of scaling the data

Changing the scale of variables in EViews is very simple. **Generate** new variables that have been redefined to suit you. To illustrate, suppose we measure *INCOME* in \$ rather than in 100\$ increments. That is, we prefer the variable  $DOLLAR\_INC = 100*INCOME$ . Create this new variable by clicking the **Genr** button, then enter

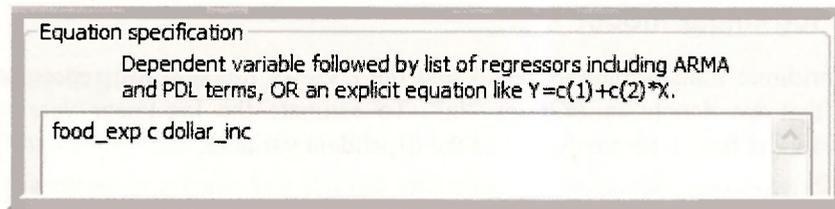
Enter equation

dollar\_inc=100\*income

and click **OK**. Alternatively, on the command line, enter

```
series dollar_inc=100*income
```

Estimate the food expenditure model using this new variable. Click **Quick/Estimate Equation** and enter



Click **OK** Alternatively, on the command line enter

**ls food\_exp c dollar\_inc**

The result is

Dependent Variable: FOOD\_EXP

	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
DOLLAR_INC	0.102096	0.020933	4.877381	0.0000
R-squared	0.385002	Mean dependent var		283.5735
Adjusted R-squared	0.368818	S.D. dependent var		112.6752
S.E. of regression	89.51700	Akaike info criterion		11.87544
Sum squared resid	304505.2	Schwarz criterion		11.95988
Log likelihood	-235.5088	Hannan-Quinn criter.		11.90597
F-statistic	23.78884	Durbin-Watson stat		1.893880
Prob(F-statistic)	0.000019			

The coefficient on income has changed, as has its standard error. Everything else in this regression is the same as earlier estimations of the food expenditure equation.

A useful feature of EViews is that the regression commands allow variables to be transformed directly. That is we could obtain the same results by entering

**ls food\_exp c (100\*income)**

The regression coefficients are now

Dependent Variable: FOOD\_EXP

	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
100*INCOME	0.102096	0.020933	4.877381	0.0000

### 4.3.2 The log-linear model

To use logarithmic transformations recall that the EViews function **log** represents the “natural logarithm” that we denote in *POE* as “ln”. To estimate the log-linear version of the food expenditure model first generate the log of the dependent variable,

**series lfood\_exp = log(food\_exp)**

Recall that transforming the dependent variable in this way fundamentally changes how the results are interpreted. In the equation  $\ln(y) = \beta_1 + \beta_2 x + e$ , a 1-unit increase in “*x*” leads to a  $100\beta_2\%$  increase in the expected value of *y*.

Now, use this dependent variable in the regression model,

**ls lfood\_exp c income**

The results are

Dependent Variable: LFOOD_EXP				
	Coefficient	Std. Error	t-Statistic	Prob.
C	4.780239	0.158959	30.07210	0.0000
INCOME	0.040030	0.007665	5.222377	0.0000
R-squared	0.417832	Mean dependent var		5.565019
Adjusted R-squared	0.402511	S.D. dependent var		0.424068
S.E. of regression	0.327793	Akaike info criterion		0.655839
Sum squared resid	4.083038	Schwarz criterion		0.740283
Log likelihood	-11.11678	Hannan-Quinn criter.		0.686371
F-statistic	27.27322	Durbin-Watson stat		1.877139
Prob(F-statistic)	0.000007			

We would interpret this by saying that an increase in income of \$100 (1-unit) leads to about a 4% increase in food expenditure. Because we have transformed the dependent variable in this way, the  $R^2$  changes and is not comparable to earlier estimations. More on this later.

Instead of creating **lfood\_exp = log(food\_exp)** we could have specified the transformation directly in the regression statement, as

**ls log(food\_exp) c income**

### 4.3.3 The linear-log model

The linear-log model transforms the  $x$  variable, but not the  $y$  variable:  $y = \beta_1 + \beta_2 \ln(x) + e$ . In this model a 1% increase in  $x$  leads to a  $\beta_2/100$  unit change in  $y$ . In the food expenditure model the commands are

```
series lincome = log(income)
ls food_exp c lincome
```

The resulting regression output is:

Dependent Variable: FOOD_EXP				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-97.18642	84.23744	-1.153720	0.2558
LINCOME	132.1658	28.80461	4.588357	0.0000
R-squared	0.356510	Mean dependent var		283.5735
Adjusted R-squared	0.339577	S.D. dependent var		112.6752
S.E. of regression	91.56711	Akaike info criterion		11.92073
Sum squared resid	318612.4	Schwarz criterion		12.00517
Log likelihood	-236.4146	Hannan-Quinn criter.		11.95126
F-statistic	21.05302	Durbin-Watson stat		1.836580
Prob(F-statistic)	0.000048			

We would interpret the results by saying that a 1% increase in income leads to about a \$1.32 increase in weekly food expenditure.

The linear-log model can be estimated directly as

```
ls food_exp c log(income)
```

### 4.3.4 The log-log model

In the log-log model  $\ln(y) = \beta_1 + \beta_2 \ln(x) + e$  the parameter  $\beta_2$  is an elasticity. For the food expenditure model, using the log-variables we have already created, the regression command is

```
ls lfood_exp c lincome
```

The result is shown on the next page. A 1% increase in income leads to about a 1/2% increase in food expenditure. Alternatively use the regression command

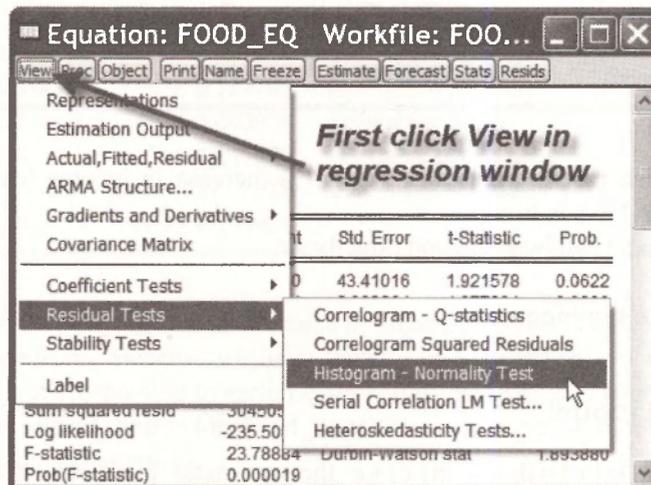
```
ls log(food_exp) c log(income)
```

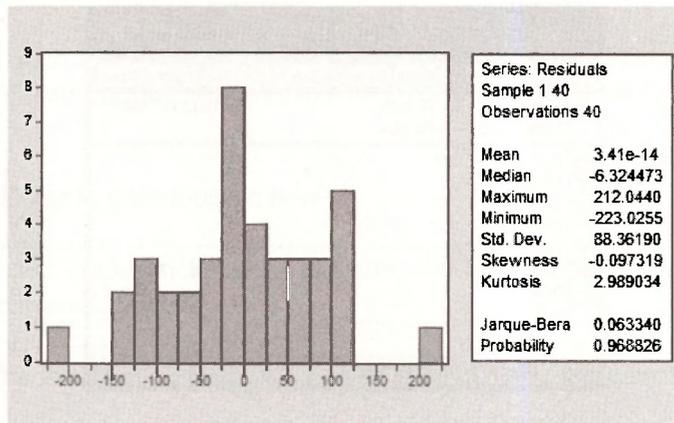
Dependent Variable: LFOOD\_EXP

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.963567	0.294373	13.46444	0.0000
LINCOME	0.555881	0.100660	5.522391	0.0000
R-squared	0.445230	Mean dependent var		5.565019
Adjusted R-squared	0.430630	S.D. dependent var		0.424068
S.E. of regression	0.319987	Akaike info criterion		0.607634
Sum squared resid	3.890883	Schwarz criterion		0.692078
Log likelihood	-10.15268	Hannan-Quinn criter.		0.638166
F-statistic	30.49680	Durbin-Watson stat		1.982420
Prob(F-statistic)	0.000003			

### 4.3.5 Are the regression errors normally distributed?

Each time a regression is estimated a certain number of regression diagnostics should be carried out. It is through the residuals of the fitted model that we may detect problems in a model's specification. One aspect of the error that we can examine is whether the errors appear normally distributed. EViews reports diagnostics for the residuals each time a model is estimated. For example, in the **FOOD\_EQ** window, select **View/Residual Tests/Histogram-Normality Test**





A **Histogram** is produced along with other summary measures. The **Mean** of the residuals is always zero for a regression that includes an intercept term. In the histogram we are looking for a general “Bell-shape”, and a value of the **Jarque-Bera** test statistic with a large  $p$ -value. This test is valid in large samples, so what it tells us in a sample of size  $N = 40$  is questionable. The test statistic has a  $\chi^2_{(2)}$  distribution under the null hypothesis that the **Skewness** is zero and **Kurtosis** is three, which are the measures for a normal distribution. The critical value for the chi-square distribution is obtained by typing into the command line

```
=@qchisq(.95,2)
```

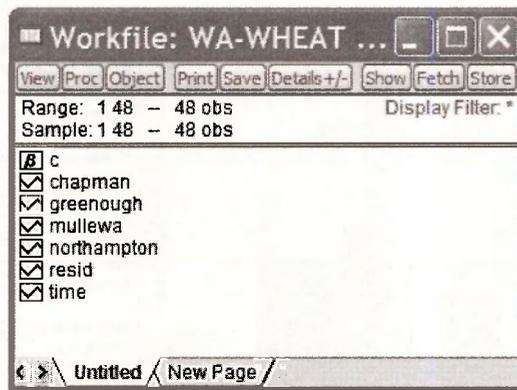
which produces the scalar value

```
Scalar = 5.99146454711
```

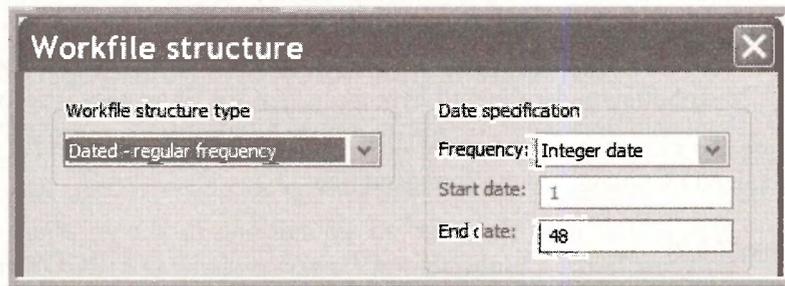
Save your workfile and close it, as we are moving to another example.

#### 4.3.6 Another example

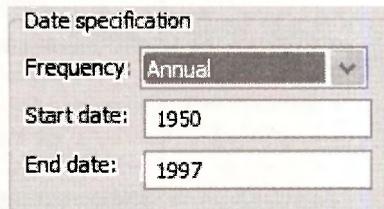
Open the workfile *wa-wheat.wfl* by selecting on the EViews menu **File/Open/EViews Workfile**. Locate *wa-wheat.wfl* and click **OK**. It contains 48 annual observations on the variables *NORTHAMPTON*, *CHAPMAN*, *MULLEWA*, *GREENOUGH*, and *TIME*. The first 4 variables are average annual wheat yields in shires of Western Australia. See the definition file *wa-wheat.def*. These are annual data from 1950 to 1997



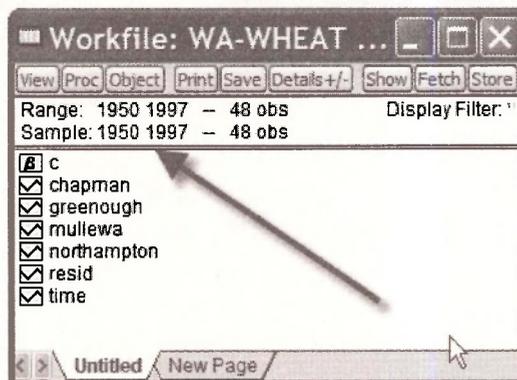
Before working with the data, double-click on **Range**. This will reveal the **Workfile structure**. When this file was created the annual nature of the data and time span were not used.



In the **Date specification** choose an **Annual** frequency with **Start date** 1950 and **End date** 1997, then **OK**.



This will not have any impact on the actual results we obtain, but it is good to take advantage of the time series features of EViews. The resulting workfile is now



Save this workfile with a new name. Select **File/Save As** to open a dialog box. We will call it *wheat\_chap04*. Estimate the linear regression of *YIELD* in GREENOUGH shire on *TIME* by entering

**equation linear.is greenough c time**

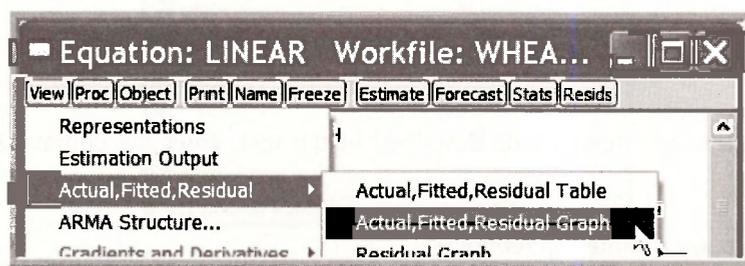
The command **equation linear.is** estimates the least squares regression AND gives it the name **LINEAR**.

Alternatively use the usual **Quick/Estimate Equation** dialog box, and then name the result.

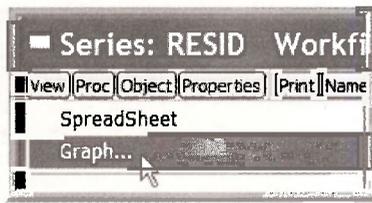
Dependent Variable: GREENOUGH  
 Method: Least Squares  
 Sample: 1950 1997  
 Included observations: 48

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.637778	0.064131	9.944999	0.0000
TIME	0.021032	0.002279	9.230452	0.0000
R-squared	0.649394	Mean dependent var		1.153060
Adjusted R-squared	0.641772	S.D. dependent var		0.365387
S.E. of regression	0.218692	Akaike info criterion		-0.161529
Sum squared resid	2.200009	Schwarz criterion		-0.083562
Log likelihood	5.876694	Hannan-Quinn criter.		-0.132065
F-statistic	85.20125	Durbin-Watson stat		1.200869
Prob(F-statistic)	0.000000			

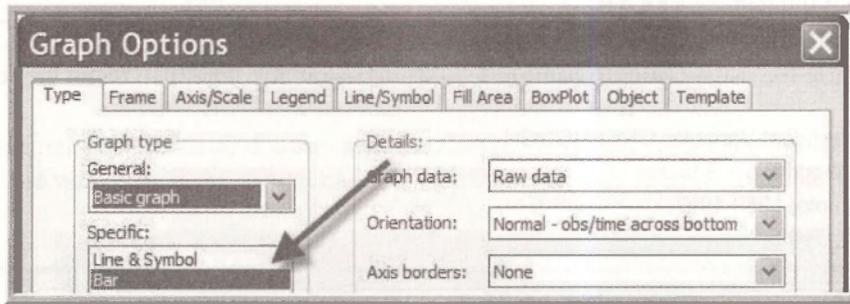
In the regression results window, click on **View/Actual,Fitted,Residual/ Actual,Fitted,Residual Graph** to construct Figure 4.8 in *POE*.



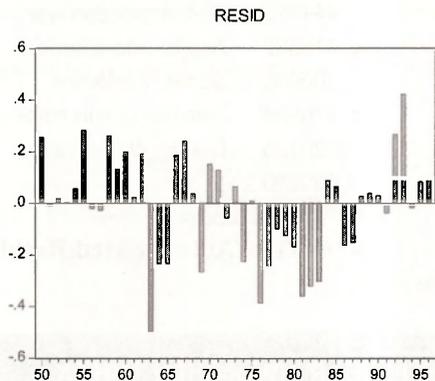
The bar graph in Figure 4.9 of *POE* is obtained by opening (double-click) the series **RESID** in the workfile window. Recall that EViews always saves the most recent regression residuals as **RESID**. In the spreadsheet view click **View/Graph**



In the **Graph Options** window choose **Bar** as the **Specific** type of graph.



The result is shown below. The advantage of specifying that the data series is **Annual** with specified dates is that EViews then labels the horizontal axis with the years.



To generate the cubic equation results described in the text, enter the commands (or use drop down boxes)

```

genr timecube = (time^3)/1000000
equation cubic.ls greenough c timecube
    
```

Or use the single command

```

equation cubic.ls greenough c (time^3)/1000000
    
```

Dependent Variable: GREENOUGH

Sample: 1950 1997

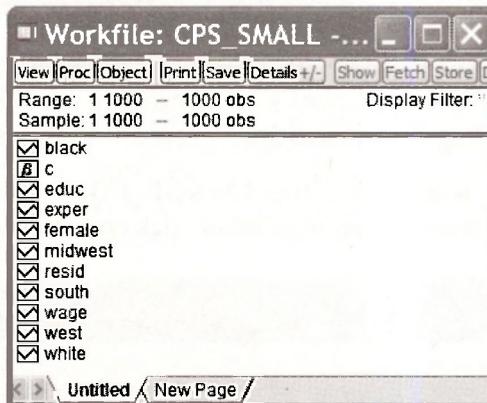
Included observations: 48

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.874117	0.035631	24.53270	0.0000
(TIME^3)/1000000	9.681516	0.822355	11.77292	0.0000
R-squared	0.750815	Mean dependent var		1.153060
Adjusted R-squared	0.745398	S.D. dependent var		0.365387
S.E. of regression	0.184368	Akaike info criterion		-0.502997
Sum squared resid	1.563604	Schwarz criterion		-0.425030
Log likelihood	14.07193	Hannan-Quinn criter.		-0.473533
F-statistic	138.6017	Durbin-Watson stat		1.659185
Prob(F-statistic)	0.000000			

This workfile (*wheat\_chap04.wf1*) can now be saved and closed.

#### 4.4 THE LOG-LINEAR MODEL

To illustrate the log-linear model we will use the workfile *cps\_small.wf1*, with data definitions in *cps\_small.def*. This data file consists of 1000 observations.



**Note:** EViews 6 Student Version has some limitations that the full version does not have. In particular it is limited to 1500 observations per series (which is not a problem here) and 15,000 total observations (series \* observations per series). This latter constraint is a problem here because we will be generating several new series in the example. For other limitations select **Help/Student Version Getting Started (pdf)** and examine Student Version Limitations.

To prevent a problem delete all the series **except** *WAGE* and *EDUC*. To do this, click on each series while holding down **Ctrl**. Right-click in the blue area and select **Delete**. Save the workfile with a new name, such as *wage\_chap04.wfl*, because we will use the data to estimate a **wage equation**.

Create a new series that is  $\ln(WAGE)$  and estimate the log-linear equation.

```
series lwage = log(wage)
equation lwage_eq.ls lwage c educ
```

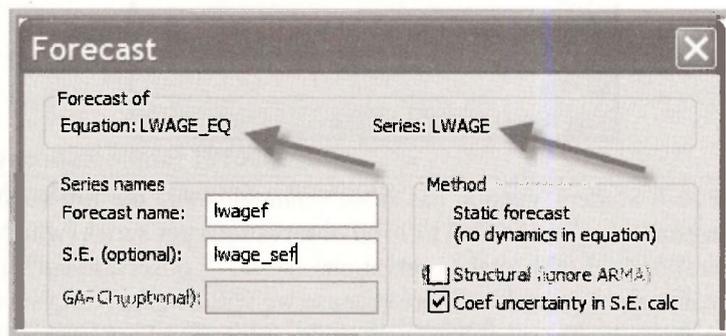
Note that we have given the **Name LWAGE\_EQ** to this equation.

Dependent Variable: LWAGE  
Method: Least Squares  
Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.788374	0.084898	9.286186	0.0000
EDUC	0.103761	0.006283	16.51434	0.0000
R-squared	0.214621	Mean dependent var		2.166837
Adjusted R-squared	0.213834	S.D. dependent var		0.552806
S.E. of regression	0.490151	Akaike info criterion		1.413792
Sum squared resid	239.7676	Schwarz criterion		1.423607
Log likelihood	-704.8960	Hannan-Quinn criter.		1.417523
F-statistic	272.7235	Durbin-Watson stat		0.411703
Prob(F-statistic)	0.000000			

#### 4.4.1 Prediction in the log-linear model

First we illustrate prediction with the equation **LWAGE\_EQ** in which we regressed the series *LWAGE* on *EDUC*. In the estimated equation window click on **Forecast**.



Select names for both the forecast and standard error.

To create a prediction interval for the predicted value of *WAGE*, we first create a 95% interval for **LWAGEF** as the forecast plus and minus the *t*-critical value times the standard error of the forecast. Then to convert it from logs to a numerical scale we take the antilog using the **exponential function**. The following commands create the *t*-value and the upper and lower bounds of predicted wage.

```
scalar tc = @qtdist(.975,998)
series w_ub = exp(lwagef + tc*lwage_sef)
series w_lb = exp(lwagef - tc*lwage_sef)
```

In repeated samples this prediction interval procedure will work 95% of the time. If however we seek a single predicted value, rather than an interval, it is possible that an alternative predictor, based on the properties of the log-normal distribution may be better. The **natural predictor** is the anti-log of the predicted **log(wage)**

```
series w_n = exp(lwagef)
```

In large samples a more precise predictor is obtained by correcting for log-normality. To do so we multiply the natural predictor by  $\exp(\hat{\sigma}^2/2)$ . The value of the estimated  $\hat{\sigma}$  is saved after a regression as **@se**. Thus the **corrected predictor** is

```
scalar sig2 = (@se)^2
series w_c = exp(lwagef)*exp(sig2/2)
```

A few values of the actual wage, the prediction interval, and the natural and corrected predictors are shown below. Note that the corrected predictor is always going to be larger than the natural predictor because the correction factor is always larger than one.

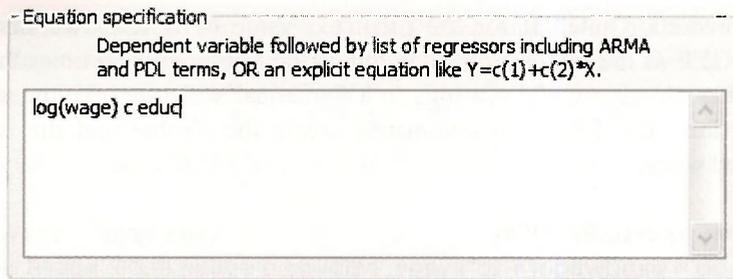
obs	WAGE	W_LB	W_N	W_C	W_UB
1	2.030000	3.237925	8.476222	9.558099	22.18901
2	2.070000	2.918436	7.640813	8.616061	20.00456
3	2.120000	2.918436	7.640813	8.616061	20.00456
4	2.540000	4.417789	11.57152	13.04848	30.30931
5	2.680000	2.918436	7.640813	8.616061	20.00456

#### 4.4.2 Alternative commands in the log-linear model

EViews allows transformations of variables to be included in the statement of the regression model, so instead of creating a new variable **LWAGE** as we did in Section 4.4.1, we can enter the statement

```
equation wage_eq.ls log(wage) c educ
```

Or, in the **Quick/Estimate Equation** dialog box enter the **Equation specification**



then name the equation **WAGE\_EQ**. Either way, the results are as shown below.

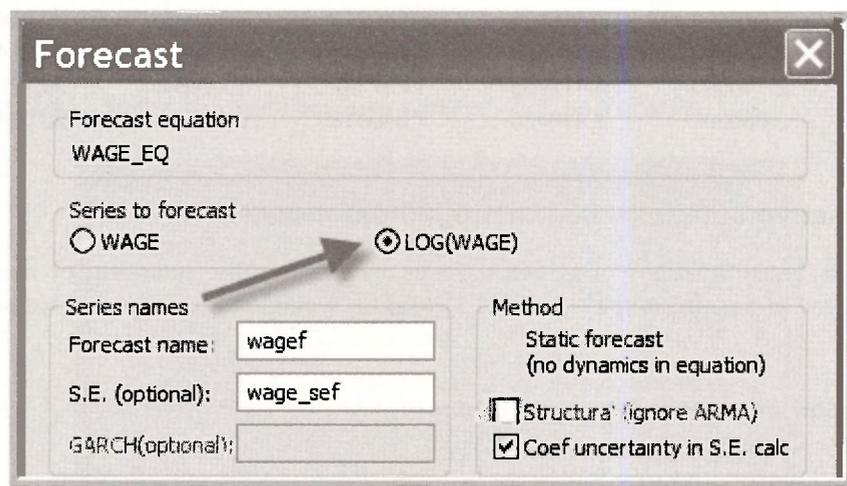
Dependent Variable: LOG(WAGE)

Method: Least Squares

Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.788374	0.084898	9.286186	0.0000
EDUC	0.103761	0.006283	16.51434	0.0000
R-squared	0.214621	Mean dependent var		2.166837
Adjusted R-squared	0.213834	S.D. dependent var		0.552806
S.E. of regression	0.490151	Akaike info criterion		1.413792
Sum squared resid	239.7676	Schwarz criterion		1.423607
Log likelihood	-704.8960	Hannan-Quinn criter.		1.417523
F-statistic	272.7235	Durbin-Watson stat		0.411703
Prob(F-statistic)	0.000000			

In the **WAGE\_EQ** window select **Forecast**. Choose the **LOG(WAGE)** radio button



The series **WAGEF** and **WAGE\_SEF** will be equal to series **LWAGEF** and **LWAGE\_SEF**, respectively. Then proceed with prediction interval calculations as in Section 4.4.1

### 4.4.3 Generalized $R^2$

A generalized goodness of fit measure is the squared correlation between the actual value of the dependent variable and its best predictor. Using the EViews function `@cor`, we obtain

```
scalar r2 = (@cor(wage,w_c))^2
```

Save and close your workfile *wage\_chap04.wfl*.

### Keywords

@cor	equation eqname.ls	ls
@qchisq	exponential function	normality test
@qtdist	forecast	prediction
@regobs	forecast standard error	prediction interval
@se	generalized $R^2$	prediction: corrected
@ssdep	goodness-of-fit	prediction: log-linear model
@ssr	histogram	prediction: natural
cubic equation	Jarque-Bera test	residual plot
data range	linear-log model	$R^2$
data sample	log function	workfile structure
data scaling	log-linear model	
elasticity	log-log model	

# CHAPTER 5

## The Multiple Regression Model

### CHAPTER OUTLINE

- 5.1 The Workfile: Some Preliminaries
  - 5.1.1 Naming the page
  - 5.1.2 Creating objects: a group
- 5.2 Estimating a Multiple Regression Model
  - 5.2.1 Using the Quick menu
  - 5.2.2 Using the Object menu
- 5.3 Forecasting from a Multiple Regression Model
  - 5.3.1 A simple forecasting procedure
  - 5.3.2 Using the forecast option
- 5.4 Interval Estimation
  - 5.4.1 The least squares covariance matrix
  - 5.4.2 Computing interval estimates
- 5.5 Hypothesis Testing
  - 5.5.1 Two-tail tests of significance
  - 5.5.2 A one-tail test of significance
  - 5.5.3 Testing nonzero values
- 5.6 Saving Commands

KEYWORDS

In the simple linear regression model the average value of a dependent variable is modeled as a linear function of a constant and a single explanatory variable. The multiple linear regression model expands the number of explanatory variables. As such it is a simple but important extension that makes linear regression quite powerful.

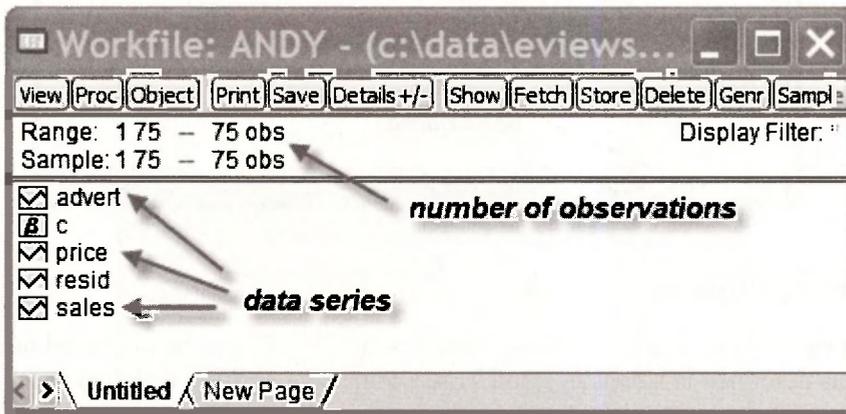
The example used in this chapter is a model of sales for Big Andy's Burger Barn. Big Andy's sales revenue depends on the prices charged for hamburgers, fries, shakes, and so on, and on the level of advertising. The prices charged in a given city are collected together into a weighted price index that is denoted by  $P = PRICE$  and measured in dollars. Monthly sales revenue for a given city is denoted by  $S = SALES$  and measured in \$1,000 units. Advertising expenditure for each city  $A = ADVERT$  is also measured in thousands of dollars. The model includes two explanatory variables and a constant and is written as

$$SALES = E(SALES) + e = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$

In this Chapter we use EViews to estimate this model, to obtain forecasts from the model, to examine the covariance matrix and standard errors of the estimates, and to compute confidence intervals and hypothesis test values for each of the coefficients. While performing these tasks we reinforce some of the EViews steps described in earlier chapters as well as introduce some new ones.

### 5.1 THE WORKFILE: SOME PRELIMINARIES

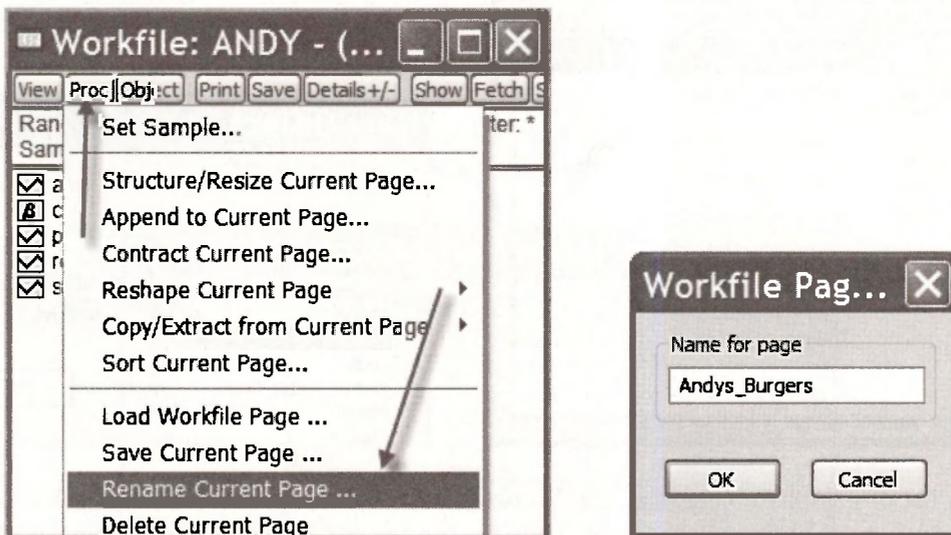
Observations on *SALES*, *PRICE* and *ADVERT* for 75 cities are available in the file *andy.wfl*. Opening this file as described in Chapters 1 and 2 yields the following screen



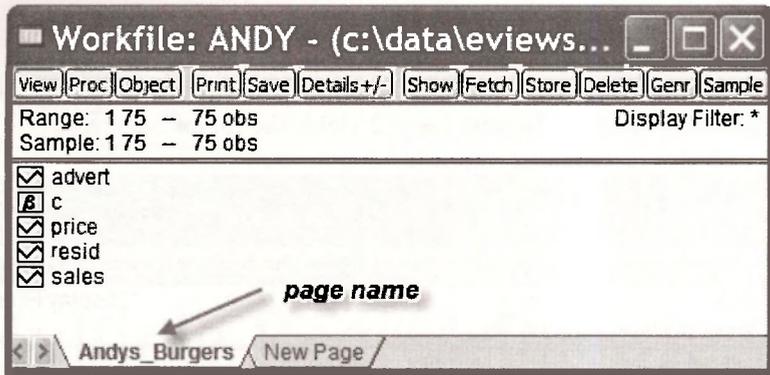
Note that the **range** and **sample** are set at 75 observations. And note the location of the data series in the workfile. The other objects **C** and **RESID** appear automatically in all EViews workfiles. We explain them as they become needed.

#### 5.1.1 Naming the page

It is possible to use a number of “pages” within the same EViews file. We will rarely use this option because most problems can fit neatly within the one page. However, if working with an **untitled** page is disconcerting for you, you can give it a name by selecting from your workfile toolbar **Proc/Rename Current Page**



A window appears in which you can name the page. After choosing the name **Andys\_Burgers**, your workfile will appear as

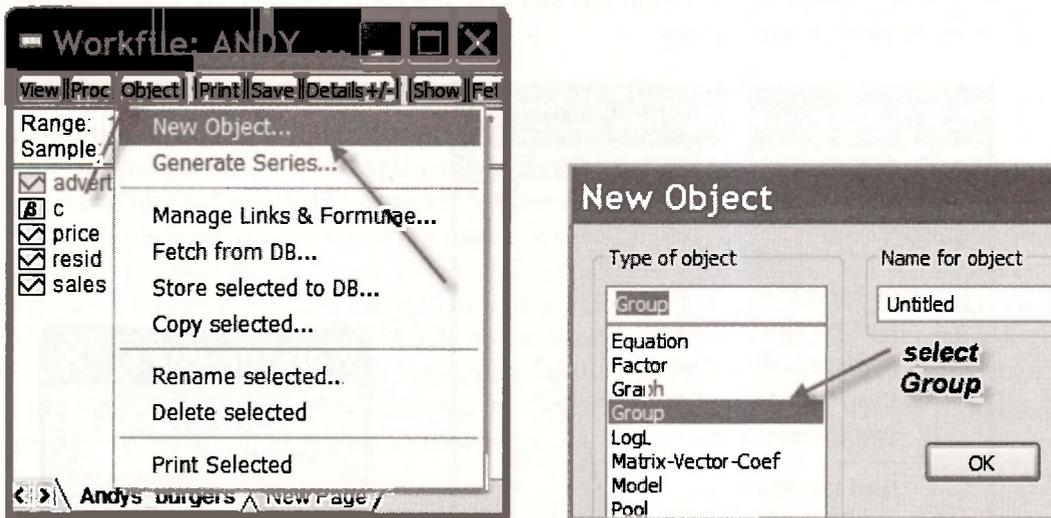


### 5.1.2 Creating objects: a group

The data on each of the variables *SALES*, *PRICE* and *ADVERT* can be examined one at a time or as a group, as described in Chapters 1 and 2. We will create a **group** and then check the data and summary statistics to make sure they match those in Table 5.1 on page 109 of the text. In Chapters 1 and 2 we created a group by (1) highlighting the series to be included in the group, (2) double clicking the highlighted area, and (3) selecting **Open Group**. To extend your knowledge of EViews, we now describe another way. This new way is more cumbersome, but it will help you understand the more general concept of an **object** and how objects are created.

A group is one of many types of objects that can be created by EViews. The concept of an object is a bit vague, but you can think of it as anything that gets stored in your workfile. As Richard Startz says in *EViews Illustrated* [QMS, 2007, p.5], “object is a computer science buzz word meaning ‘thingie’.” Several chapters of the Startz book can be found under **Help**.

To see a list of possible objects, select **Object/New Object** from the workfile toolbar.



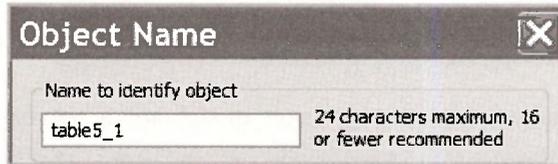
A long list of possible objects appears. We will encounter many of these objects (but not all of them) as we proceed through the book. Only the top few are displayed in the above screen shot. At present we select **Group** as the relevant object, and then click **OK**. We have left **Name for object** as **Untitled**. We will name it later. In the following window that appears, we type the names of the series to be included in the group, and then click **OK**.



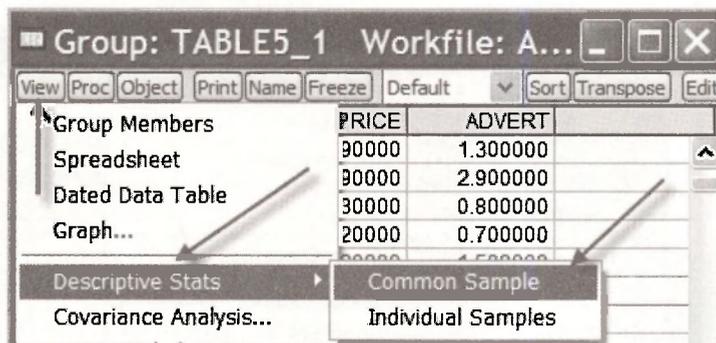
The following screen appears. Note that the first 5 observations are the same as those in Table 5.1 on page 109 of the text. The last 3 observations can be checked by scrolling down.

obs	SALES	PRICE	ADVERT
1	73.20000	5.690000	1.300000
2	71.80000	6.490000	2.900000
3	62.40000	5.630000	0.800000
4	67.40000	6.220000	0.700000
5	89.30000	5.020000	1.500000
6	70.30000	6.410000	1.300000
7			

By selecting **Name** we can name the group object in the following window. In line with the text, we call it **table5\_1**.



One of the advantages of creating a group of variables is that we can view a variety of information on the collection of variables in that group. The list of observations that we checked against Table 5.1 is one type of information; it is called the **spreadsheet** view of the group. Another useful view is the **Descriptive Stats** view that gives summary statistics that can be checked against those that appear in the lower panel of Table 5.1. To obtain this view we open the group and then select **View/Descriptive Stats/Common Sample**.



The summary statistics appear in the following window. In addition to the sample **mean**, **median**, **maximum**, **minimum** and **standard deviation** for each series, the table presents **skewness** and

**kurtosis** measures (see pages 490, 511 and 512 of the text), the value of the **Jarque-Bera statistic** for testing whether a series is normally distributed and its corresponding  $p$ -value (see pages 89-90 of the text), and the **sum**  $\sum x_i$  and **sum of squared deviations**  $\sum (x_i - \bar{x})^2$  for each of the series. Stop and check to see if you know how to obtain the standard deviation values from the sum of squared deviations.

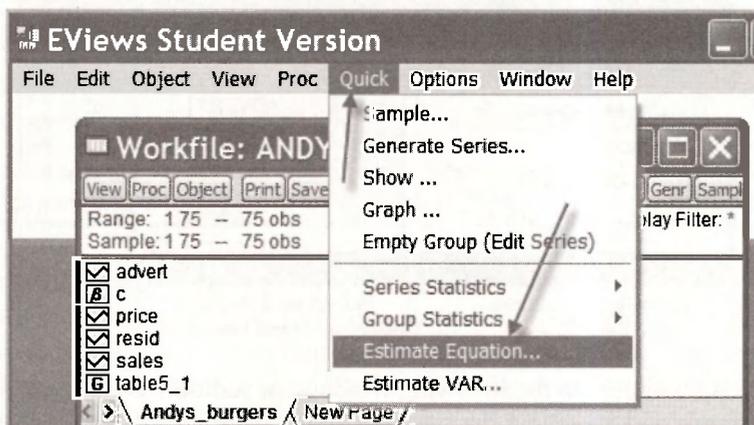
Group: TABLE5_1 Workfile: ...									
View	Proc	Object	Print	Name	Freeze	Sample	Sheet	Stats	Spec
				SALES		PRICE		ADVERT	
Mean				77.37467		5.687200		1.844000	
Median				76.50000		5.690000		1.800000	
Maximum				91.20000		6.490000		3.100000	
Minimum				62.40000		4.830000		0.500000	
Std. Dev.				6.488537		0.518432		0.831677	
Skewness				-0.010631		0.061846		0.037087	
Kurtosis				2.255328		1.667162		1.704890	
Jarque-Bera				1.734340		5.599242		5.258786	
Probability				0.420139		0.060833		0.072122	
Sum				5803.100		426.5400		138.3000	
Sum Sq. Dev.				3115.482		19.88911		51.18480	
Observations				75		75		75	

## 5.2 ESTIMATING A MULTIPLE REGRESSION MODEL

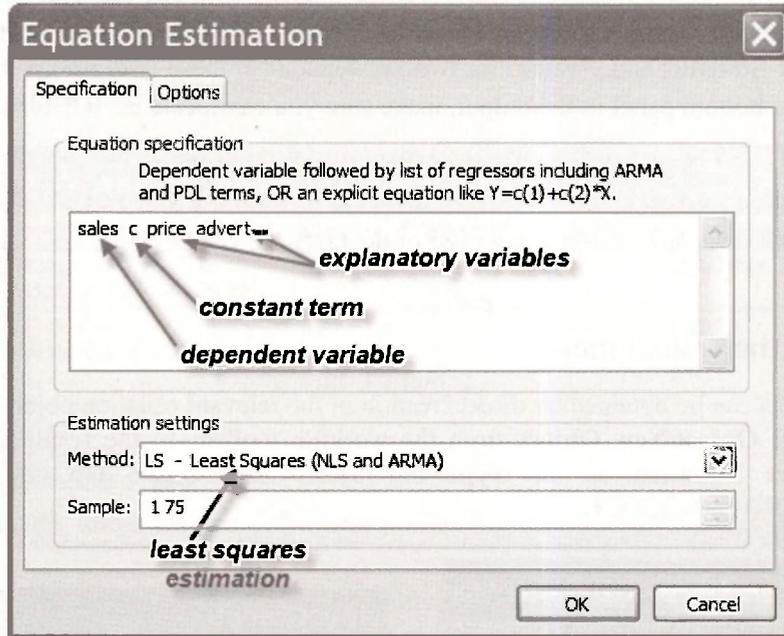
The steps for estimating a multiple regression model are a natural extension of those for estimating a simple regression. We will consider two alternative ways. One is using the **Quick** menu considered in earlier chapters. The other is via the **Object** menu that we used in the previous section to define a **group**.

### 5.2.1 Using the Quick menu

To use the Quick menu for estimating an equation go to the upper EViews window and select **Quick/Estimate Equation**.



An **Equation Estimation** window appears. We add to it in the following way.



The **Equation Specification** dialog box is where you tell EViews what model you would like to estimate. Our equation is

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$

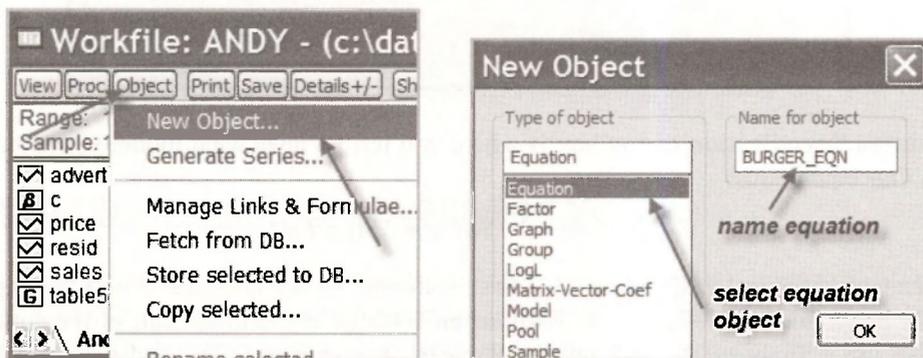
The dependent variable *SALES* is inserted first, followed by the constant *C* and the explanatory variables *PRICE* and *ADVERT*. Under **Estimation settings** in the lower half of the window, you can choose the estimation **Method** and the **Sample** observations to be used for estimation. The **least-squares** method is the one we want, and the one that is automatically used unless another one is selected. A **sample of 1 75** means that all observations in our sample are being used to estimate the equation. Clicking **OK** yields the regression output.

Equation: UNTITLED Workfile: A...									
View	Proc	Object	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: SALES					<i>click to name equation</i>				
Method: Least Squares									
Sample: 1 75									
Included observations: 75									
Variable	Coefficient	Std. Error	t-Statistic	Prob.					
C	118.9136	6.351638	18.72172	0.0000					
PRICE	-7.907854	1.095993	-7.215241	0.0000					
ADVERT	1.862584	0.683195	2.726283	0.0080					
R-squared	0.448258	Mean dependent var	77.37467						
Adjusted R-squared	0.432932	S.D. dependent var	6.488537						
S.E. of regression	4.886124	Akaike info criterion	6.049854						
Sum squared resid	1718.943	Schwarz criterion	6.142553						
Log likelihood	-223.8695	Hannan-Quinn criter.	6.086868						
F-statistic	29.24786	Durbin-Watson stat	2.183037						
Prob(F-statistic)	0.000000								

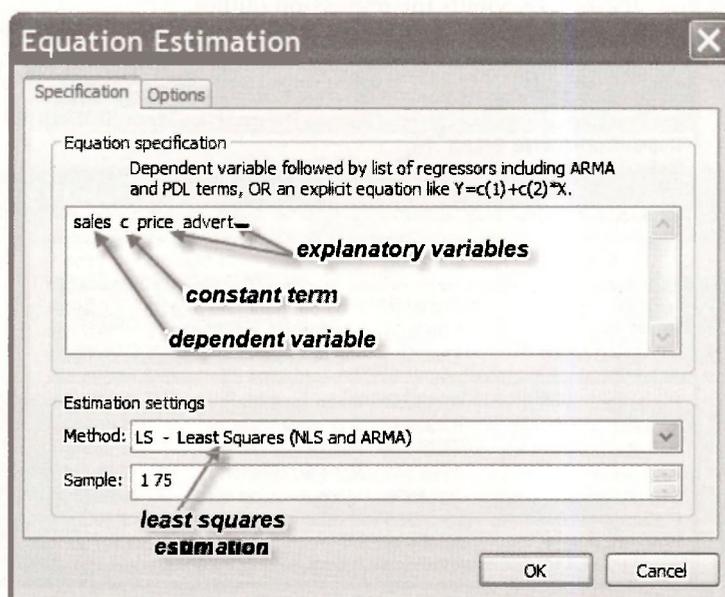
Compare this output with Table 5.2 on page 112 of the text. Least squares estimates of the coefficients  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  appear in the column **Coefficient**; their standard errors are in the column **Std. Error**;  $t$ -values for testing a zero null hypothesis for each of the coefficients appear in the column **t-Statistic**; and  $p$ -values for two-tail versions of these tests are given in the column **Prob.** From the bottom panel in the output, make sure you can locate  $R^2 = 0.4483$ ,  $\hat{\sigma}_y = 6.4885$ ,  $SSE = \sum \hat{e}_i^2 = 1718.943$  and the estimated standard deviation of the error term  $\hat{\sigma} = 4.8861$ . Also, you should make sure you know how to compute (a)  $\hat{\sigma}$  from the value of  $SSE$  (page 114 of the text), and (b)  $R^2$  from  $SSE$  and  $\hat{\sigma}_y$  (page 125 of the text).

### 5.2.2 Using the Object menu

The same results can be obtained by direct creation of the relevant equation object. To proceed in this way select **Object/New Object** from the workfile toolbar. In the resulting **New Object** window, select **Equation** as the **Type of object**. Then, name the object (we chose **BURGER\_EQN**) and click **OK**.

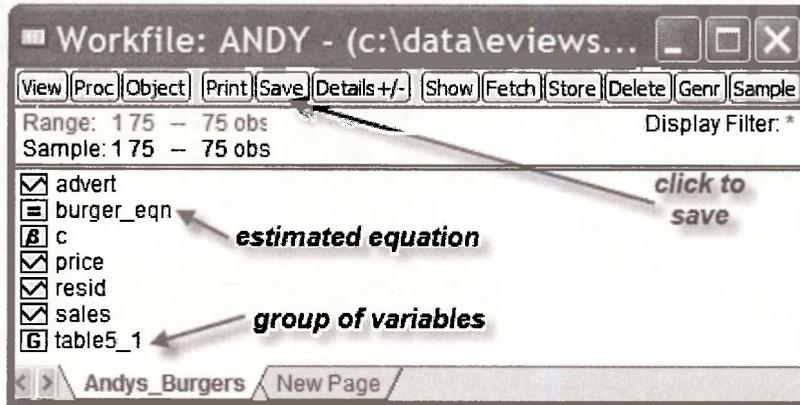


The **Equation Estimation** window will appear. It can be filled in as described earlier.



Clicking **OK** yields the same regression output that we illustrated earlier using the quick menu.

After closing the output window, check the EViews workfile and note that, since you opened the workfile, two new objects have been added. These objects are the group of variables **TABLE5\_1** and the estimated equation **BURGER\_EQN**. To reopen one of these objects, highlight it, and then double click.



To ensure the estimated equation and the group of variables are retained for future use, click **save**. If you wish to save the file under another name so that the original workfile with data only is preserved, go to the upper EViews toolbar and select **File/Save As**.

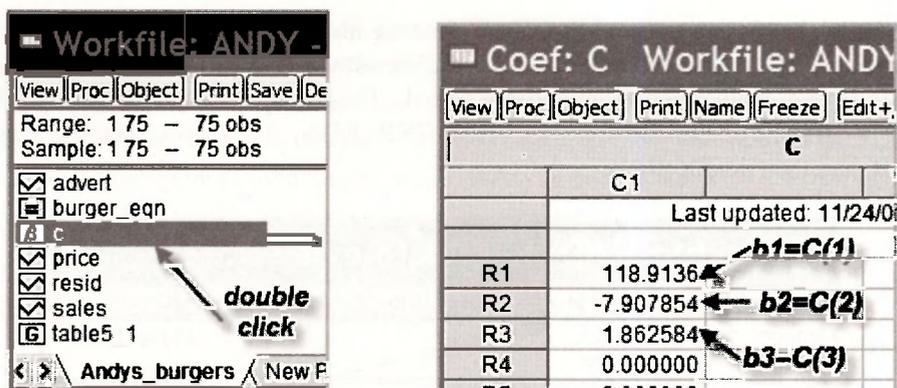
### 5.3 FORECASTING FROM A REGRESSION MODEL

How to use EViews to obtain forecasts was considered extensively in Chapter 4 for both linear and log-linear models. Those procedures carry over directly to the multiple regression model. In this section we reinforce those procedures by showing how EViews can be used to forecast (or predict) hamburger sales revenue for  $PRICE = 5.5$  and  $ADVERT = 1.2$ , as is done on page 113 of the text. Two preliminary explanatory remarks are in order. First, note that we are using the terms “forecast” and “predict” interchangeably; each one has no special significance. Second, the steps we follow do not mimic exactly those in Chapter 4. The variations are deliberate. They are designed to expose you to more of the features of EViews. As in Chapter 4, we consider a simple forecasting procedure and one using EViews special forecasting capabilities. While the simple one is ideal for obtaining a single static forecast, it is not convenient for obtaining a forecast standard error, and it less than ideal for dynamic forecasting, a topic considered in Chapter 9.

#### 5.3.1 A simple forecasting procedure

After you use EViews to estimate a regression model the estimated coefficients are stored in the object **C** that appears in your workfile. You can check this fact out by highlighting **C** and double clicking it. A spreadsheet will appear with the estimates stored in a column called **C1**. In further commands that you might supply to EViews, the three values in that column can be used by referring to them as  $C(1)$ ,  $C(2)$  and  $C(3)$ , respectively. That is, in terms of notation used in the book, the least squares estimates are

$$b_1 = C(1) \quad b_2 = C(2) \quad b_3 = C(3)$$



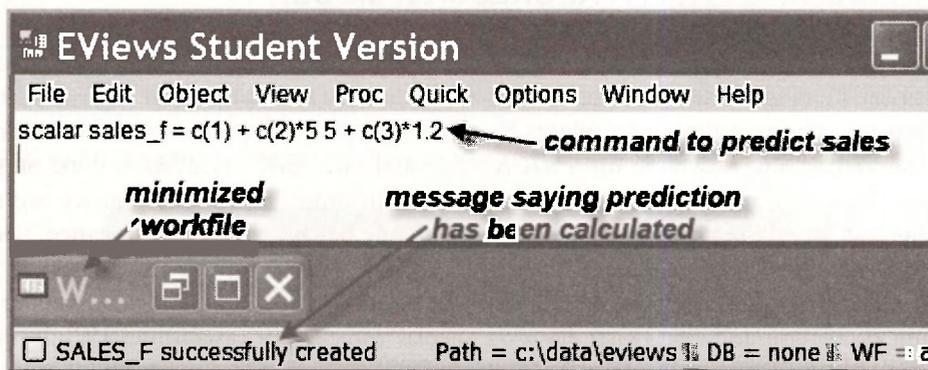
Our objective is to get EViews to perform the calculation

$$\widehat{SALES} = b_1 + b_2 \times 5.5 + b_3 \times 1.2$$

The corresponding EViews command is

**scalar sales\_f = c(1) + c(2)\*5.5 + c(3)\*1.2**

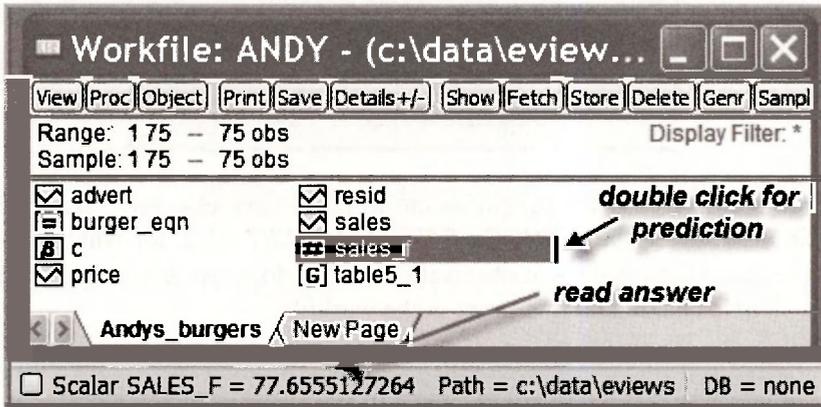
The first word **scalar** tells EViews that we are computing a scalar object (a single number) to be stored in the workfile. The second word **sales\_f** is the name we are giving to that scalar object which is our predicted value. The right side of the equation performs the calculation. The command is placed in the upper EViews window as shown below.



This window might look a little strange to you. We have compressed the typical EViews window so that we can show you all the information in a convenient space. The workfile has been temporarily minimized to move it out of the way. Then the bottom of the window has been moved upwards. Notice two things. The command to predict sales has been typed in the upper window. And, there is a message at the bottom indicating that the scalar object **SALES\_F** has been successfully calculated. Providing you have not done something wrong that offends EViews, this message will appear after you type in the command and push the **enter** key.

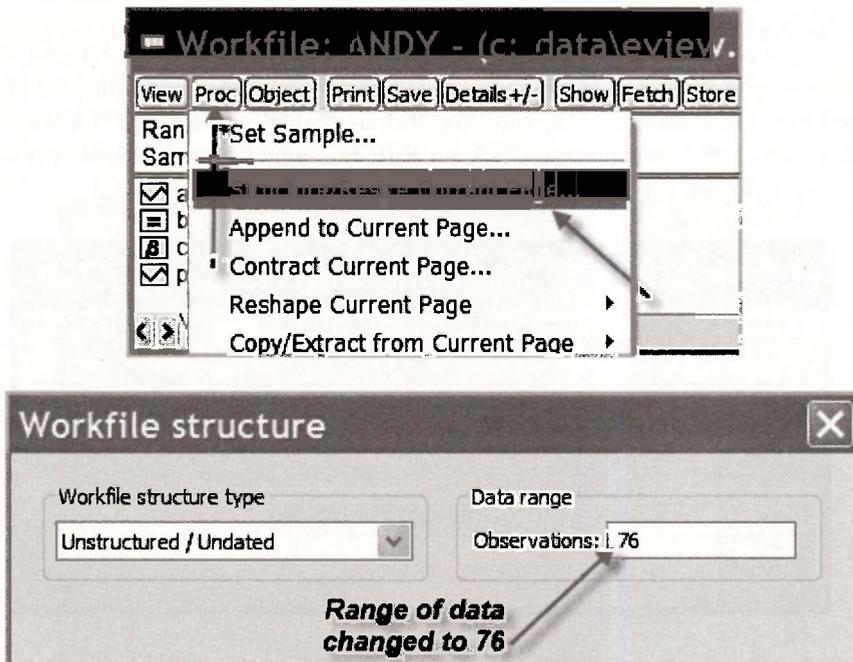
A word of warning: The values C(1), C(2) and C(3) will always be the coefficient estimates for the model most recently estimated. If you have only estimated one equation, there will be no confusion. However, if you have estimated another model, successfully or not, the values will change. Make sure you are using the correct ones.

You have now calculated the forecast. How do you read off the answer? Go to the **SALES\_F** object in your workfile and double click it. The answer appears in the bottom panel of your workfile.

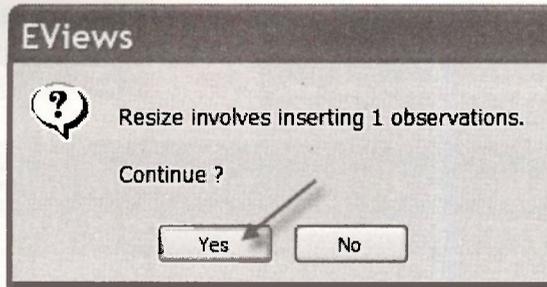


### 5.3.2 Using the forecast option

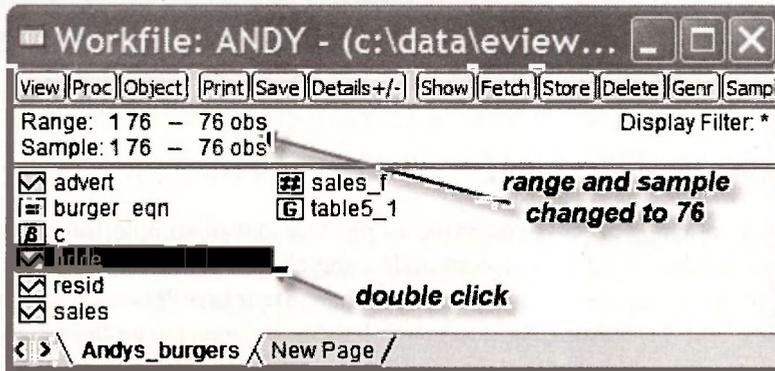
To use EViews automatic **forecast** command to produce out-of-sample forecasts, it is necessary to extend the size of the workfile to accommodate the observations for which we want forecasts. To do so you select, from the workfile toolbar, **Proc/Structure/Resize Current Page**. In this case, since we are only forecasting for one extra observation, we change the range from 75 to 76.



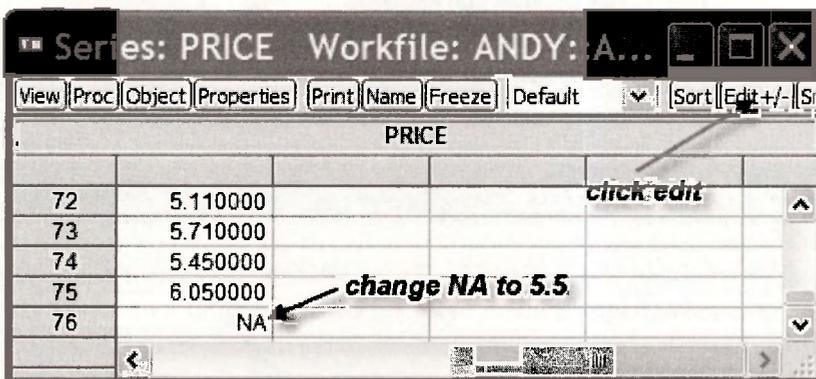
EViews will ask you whether you are sure you wish to make this change.



Notice that both the range and the sample in the workfile have changed from 1 75 to 1 76. The next task is to insert the values  $PRICE = 5.5$  and  $ADVERT = 1.2$  for which we want to make the forecast. We insert these values at observation 76. To do so we begin by opening the *PRICE* spreadsheet by double clicking on this series in the workfile.

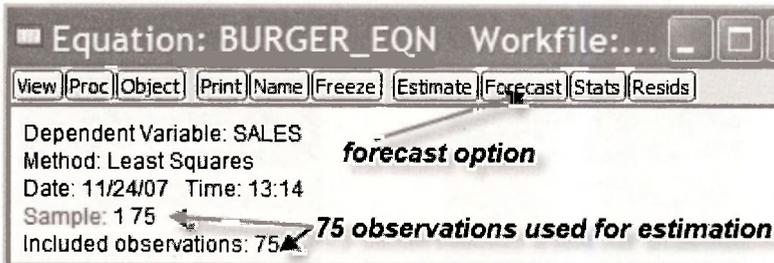


The lower portion of the *PRICE* spreadsheet appears below. Notice how EViews responded to your request to extend the range from 75 to 76. It did not have an observation for observation 76 so it specified this observation as NA, short for “not available”. To replace NA with 5.5, click on **Edit +/-**, and change the spreadsheet. Click on **Edit +/-** again after you have made the change. Similar steps are followed for the *ADVERT* spreadsheet to insert the value 1.2.



Now you are ready to compute the forecast. Go to your workfile and open the equation **BURGER\_EQN** by double clicking on it. Then click on the **forecast** button in the toolbar. Before doing so, note that the number of observations used to estimate the equation is still 75. We are using the first 75 observations to estimate an equation which is then used to forecast sales for

observation 76. Increasing the range of observations in a workfile does not change equations in the workfile that have already been estimated with fewer observations.



The following forecast dialog box appears. Let us consider the various items in this box.

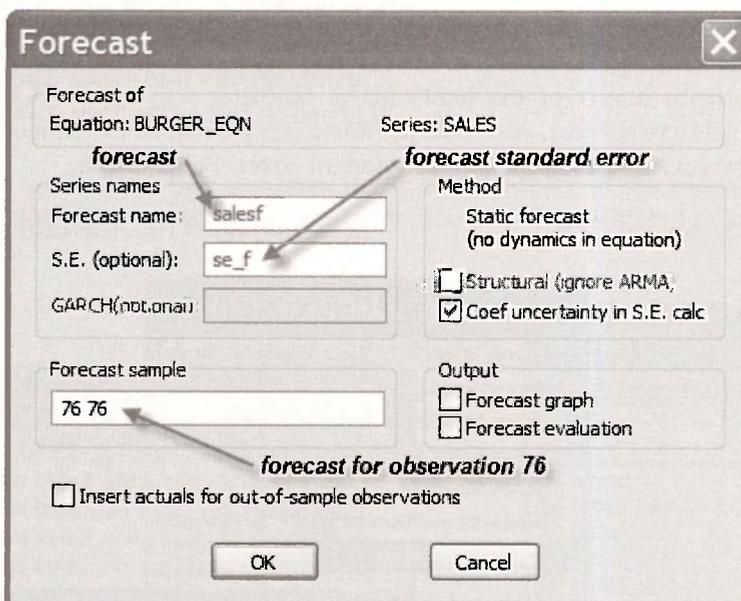
**Series names:** The forecasts and their standard errors will appear in the workfile under the names SALESF and SE\_F, respectively. The forecast standard error is computed using the formula on page 157 of the text. This formula includes what EViews calls **Coef uncertainty in S.E. calculation**. In this particular case, not including this uncertainty would mean the forecast standard error is the same as the standard deviation of the error term.

**Forecast sample:** We have chosen to forecast for just observation 76. We could have defined the **forecast sample** as 1 76, in which case EViews would produce both the in-sample forecasts as well as the out-of-sample forecast.

**Method:** There are no **dynamics** in the equation because we do not have time series observations with lagged variables. These issues are considered in Chapter 9.

**Output:** At this point we are not concerned with a **Forecast graph**, or a **Forecast evaluation**.

**Insert actuals for out-of-sample observations:** A tick in this box asks EViews to insert actual values for SALES for the observations that lie outside your **Forecast sample** – in this case that would be observations 1 to 75. We did not choose this option.



After clicking **OK** and closing the equation, you will be returned to the workfile where you will discover that **SALESF** and **SE\_F** appear as two new series in the workfile. On opening these series by double clicking on them, you will further discover that the forecast and its standard error appear at observation 76, with the forecasts and standard errors at observations 1 to 75 being listed as **NA**, a consequence of the **Forecast sample** that we specified in the dialog box.

sales	forecast
73	NA
74	NA
75	NA
76	77.65551

forecast	standard error
73	NA
74	NA
75	NA
76	4.942008

The forecast and its standard error are  $\widehat{SALES} = 77.6555$  and  $se(f) = 4.942$ . These values can be used to compute a forecast interval as  $\widehat{SALES} \pm t_{(1-\alpha/2, 72)} \times se(f)$ .

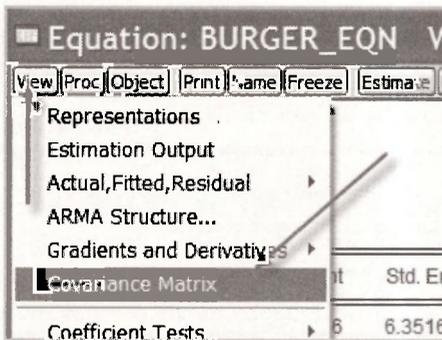
## 5.4 INTERVAL ESTIMATION

After obtaining least squares estimates of an equation we can proceed to use it for forecasting as we have done in the preceding section. In addition, we may be interested in obtaining interval estimates that reflect the precision of our estimates, or testing hypotheses about the unknown coefficients. The covariance matrix of the least squares estimates is a useful tool for these purposes, and one we will return to in Chapter 6. We begin by explaining how it can be viewed.

### 5.4.1 The least squares covariance matrix

To examine the least squares covariance matrix go to the **BURGER\_EQN** in your workfile and open it by double clicking. Select **View/Covariance Matrix** from the toolbar and drop-down menu. The covariance matrix of the least-squares estimates will appear. Check these values against those on p.116 of the text. Also note the relationship between the variances that appear on the diagonal of the covariance matrix and the standard errors. For example,

$$\text{cov}(b_2, b_3) = -0.74842 \qquad se(b_2) = \sqrt{\text{var}(b_2)} = \sqrt{1.201201} = 1.09599$$



Coefficient Covariance Matrix			
	C	PRICE	ADVERT
C	40.34330	-6.795064	-0.748421
PRICE	-6.795064	1.201201	-0.019742
ADVERT	-0.748421	-0.019742	0.466756

### 5.4.2 Computing interval estimates

A  $100(1-\alpha)\%$  confidence interval for one of the unknown parameters, say  $\beta_k$ , is given by

$$b_k \pm t_{(1-\alpha/2, N-K)} \times se(b_k)$$

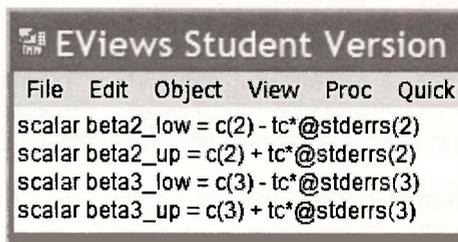
Thus, to get EViews to compute a confidence interval, we need to locate values for  $b_k$ ,  $se(b_k)$  and  $t_{(1-\alpha/2, N-K)}$ , and then do the calculations. As we noted earlier in this chapter, the least squares estimates  $b_k$  will be stored in the object **C** in the workfile. Alternatively, they are stored in the array **@coefs** which was used for computing interval estimates in Chapter 3. That is, **C** = **@coefs**. If we are interested in one particular  $b_k$ , say  $b_2$ , then  $C(2)$  = **@coefs(2)** = -7.907854. Similarly, the standard errors are stored in the array **@stderrs**, so that **@stderrs(2)** = 1.09599. Note that **C**, **@coefs** and **@stderrs** will contain values from the most recently estimated equation. If you are in doubt about their contents, quickly re-estimate the equation of interest. The remaining value that is required is  $t_{(1-\alpha/2, N-K)}$ . It can be found using the EViews function **@qtdistn(p,v)** where **p** is equal to  $1-\alpha/2$  and **v** is the number of degrees of freedom, in this case  $N-K=75-3=72$ . Putting all these ingredients together, upper and lower bounds for 95% interval estimates for  $\beta_2$  and  $\beta_3$  can be found from the following sequence of commands.

```

scalar tc = @qtdist(0.975,72)
scalar beta2_low = c(2) - tc*@stderrs(2)
scalar beta2_up = c(2) + tc*@stderrs(2)
scalar beta3_low = c(3) - tc*@stderrs(3)
scalar beta3_up = c(3) + tc*@stderrs(3)

```

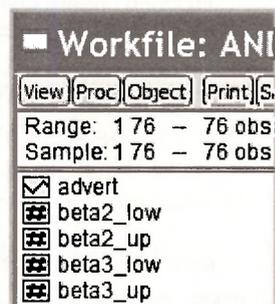
These commands are entered, one at a time, in the upper display of the EViews window. Each command is executed after you push the **enter** key. The answers are stored as scalars marked by  in the workfile.



```

EViews Student Version
File Edit Object View Proc Quick
scalar beta2_low = c(2) - tc*@stderrs(2)
scalar beta2_up = c(2) + tc*@stderrs(2)
scalar beta3_low = c(3) - tc*@stderrs(3)
scalar beta3_up = c(3) + tc*@stderrs(3)

```



```

Workfile: ANI
View Proc Object Print S...
Range: 1 76 -- 76 obs
Sample: 1 76 -- 76 obs
 advert
 beta2_low
 beta2_up
 beta3_low
 beta3_up

```

To view the upper and lower bounds of the interval estimates double click each of the scalars in the workfile. Each time the answer will appear in the bottom of the EViews window. Collecting these values one at a time, we obtain

- Scalar BETA2\_LOW = -10.0926764859
- Scalar BETA2\_UP = -5.72303216832
- Scalar BETA3\_LOW = 0.500658984708
- Scalar BETA3\_UP = 3.22450955651

Apart from a small amount of rounding in the text values, you will discover that these interval estimates coincide with those on page 119 of the text.

### 5.5 HYPOTHESIS TESTING

In this Chapter we are concerned with hypothesis tests on a single coefficient in the multiple regression model. More complex tests are deferred until Chapter 6. The most common single coefficient tests are two-tail tests of significance where, in the context of Andy’s Burger Barn, we are testing whether price effects sales and whether advertising expenditure effects sales.

#### 5.5.1 Two-tail tests of significance

Two-tail tests of significance for the effect of price and the effect of advertising are considered on pages 121-2 of the text. The hypotheses for these tests are

$$H_0 : \beta_2 = 0 \quad (\text{no price effect})$$

$$H_1 : \beta_2 \neq 0 \quad (\text{there is a price effect})$$

$$H_0 : \beta_3 = 0 \quad (\text{no advertising effect})$$

$$H_1 : \beta_3 \neq 0 \quad (\text{there is an advertising effect})$$

Using EViews to calculate the *t*-values and *p*-values for these tests is trivial. They are automatically computed when you estimate the equation. To see where they are reported, we return to the least squares output for **BURGER\_EQN**.

Dependent Variable: SALES		<b><i>t</i>-values and <i>p</i>-values for two-tail tests of significance</b>		
Method: Least Squares				
Date: 11/24/07 Time: 13:14				
Sample: 1 75				
Included observations: 75				
	Coefficient	Std. Error	t-Statistic	Prob.
C	118.9136	6.351638	18.72172	0.0000
PRICE	-7.907854	1.095993	-7.215241	0.0000
ADVERT	1.862584	0.683195	2.726283	0.0080

Do you know where these numbers come from? Consider the test for the effect of advertising. The *t*-value is given by  $t = 1.8626/0.6832 = 2.726$ . The *p*-value is given by

$$p\text{-value} = P(t_{(72)} > 2.726) + P(t_{(72)} < -2.726) = 2 \times P(t_{(72)} < -2.726) = 0.0080$$

We can confirm the above result by asking EViews to compute the above probability using the command

```
scalar pee = 2*@ctdist(-2.726,72)
```

The function **@ctdist(x, v)** computes the distribution function value  $P(t_{(v)} < x)$ . The command can be entered in the top display of the EViews window. If you are unsure of how to do so, or how to read off the result, go back and check the earlier part of this chapter where we introduced a simple forecasting procedure, or the section where we computed interval estimates.

Knowing the  $p$ -value is sufficient information for rejecting or not rejecting  $H_0$ . In the case of advertising expenditure we reject  $H_0 : \beta_3 = 0$  at a 5% significance level because the  $p$ -value of 0.0080 is less than 0.05. Suppose, however, that we wanted to make a decision about  $H_0$  by comparing the calculated value  $t = 2.726$  to a 5% critical value. How do we find that critical value? We need values  $t_c$  and  $-t_c$  such that  $P(t_{(72)} < t_c) = 0.975$ . Table 2 at the end of the book is not sufficiently detailed to provide this value. It can be obtained using the EViews command

```
scalar tc = @qtdist(0.975,72)
```

The answer is  $t_c = 1.993$ , a value that leads us to reject  $H_0 : \beta_3 = 0$  because  $2.726 > 1.993$ .

The  $p$ -value for testing  $H_0 : \beta_2 = 0$  against  $H_1 : \beta_2 \neq 0$  is given as 0.0000 in the EViews regression output. As an exercise, use EViews to show that, using more decimal places, the value is  $4.424 \times 10^{-10}$ .

### 5.5.2 A one-tail test of significance

To collect evidence on whether or not the demand for burgers is price elastic, on pages 122-3 of the text we test  $H_0 : \beta_2 > 0$  against the alternative  $H_1 : \beta_2 < 0$ . In this case we are not particularly interested in the single point  $\beta_2 = 0$ , but, nevertheless, for testing  $H_0 : \beta_2 > 0$  we act as if the null hypothesis is  $H_0 : \beta_2 = 0$ . Thus, this test can be viewed as a one-tail test of significance. The  $p$ -value for this test is  $P(t_{(72)} < -7.215241) = 2.212 \times 10^{-10}$ . Because the calculated value  $t$ -value  $t = -7.215241$  is negative, and the rejection region is in the left tail (as suggested by the direction of the alternative hypothesis  $H_1 : \beta_2 < 0$ ), we can compute the  $p$ -value by taking half of the  $p$ -value given in the EViews regression output. However, since half of 0.0000 is 0.0000, this example is not a very interesting one. If we considered a one-tail test for advertising of the form  $H_0 : \beta_3 < 0$  against  $H_1 : \beta_3 > 0$ , we could calculate its  $p$ -value as 0.0040, half of 0.0080.

If the calculated  $t$ -value is positive, and the rejection region is the left tail (or the calculated  $t$ -value is negative, and the rejection region is the right tail), the  $p$ -value will be greater than 0.5 and is not simply half of the EViews  $p$ -value. In such instances the  $p$ -value is given by  $p = 1 - p^*/2$  where  $p^*$  is the EViews supplied  $p$ -value. Do you understand why? Check it out!

What is the 5% critical value for a one-tail test? For the case of  $\beta_2$  where the critical value is a negative one in the left tail of the distribution, we can obtain it using the EViews command

```
scalar tc_1tail = @qtdist(0.05,72)
```

The value obtained is  $t_c = -1.666$ . Thus, making the test decision by reference to the critical value, we reject  $H_0 : \beta_2 \geq 0$  in favor of  $H_1 : \beta_2 < 0$  because  $-7.215 < -1.666$ .

### 5.5.3 Testing nonzero values

#### 5.5.3a One-tail test

For advertising to be effective  $\beta_3$  must be greater than 1. Thus we test  $H_0 : \beta_3 < 1$  against  $H_1 : \beta_3 > 1$ . On page 123-4 of the text, we compute the key quantities for performing this test. They are the calculated  $t$ -value

$$t = \frac{b_3 - 1}{\text{se}(b_3)} = \frac{1.8626 - 1}{0.6832} = \frac{0.8626}{0.6832} = 1.263$$

and its corresponding  $p$ -value

$$P(t_{(72)} > 1.263) = 1 - P(t_{(72)} \leq 1.263) = 0.105$$

You can compute these quantities using the following commands in the upper display of the EViews window.

```
scalar t3 = (c(3) - 1)/@stderrs(3)
scalar pee3 = 1 - @ctdist(t3,72)
```

#### 5.5.3b Two-tail test

For two-tail tests there is an easier way to get the results from EViews. You can tell EViews the hypothesis that you want to test and it will do the rest. There is one temporary complication. EViews computes an  $F$ -value and a  $\chi^2$ -value but not a  $t$ -value. This complication will disappear once you have the extra background covered in Chapter 6. However, given that you are likely to be eagerly waiting to find out how EViews automatic testing commands work, we will give you some exposure now. If you are struggling with our explanations, please come back again after you have finished Chapter 6.

To illustrate we turn the recent hypothesis about the effect of advertising expenditure into a two-tail test, namely

$$H_0 : \beta_3 = 1 \quad \text{against} \quad H_1 : \beta_3 \neq 1$$

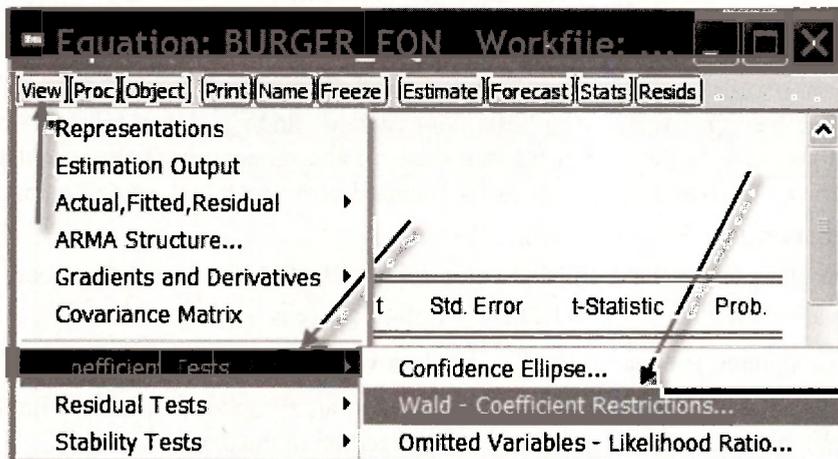
For testing this hypothesis, the calculated  $t$ -value is the same as before.

$$t = \frac{b_3 - 1}{se(b_3)} = \frac{1.8626 - 1}{0.6832} = \frac{0.8626}{0.6832} = 1.2626$$

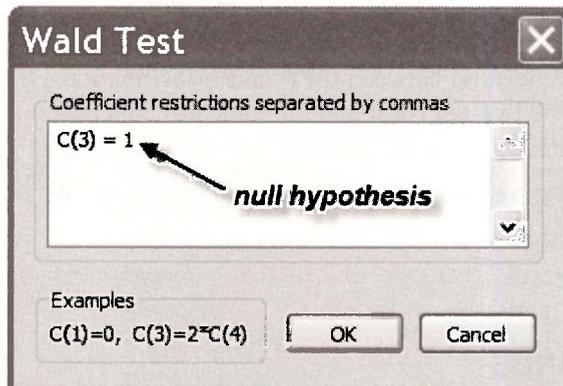
The  $p$ -value will be different, however. It is

$$P(t_{(72)} > 1.2626) + P(t_{(72)} < -1.2626) = 2 \times 0.1054 = 0.2108$$

We have included more digits after the decimal so that we can match the accuracy of EViews. To get EViews to automatically compute these values, we proceed as follows. Open the equation object **BURGER\_EQN** and then select **View/Coefficient Tests/Wald Coefficient Restrictions**.



In the resulting dialog box type in the null hypothesis using the notation  $C(1) = \beta_1$ ,  $C(2) = \beta_2$ ,  $C(3) = \beta_3$ , and so on. For the null hypothesis  $H_0 : \beta_3 = 1$ , we type **c(3) = 1**. Then click **OK**.



The following test results appear:

Wald Test:			
Equation: BURGER_EQN			
Test Statistic	Value	df	Probability
F-statistic	1.594091	(1, 72)	0.2108
Chi-square	1.594091	1	0.2067
Null Hypothesis Summary:			
Normalized Restriction (= 0)	t-value numerator	t-value denominator	Std. Err.
-1 + C(3)	0.862584	0.683195	

Note the following points:

1. The test is called a **Wald test**. The  $t$ -test and  $F$ -tests on the coefficients of regression equations belong to this class of tests. More details can be found on page 538 of the text.
2. In the last row of the output we can read off the numerator of the calculated  $t$ -value, namely  $b_3 - 1 = 0.8626$ , as well as its standard error  $se(b_3 - 1) = 0.6832$  that appears in the denominator. Of course,  $se(b_3 - 1) = se(b_3)$ .
3. Instead of reporting the calculated value  $t = 1.2626$ , EViews reports its square and calls it an  $F$ -value. That is,  $F = t^2 = 1.2626^2 = 1.594$ . There is a theorem that says a  $t_{(v)}$  random variable squared is equal to an  $F_{(1,v)}$  random variable. In words, the square of a  $t$  random variable with  $v$  degrees of freedom is equal to an  $F$  random variable with 1 degree of freedom in the numerator and  $v$  degrees of freedom in the denominator.
4. It is equally valid to perform a two-tail test with an  $F$ -distribution or a  $t$ -distribution. With one-tail tests squaring the  $t$ -value complicates matters. To avoid confusion, use the  $t$ -distribution.
5. The  $p$ -value for the  $F$ -test is obtained from the right tail of its distribution. Specifically,

$$P(F_{(1,72)} > 1.594) = 0.2108$$

Because of the relationship between the  $t$ - and  $F$ -distributions, this  $p$ -value is identical to that obtained for the two-tail  $t$ -test and the same test conclusion is reached, namely, there is insufficient evidence to reject  $H_0$  at a 5% level of significance.

6. The EViews output also reports a  $\chi^2$  (chi-square) value. We will say more about this value in Chapter 6.

## 5.6 SAVING COMMANDS

Throughout this chapter we have entered a number of commands in the upper display of the EViews window. It is a good idea to save these commands so that you have a record of them when you return to your work. To do so, highlight the commands and push **Ctrl+C**

```

EViews Student Version
File Edit Object View Proc Quick Options Window Help
scalar sales_f = c(1) + c(2)*5.5 + c(3)*1.2
scalar tc = @qtdist(0.975,72)
scalar beta2_low = c(2) - tc*@stderrs(2)
scalar beta2_up = c(2) + tc*@stderrs(2)
scalar beta3_low = c(3) - tc*@stderrs(3)
scalar beta3_up = c(3) + tc*@stderrs(3)
scalar pee = 2*@ctdist(-2.726,72)
scalar tc_1tail = @qtdist(0.05,72)
scalar t3 = (c(3) - 1)/@stderrs(3)
scalar pee3 = 1 - @ctdist(t3,72)

```

Then go **Object/New Object** and select the object **Text**. As a name for the object, enter **CHAP05\_CMDS**. After positioning the cursor within the **Text** dialog box push **Ctrl+V**. The following **Text** object will then be stored in your workfile

```

Text: CHAP05_CMDS Workfile: ANDY...
View Object Print Name Cut Copy Paste Find Replace
scalar sales_f = c(1) + c(2)*5.5 + c(3)*1.2
scalar tc = @qtdist(0.975,72)
scalar beta2_low = c(2) - tc*@stderrs(2)
scalar beta2_up = c(2) + tc*@stderrs(2)
scalar beta3_low = c(3) - tc*@stderrs(3)
scalar beta3_up = c(3) + tc*@stderrs(3)
scalar pee = 2*@ctdist(-2.726,72)
scalar tc_1tail = @qtdist(0.05,72)
scalar t3 = (c(3) - 1)/@stderrs(3)
scalar pee3 = 1 - @ctdist(t3,72)

```

## Keywords

@coefs	F-value	range: change
@ctdist	group: naming	regression output
@qtdistn	group: open	S.D. dependent variable
@stderrs	hypothesis testing	S.E. of regression
c object	interval estimates	sample
coefficient	least squares	sample: change
coefficient tests	NA	scalar
coefficient uncertainty in S.E	object: creating	spreadsheet
commands: saving	object: equation	standard errors
covariance matrix	object: group	Std. Error
descriptive statistics	object: name	test of significance
edit +/-	object:text	test: nonzero value
equation specification	page: naming	test: one-tail
estimate equation	page: resize	test: two-tails
forecast	proc	t-test
forecast sample	p-value (Prob.)	t-value (t-Statistic)
forecast standard error	quick/estimate equation	Wald test
F-test	range	workfile

# CHAPTER 6

## Further Inference in the Multiple Regression Model

### CHAPTER OUTLINE

- 6.1 *F* and Chi-Square Tests
  - 6.1.1 Testing significance: a coefficient
  - 6.1.2 Testing significance: the model
- 6.2 Testing in an Extended Model
  - 6.2.1 Estimating the model
  - 6.2.2 Testing: a joint  $H_0$ , 2 coefficients
  - 6.2.3 Testing: a single  $H_0$ , 2 coefficients
  - 6.2.4 Testing: a joint  $H_0$ , 4 coefficients
- 6.3 Including Nonsample Information
- 6.4 The RESET Test
  - 6.4.1 The short way
  - 6.4.2 The long way
- 6.5 Viewing the Correlation Matrix
  - 6.5.1 Collinearity: An exercise

### KEYWORDS

### 6.1 F AND CHI-SQUARE TESTS

In Chapter 5 we saw how to use EViews to test a null hypothesis about a single coefficient in a regression model. This test can be extended in two ways. We may want to test a **single null hypothesis** that involves two or more coefficients, or we might want to test a **joint null hypothesis** that specifies two or more restrictions on two or more coefficients. The choice of test statistic depends on whether the null hypothesis is single or joint and on whether the test is a one-tail test or a two-tail test. One-tail tests are only considered for single null hypotheses. In this case the relevant test statistic is  $t_{(N-K)}$ . For two-tail tests of single hypotheses, we can use either the test statistic  $t_{(N-K)}$  or the statistic  $F_{(1, N-K)}$ . The tests from each are equivalent because  $t_{(N-K)}^2 = F_{(1, N-K)}$ . An illustration was given in Section 5.5.3. Another test that can be used is a **chi-square test** that uses a **chi-square statistic** with one degree of freedom  $\chi_{(1)}^2$ . The value of this statistic is identical to  $F_{(1, N-K)}$ , but the test is different because a different distribution is used to compute the  $p$ -value. The  $F$ -test is an exact finite sample test suitable when the equation errors are normally distributed. The  $\chi^2$ -test is an approximate large sample test that does not require the normality assumption. For joint null hypotheses the  $t$ -test is no longer suitable, nor do we

consider one-tail tests. The alternative hypothesis  $H_1$  is that one or more of the restrictions in  $H_0$  does not hold. We can use the  $F$ -test or the  $\chi^2$ -test depending on whether or not we are invoking the normality assumption. The number of restrictions in  $H_0$  gives the numerator degrees of freedom for the  $F$ -statistic and the degrees of freedom for the  $\chi^2$ -statistic. The two tests are different, but the value of one statistic can be calculated from the other using the relationship  $F_{(J, N-K)} = \chi_{(J)}^2 / J$ . All these different possibilities are summarized in the following table.

**Test Statistics for Testing Coefficients in a Multiple Regression Model**

Test Type	Null Hypothesis		Statistic	Relationship
	Coefficients	Restrictions		
Single $H_0$ , 1 tail	$> 1$	1	$t_{(N-K)}$	
Single $H_0$ , 2 tails	$\geq 1$	1	$t_{(N-K)}$ or $F_{(1, N-K)}$ or $\chi_{(1)}^2$	$t_{(N-K)}^2 = F_{(1, N-K)} = \chi_{(1)}^2$
Joint $H_0$	$> 2$	$J \geq 2$	$F_{(J, N-K)}$ or $\chi_{(J)}^2$	$F_{(J, N-K)} = \chi_{(J)}^2 / J$

In the first part of this Chapter we use Andy's Burger Barn example to demonstrate how these various testing scenarios can be handled within EViews. Our main focus will be on  $F$ - and  $\chi^2$ -tests. You should check Chapter 5 for an introduction to the  $t$ -test.

The general formula for the  $F$ -value is

$$F = \frac{(SSE_R - SSE_U) / J}{SSE_U / (N - K)}$$

where  $SSE_R$  is the sum of squared errors from the model estimated assuming the restrictions in  $H_0$  hold and  $SSE_U$  is the sum of squared errors from the unrestricted model. The corresponding  $\chi^2$ -value is given by  $\chi^2 = J \times F$ . We can use EViews to compute  $F$  and  $\chi^2$  and their  $p$ -values automatically or we can use EViews to compute the restricted and unrestricted models, locate  $SSE_R$  and  $SSE_U$  on the output, and then calculate  $F$  and  $\chi^2$ .

### 6.1.1 Testing significance: a coefficient

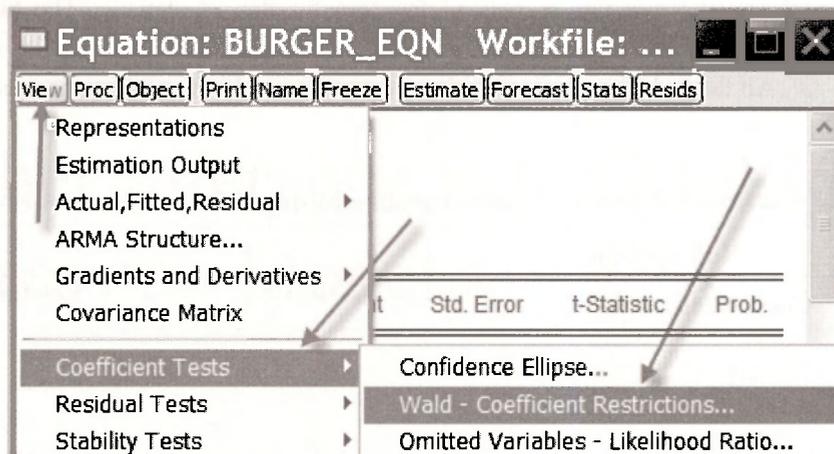
Our first example is to test  $H_0 : \beta_2 = 0$  against the alternative  $H_1 : \beta_2 \neq 0$  in the model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$

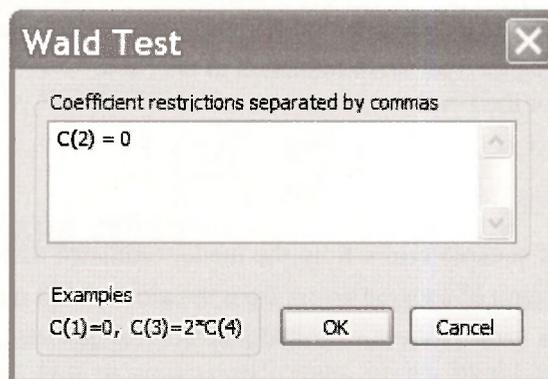
In other words, should  $PRICE$  be included in the equation? We used a  $t$ -test to perform this test in Section 5.5.1; we discovered we could read the result directly from the regression output. Let us see how we can do it using  $F$ - and  $\chi^2$ -tests.

### 6.1.1a Using EViews test option

Return to the workfile *andy.wf1* and open the equation object **BURGER\_EQN**. Select **View/Coefficient Tests/Wald Coefficient Restrictions**



In the dialog box that appears we type the null hypothesis  $H_0 : \beta_2 = 0$  as **C(2) = 0**. EViews uses the notation  $C(k)$  to denote the coefficient  $\beta_k$ . The order of the coefficients  $C(1)$ ,  $C(2)$ ,  $C(3)$ ,... is the order that they were specified in the **Equation Estimation** dialog box (and the order in which they appear in the regression output).



Clicking **OK** yields the output that appears below. You should note the following:

1. The test is called a **Wald test**. The  $t$ -test and  $F$ -tests on the coefficients of regression equations belong to this class of tests. More details can be found on page 538 of the text.
2. The **Normalized Restriction (=0)** in the bottom part of the table refers to the null hypothesis rearranged so that the right-hand side of the restriction in  $H_0$  is zero. In this particular example no rearrangement is necessary because the right-hand side of  $H_0 : \beta_2 = 0$  is already zero.
3. **Value** and **Std. Err.** of the **Normalized Restriction** refer to the estimated value of the left-hand side of the rearranged  $H_0$  and its standard error. In this case these values are  $b_2 = -7.907854$  and  $se(b_2) = 1.095993$ .

4. The calculated  $F$ - and  $\chi^2$ -values are approximately  $F = \chi^2 = 52.06$ . They are identical because there is only one restriction in  $H_0$  ( $J = 1$ ). And they are equal to the square of the  $t$ -value for testing this hypothesis. That is,

$$\left( \frac{b_2}{se(b_2)} \right)^2 = \left( \frac{-7.907854}{1.095993} \right)^2 = 52.06$$

5. The degrees of freedom (**df**) are (1,72) for the  $F$ -test and 1 for the  $\chi^2$ -test.  
 6. The reported  $p$ -values for each of the tests are both 0.0000. Thus, we reject  $H_0 : \beta_2 = 0$  at all reasonable significance levels.

Wald Test		test values		p-values	
Equation: BURGER_EQN					
Test Statistic	Value	df	Probability		
F-statistic	52.05971	(1, 72)	0.0000		
Chi-square	52.05971	1	0.0000		
Null Hypothesis Summary:		$b_2$	$se(b_2)$		
Normalized Restriction (= 0)		t Value	Std. Err.		
C(2)		-7.907854	1.095993		

**6.1.1b Using the formula for F**

To perform the test using the formula for  $F$  we need the quantities  $SSE_U$  and  $SSE_R$ . We can read  $SSE_U = 1718.943$  from the regression output:

R-squared	0.448258	Mean dependent var	77.37467
Adjusted R-squared	0.432932	S.D. dependent var	6.488537
S.E. of regression	4.886124	Akaike info criterion	6.049854
<b>Sum squared resid</b>	<b>1718.943</b>	Schwarz criterion	6.142553
Log likelihood	-223.8695	Hannan-Quinn criter.	6.086868
F-statistic	29.24786	Durbin-Watson stat	2.183037

After estimation EViews stores this quantity as **@ssr**, short for “sum of squared residuals”. Since the text uses  $SSR$  for “regression sum of squares”, this notation can be confusing. Be careful! We can call it something more familiar by using the EViews command

```
scalar sse_u = @ssr
```

To find  $SSE_R$  we estimate the model under the assumption that  $H_0 : \beta_2 = 0$  is true. This model is

$$SALES = \beta_1 + \beta_3 ADVERT + e$$

Using EViews to estimate this model we find  $SSE_R = 2961.827$  which can be read directly from the regression output.

S.E. of regression	6.369692	Akai
Sum squared resid	2961.827	Sc
Log likelihood	-244.2731	Ha

To save this quantity using a convenient name, we use the EViews command

```
scalar sse_r = @ssr
```

Then, the required  $F$ -value is given by

```
scalar f_val = (sse_r - sse_u)/(sse_u/(75-3))
```

A check of this calculated value shows it is the same as that obtained using EViews' test option.

```
Scalar F_VAL = 52.05970978
```

To finalize the test we need either the  $p$ -value or the critical value. These values can be obtained using EViews commands for the  $F$  distribution function and the  $F$  quantile function. For the  $p$ -value we have

```
scalar pp = 1 - @cfdist(f_val,1,72)
```

```
Scalar PP = 4.42399783473e-010
```

For the critical value we have

```
scalar fc = @qfdist(0.95,1,72)
```

```
Scalar FC = 3.97389699161
```

Again, we are led to reject  $H_0$ . Note that that the  $p$ -value is the same as that on page 121 of the text where a  $t$ -test was performed. Also the critical value is equal to the square of the  $t$  critical value:  $F_c = 3.9739 = t_c^2 = 1.99346^2$

What about the  $\chi^2$ -test? The  $\chi^2$ -value is the same as the  $F$ , that is,  $\chi^2 = 52.06$ . Its  $p$  and critical values can be found using EViews commands for the  $\chi^2$  distribution function and the  $\chi^2$  quantile function.

```
scalar chi_pee = 1 - @cchisq(f_val,1)
```

```
Scalar CHI_PEE = 5.38347144641e-013
```

```
scalar chic = @qchisq(0.95,1)
```

```
Scalar CHIC = 3.84145882061
```

Note that the  $p$ -values from the  $F$ - and  $\chi^2$ -tests are different, although the test conclusion is clearly the same.

### 6.1.2 Testing significance: the model

The  $F$ -test for testing the significance of a model is given special prominence in the regression output. In the context of Andy's Burger Barn the hypotheses for this test are

$$H_0 : \beta_2 = 0 \text{ and } \beta_3 = 0 \quad H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

The null hypothesis is a joint one because there are 2 restrictions  $\beta_2 = 0$  and  $\beta_3 = 0$ . The restricted model that assumes  $H_0$  is true is

$$SALES = \beta_1 + e$$

This model has no explanatory variables. Testing the significance of a model is equivalent to testing whether any of the explanatory variables influences the dependent variable. The sum of squared errors for the unrestricted model is the same as before,  $SSE_U = 1718.943$ . The sum of squared errors for the restricted model is equal to the sum of squared deviations of  $SALES$  around its mean, also known as the total sum of squares ( $TSS$ ). This result holds because the restricted least squares estimator for  $\beta_1$  is the sample mean for  $SALES$ . Note that  $TSS$  for a series  $y$  is given by

$$TSS = \sum (y_i - \bar{y})^2 = \sum y_i^2 - N\bar{y}^2$$

The EViews functions for  $\bar{y}$  and  $\sum y_i^2$  are `@mean(.)` and `@sumsq(.)`, respectively. Using this information, a sequence of EViews commands that computes the required  $F$ -value and its  $p$ -value are

```
scalar tss = @sumsq(sales) - 75*(@mean(sales))^2
scalar f_model = ((tss - sse_u)/2)/(sse_u/(75-3))
scalar p_model = 1 - @cdfist(f_model,2,72)
```

which yield

```
Scalar F_MODEL = 29.2478594797
```

```
Scalar P_MODEL = 5.04085662101e-010
```

In practice there is no need to go through this sequence of calculations. The  $F$ - and  $p$ -values are automatically reported on the **BURGER\_EQN** regression output.

R-squared	0.448258	Mean dependent var	77.37467
Adjusted R-squared	0.432932	S.D. dependent var	6.488537
S.E. of regression	4.886124	Akaike info criterion	6.049854
Sum squared resid	1718.943	Schwarz criterion	6.142553
Log likelihood	-223.8695	Hannan-Quinn criter.	6.086868
F-statistic	29.24786	Durbin-Watson stat	2.183037
Prob(F-statistic)	0.000000		

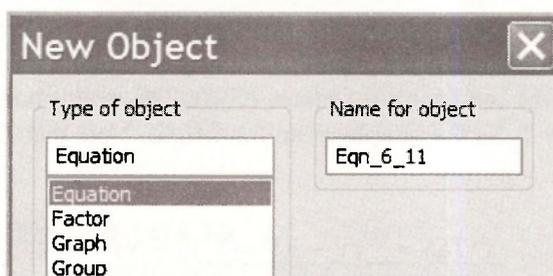
## 6.2 TESTING IN AN EXTENDED MODEL

### 6.2.1 Estimating the model

On page 140 of the text Andy's Burger Barn model is extended to also include the square of advertising expenditure as one of the explanatory variables. The new model is

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$$

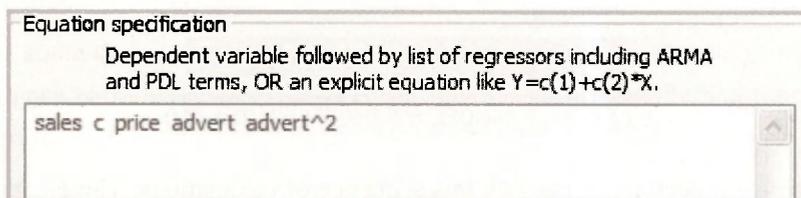
To estimate this model we begin by selecting **Object/New Object** and choose **Equation** from the menu of objects. We have named the equation **EQN\_6\_11** in line with equation number on page 141 of the text.



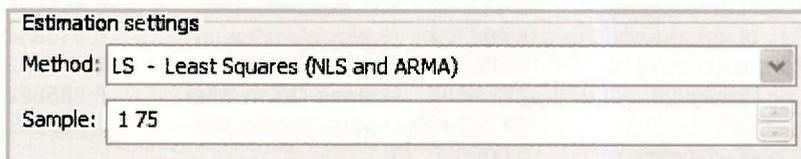
The names of the series are entered in the **Equation specification** dialog box, with the dependent variable *SALES*, appearing first, followed by the constant *C*, then the explanatory variables *PRICE*, *ADVERT* and *ADVERT*<sup>2</sup>. Notice that it is legitimate to write simply **advert^2** for *ADVERT*<sup>2</sup>. An alternative way is to define a new series, say

**series advert2 = advert^2**

and include **advert2** as one of the explanatory variables.



The **Estimation settings** remain as before with **Least Squares** being the **Method** and **1 75** for the **Sample**.

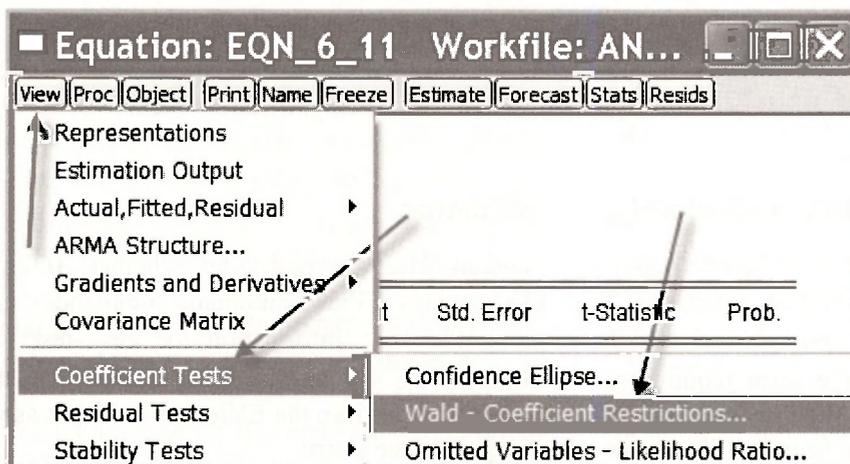


The following results appear. Check them against those on page 141 of the text.

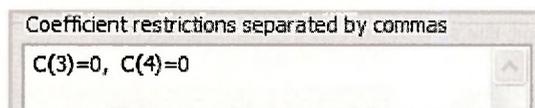
Dependent Variable: SALES				
Method: Least Squares				
Date: 11/27/07 Time: 12:55				
Sample: 1 75				
Included observations: 75				
	Coefficient	Std. Error	t-Statistic	Prob.
C	109.7190	6.799045	16.13742	0.0000
PRICE	-7.640000	1.045939	-7.304442	0.0000
ADVERT	12.15124	3.556164	3.416950	0.0011
ADVERT^2	-2.767963	0.940624	-2.942688	0.0044

### 6.2.2 Testing: a joint $H_0$ , 2 coefficients

Since advertising appears twice in the equation, as  $ADVERT$  and as  $ADVERT^2$ , to test whether advertising has an effect on sales we need to test  $H_0: \beta_3 = 0$  and  $\beta_4 = 0$  against the alternative  $H_1: \beta_3 \neq 0$  and/or  $\beta_4 \neq 0$ , as described on page 142 of the text. The null hypothesis is called a joint null hypothesis because it contains two restrictions. To get EViews to perform the test go to the EViews workfile and open the equation object **EQN\_6\_11**. Then, select **View/Coefficient Tests/Wald Coefficient Restrictions**



In the resulting **Wald Test** dialog box enter the two restrictions **C(3) = 0, C(4) = 0** and click **OK**.



Wald Test:			
Equation: EQN_6_11			
Test Statistic	Value	df	Probability
F-statistic	8.441360	(2, 71)	0.0005
Chi-square	16.88272	2	0.0002

$\chi^2 = 2 \times F$

<b>Estimates and st. errors of LHS of null restrictions</b>		
Null Hypothesis Summary:		
Normalized Restriction (= 0)	Value	Std. Err.
C(3)	12.15124	3.556164
C(4)	-2.767963	0.940624

The above output contains the following information.

1. The **Normalized Restriction (=0)** in the bottom part of the table refers to the two restrictions in the null hypothesis rearranged so that their right-hand sides are zero. In this particular example no rearrangement is necessary because the right-hand sides are already zero and the left-hand sides are simply  $\beta_3$  and  $\beta_4$ .
2. **Value** and **Std. Err.** of the **Normalized Restriction** refer to the estimated values of the left-hand sides of the rearranged restrictions and their standard errors. The values are  $b_3 = 12.151$  and  $b_4 = -2.7680$ , with standard errors  $se(b_3) = 3.556$  and  $se(b_4) = 0.9406$ .
3. The calculated  $F$ - and  $\chi^2$ -values for testing  $H_0$  are  $F = 8.441$  and  $\chi^2 = 16.883$ . Because there are 2 restrictions,  $\chi^2 = 2 \times F$ .
4. The degrees of freedom (**df**) are (2,71) for the  $F$ -test and 2 for the  $\chi^2$ -test.
5. The reported  $p$ -values for  $F$ - and  $\chi^2$ -tests are 0.0005 and 0.0002, respectively. Thus, we reject  $H_0 : \beta_3 = 0$  and  $\beta_4 = 0$  at all conventional significance levels.

### 6.2.3 Testing: a single $H_0$ , 2 coefficients

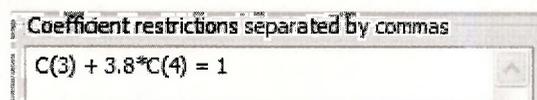
On page 143-4 of the text both a  $t$ -test and an  $F$ -test are used to test whether  $ADVERT = 1.9$  is the optimal level of advertising. We will show how EViews automatic commands can be used to perform  $F$ - and  $\chi^2$ -tests and how, along the way, information for the  $t$ -test is produced. Performing the  $t$ -test requires one to compute the standard error for a linear function of two coefficients. We illustrate how this value can be read from the EViews test output as well as how to calculate it from the EViews coefficient covariance matrix.

The null and alternative hypotheses are

$$H_0 : \beta_3 + 3.8\beta_4 = 1$$

$$H_1 : \beta_3 + 3.8\beta_4 \neq 1$$

In the **Wald Test** dialog box the restriction in  $H_0$  is entered as



which produces the following test output

Wald Test			
Equation: EQN_6_11			
Test Statistic	Value	df	Probability
F-statistic	0.936195	(1, 71)	0.3365
Chi-square	0.936195	1	0.3333
<i>Estimate and st error of LHS of normalized null</i>			
Null Hypothesis Summary:			
Normalized Restriction (= 0)	Value	Std. Err.	
-1 + C(3) + 3.8*C(4)	0.632976	0.654190	

To write  $H_0: \beta_3 + 3.8\beta_4 = 1$  as a **Normalized Restriction (=0)**, EViews moves 1 from the right-hand side to the left-hand side, giving the normalization  $H_0: -1 + \beta_3 + 3.8\beta_4 = 0$ . **Value** is an estimate of the left-hand side, namely,  $-1 + b_3 + 3.8b_4 = 0.632976$ . **Std. Err.** refers to  $se(-1 + b_3 + 3.8b_4) = 0.65419$ . It is calculated by EViews using the formula

$$\begin{aligned} se(-1 + b_3 + 3.8b_4) &= \sqrt{\text{var}(b_3 + 3.8b_4)} \\ &= \sqrt{\text{var}(b_3) + 3.8^2 \times \text{var}(b_4) + 2 \times 3.8 \times \text{cov}(b_3, b_4)} \end{aligned}$$

The calculated  $F$ - and  $\chi^2$ -values for testing  $H_0$  are  $F = \chi^2 = 0.9362$ . They are both the same because there is only one restriction. The degrees of freedom (**df**) are (1,71) for the  $F$ -test and 1 for the  $\chi^2$ -test. The reported  $p$ -values for  $F$ - and  $\chi^2$ -tests are 0.3363 and 0.3333, respectively. Thus, we do not reject  $H_0$  at a 5% significance level. The  $p$ -values can be confirmed with the commands

```
scalar p_f_o = 1 - @cfdist(0.936195, 1,71)
scalar p_chi_o = 1 - @cchisq(0.936195, 1)
```

Notice that the EViews output also gives enough information to perform a  $t$ -test. The required test value is given by

$$t = \frac{\text{Value}}{\text{Std. Err.}} = \frac{0.632976}{0.654190} = 0.9676$$

Because  $t^2 = 0.9676^2 = 0.936 = F$ , for a two-tail test there is no need to consider both  $t$ - and  $F$ -tests. Both give the same result. However, the information for the  $t$ -test is useful for one-tail tests as described on page 145 of the text.

### 6.2.3a Standard error for a linear function of coefficients

It is instructive to see how to compute the standard error  $se(b_3 + 3.8b_4) = 0.65419$  from the least squares covariance matrix. After estimating equation (6.11) and saving it as the equation object **EQN\_6\_11**, the covariance matrix for the least squares estimates is stored as a symmetric matrix called **eqn\_6\_11.@cov**. A symmetric matrix is a square array of numbers where the values above the diagonal are equal to the corresponding ones below the diagonal. If the columns are made rows and the rows are made columns, we get the same array. A covariance matrix is always symmetric because  $cov(b_k, b_\ell) = cov(b_\ell, b_k)$  for any two coefficients  $b_\ell$  and  $b_k$ . EViews refers to symmetric matrix objects as **sym**. Thus, to list the least squares covariance matrix in our workfile with the name **covb**, we use the command

```
sym covb = eqn_6_11.@cov
```

The command to compute  $\widehat{var}(b_3 + 3.8b_4) = \widehat{var}(b_3) + 3.8^2 \times \widehat{var}(b_4) + 2 \times 3.8 \times \widehat{cov}(b_3, b_4)$  and save it in the workfile with name **vee** is

```
scalar vee = covb(3,3) + 3.8^2*covb(4,4) + 2*3.8*covb(3,4)
```

and the standard error, called **se\_o**, is

```
scalar se_o = @sqrt(vee)
```

Following these steps will give the value  $se(b_3 + 3.8b_4) = 0.65419$ .

### 6.2.3b Using restricted and unrestricted SSE

On page 144 of the text, the  $F$ -value for testing the optimality of advertising expenditure is computed using  $SSE_U$  and  $SSE_R$ . As we have seen, it is more easily computed using EViews automatic test option. Nevertheless, we will show you how the values for  $SSE_U$  and  $SSE_R$  can be obtained. The value  $SSE_U = 1532.084$  is located from the output for **EQN\_6\_11**.

S.E. of regression	4.645283	Akaike in
Sum squared resid	1532.084	Schwarz
Log likelihood	-219.5540	Hannan

The value  $SSE_R = 1552.286$  is obtained by estimating the model

$$(SALES - ADVERT) = \beta_1 + \beta_2 PRICE + \beta_3 (ADVERT^2 - 3.8 \times ADVERT) + e$$

To estimate this model we use the following **Equation specification**.

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

(sales-advert) = c(1) + c(2)\*price + c(4)\*(advert^2 - 3.8\*advert)

Take another look at this box. The way the equation is entered is very different from what we have seen so far. Before when we specified the equation we simply listed the dependent variable

followed by the constant and the explanatory variables. Here we have written out the equation in full using C(1), C(2) and C(4) to denote  $\beta_1$ ,  $\beta_2$  and  $\beta_4$ . This is another way that an equation can be specified in the **Equation specification** dialog box. It is convenient in this instance because of the way we have rearranged the equation. It produces the following output.

Dependent Variable: SALES-ADVERT				
Method: Least Squares				
Sample: 1 75				
Included observations: 75				
(SALES-ADVERT) = C(1) + C(2)*PRICE + C(4)*(ADVERT^2 - 3.8*ADVERT)				
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	110.3590	6.763803	16.31611	0.0000
C(2)	-7.603104	1.044780	-7.277227	0.0000
C(4)	-2.876515	0.933496	-3.081445	0.0029
R-squared	0.480719	Mean dependent var	75.53067	
Adjusted R-squared	0.466295	S.D. dependent var	6.355780	
S.E. of regression	4.643224	Akaike info criterion	5.947873	
Sum squared resid	1552.286	Schwarz criterion	6.040573	
Log likelihood	-220.0432	Hannan-Quinn criter.	5.984887	

For testing purposes, the value  $SSE_R = 1552.286$  is of interest. However, notice also that the coefficients are listed as C(1), C(2) and C(4) instead of by the names of the variables to which they are attached. In this case there are not unambiguous variable names that can be attached to the coefficients.

### 6.2.4 Testing: a joint $H_0$ , 4 coefficients

The final example of a test using the extended hamburger model is on page 145 of the text. Here we are concerned with testing the joint null hypothesis

$$H_0: \beta_3 + 3.8\beta_4 = 1 \quad \text{and} \quad \beta_1 + 6\beta_2 + 1.9\beta_3 + 3.61\beta_4 = 80$$

They are entered in the **Wald Test** dialog box in the following way.

Coefficient restrictions separated by commas	
C(1) + 6*C(2) + 1.9*C(3) + 3.61*C(4) = 80,	
C(3) + 3.8*C(4) = 1	

In the output that follows EViews has written these restrictions in the normalized formats  $-1 + \beta_3 + 3.8\beta_4 = 0$  and  $-80 + \beta_1 + 6\beta_2 + 1.9\beta_3 + 3.61\beta_4 = 0$ . Note that the EViews output has abbreviated the latter of these two restrictions. Their estimated values and the corresponding standard errors found in the bottom part of the output are

$$-1 + b_3 + 3.8b_4 = 0.632976 \quad \text{se}(b_3 + 3.8b_4) = 0.65419$$

$$-80 + b_1 + 6b_2 + 1.9b_3 + 3.61b_4 = -3.025963 \quad \text{se}(b_1 + 6b_2 + 1.9b_3 + 3.61b_4) = 0.917713$$

As expected, the values for the restriction considered in the previous section have not changed.

Wald Test: Equation: EQN_6_11			
Test Statistic	Value	df	Probability
F-statistic	5.741229	(2, 71)	0.0049
Chi-square	11.48246	2	0.0032
Null Hypothesis Summary:			
Normalized Restriction (= 0)	Value	Std. Err.	
-80 + C(1) + 6*C(2) + 1.9*C(3) ...	-3.025963	0.917713	
-1 + C(3) + 3.8*C(4)	0.632976	0.654190	

The test values  $F = 5.7413$  and  $\chi^2 = 11.482$ , and their respective  $p$ -values of 0.0049 and 0.0032, lead to rejection of  $H_0$  at a 5% significance level.

### 6.3 INCLUDING NONSAMPLE INFORMATION

The model used on page 146 of the text to illustrate the inclusion of nonsample information is the demand for beer equation

$$\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) + e$$

where  $Q$  is quantity demanded,  $PB$  is the price of beer,  $PL$  is the price of liquor,  $PR$  is the price of remaining goods and services and  $I$  is income. The data are stored in the file *beer.wfl*. Before proceeding with estimation, we check the summary statistics in Table 6.1. Open the file, create a group of variables as described in Chapter 5, and select **View/Descriptive Stats/Common Sample**.

Group: BEER_VARS Workfile: BEER::Untitled\									
View	Proc	Object	Print	Name	Freeze	Sample	Sheet	Stats	Spec
	Q	PB	PL	PR	I				
Mean	56.11333	3.080000	8.367333	1.251333	32601.80				
Median	54.90000	3.110000	8.385000	1.180000	32457.00				
Maximum	81.70000	4.070000	9.520000	1.730000	41593.00				
Minimum	44.30000	1.780000	6.950000	0.670000	25088.00				
Std. Dev.	7.857381	0.642195	0.769635	0.298314	4541.966				

The nonsample information, that economic agents do not suffer from “money illusion”, can be expressed as

$$\beta_4 = -\beta_2 - \beta_3 - \beta_5$$

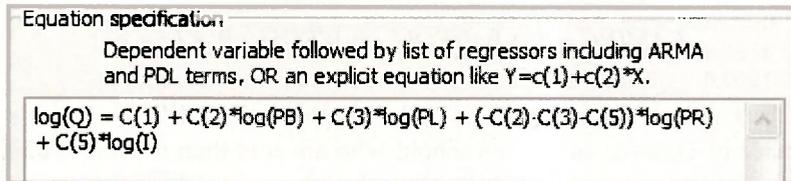
Restricted least squares estimates of the coefficients that satisfy this restriction incorporate the nonsample information.

Several examples of restricted least squares estimation were given in the previous section. Each time we estimate a model assuming a null hypothesis is true we are finding restricted least squares estimates. In Section 6.2.4 we used the restrictions to rearrange the equation, and estimated the rearranged equation. The same thing can be done in this case. Indeed, the rearranged equation appears as (6.18) and (6.19) in the text. As an exercise, we recommend that you use EViews to estimate (6.18) and confirm the results presented in (6.19).

To broaden your EViews experience, we will do it another way. Instead of estimating the rearranged equation, it is possible to simply substitute the restriction into the equation. EViews is smart enough to estimate it without you worrying about how to rearrange it. Substituting the restriction in to the equation yields

$$\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + (-\beta_2 - \beta_3 - \beta_5) \ln(PR) + \beta_5 \ln(I) + e$$

This equation can be written into the **Equation specification** dialog box as follows



Notice that the equation has been written in full. It is not just a list of variables. The resulting output follows. The values are consistent with those in equation (6.19) of the text.

Dependent Variable: LOG(Q)				
Method: Least Squares				
Sample: 1 30				
Included observations: 30				
LOG(Q) = C(1) + C(2)*LOG(PB) + C(3)*LOG(PL) + (-C(2)-C(3)-C(5))				
*LOG(PR) + C(5)*LOG(I)				
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-4.797798	3.713905	-1.291847	0.2078
C(2)	-1.299386	0.165738	-7.840022	0.0000
C(3)	0.186816	0.284383	0.656916	0.5170
C(5)	0.945829	0.427047	2.214813	0.0357

The value for  $b_4^*$  can be retrieved using the command

$$c(4) = -c(2) - c(3) - c(5)$$

Checking the C object yields the complete set of estimates

	C1
	La
R1	-4.797798
R2	-1.299386
R3	0.186816
R4	0.166742
R5	0.945829

## 6.4 THE RESET TEST

In Section 6.6 of the text, an example that relates family income to husband's education, wife's education and number of children is used to illustrate the effects of omitted and irrelevant variables. Various equations are estimated, summary statistics are given, including the correlation matrix of the variables, and the RESET test is introduced as a device for discriminating between models. We will not dwell on how to estimate the various equations. To do so is straightforward given the material you have covered so far in Chapters 5 and 6. Finding the correlation matrix for the variables is new, and important. It helps explain the effect of omitted and irrelevant variables and it is useful for detecting collinearity, a topic considered in Section 6.7. However, at this point it is convenient to defer reproducing Table 6.2 on page 149 until the next section where we also consider the correlation matrix for the variables in a gasoline consumption example. In this section our current focus is on how to get EViews to compute test statistic values for the RESET test. The model we consider is

$$FAMINC = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + \beta_4 KL6 + e$$

where *FAMINC* is family income, *HEDU* is husband's education, *WEDU* is wife's education and *KL6* is the number of children in the household who are less than 6 years old. To perform the RESET test we estimate this equation, obtain the predictions  $\widehat{FAMINC}$ , then estimate one or both of the following models

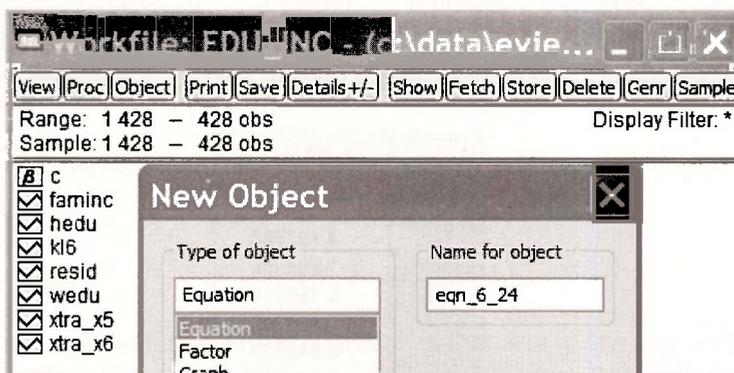
$$FAMINC = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + \beta_4 KL6 + \gamma_1 \widehat{FAMINC}^2 + e$$

$$FAMINC = \beta_1 + \beta_2 HEDU + \beta_3 WEDU + \beta_4 KL6 + \gamma_1 \widehat{FAMINC}^2 + \gamma_2 \widehat{FAMINC}^3 + e$$

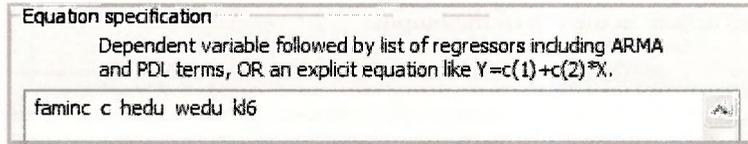
RESET tests are *F*-tests for  $H_0 : \gamma_1 = 0$  or  $H_0 : \gamma_1 = 0$  and  $\gamma_2 = 0$ . Rejection of either  $H_0$  implies the specification of the equation can be improved. The tests can be performed in the same way as the *F*-tests described earlier in this Chapter, but in this case EViews has special capabilities which require less effort. We will consider the special capabilities (the short way) as well as a long way that reinforces the fundamentals of the test.

### 6.4.1 The short way

Open the workfile *edu\_inc.wf1*. Create an equation object called *EQN\_6\_24*.



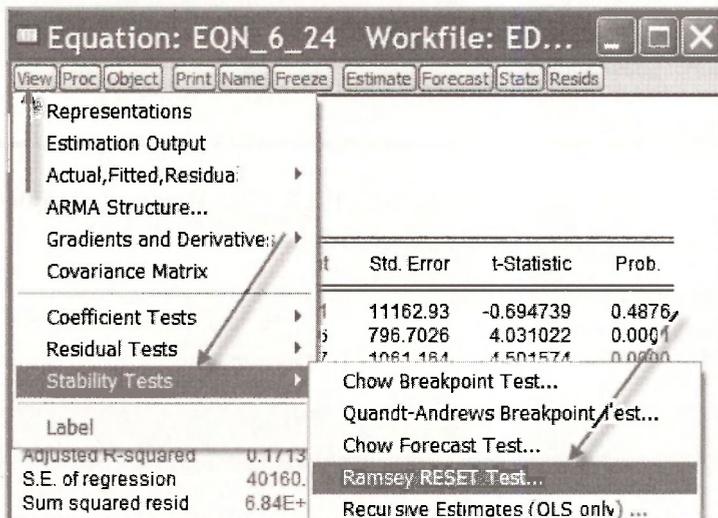
Enter the variables in the **Equation specification** dialog box.



The estimated equation, in line with (6.24) on page 150 of the text, is

Dependent Variable: FAMINC				
Method: Least Squares				
Date: 11/28/07 Time: 12:55				
Sample: 1 428				
Included observations: 428				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-7755.331	11162.93	-0.694739	0.4876
HEDU	3211.526	796.7026	4.031022	0.0001
WEDU	4776.907	1061.164	4.501574	0.0000
KL6	-14310.92	5003.928	-2.859937	0.0044

With this equation open, go to **View/Stability Tests/Ramsey RESET Test**.



A dialog box will ask you for the **number of fitted terms**. Inserting **1** leads to the model with  $\widehat{FAMINC}^2$ . Inserting **2** gives you the model with both  $\widehat{FAMINC}^2$  and  $\widehat{FAMINC}^3$  included.



Clicking **OK** gives detailed output from estimating the specified test equation. Because most of this output should be meaningful to you by now, we will focus just on the *F*- and *p*-values for the tests. These values appear at the top of the output.

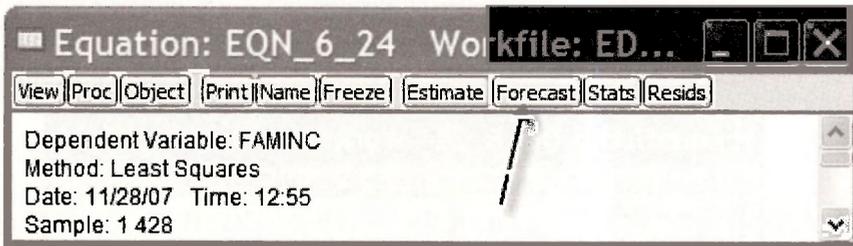
Ramsey RESET Test <span style="float: right;"><i>with 1 fitted term</i></span>			
F-statistic	5.983983	Prob. F(1,423)	0.0148

Ramsey RESET Test <span style="float: right;"><i>with 2 fitted terms</i></span>			
F-statistic	3.122582	Prob. F(2,422)	0.0451

In both cases the null hypothesis of no specification error is rejected at a 5% level of significance. Improvements to the model should be possible.

### 6.4.2 The long way

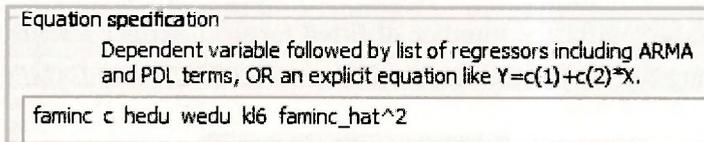
After estimating the basic equation go to **Forecast**.



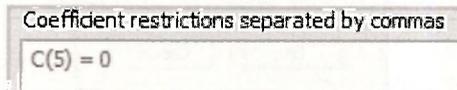
Give the forecasts a name such as **FAMINC\_HAT**. The **Forecast sample** is the same as the sample used for estimation, **1 428**.



The series **FAMINC\_HAT** will appear in your workfile. Estimate the equation with one fitted term.



In the output that follows, go to **View/Coefficient Tests/Wald – Coefficient Restrictions**. Insert **c(5) = 0** as the hypothesis to test. The test result will agree with that obtained the short way.



Now estimate the equation with two fitted terms.

<b>Equation specification</b> Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$ . faminc c hedu wedu kl6 faminc_hat^2 faminc_hat^3
---

In the output that follows, go to **View/Coefficient Tests/Wald – Coefficient Restrictions**. Insert  $c(5) = 0$ ,  $c(6) = 0$  as the hypothesis to test. The test result will agree with that obtained the short way.

Coefficient restrictions separated by commas $C(5) = 0, C(6) = 0$
--

## 6.5 VIEWING THE CORRELATION MATRIX

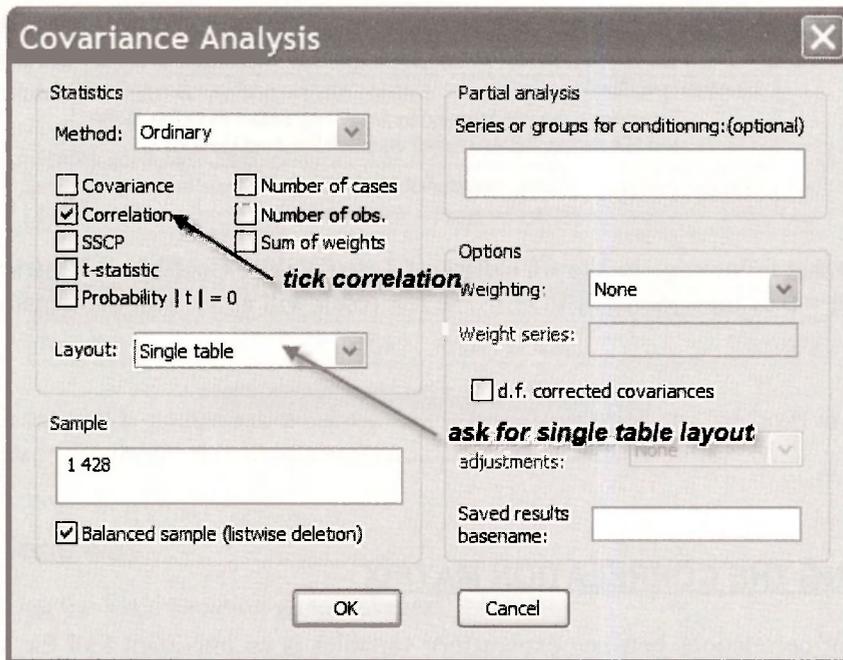
The matrix of correlations between explanatory variables is an important tool for assessing the sensitivity of results to inclusion or exclusion of variables and the likely causes of imprecise estimates. To obtain the correlation matrix for the variables in the file *edu\_inc.wfl*, we begin by creating a **group** object containing those variables. Suppose that group has been created and, in line with page 149 of the text, we call it **TABLE\_6\_2**.

List of series, groups, and/or series expressions faminc hedu wedu kl6 xtra x5 xtra x6
Name to identify object table_6_2 <span style="float: right;">24 characters maximum, 16 or fewer recommended</span>

To view the correlation matrix of the variables in the group, go to **View/Covariance Analysis**.

Group: TABLE_6_2 Workfile: EDU_INC::Untitled\												
View	Proc	Object	Print	Name	Freeze	Default	Sort	Transpose	Edit +/-	Smpl +/-	Title	Sample
Group Members				HEDU								
Spreadsheet				00000		12.00000		1.000000		11.01355		23.44492
Dated Data Table				00000		12.00000		0.000000		9.372190		22.59274
Graph...				00000		12.00000		1.000000		12.42620		23.16608
Descriptive Stats				00000		14.00000		1.000000		11.79830		25.61441
Covariance Analysis...				00000		12.00000		0.000000		11.44620		24.16108
N-Way Tabulation...				00000		16.00000		0.000000		11.69595		26.28512
				00000		12.00000		0.000000		5.067864		16.52149

In the **Covariance Analysis** dialog box that follows, you will find a large number of options. At present we are only interested in **correlation** presented as a **single table**. Our **method** is **ordinary** and we have a **balanced sample**.



Clicking **OK** produces the following table. Check it against Table 6.2 on page 149 of the text.

Covariance Analysis: Ordinary  
 Date: 11/29/07 Time: 02:00  
 Sample: 1 428  
 Included observations: 428

Correlation	FAMINC	HEDU	WEDU	KL6	XTRA X5	XTRA X6
FAMINC	1.000000					
HEDU	0.354684	1.000000				
WEDU	0.362328	0.594343	1.000000			
KL6	-0.071956	0.104877	0.129340	1.000000		
XTRA X5	0.289817	0.836168	0.517798	0.148742	1.000000	
XTRA_X6	0.351366	0.820563	0.799306	0.159522	0.900206	1.000000

### 6.5.1 Collinearity: an exercise

The final example in Chapter 6 is described on pages 154-5 of the text. It involves a model for gasoline consumption, used to illustrate the effects of collinearity. The data are stored in the workfile *cars.wfl*. Because the information provided in the text can all be obtained using EViews commands that we have covered earlier, this example is a good candidate for an exercise. Check your EViews skills by answering the following questions.

1. Estimate the two equations on page 155 of the text. Check your estimates, standard errors and *p*-values against those that are reported.

2. Consider the model

$$MPG = \beta_1 + \beta_2 CYL + \beta_3 ENG + \beta_4 WGT + e$$

Show that the test results for testing  $H_0 : \beta_2 = 0$  and  $\beta_3 = 0$  are

Test Statistic	Value	df	Probability
F-statistic	4.298023	(2, 388)	0.0142
Chi-square	8.596046	2	0.0136

3. Show that the RESET test result (with two-fitted terms) for this model is

Ramsey RESET Test:			
F-statistic	18.26092	Prob. F(2,386)	0.0000

What do you conclude?

4. Show that the correlation matrix for the variables is

Correlation	MPG	CYL	ENG	WGT
MPG	1.000000			
CYL	-0.777618	1.000000		
ENG	-0.805127	0.950823	1.000000	
WGT	-0.832244	0.897527	0.932994	1.000000

### Keywords

- |                      |                         |                               |
|----------------------|-------------------------|-------------------------------|
| @cchisq              | covariance matrix       | Prob(F-statistic)             |
| @cfdist              | descriptive statistics  | p-value (Prob.)               |
| @cov                 | df                      | RESET test                    |
| @mean                | fitted terms            | restricted least squares      |
| @qchisq              | forecast                | SSE: restricted               |
| @qfdist              | F-statistic             | SSE: unrestricted             |
| @sqrt                | F-test                  | stability tests               |
| @ssr                 | F-value                 | sum squared resid             |
| @sumsq               | group                   | sym                           |
| chi-square statistic | nonsample information   | symmetric matrix              |
| chi-square test      | normalized restriction  | testing significance          |
| collinearity         | null hypothesis: joint  | Wald coefficient restrictions |
| correlation matrix   | null hypothesis: single | Wald test                     |
| covariance analysis  |                         |                               |

# CHAPTER 7

## Nonlinear Relationships

### CHAPTER OUTLINE

- 7.1 Polynomials
- 7.2 Dummy Variables
  - 7.2.1 Creating dummy variables
- 7.3 Interacting Dummy Variables
- 7.4 Dummy Variables with Several Categories
- 7.5 Testing the Equivalence of Two Regressions
- 7.6 Interactions Between Continuous Variables
- 7.7 Log-Linear Models

### KEYWORDS

### 7.1 POLYNOMIALS

In microeconomics you studied “cost” curves and “product” curves that describe a firm. Total cost and total product curves are mirror images of each other, taking the standard “cubic” shapes shown in *POE* Figure 7.1. Average and marginal cost curves, and their mirror images, average and marginal product curves, take quadratic shapes, usually represented as shown in *POE* Figure 7.2. The slopes of these relationships are not constant and cannot be represented by regression models that are “linear in the variables.” However, these shapes are easily represented by polynomials. For example, if we consider the average cost relationship a suitable regression model is:

$$AC = \beta_1 + \beta_2 Q + \beta_3 Q^2 + e$$

This quadratic function can take the “U” shape we associate with average cost functions.

To illustrate we use a wage equation with wages a function of education and the worker’s years of experience. What we expect is that young, inexperienced workers will have relatively low wages; with additional experience their wages will rise, but the wages will begin to decline after middle age, as the worker nears retirement. To capture this life-cycle pattern of wages we introduce experience and experience squared to explain the level of wages

$$WAGE = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

To obtain the inverted-U shape, we expect  $\beta_3 > 0$  and  $\beta_4 < 0$ .

In EViews open the workfile *cps\_small.wf1*. Save it under a new name, *wage\_chap07.wf1*. To estimate the wage equation with quadratic experience we enter the command

**ls wage c educ exper exper^2**

This leads to the estimates

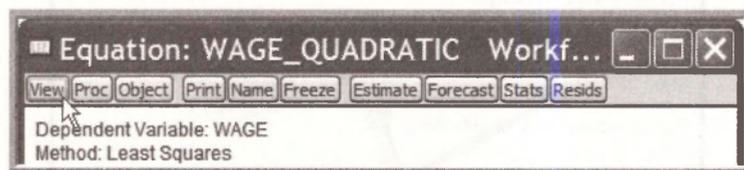
Dependent Variable: WAGE				
Method: Least Squares				
Sample: 1 1000				
Included observations: 1000				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.817697	1.054964	-9.306195	0.0000
EDUC	1.210072	0.070238	17.22821	0.0000
EXPER	0.340949	0.051431	6.629208	0.0000
EXPER^2	-0.005093	0.001198	-4.251513	0.0000
R-squared	0.270934	Mean dependent var		10.21302
Adjusted R-squared	0.268738	S.D. dependent var		6.246641
S.E. of regression	5.341743	Akaike info criterion		6.192973
Sum squared resid	28420.08	Schwarz criterion		6.212604
Log likelihood	-3092.487	Hannan-Quinn criter.		6.200434
F-statistic	123.3772	Durbin-Watson stat		0.491111

Interpretation in a model that is nonlinear in the variables requires some work. The effect of *EDUC* on expected *WAGE* is given by the coefficient 1.21. Each additional year of education is estimated to increase hourly wage by \$1.21, holding all else constant.

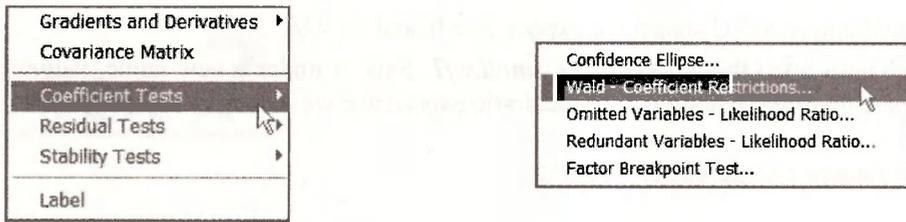
For experience, we must make use of *POE* equation (7.6). The marginal effect of experience on wage, holding education and other factors constant, is

$$\frac{\partial E(WAGE)}{\partial EXPER} = \beta_3 + 2\beta_4 EXPER$$

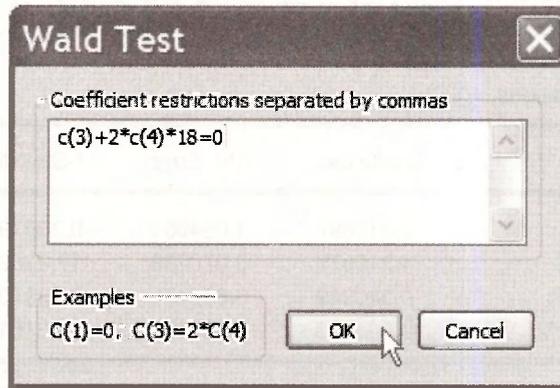
We can evaluate this marginal effect at a particular level of *EXPER*, such as *EXPER* = 18. To do this in EViews, from within the regression (which we named **WAGE\_QUADRATIC**) window, select **View/Coefficient Tests/Wald-Coefficient Restrictions**



Then



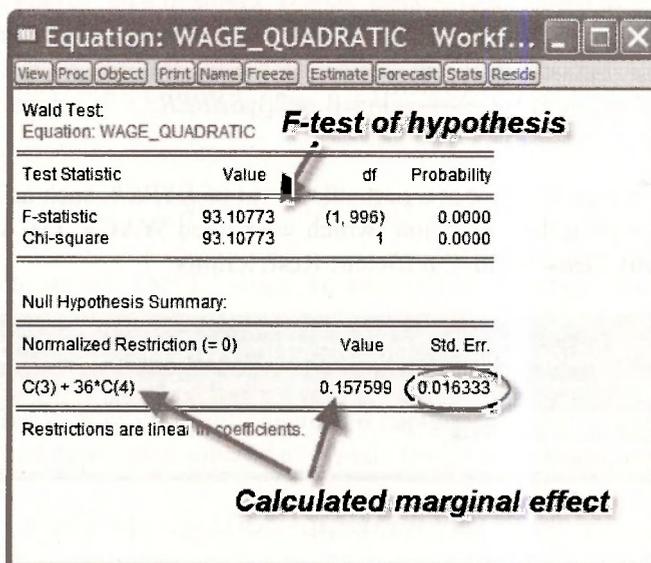
Into the test dialog box enter the equation for the marginal effect of experience



Recall that EViews saves the most recent regression results in the coefficient vector called **C**, and denoted in the EViews workfile by the object

**@ c**

Thus  $C(3) = b_3$  and  $C(4) = b_4$ . The **Coefficient restriction** we have entered is the marginal effect set equal to zero. This command will not only test the hypothesis that the marginal effect is zero, but it also computes the marginal effect and also computes the standard error of the marginal effect so that interval estimates can easily be created.



It may be useful to have a “picture” of the effect of experience on wage. Open a **Group** consisting of the variables *EDUC*, *EXPER* and *WAGE*. Do this by holding the **Ctrl**-key and clicking the series. Then double-click in the blue area. From the spreadsheet, click **View/Descriptive Stats/Common Table**

	EDUC	EXPER	WAGE
Mean	13.28500	18.78000	10.21302
Median	13.00000	18.00000	8.790000
Maximum	18.00000	52.00000	60.19000
Minimum	1.000000	0.000000	2.030000
Std. Dev.	2.468171	11.31882	6.246641

Note that experience ranges from 0 to 52 years. In the main EViews window click on **Sample** and in the dialog window enter

Sample range pairs (or sample object to copy)

1 53

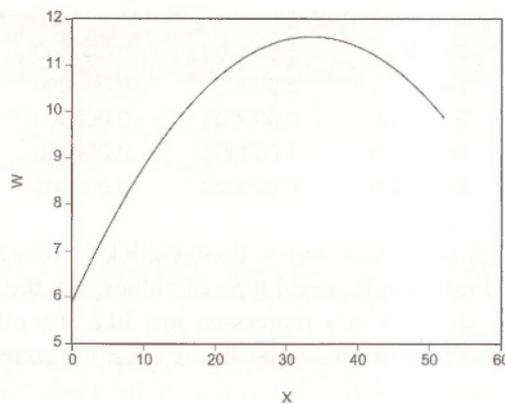
Create a series **X** that will represent years of experience in a plot.

**series x = @trend**

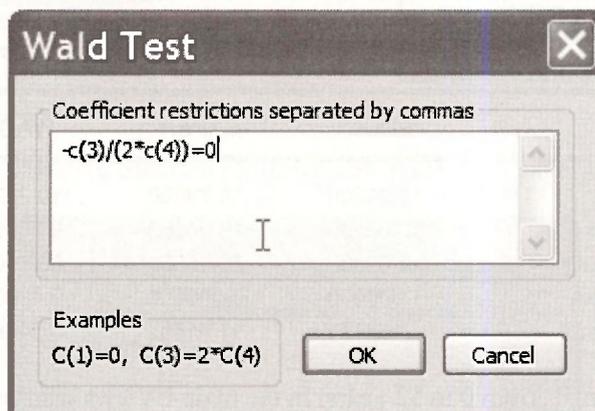
Using the estimated regression coefficients we can calculate the predicted wages of a person with 13 years of education ( $EDUC = 13$  is the median value) and experience **X**. The command is

**series w = c(1) + c(2)\*13+c(3)\*x+c(4)\*x^2**

Now graph the series **W** against the series **X**. Select from the main menu **Quick/Graph**. Then in the **Graph Options** dialog choose **XY Line**. The result is a nice visual.



The maximum wage occurs when  $\text{experience} = -\beta_3/2\beta_4$ . Open the saved regression **WAGE\_QUADRATIC**. Select **View/Coefficient Tests/Wald-Coefficient Restrictions**



This results in the calculation shown at the top of page 170 in *POE*. We estimate that the turning point in wages occurs at 33.47 years. Using the standard error **Std. Err.** we can compute an interval estimate if we choose.

Normalized Restriction (= 0)	Value	Std. Err.
$-1 / 2 * C(3) / C(4)$	33.47192	3.393876

Save your workfile and close.

## 7.2 DUMMY VARIABLES

Dummy variables are binary 0-1 variables indicating the presence or absence of some condition. Open workfile *utown.wfl* containing real estate transaction data from “University Town.”

obs	PRICE	SQFT	AGE	UTOWN	POOL	FPLACE
1	205.4520	23.46000	6.000000	0.000000	0.000000	1.000000
2	185.3280	20.03000	5.000000	0.000000	0.000000	1.000000
3	248.4220	27.77000	6.000000	0.000000	0.000000	0.000000
4	154.6900	20.17000	1.000000	0.000000	0.000000	0.000000
5	221.8010	26.45000	0.000000	0.000000	0.000000	1.000000

Opening a **Group** (hold **Ctrl** click each series, double-click in blue) with the variables we see that *PRICE*, *SQFT* and *AGE* contain the usual type of values, but the rest are 0’s and 1’s. These are dummy variables. They are used in a regression just like any other variables. Estimate the *POE* equation (7.13). Use **Quick/Estimate Equation** or enter the command

**ls price c utown sqft sqft\*utown age pool fplace**

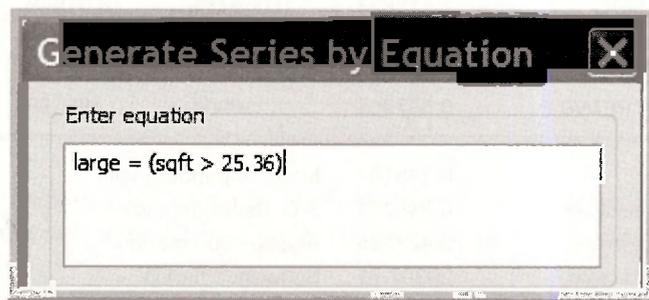
The result is

Dependent Variable: PRICE  
 Method: Least Squares  
 Sample: 1 1000  
 Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	24.49998	6.191721	3.956894	0.0001
UTOWN	27.45295	8.422582	3.259446	0.0012
SQFT	7.612177	0.245176	31.04775	0.0000
SQFT*UTOWN	1.299405	0.332048	3.913307	0.0001
AGE	-0.190086	0.051205	-3.712291	0.0002
POOL	4.377163	1.196692	3.657720	0.0003
FPLACE	1.649176	0.971957	1.696758	0.0901
R-squared	0.870570	Mean dependent var		247.6557
Adjusted R-squared	0.869788	S.D. dependent var		42.19273
S.E. of regression	15.22521	Akaike info criterion		8.290758
Sum squared resid	230184.4	Schwarz criterion		8.325112
Log likelihood	-4138.379	Hannan-Quinn criter.		8.303815
F-statistic	1113.183	Durbin-Watson stat		1.986480
Prob(F-statistic)	0.000000			

### 7.2.1 Creating dummy variables

Creating dummy variables is not exactly like creating any other variable. To create a dummy variable that is 1 for large houses, and zero otherwise we must decide what a large house is. The summary statistics for *SQFT* shows that the median house size in the sample is 2536 square feet. Because *SQFT* is measured in 100's of square feet, this is  $SQFT = 25.36$ . Suppose that houses larger than this size we take to be "large". On the workfile window click the **Genr** button and enter



What this does is create a new variable, **LARGE**, that takes the value 1 if the statement (**sqft > 25.36**) is true for a particular observation, and zero otherwise. Looking the first few observations we can see that this has worked.

obs	SQFT	LARGE
1	23.46000	0.000000
2	20.03000	0.000000
3	27.77000	1.000000
4	20.17000	0.000000
5	26.45000	1.000000

Search for **Help on Operators** for more information.

Save the workfile as *utown\_chap07.wfl* to maintain the original workfile, and close.

### 7.3 INTERACTING DUMMY VARIABLES

To illustrate further aspects of dummy variables, open *cps\_small.wfl*. Save the file under the name *cps\_small\_chap07.wfl*. Estimate the wage equation given in *POE* on page 176.

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma(BLACK \times FEMALE) + e$$

Use **Quick/Estimate Equation** or enter the command

**ls wage c educ black female black\*female**

Save the regression results by naming them **WAGE\_EQN**.

Dependent Variable: WAGE  
Method: Least Squares  
Sample: 1 1000  
Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.230327	0.967499	-3.338841	0.0009
EDUC	1.116823	0.069714	16.01998	0.0000
BLACK	-1.831240	0.895726	-2.044418	0.0412
FEMALE	-2.552070	0.359686	-7.095280	0.0000
BLACK*FEMALE	0.587905	1.216954	0.483096	0.6291
R-squared	0.248164	Mean dependent var		10.21302
Adjusted R-squared	0.245141	S.D. dependent var		6.246641
S.E. of regression	5.427245	Akaike info criterion		6.225728
Sum squared resid	29307.71	Schwarz criterion		6.250266
Log likelihood	-3107.864	Hannan-Quinn criter.		6.235054
F-statistic	82.10655	Durbin-Watson stat		0.480319
Prob(F-statistic)	0.000000			

In the regression results window, select **View/Representations**. The estimation equation is

Estimation Equation:

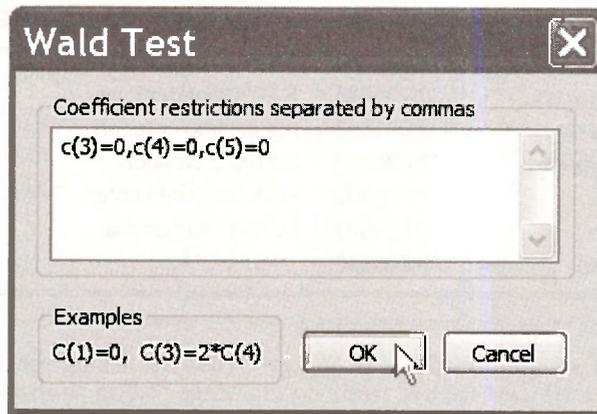
=====

$$WAGE = C(1) + C(2)*EDUC + C(3)*BLACK + C(4)*FEMALE + C(5)*BLACK*FEMALE$$

To test the hypothesis that neither race nor gender affects wage we formulate the null hypothesis

$$H_0 : \delta_1 = 0, \delta_2 = 0, \gamma = 0$$

In the regression window, select **View/Coefficient Tests/Wald - Coefficient Restrictions**. Using the equation representation we see that it is the coefficients C(3), C(4) and C(5) that we wish to test.



The result is

Test Statistic	Value	df	Probability
F-statistic	20.20346	(3, 995)	0.0000

Alternatively, to directly use the *F*-statistic,

$$F = \frac{(SSE_R - SSE_U) / J}{SSE_U / (N - K)}$$

we require the sum of squared least squares residuals from the unrestricted model and the model that is restricted by the null hypothesis. The WAGE regression is the unrestricted model in this case, and the *SSE<sub>U</sub>* is

**Sum squared resid      29307.71**

To obtain the restricted model we omit the variables *BLACK*, *FEMALE* and their interaction. Use the command

**equation wage\_r.ls wage c educ**

Recall that this notation assigns the name **WAGE\_R** to the regression object. Alternatively use **Quick/Estimate Equation** and then assign a name. The result is

Dependent Variable: WAGE  
 Method: Least Squares  
 Sample: 1 1000  
 Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.912181	0.966788	-5.080931	0.0000
EDUC	1.138517	0.071550	15.91225	0.0000
R-squared	0.202366	Mean dependent var		10.21302
Adjusted R-squared	0.201566	S.D. dependent var		6.246641
S.E. of regression	5.581693	Akaike info criterion		6.278859
Sum squared resid	31092.99	Schwarz criterion		6.288675
Log likelihood	-3137.430	Hannan-Quinn criter.		6.282590
F-statistic	253.1997	Durbin-Watson stat		0.391052
Prob(F-statistic)	0.000000			

The “restricted” sum of squared residuals is  $SSE_R = 31092.99$ . Using these values the  $F = 20.20$  can be calculated.

The critical value for the test comes from an  $F$ -distribution with 3 numerator degrees of freedom and 995 denominator degrees of freedom. The critical value is computed using **@qfdist**.

**scalar fc = @qfdist(.99,3,995)**

Scalar FC = 3.80134470284

## **7.4 DUMMY VARIABLES WITH SEVERAL CATEGORIES**

Open *cps\_small.wfl*. Save the file as *wage\_regions\_chap07.wfl*. Estimate the regression shown in Table 7.5 on *POE* page 178. On the command line enter

**equation regions.ls wage c educ black black\*female south midwest west**

The result is on the next page.

Dependent Variable: WAGE  
 Method: Least Squares  
 Sample: 1 1000  
 Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.455685	1.050990	-2.336544	0.0197
EDUC	1.102462	0.069986	15.75256	0.0000
BLACK	-1.607664	0.903432	-1.779507	0.0755
FEMALE	-2.500920	0.359975	-6.947490	0.0000
BLACK*FEMALE	0.646463	1.215208	0.531977	0.5949
SOUTH	-1.244281	0.479427	-2.595348	0.0096
MIDWEST	-0.499562	0.505628	-0.988003	0.3234
WEST	-0.546183	0.515398	-1.059732	0.2895
R-squared	0.253458	Mean dependent var		10.21302
Adjusted R-squared	0.248190	S.D. dependent var		6.246641
S.E. of regression	5.416272	Akaike info criterion		6.224660
Sum squared resid	29101.31	Schwarz criterion		6.263922
Log likelihood	-3104.330	Hannan-Quinn criter.		6.239583
F-statistic	48.11340	Durbin-Watson stat		0.495517
Prob(F-statistic)	0.000000			

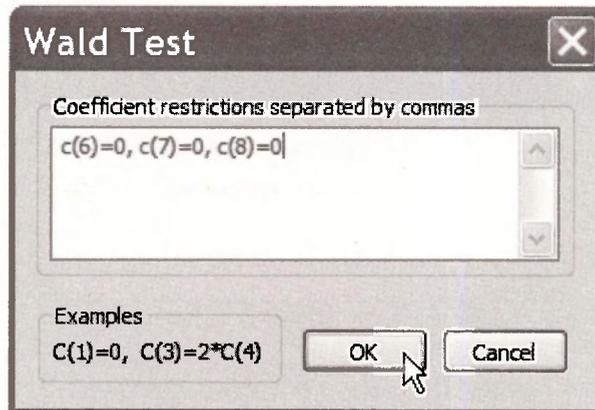
Select **View/Representations** in the regression window. We see that the estimation equation is

Estimation Equation:

$$\text{WAGE} = C(1) + C(2)*\text{EDUC} + C(3)*\text{BLACK} + C(4)*\text{FEMALE} + C(5)*\text{BLACK*FEMALE} + C(6)*\text{SOUTH} + C(7)*\text{MIDWEST} + C(8)*\text{WEST}$$

Test the null hypothesis that there are no regional differences by selecting **View/Coefficient Tests/Wald – Coefficient Restrictions**.

Enter the null hypothesis that coefficients 6, 7 and 8 are zero.



The result is

Wald Test:  
Equation: REGIONS

Test Statistic	Value	df	Probability
F-statistic	2.345176	(3, 992)	0.0714
Chi-square	7.035529	3	0.0708

The  $F$ -statistic's  $p$ -value shows that we reject the null hypothesis of no regional differences at the 10% level of significance, but not at the 5% level. The Chi-square statistic is an alternative approach to the test.

To construct the  $F$ -statistic directly we need the “unrestricted” sum of squared residuals in **REGIONS** model above. The “restricted” sum of squared errors comes from the model omitting the regional dummies. We obtained this result in Section 7.3. But it is easy to replicate them using

**equation regions\_rest.ls wage c educ black female black\*female**

## **7.5 TESTING THE EQUIVALENCE OF TWO REGRESSIONS**

The Chow test is illustrated using *cps\_small.wf1* in Section 7.3.3 of *POE*. The *WAGE* model in *POE* equation (7.16) is obtained by interacting the variable *SOUTH* with the variables *EDUC*, *BLACK*, *FEMALE* and *BLACK*×*FEMALE*

The estimation can be carried out using the command

**ls wage c educ black female black\*female south educ\*south black\*south  
female\*south black\*female\*south**

The result is shown on the next page.

Dependent Variable: WAGE  
 Method: Least Squares  
 Sample: 1 1000  
 Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.577536	1.151332	-3.107301	0.0019
EDUC	1.165847	0.082408	14.14719	0.0000
BLACK	-0.431165	1.348249	-0.319796	0.7492
FEMALE	-2.754044	0.425705	-6.469368	0.0000
BLACK*FEMALE	0.067320	1.906318	0.035314	0.9718
SOUTH	1.302260	2.114735	0.615803	0.5382
EDUC*SOUTH	-0.191725	0.154240	-1.243036	0.2141
BLACK*SOUTH	-1.744432	1.826695	-0.954966	0.3398
FEMALE*SOUTH	0.911939	0.795976	1.145686	0.2522
BLACK*FEMALE*SOUTH	0.542833	2.511154	0.216169	0.8289
R-squared	0.255731	Mean dependent var		10.21302
Adjusted R-squared	0.248965	S.D. dependent var		6.246641
S.E. of regression	5.413481	Akaike info criterion		6.225611
Sum squared resid	29012.71	Schwarz criterion		6.274689
Log likelihood	-3102.806	Hannan-Quinn criter.		6.244264
F-statistic	37.79605	Durbin-Watson stat		0.499707
Prob(F-statistic)	0.000000			

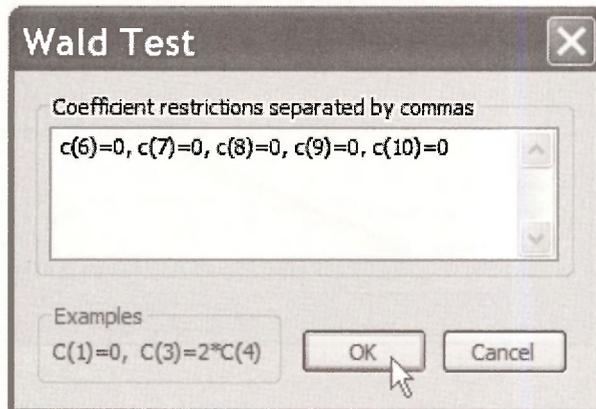
Select **View/Representations** to see the estimation equation

Estimation Equation:

=====

$$WAGE = C(1) + C(2)*EDUC + C(3)*BLACK + C(4)*FEMALE + C(5)*BLACK*FEMALE + C(6)*SOUTH + C(7)*EDUC*SOUTH + C(8)*BLACK*SOUTH + C(9)*FEMALE*SOUTH + C(10)*BLACK*FEMALE*SOUTH$$

To test the null hypothesis that wages for the *SOUTH* are no different from the rest of the country we test that coefficients 6-10 are zero.

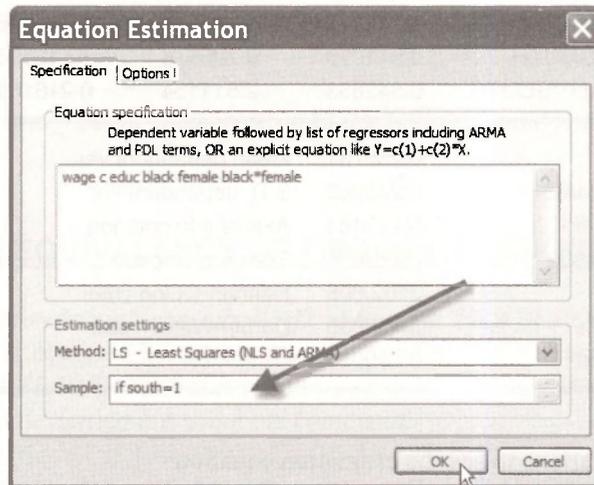


The *F*-test statistic value is

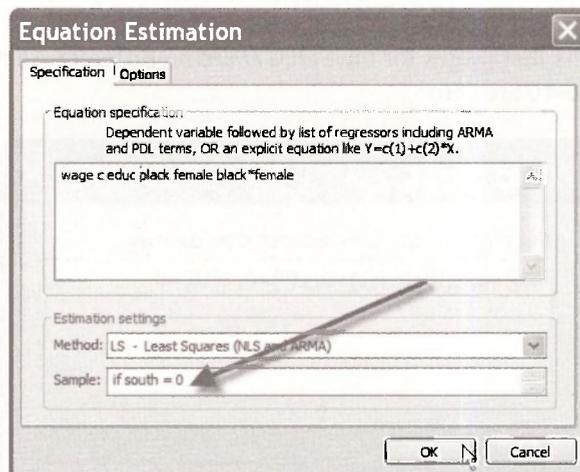
Wald Test:  
Equation: CHOW

Test Statistic	Value	df	Probability
F-statistic	2.013212	(5, 990)	0.0744

Obtaining the regression for the *SOUTH* observations is obtained by selecting **Quick/Estimate Equation**. In the dialog box enter the equation and modify the **Sample** to include observations for which *SOUTH* = 1.



To obtain the results for the *NON SOUTH* estimate the equation using the observations for which *SOUTH* = 0.



## 7.6 INTERACTIONS BETWEEN CONTINUOUS VARIABLES

Open the workfile *pizza.wfl*. Estimate the least squares regression of *PIZZA* on *AGE* and *INCOME*.

Is pizza c age income

Now add the interaction of *AGE* and *INCOME*

Is pizza c age income age\*income

Dependent Variable: PIZZA

Method: Least Squares

Sample: 1 40

Included observations: 40

	Coefficient	Std. Error	t-Statistic	Prob.
C	161.4654	120.6634	1.338147	0.1892
AGE	-2.977423	3.352101	-0.888226	0.3803
INCOME	0.009074	0.003670	2.472717	0.0183
AGE*INCOME	-0.000160	8.67E-05	-1.847148	0.0730
R-squared	0.387319	Mean dependent var		191.5500
Adjusted R-squared	0.336262	S.D. dependent var		155.8806
S.E. of regression	126.9961	Akaike info criterion		12.62083
Sum squared resid	580608.7	Schwarz criterion		12.78972
Log likelihood	-248.4166	Hannan-Quinn criter.		12.68189
F-statistic	7.586038	Durbin-Watson stat		0.932029
Prob(F-statistic)	0.000468			

The marginal effect of *AGE* is

$$\partial E(\text{PIZZA})/\partial \text{AGE} = \beta_2 + \beta_4 \text{INCOME}$$

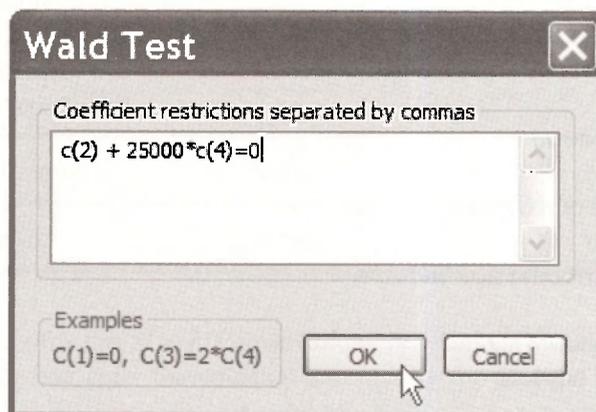
To evaluate this marginal effect at *INCOME* = \$25,000, select in the regression window **View/Representations** to see

Estimation Equation:

=====

$$\text{PIZZA} = \text{C}(1) + \text{C}(2)*\text{AGE} + \text{C}(3)*\text{INCOME} + \text{C}(4)*\text{AGE}*\text{INCOME}$$

Then select **View/Coefficient Tests/Wald – Coefficient Restrictions**.



The Wald test results include the marginal effect, which we posed to EViews as a hypothesis, and its standard error.

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(2) + 25000*C(4)	-6.982702	2.267797

Save this workfile and close.

## **7.7 LOG-LINEAR MODELS**

Regression equations with log-transformed dependent variables are common. To illustrate, open the workfile *cps\_small.wf1*. In EViews the function **log** creates the natural logarithm. The estimation equation can be represented as

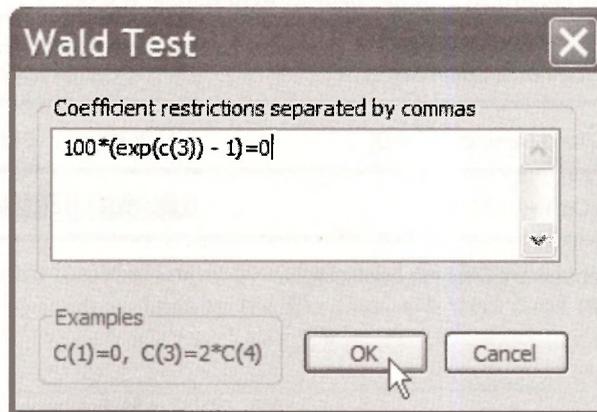
**ls log(wage) c educ female**

The result (next page) is as shown on page 185 of *POE*.

Dependent Variable: LOG(WAGE)  
 Method: Least Squares  
 Sample: 1 1000  
 Included observations: 1000

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.929036	0.083748	11.09319	0.0000
EDUC	0.102566	0.006075	16.88240	0.0000
FEMALE	-0.252603	0.029977	-8.426577	0.0000
R-squared	0.266837	Mean dependent var		2.166837
Adjusted R-squared	0.265366	S.D. dependent var		0.552806
S.E. of regression	0.473814	Akaike info criterion		1.346993
Sum squared resid	223.8265	Schwarz criterion		1.361716
Log likelihood	-670.4965	Hannan-Quinn criter.		1.352589
F-statistic	181.4308	Durbin-Watson stat		0.524339
Prob(F-statistic)	0.000000			

The exact calculation of the effect of gender on wages looks complicated, but it is simple in EViews. Select **View/Coefficient Tests/Wald – Coefficient Restrictions**. In EViews the **exponential** function is **exp**. To make the nonlinear calculation enter it as a hypothesis.



The result is

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
100 * (-1 + EXP(C(3)))	-22.32240	2.328539

Delta method computed using analytic derivatives.

The calculated percentage difference in wages is -22.32% and EViews computes a standard error

for this quantity by the “Delta method,” which you will study in your graduate econometrics courses.

The next example includes an interaction term

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \gamma(EDUC \times EXPER)$$

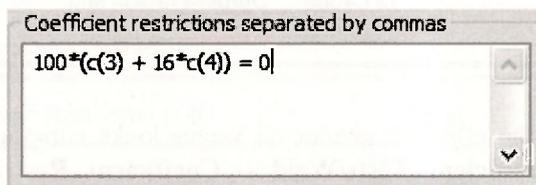
To estimate this model the command is

**ls log(wage) c educ exper educ\*exper**

The approximate effect of another year of experience, holding education constant, is

$$100(\beta_3 + \gamma EDUC)\%$$

Using the same approach, select **View/Coefficient Tests/Wald – Coefficient Restrictions** and enter



This yields the estimated benefit of another year of experience, 0.95%.

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
100 * (C(3) + 16*C(4))	0.951838	0.215985

Restrictions are linear in coefficients.

## Keywords

@trend	estimation equation	operators
Chow test	exponential function	polynomials
coefficient vector	interactions	representations
delta method	log function	sample range
dummy variables	marginal effect	series
equation name.ls	nonlinear hypothesis	Wald test

# CHAPTER 8

## Heteroskedasticity

### CHAPTER OUTLINE

- 8.1 Examining Residuals
    - 8.1.1 Plot against observation number
    - 8.1.2 Plot against an explanatory variable
    - 8.1.3 Plot of least squares line
  - 8.2 Heteroskedasticity-Consistent Standard Errors
  - 8.3 Weighted Least Squares
    - 8.3.1 A short way
    - 8.3.2 A long way
  - 8.4 Estimating a Variance Function
    - 8.4.1 Variance function estimates
    - 8.4.2 Generalized least squares
  - 8.5 A Heteroskedastic Partition
    - 8.5.1 Least-squares estimates: one equation
    - 8.5.2 Least-squares estimates: two equations
    - 8.5.3 Generalized least-squares estimates
  - 8.6 The Goldfeld-Quandt Test
    - 8.6.1 The wage equation
    - 8.6.2 The food expenditure equation
  - 8.7 Testing the Variance Function
    - 8.7.1 The Breusch-Pagan test
    - 8.7.2 The White test
- KEYWORDS

### 8.1 EXAMINING RESIDUALS

In this chapter we return to the example considered in Chapters 2 to 4 where weekly expenditure on food was related to income. Data in the file *food.wfl* were used to find the following least squares estimates.

Dependent Variable: FOOD_EXP				
Included observations: 40				
	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

We are now concerned with whether the error variance for this equation is likely to vary over observations, a characteristic called heteroskedasticity. To carry out a preliminary investigation of this question, we examine the least squares residuals. If they increase with increasing income, that suggests the error variance increases with income.

### 8.1.1 Plot against observation number

There are a variety of ways in which EViews can be used to examine least squares residuals. Let us begin by checking the obvious ones. After estimating the equation and naming it `ls_eqn`, go to **View/Actual, Fitted, Residual**. At that point you will see a menu with the following options.

**Actual, Fitted, Residual Table**  
**Actual, Fitted, Residual Graph**  
**Residual Graph**  
**Standardized Residual Graph**

Check each of these options to get a feel for the different ways in which they convey information. As you might expect from the names of the options, each alternative presents information on one or more of the series **actual**, **fitted** and **residual**. In terms of the names of the series in your workfile

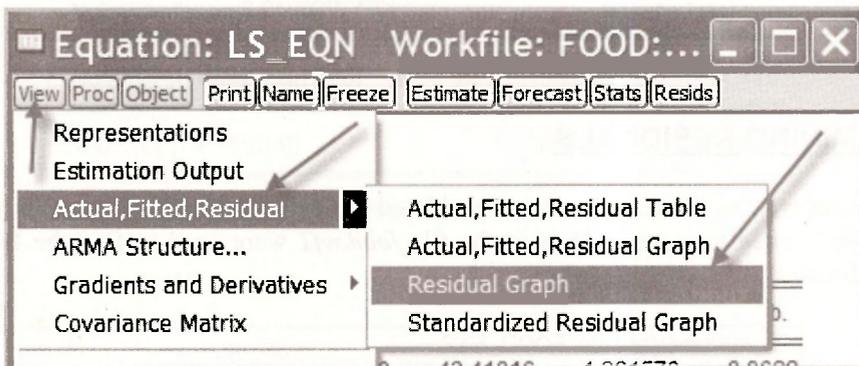
$$\mathbf{actual} = FOOD\_EXP$$

$$\mathbf{fitted} = \widehat{FOOD\_EXP} = b_1 + b_2 INCOME$$

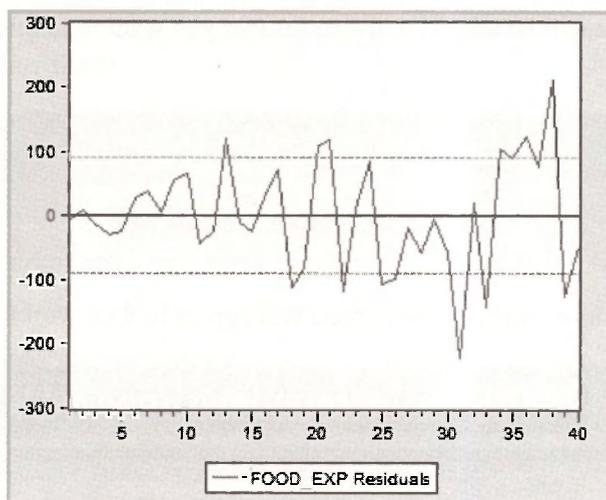
$$\mathbf{resid} = \hat{e} = FOOD\_EXP - \widehat{FOOD\_EXP}$$

The **Standardized Residual Graph** is a graph of  $\hat{e}/\hat{\sigma}$ ; the residuals have been standardized (made free of units of measurement) by dividing by the estimated standard deviation of the error term.

In each case the series are graphed against the observation number. As an example, consider the **Residual Graph** selected in the following way.



In the residual graph that follows it is clear that the absolute magnitude of the residuals has a tendency to be larger as the observation number gets larger. The reason such is the case is that the observations are ordered according to increasing values of *INCOME*, and the absolute magnitude of the residuals increases as *INCOME* increases. Given it is this latter relationship that we are really interested in, it is preferable to graph the residuals against income. Nevertheless, residual graphs like the one below are important for examining which observations are not well captured by the estimated model (outliers), and, in the case of time series data, for discerning patterns in the residuals. To help you assess which observations could be viewed as outliers, dotted lines are drawn at points one standard deviation ( $\hat{\sigma} = 89.517$ ) either side of zero.

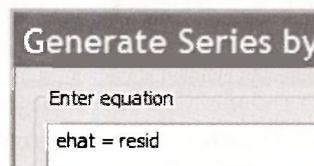
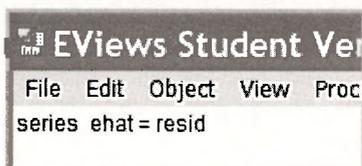


### 8.1.2 Plot against an explanatory variable

To graph the residuals against income we begin by naming the residuals and the fitted values.

```
series ehat = resid
series foodhat = food_exp - ehat
```

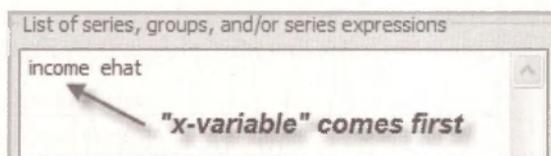
Recall that these commands can be executed by typing them in the upper EViews window or by clicking on **Genr** and writing the equation to generate the series in the resulting box. Examples of these two alternatives for the first command follow.



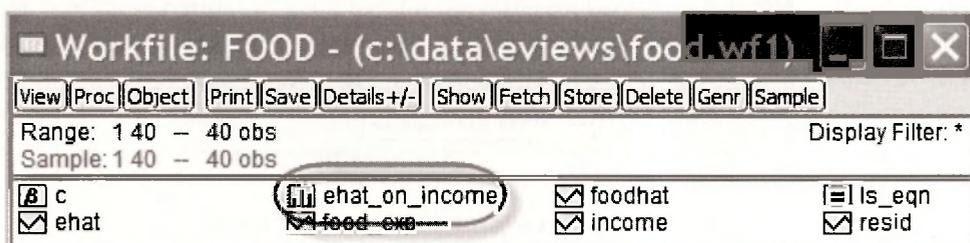
Returning to our task of graphing the residuals, we create a **graph object** by going to **Object/ New Object** and selecting **Graph**. As a name for the graph, we chose **EHAT\_ON\_INCOME**.



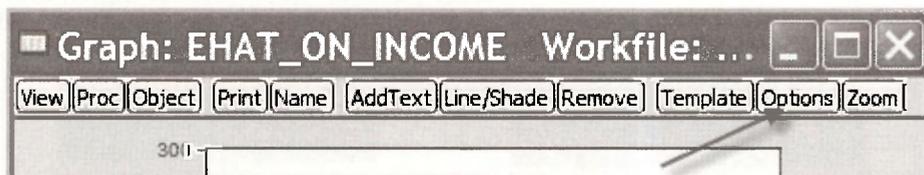
After clicking **OK**, you will be asked for the series that you want to graph. The one that is to go on the  $x$ -axis comes first.



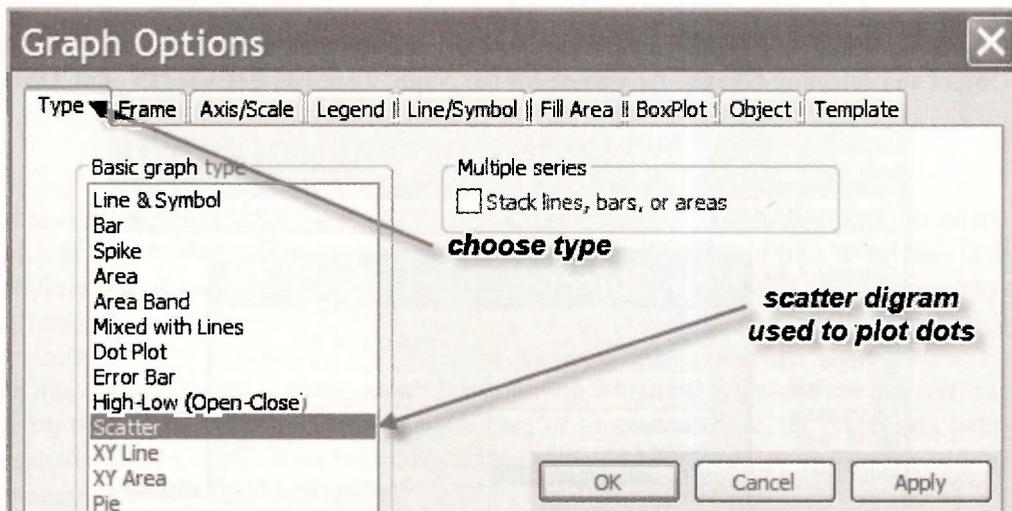
After clicking **OK** one more time, a graph object will appear in your workfile.



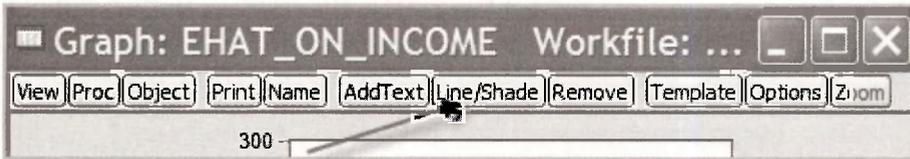
Double clicking on this object will open it. Be careful, however. It may not look like you expected! Unless told otherwise, EViews will assume you want both *INCOME* and *EHAT* graphed against the observation number. You need to tell EViews to change the graph so that *INCOME* is on the  $x$ -axis and *EHAT* is on the  $y$ -axis. Also, given that income is not measured in equally spaced intervals, dots are preferred to a line graph. With these factors in mind, open the graph and select **Options**.



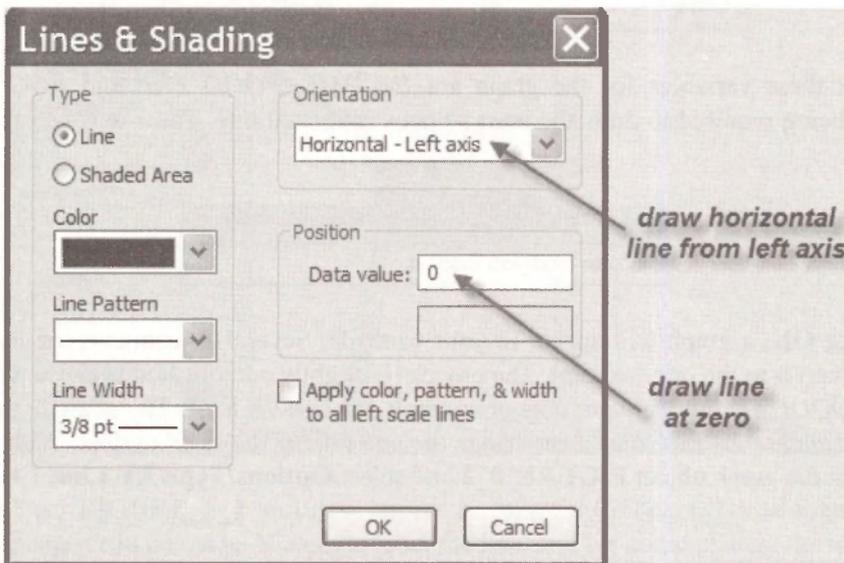
Then select **Type/Scatter**, click **Apply**, and click **OK**.



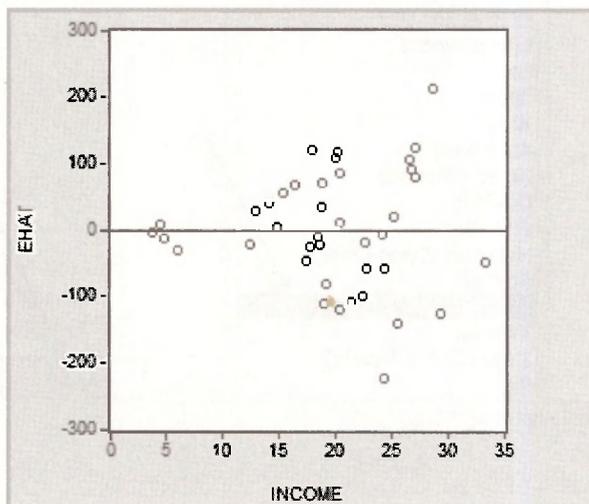
A nice looking scatter plot will appear. You can make it look even nicer by drawing a horizontal line at zero. Select **Line/Shade**.



Fill in the resulting dialog box as follows.

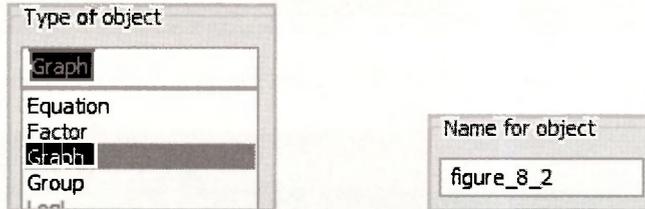


Clicking **OK** gives the required graph. Notice how the absolute magnitude of the residuals is larger for larger values of income, an indication of heteroskedasticity.

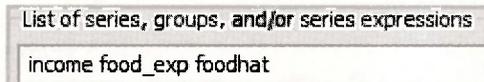


### 8.1.3 Plot of least squares line

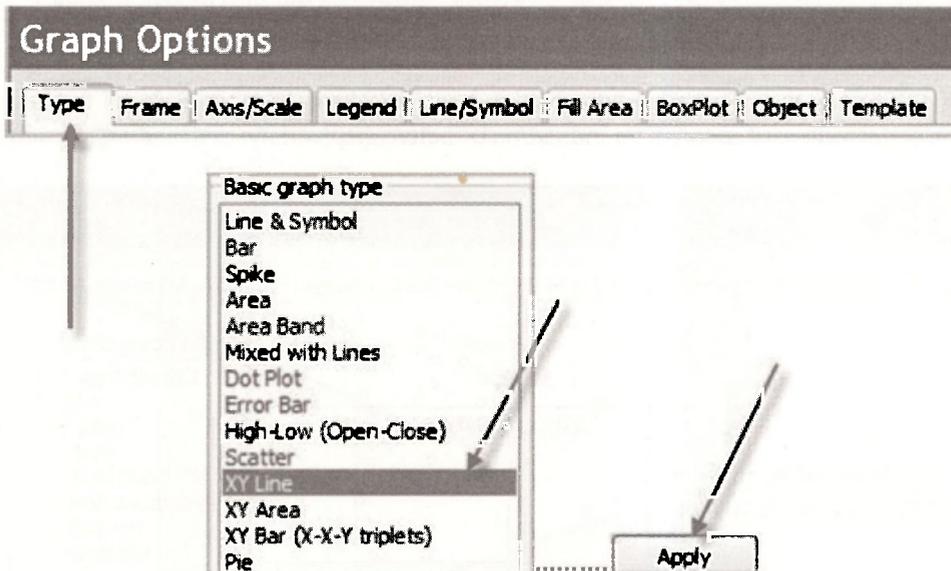
Another way to illustrate the dependence of the magnitude of the residuals on *INCOME* is to plot *FOOD\_EXP* and the least-squares estimated line against *INCOME*, as is displayed in Figure 8.2 on page 200 of the text. To reproduce this figure, select **Object/New Object/Graph** and give the graph a name, say **FIGURE\_8\_2**.



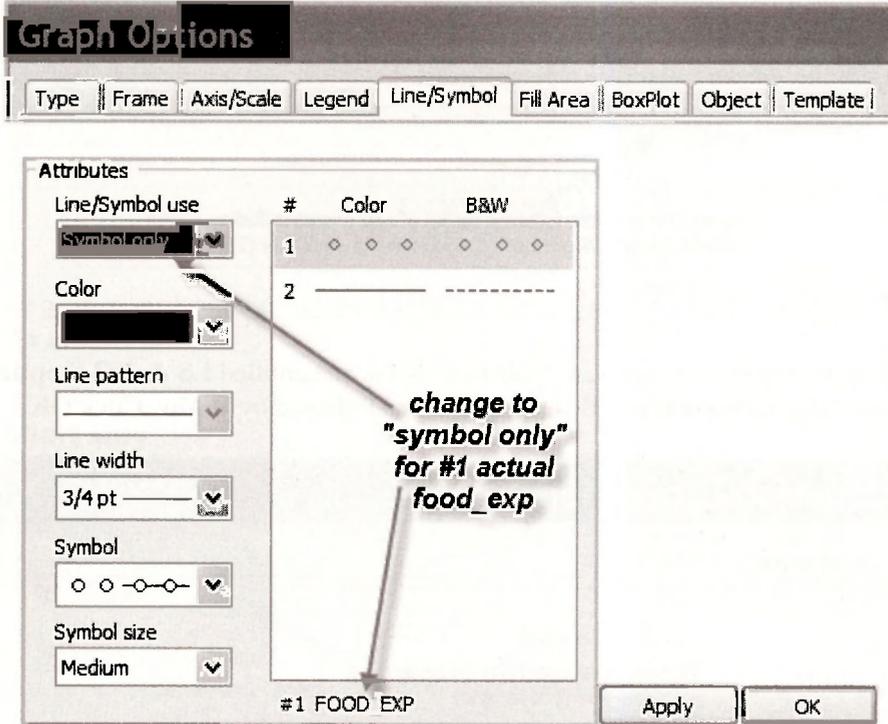
The relevant three variables for the graph are *INCOME*, *FOOD\_EXP* and *FOODHAT*, with *FOODHAT* being required to draw the least-squares estimated line. The *x*-axis variable *INCOME* is listed first.



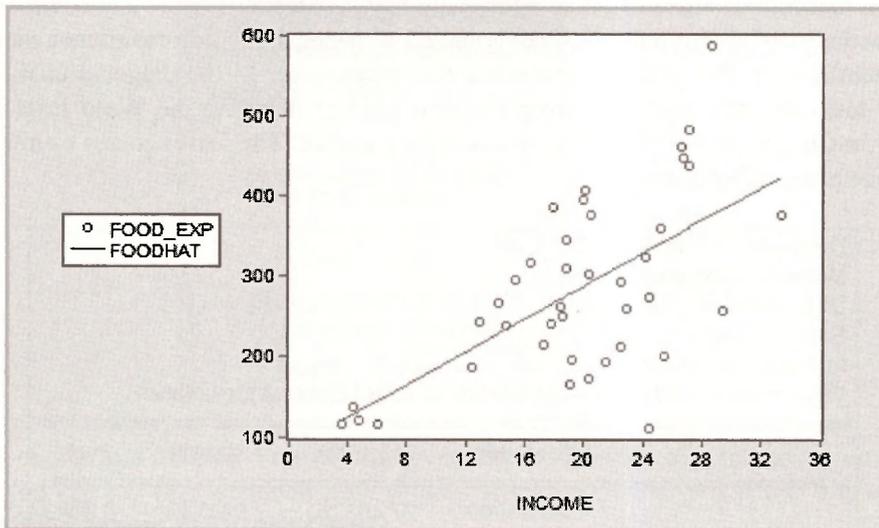
After clicking **OK**, a graph will appear in your workfile. Several adjustments are needed to this graph to convert it to the one we want. The process is slightly complicated because we want a line graph for *FOODHAT*, but we want dots or symbols for *FOOD\_EXP*. The strategy that we adopt is to ask for line graphs first, and then change the one for *FOOD\_EXP* to dots. With these points in mind, open the graph object **FIGURE\_8\_2** and select **Options/Type/XY Line**. Click **Apply**.



Then select **Line/Symbol**. The snapshot below shows you how you change the line for *FOOD\_EXP* to dots (symbols). No changes are needed for series #2, *FOODHAT*. It is already represented by the required line. Click **Apply**. Click **OK**.

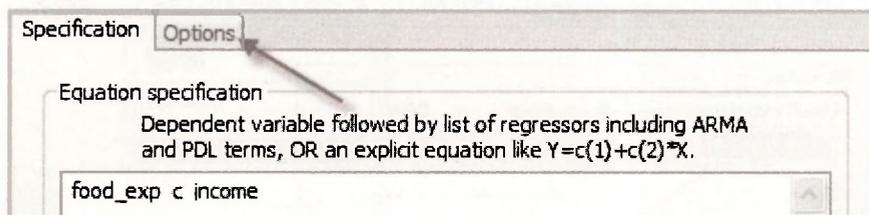


The graph object **FIGURE\_8\_2** will now appear as follows. Compare it with Figure 8.2 in the text. Other changes can be made. We could label the line, and we could change the **legend**, that at present appears in a box on the left of the graph. We suggest you experiment with these options. For labeling, select the button **AddText**. For changing the legend, go to **Options/ Legend**. You can also cut and paste it into a document using **Ctrl+C** followed by **Ctrl+V**.

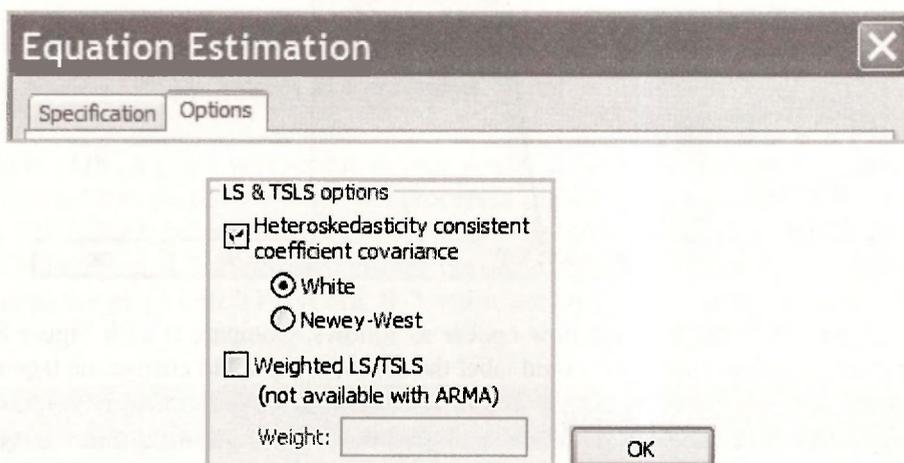


## 8.2 HETEROSKEDASTICITY-CONSISTENT STANDARD ERRORS

One option for correcting conventional least-squares interval estimates and hypothesis tests that are no longer appropriate under heteroskedasticity is to use what are known as White's heteroskedasticity-consistent standard errors. These standard errors are obtained in EViews by choosing an estimation option. In the **Equation Estimation** box, click on the **Options** tab.



In the **Options** dialog box, the relevant “sub-box” is the one entitled **LS & TSLS options**. Select **Heteroskedasticity consistent coefficient covariance** followed by **White**. Click **OK**.



In the output that follows there is a note telling you that the standard errors and covariance are the heteroskedasticity-consistent ones. By “covariance”, it means the whole covariance matrix for the estimated coefficients. The standard errors are the square roots of the diagonal elements of this matrix. All test outcomes computed from this new object, including the **Wald tests** considered extensively in Chapter 6, will use the new covariance matrix. The least squares estimates remain the same. See page 202 of the text.

Dependent Variable: FOOD_EXP				
Method: Least Squares				
Date: 11/30/07 Time: 00:31				
Sample: 1 40				
Included observations: 40				
White Heteroskedasticity-Consistent Standard Errors & Covariance				
	Coefficient	Std. Error	t-Statistic	Prob.
C	83.41600	27.46375	3.037313	0.0043
INCOME	10.20964	1.809077	5.643565	0.0000

### 8.3 WEIGHTED LEAST SQUARES

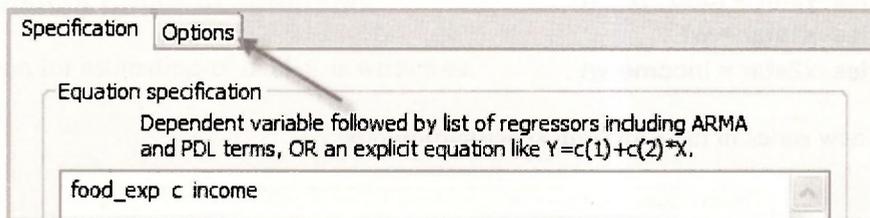
In Section 8.3.1 of the text the regression error variance is assumed to be heteroskedastic of the form  $\sigma_i^2 = \sigma^2 x_i$  where  $x_i = INCOME_i$ . Under this specification the minimum variance unbiased estimator for the regression coefficients  $\beta_1$  and  $\beta_2$  is the generalized least squares estimator. This estimator is also known as the weighted least squares estimator where, in this case, each observation is weighted by

$$\frac{1}{\sqrt{x_i}} = \frac{1}{\sqrt{INCOME_i}}$$

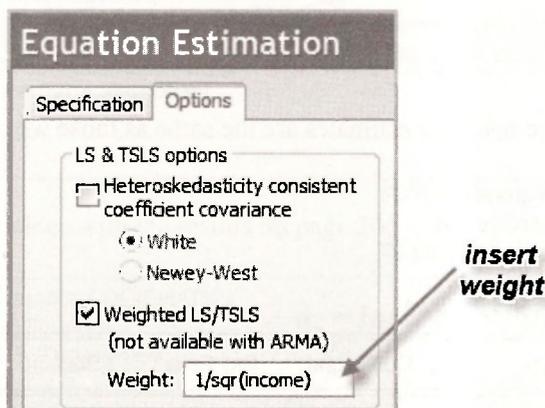
There are two ways to obtain the weighted least squares estimator, a short way and a long way. It is instructive to consider both.

#### 8.3.1 A short way

Weighted least squares is another **Equation Estimation** option, so our starting point is the same as that for the White standard errors, namely



In this case, however, we select **Weighted LS/TSLS** from the **LS & TSLS** options. In the **Weight** box we type **1/sqr(income)**, where **sqr** is the EVIEWS function for square root



Check the output that follows. You will discover that it coincides with that given on page 204 of the text. You can tell that weighted least squares is the estimation procedure from the line that says **Weighting series: 1/SQR(INCOME)**.

Dependent Variable: FOOD\_EXP  
 Method: Least Squares  
 Sample: 1 40  
 Included observations: 40  
 Weighting series: 1/SQR(INCOME) *weight*

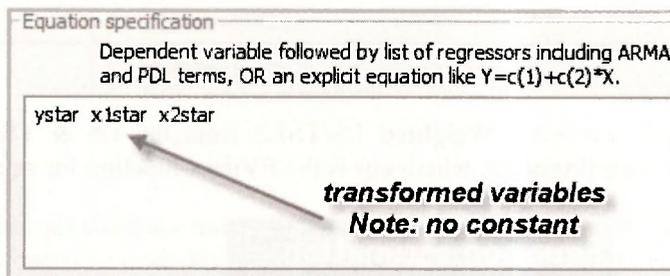
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	78.68408	23.78872	3.307621	0.0021
INCOME	10.45101	1.385891	7.541002	0.0000

### 8.3.2 A long way

The long way to obtain weighted least squares estimates is to transform each of your variables by dividing by  $\sqrt{INCOME}$  as described on page 203 of the text, and to then apply least squares without the weights. The variables can be transformed by creating new series, or by dividing each variable by  $\sqrt{INCOME}$  in the equation specification. If you choose to create new series, the following commands are suitable.

```
series wt = 1/sqr(income)
series ystar = food_exp*wt
series x1star = wt
series x2star = income*wt
```

Enter these new series in the **Equation specification** box

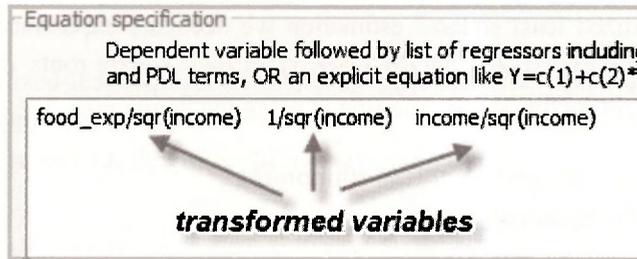


Click **OK**. Observe that the resulting estimates are the same as those we obtained the short way.

Dependent Variable: YSTAR  
 Method: Least Squares  
 Date: 11/30/07 Time: 02:27  
 Sample: 1 40  
 Included observations: 40

	Coefficient	Std. Error	t-Statistic	Prob.
X1STAR	78.68408	23.78872	3.307621	0.0021
X2STAR	10.45101	1.385891	7.541002	0.0000

An alternative way that avoids the need to define new series is to transform the variables within the **Equation specification**, as illustrated below. Try it. Check your output.



### 8.4 ESTIMATING A VARIANCE FUNCTION

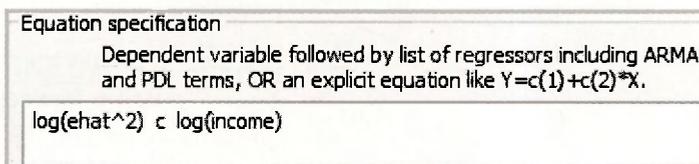
The heteroskedastic assumption made in the previous section ( $\sigma_i^2 = \sigma^2 x_i$ ) can be viewed as a special case of the more general assumption  $\sigma_i^2 = \sigma^2 x_i^\gamma$  where  $\gamma$  is an unknown parameter. Under this more general assumption  $\gamma$  must be estimated before we can proceed with weighted or generalized least squares estimation. In line with Section 8.3.2 of the text (page 205), we first estimate  $\sigma^2$  and  $\gamma$  and then proceed to generalized least squares estimation.

#### 8.4.1 Variance function estimates

The equation for estimating  $\sigma^2$  and  $\gamma$  is written as

$$\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 \ln(x_i) + v_i$$

where  $\hat{e}_i$  are the least squares residuals,  $\alpha_1 = \ln(\sigma^2)$  and  $\alpha_2 = \gamma$ . Recognizing that the  $\hat{e}_i$  were previously saved in the workfile *food.wfl* under the name *EHAT*, and that  $x_i = INCOME$ , least squares estimates for  $\alpha_1$  and  $\alpha_2$  are obtained using the following **Equation specification**.



The resulting output coincides with the results on page 206 of the text.

Dependent Variable: LOG(EHAT^2)  
 Method: Least Squares  
 Date: 12/01/07 Time: 04:54  
 Sample: 1 40  
 Included observations: 40

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.937796	1.583106	0.592377	0.5571
LOG(INCOME)	2.329239	0.541336	4.302761	0.0001

$\hat{\alpha}_1 = \ln(\hat{\sigma}^2)$        $\hat{\alpha}_2 = \hat{\gamma}$

To proceed to generalized least squares estimation we need the exponential of the predictions from this equation,  $\hat{\sigma}_i^2 = \exp(\hat{\alpha}_1 + \hat{\alpha}_2 \ln(INCOME_i))$  or their square roots  $\hat{\sigma}_i$ . It is instructive to consider two ways of computing them. The first is using the commands

```
series sig2hat = exp(c(1) + c(2)*log(income))
series sighat = @sqrt(sig2hat)
```

As long as the variance equation is the most recent regression model that has been estimated, in the first of these commands  $C(1) = \hat{\alpha}_1$  and  $C(2) = \hat{\alpha}_2$

An alternative way of obtaining  $\hat{\sigma}_i = \mathbf{sighat}$  is to use the **forecast** option. In the window displaying the output from estimating the variance equation, click on **Forecast**. In the resulting forecast window, you will see two possible series that can be forecast, **EHAT** and **LOG(EHAT^2)**. This choice has arisen because the dependent variable in the **Equation specification** was written as  $\log(\hat{e}^2)$ , a transformation of the series  $\hat{e}$ . EVIEWS is giving you the option of forecasting the original series  $\hat{e}$  or its transformed version  $\log(\hat{e}^2)$ . If you had defined the dependent variable as  $q = \log(\hat{e}^2)$  via a **series** command, and then specified your dependent variable as  $q$ , EVIEWS would not have been given you a choice. It would assume you want to forecast  $q$ . Writing the transformation as part of the equation specification is what leads to the choice.

Now consider the two options. If you click **LOG(EHAT^2)**, EVIEWS will give you the forecasts  $\hat{q}_i = \ln(\hat{\sigma}_i^2)$ . If you click **EHAT**, it will invert the transformation given in the **Equation specification** and give you the forecasts  $\hat{\sigma}_i = \sqrt{\exp(\hat{q}_i)}$ . Since it is  $\hat{\sigma}_i$  that is needed to transform the variables in the generalized least squares procedure, we choose **EHAT**. We call the forecast series **SIGHAT**.

Forecast

Forecast equation: VAR\_EQN

Series to forecast:  EHAT  LOG(EHAT^2)

Series names: Forecast name: sighat

S.E. (optional):

VARCOV (optional):

Forecast sample: 1 40

Method:  Static forecast (no dynamics in equation)  Structural (ignore ARMA)  Coef uncertainty in S.E. calc

Output:  Forecast graph  Forecast evaluation

Insert actuals for out-of-sample observations

OK Cancel

### 8.4.2 Generalized least squares

To obtain the generalized least squares estimates in equation (8.27) on page 207 of the text we can use EViews weighted least squares option, with weighting series  $\hat{\sigma}_i^{-1} = 1/\text{SIGHAT}$ . The **Equation specification** and **LS & TSLS options** are given by

<p>Equation specification</p> <p>Dependent variable f and PDL terms, OR a</p> <p>food_exp c income</p>	<p>LS &amp; TSLS options</p> <p><input type="checkbox"/> Heteroskedasticity consistent coefficient covariance</p> <p><input type="radio"/> White</p> <p><input type="radio"/> Newey-West</p> <p><input checked="" type="checkbox"/> Weighted LS/TSLS (not available with ARMA)</p> <p>Weight: 1/sighat</p>
--	--

These selections yield the following output.

Dependent Variable: FOOD_EXP				
Method: Least Squares				
Date: 12/02/07 Time: 10:35				
Sample: 1 40				
Included observations: 40				
Weighting series: 1/SIGHAT				
	Coefficient	Std. Error	t-Statistic	Prob.
C	76.05379	9.713489	7.829709	0.0000
INCOME	10.63349	0.971514	10.94528	0.0000

### 8.5 A HETEROSKEDASTIC PARTITION

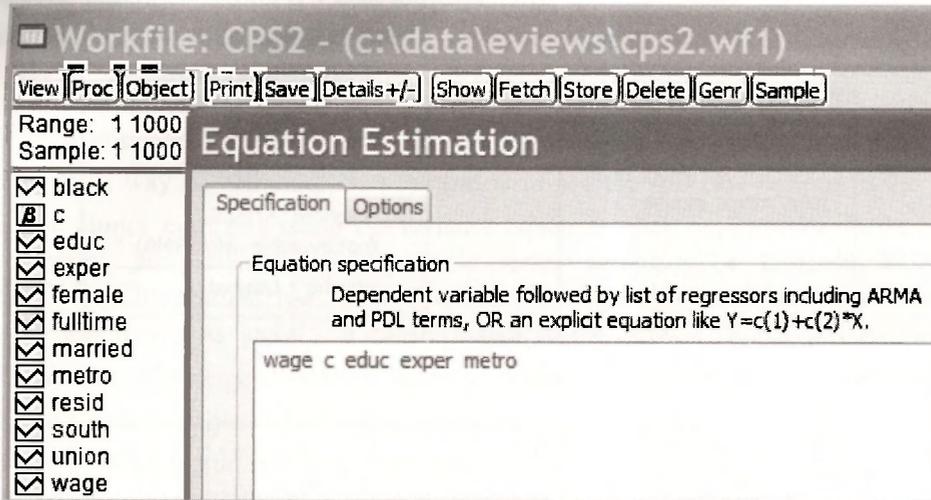
In Section 8.3.3 of the text we use data from the file *cps2.dat* to estimate the equation

$$WAGE = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 METRO + e$$

We are hypothesizing that *WAGE* depends on education (*EDUC*), experience (*EXPER*), and whether a worker lives in a metropolitan area (*METRO* = 1 for metropolitan area, *METRO* = 0 for rural area). Three sets of estimates are obtained: (1) a least-squares regression on all observations, (2) two separate least-squares regressions, one for metropolitan workers and one for rural workers, and (3) a generalized least-squares regression that uses all observations, but which assumes the error variances for metropolitan and rural workers are different. This latter assumption is referred to as a “heteroskedastic partition”.

### 8.5.1 Least-squares estimates: one equation

No new features of EViews are required for single equation least-squares estimation of the wage equation. However, we report the equation specification and the output for completeness.



Dependent variable: WAGE				
Method: Least Squares				
Date: 11/30/07 Time: 14:21				
Sample: 1 1000				
Included observations: 1000				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.913984	1.075663	-9.216631	0.0000
EDUC	1.233964	0.069961	17.63782	0.0000
EXPER	0.133244	0.015232	8.747835	0.0000
METRO	1.524104	0.431091	3.535459	0.0004

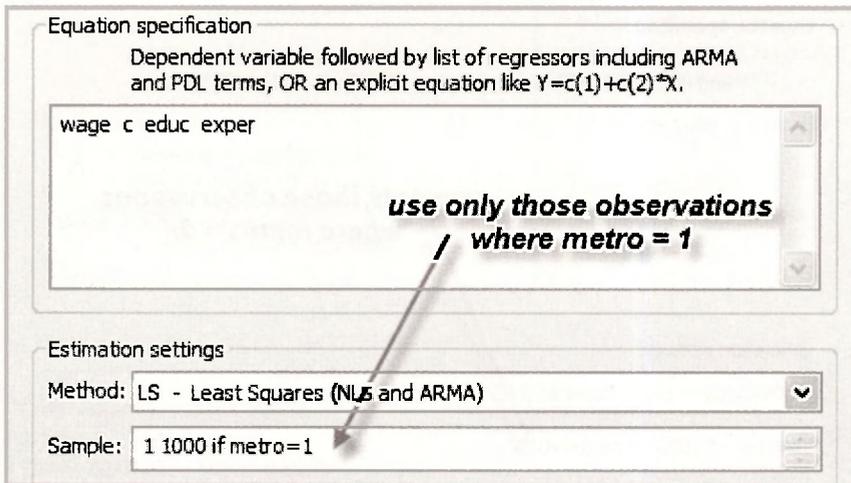
Check these results against those in equation (8.28) on page 208 of the text.

### 8.5.2 Least-squares estimates: two equations

To estimate two separate equations, one for metropolitan workers and one for rural workers, we use EViews to restrict the sample to the relevant observations. For the metropolitan observations, we change the sample by going to the **Estimation settings** box and specifying

**sample 1 1000 if metro = 1**

This instruction tells EViews to consider all 1000 observations, but to restrict estimation to those where **metro = 1**.



The results follow. Notice that EViews reminds you about the sample you have chosen and it also tells you how many observations satisfy the restriction that you imposed for their inclusion. We have 808 metropolitan observations. The estimates are consistent with those on page 209 of the text. Of particular interest are the standard deviation and variance of the error term,  $\hat{\sigma}_M = 5.641$  and  $\hat{\sigma}_M^2 = 31.824$ . They are needed for the generalized least squares estimates in the next section. The value  $\hat{\sigma}_M = 5.641$  is stored temporarily as **@se**; we can save it as **se\_metro** using the command

```
scalar se_metro = @se
```

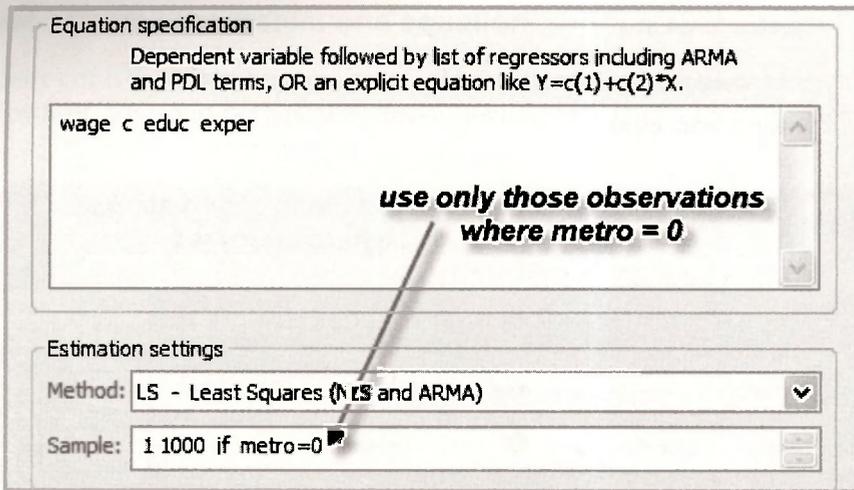
Dependent Variable: WAGE  
 Method: Least Squares  
 Date: 11/30/07 Time: 14:26  
 Sample: 1 1000 IF METRO=1  
 Included observations: 808

*reduced sample for metro = 1*

$\hat{\sigma}_M^2 = (5.641253)^2 = 31.824$

	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.052478	1.189456	-7.610603	0.0000
EDUC	1.281714	0.079763	16.06910	0.0000
EXPER	0.134560	0.017948	7.497370	0.0000
R-squared	0.258183	Mean dependent var		10.57802
Adjusted R-squared	0.256340	S.D. dependent var		6.541667
S.E. of regression	5.641253	Akaike info criterion		6.301795
Sum squared resid	25619.10	Schwarz criterion		6.210226

The same steps are followed for the rural observations, but in this case we restrict the sample to those observations where **metro = 0**.



The results below show that there are 192 rural observations. The standard deviation and variance of the error term are, respectively,  $\hat{\sigma}_R = 3.904$  and  $\hat{\sigma}_R^2 = 15.243$ . We save the value for  $\hat{\sigma}_R$  using the command

```
scalar se_rural = @se
```

Dependent Variable: WAGE		<i>reduced sample for metro = 0</i>		
Method: Least Squares				
Date: 11/30/07 Time: 23:27				
Sample: 1 1000 IF METRO=0				
Included observations: 192				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.165855	1.898511	-3.247732	0.0014
EDUC	0.955585	0.133190	7.174608	0.0000
EXPER	0.125974	0.024771	5.085538	0.0000
R-squared	0.258748	Mean dependent var	8.676979	
Adjusted R-squared	0.250906	S.D. dependent var	4.510933	
S.E. of regression	3.904227	Akaike info criterion	5.577498	
Sum squared resid	2880.924	Schwarz criterion	5.628397	

$$\hat{\sigma}_R^2 = (3.904227)^2 = 15.243$$

### 8.5.3 Generalized least-squares estimates

For generalized least squares estimation we can use EViews weighted least squares option with the weighting series equal to  $\hat{\sigma}_M^{-1}$  for the metropolitan observations and  $\hat{\sigma}_R^{-1}$  for the rural observations. Thus, we create the series

```
series weight = metro*(1/se_metro) + (1-metro)*(1/se_rural)
```

The relevant equation and option specifications are

<b>Equation specification</b> Dependent variable follow and PDL terms, OR an eq wage c educ exper metro	<b>LS &amp; TSLS options</b> <input type="checkbox"/> Heteroskedasticity consistent coefficient covariance <input checked="" type="radio"/> White <input type="radio"/> Newey-West <input checked="" type="checkbox"/> Weighted LS/TSLS (not available with ARMA) Weight: <input type="text" value="weight"/>
--	--

The following results can be checked against those on page 210 of the text.

Dependent Variable: WAGE Method: Least Squares Date: 12/01/07 Time: 04:27 Sample: 1 1000 Included observations: 1000 Weighting series: WEIGHT				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.398362	1.019673	-9.217038	0.0000
EDUC	1.195721	0.068508	17.45375	0.0000
EXPER	0.132209	0.014549	9.087448	0.0000
METRO	1.538803	0.346286	4.443740	0.0000

## 8.6 THE GOLDFELD-QUANDT TEST

As tests for heteroskedasticity we consider a test known as the Goldfeld-Quandt test and a general class of tests based on an estimated variance function. The statistic for the Goldfeld-Quandt test is the ratio of error variance estimates from two sub-samples of the observations. If those estimates are  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$ , obtained from sub-sample regressions with  $(N_1 - K_1)$  and  $(N_2 - K_2)$  degrees of freedom, respectively, then, under the null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2$ ,

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} \sim F_{(N_2 - K_2, N_1 - K_1)}$$

If the alternative hypothesis is  $H_1 : \sigma_1^2 \neq \sigma_2^2$ , and a 5% significance level is used, the test is a two-tail one with critical values  $F_{(0.975, N_2 - K_2, N_1 - K_1)}$  and  $F_{(0.025, N_2 - K_2, N_1 - K_1)}$ . For a 5% one-tail test with  $H_1 : \sigma_2^2 > \sigma_1^2$ , the critical value is  $F_{(0.95, N_2 - K_2, N_1 - K_1)}$ . For  $H_1 : \sigma_2^2 < \sigma_1^2$ , the numerator and denominator and degrees of freedom for the test can be reversed, or the critical value  $F_{(0.05, N_2 - K_2, N_1 - K_1)}$  can be used. We consider application of this test to the wage equation and to the food expenditure equation as found on page 212 of the text.

### 8.6.1 The wage equation

For the wage equation the two sub-samples are those for metropolitan and rural workers. Thus, we write  $\sigma_2^2 = \sigma_M^2$  and  $\sigma_1^2 = \sigma_R^2$  and test  $H_0 : \sigma_R^2 = \sigma_M^2$  against the alternative  $H_1 : \sigma_R^2 \neq \sigma_M^2$ . Given that  $\hat{\sigma}_M$  and  $\hat{\sigma}_R$  have been earlier saved as **SE\_METRO** and **SE\_RURAL**, respectively, the value  $F = \hat{\sigma}_M^2 / \hat{\sigma}_R^2 = 31.824/15.243 = 2.09$ , and its 5% upper and lower critical values  $F_{Uc} = 1.26$  and  $F_{Lc} = 0.81$ , can be computed from the commands below. Note that  $N_M - K_M = 805$  and that  $N_R - K_R = 189$ .

```
scalar f_val = (se_metro)^2/(se_rural)^2
scalar fcrit_up = @qfdist(0.975,805,189)
scalar fcrit_low = @qfdist(0.025,805,189)
```

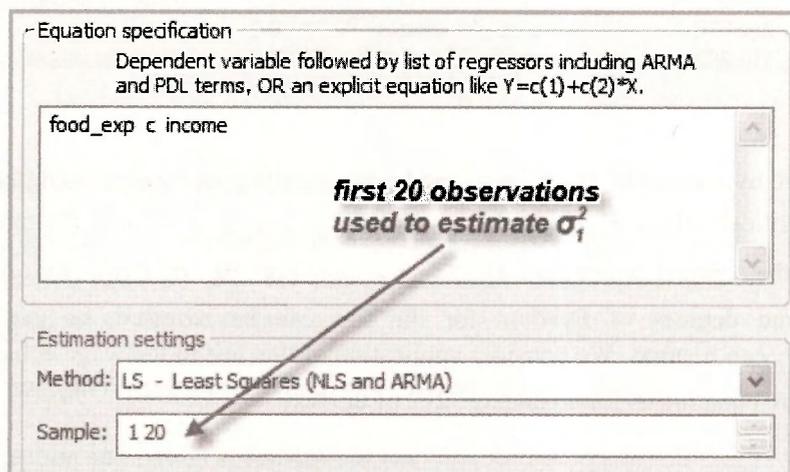
### 8.6.2 The food expenditure equation

For the food expenditure example there are no two well defined sub-samples for  $\sigma_1^2$  and  $\sigma_2^2$ . For convenience, and to improve our chances of rejecting  $H_0$  when  $H_1$  is true, we take  $\sigma_1^2$  as the variance for the first 20 observations and  $\sigma_2^2$  as the variance of the second 20 observations. Since our alternative hypothesis is that  $\sigma_1^2$  increases as *INCOME* increases,  $\sigma_1^2$  and  $\sigma_2^2$  are not actual variances, but devices to aid the testing procedure. In our sample the observations are ordered according to the values of *INCOME*. Values of *INCOME* in the second half of the sample are larger than those in the first half of the sample. Thus,  $\hat{\sigma}_2^2$  will tend to be greater than  $\hat{\sigma}_1^2$  when  $H_1$  is true, but similar when  $H_0$  is true. If your data are not ordered according to increasing values of *INCOME*, you can reorder them using the command

```
sort income
```

This command reorders all series in your workfile according the magnitude of *INCOME*.

To use the first 20 observations to estimate  $\sigma_1^2$ , we restrict the **Sample** for estimation as shown below.



The value  $\hat{\sigma}_1^2$  is obtained by squaring the value of S.E. of regression in the resulting output. We give it the name SIG1\_SQ.

```
scalar sig1_sq = @se^2
```

*first half of sample*

Dependent Variable: FOOD\_EXP  
 Method: Least Squares  
 Date: 12/03/07 Time: 13:04  
 Sample: 1 20  
 Included observations: 20

$\hat{\sigma}_1^2 = (59.78939)^2 = 3574.8$

	Coefficient	Std. Error	t-Statistic	Prob.
C	72.96174	38.93435	1.878794	0.0766
INCOME	11.50038	2.507514	4.586367	0.0002
R-squared	0.538873	Mean dependent var		240.1830
Adjusted R-squared	0.513254	S.D. dependent var		85.69849
S.E. of regression	59.78939	Akaike info criterion		11.11417
Sum squared resid	64345.89	Schwarz criterion		11.21375

A similar exercise is followed for the second half of the sample.

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

food\_exp c income

*last 20 observations used to estimate  $\hat{\sigma}_2^2$*

Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 21 40

*second half of sample*

Dependent Variable: FOOD\_EXP  
 Method: Least Squares  
 Date: 12/03/07 Time: 13:20  
 Sample: 21 40  
 Included observations: 20

$\hat{\sigma}_2^2 = (113.6747)^2 = 12921.9$

	Coefficient	Std. Error	t-Statistic	Prob.
C	-24.91465	184.8249	-0.134729	0.8943
INCOME	14.26400	7.425093	1.921054	0.0707
R-squared	0.170142	Mean dependent var		326.9640
Adjusted R-squared	0.124028	S.D. dependent var		121.4566
S.E. of regression	113.6747	Akaike info criterion		12.39920
Sum squared resid	232594.7	Schwarz criterion		12.49877

```
scalar sig2_sq = @se^2
```

Then, the following two commands yield the required  $F$ -value as well as the 5% critical value to compare it against.

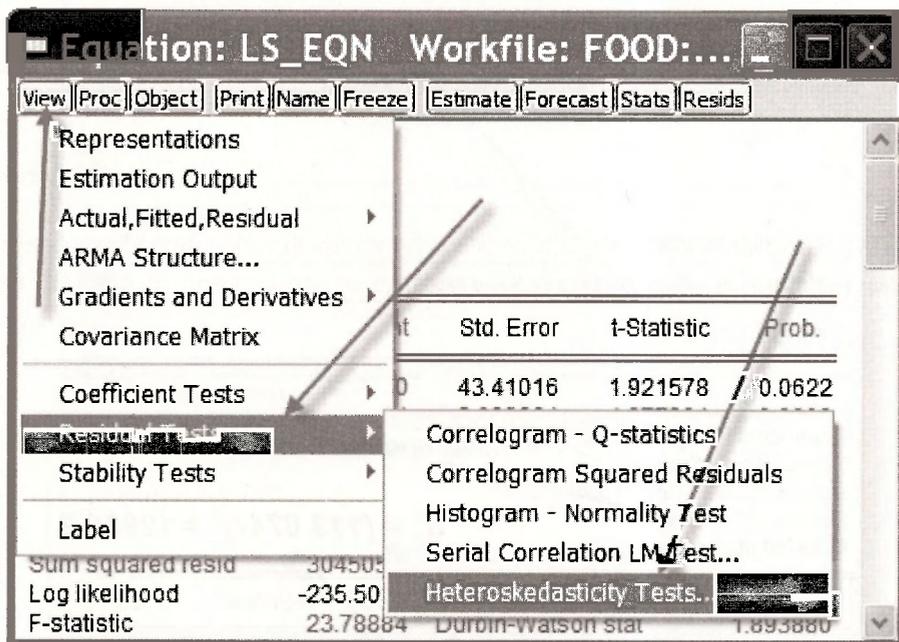
```
scalar f_val = sig2_sq/sig1_sq
scalar f_crit = @qfdist(0.95, 18, 18)
```

## 8.7 TESTING THE VARIANCE FUNCTION

There are a large number of alternative tests for heteroskedasticity based on an estimated variance function of the form

$$\hat{e}^2 = \alpha_1 + \alpha_2 z_2 + \alpha_3 z_3 + \dots + \alpha_S z_S + v$$

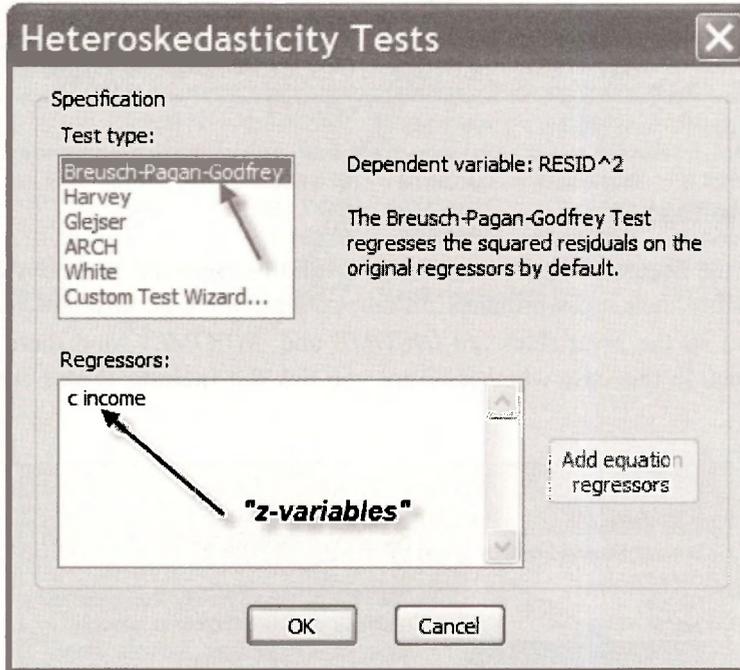
where  $\hat{e}^2$  are the squared least-squares residuals and  $z_2, z_3, \dots, z_S$  are the variance equation regressors. EViews has the capability to automatically compute test statistic values for these tests as well as their corresponding  $p$ -values. To locate this facility open the least-squares estimated equation and then select **View/Residual Tests/Heteroskedasticity Tests**.



A large number of possibilities – more than you have ever dreamed of – will appear. In line with p. 215 of the text, we will consider just two, the Breusch-Pagan test and the White test. We will also indicate where values for the tests described in Appendix 8B of the text can be found.

### 8.7.1 The Breusch-Pagan test

The Breusch-Pagan test (also called Breusch-Pagan-Godfrey test to recognize that Godfrey independently derived the test at about the same time as Breusch and Pagan) can be selected from the **Heteroskedasticity Tests** dialog box as indicated below. You have the option of selecting the “z-variables”. If you do nothing, EVIEWS will automatically insert those in the mean regression equation. For future reference, inserting *INCOME* and *INCOME*<sup>2</sup> leads to the White test example of the next section.



Heteroskedasticity Test: Breusch-Pagan-Godfrey				
F-statistic	8.603501	Prob. F(1,38)	0.0057	
Obs*R-squared	7.384424	Prob. Chi-Square(1)	0.0066	
Scaled explained SS	6.627901	Prob. Chi-Square(1)	0.0100	
Test Equation:				
Dependent Variable: RESID^2				
Method: Least Squares				
Date: 12/03/07 Time: 23:22				
Sample: 1 40				
Included observations: 40				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-5762.370	4823.501	-1.194645	0.2396
INCOME	682.2326	232.5920	2.933172	0.0057
R-squared	0.184611	Mean dependent var	7612.629	

Annotations on the table:

- Arrow from **equation (8B.2)** points to the **Obs\*R-squared** value (7.384424).
- Equation:  $7.3844 = 40 \times 0.1846$
- Arrow from **equation (8B.6)** points to the **INCOME** coefficient (682.2326).

The value of the chi-square statistic considered in the text is  $\chi^2 = N \times R^2 = 40 \times 0.1846 = 7.38$ . Its corresponding  $p$ -value is 0.0066, leading to rejection of  $H_0$  at a 5% significance level. The above screen shot shows where these values can be located on the output. There are also values for another two tests statistics on this output. If you are curious about where these values come from, read Appendix 8B. Using equations (8B.2) and (8B.6), you will discover



$$\bar{F} = \frac{(SST - SSE)/(S - 1)}{SSE/(N - S)} = \frac{(4.61075 \times 10^9 - 3.75956 \times 10^9)/1}{3.75956 \times 10^9/38} = 8.6035$$

$$\chi^2 = \frac{SST - SSE}{2\sigma_e^4} = \frac{4.61075 \times 10^9 - 3.75956 \times 10^9}{2 \times 89.517^4} = 6.628$$

### 8.7.2 The White test

The White test is the Breusch-Pagan test with the  $z$ -variables selected as the  $x$ -variables and their squares, and possibly their cross-products. In our example there is only one  $x$ -variable, namely  $x = INCOME$  and so the  $z$ -variables are  $INCOME$  and  $INCOME^2$ , and there are no possible cross product terms. In this case whether or not you tick the **Include White cross terms** box is irrelevant.

Specification	
Test type:	Dependent variable: RESID^2
<input type="checkbox"/> Breusch-Pagan-Godfrey <input type="checkbox"/> Harvey <input type="checkbox"/> Glejser <input type="checkbox"/> ARCH <input type="checkbox"/> Custom Test Wizard.	The White Test regresses the squared residuals on the the cross product of the original regressors and a constant.
	<input checked="" type="checkbox"/> Include White cross terms

After selecting **White** and clicking **OK**, the output below appears. The value of the chi-square statistic considered in the text is  $\chi^2 = N \times R^2 = 40 \times 0.18888 = 7.555$ . Its corresponding  $p$ -value is 0.0229, leading to rejection of  $H_0$  at a 5% significance level. Can you see where these values can be located on the output? The values for the other two tests statistics on this output come from equations (8B.2) and (8B.6) in Appendix 8B. After a little detective work, you will discover they are calculated as

$$F = \frac{(SST - SSE)/(S - 1)}{SSE/(N - S)} = \frac{(4.61075 \times 10^9 - 3.73989 \times 10^9)/2}{3.73989 \times 10^9/37} = 4.308$$

$$\chi^2 = \frac{SST - SSE}{2\sigma_e^4} = \frac{4.61075 \times 10^9 - 3.73989 \times 10^9}{2 \times 89.517^4} = 6.781$$

Heteroskedasticity Test: White				
F-statistic	4.307884	Prob. F(2,37)	0.0208	
Obs*R-squared	7.555079	Prob. Chi-Square(2)	0.0229	
Scaled explained SS	6.781072	Prob. Chi-Square(2)	0.0337	
Test Equation:				
Dependent Variable: RESID^2				
Method: Least Squares				
Date: 12/03/07 Time: 23:35				
Sample: 1 40				
Included observations: 40				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-2908.787	8100.109	-0.359104	0.7216
INCOME	291.7477	915.8462	0.318553	0.7519
INCOME^2	11.16829	25.30953	0.441150	0.6617
R-squared	0.188877	Mean dependent var	7612.629	

*equation (8B.2)*

$$7.555 = 40 \times 0.18888$$

*equation (8B.6)*

### Keywords

- |                           |                           |                             |
|---------------------------|---------------------------|-----------------------------|
| @qfdist                   | graph object              | sample                      |
| @se                       | heteroskedastic partition | scatter                     |
| @sqrt                     | heteroskedasticity tests  | sort                        |
| actual                    | least squares line: plot  | standard errors: White      |
| add text                  | legend                    | standardized residual graph |
| apply                     | line/shade                | transformed variables       |
| Breusch-Pagan test        | line/symbol               | type: graph                 |
| covariance matrix         | LS & TSLS options         | variance function           |
| estimation settings       | options: graph            | variance function: testing  |
| fitted                    | outliers                  | weight                      |
| forecast                  | plots                     | weighted least squares      |
| forecast name             | resid                     | weighted LS/TSLS            |
| F-test                    | residual graph            | White cross terms           |
| generalized least squares | residual tests            | White test                  |
| Goldfeld-Quandt test      | residuals                 | XY line                     |

# CHAPTER 9

## Dynamic Models, Autocorrelation and Forecasting

### CHAPTER OUTLINE

- 9.1 Least-Squares Residuals: Sugarcane Example
  - 9.1.1 Correlation between  $\hat{e}_t$  and  $\hat{e}_{t-1}$
- 9.2 Newey-West Standard Errors
- 9.3 Estimating an AR(1) Error Model
  - 9.3.1 A short way
  - 9.3.2 A long way
  - 9.3.3 A more general model
  - 9.3.4 Testing the AR(1) error restriction
- 9.4 Testing for Autocorrelation
  - 9.4.1 Residual correlogram
  - 9.4.2 Lagrange multiplier (*LM*) test
  - 9.4.3 Durbin-Watson test
- 9.5 Autoregressive Models
  - 9.5.1 Workfile structure for time series data
  - 9.5.2 Estimating an AR model
  - 9.5.3 Forecasting with an AR model
- 9.6 Finite Distributed Lags
- 9.7 Autoregressive Distributed Lag Models
  - 9.7.1 Graphing the lag weights

### KEYWORDS

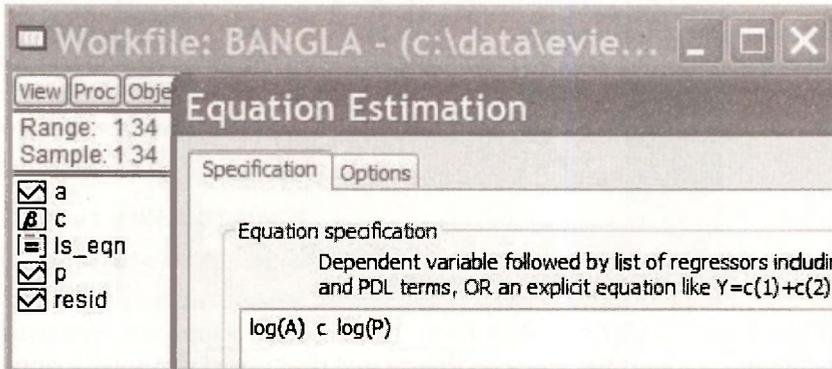
Chapter 9 is the first chapter in the text devoted to some of the special issues that are considered when estimating relationships with **time series data**. Time provides a natural ordering of the observations not present when using random cross-section observations, and it leads to dynamic features in regression equations. EViews has many options for handling such features. This chapter introduces some of those features.

### 9.1 LEAST-SQUARES RESIDUALS: SUGARCANE EXAMPLE

The first example considered in Chapter 9 is an area response model for sugarcane in Bangladesh where area sown to sugarcane  $A$  is related to price  $P$  by the equation

$$\ln(A_t) = \beta_1 + \beta_2 \ln(P_t) + e_t$$

In contrast to earlier chapters where the index used for the observations was mainly  $i$ , here we use the index  $t$  to denote time-series observations. We have 34 annual observations stored in the file *bangla.wfl*. The **Equation specification** and resulting least-squares output are

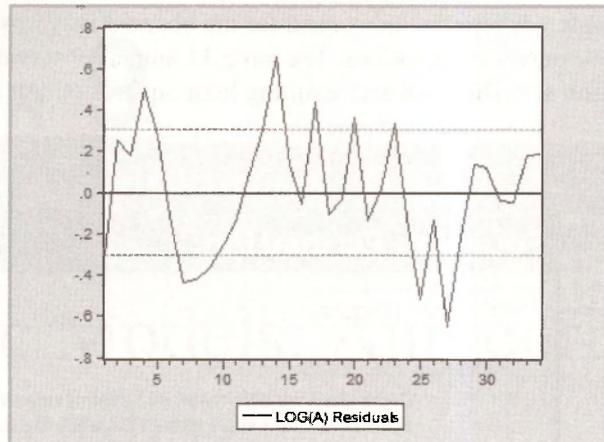


Dependent Variable: LOG(A)				
Method: Least Squares				
Sample: 1 34				
Included observations: 34				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.893256	0.061345	63.46486	0.0000
LOG(P)	0.776119	0.277467	2.797154	0.0087

We are interested in examining the residuals from this estimated equation, as displayed in Table 9.1 on page 233 of the text. Various ways of examining the residuals were described at the beginning of Chapter 8. As a first step for this example, we save them and then check them against the values that appear in Table 9.1. The command **series ehata = resid** saves the residuals as **EHAT**. To view them double click on **EHAT** and select **View/SpreadSheet**. The first and last 8 values are as follows.

EHAT	
1	-0.303029
2	0.254437
3	0.181515
4	0.503053
5	0.275078
6	-0.115483
7	-0.437147
8	-0.423488
26	-0.137079
27	-0.651414
28	-0.218325
29	0.136647
30	0.121095
31	-0.039715
32	-0.048179
33	0.182829
34	0.183576

A plot of the residuals against time can indicate whether positive residuals tend to follow positive residuals and negative residuals tend to follow negative residuals – a sign of positive autocorrelation. To obtain the plot in Figure 9.3 open the least-squares estimated equation and go to **View/Actual, Fitted, Residual/Residual Graph**. The following graph appears.



To save this graph in your workfile, click on **Freeze** and then **Name** and enter a suitable name.

Name to identify object

figure\_9\_3 24 characters maximum, 16 or fewer recommended

There are other ways to create this graph. For example, you could open the series **EHAT** and then select **View/Graph/Basic graph/Line & Symbol**. Clicking **OK** will produce the graph. If you then follow up by clicking **Freeze**, you will be able to edit and save the graph.

### 9.1.1 Correlation between $\hat{e}_t$ and $\hat{e}_{t-1}$

The sample correlation between the least squares residuals  $\hat{e}_t$  and their lagged values  $\hat{e}_{t-1}$ , is an important quantity for assessing whether or not the equation errors are autocorrelated. To compute this quantity we begin by creating the variable  $\hat{e}_{t-1}$  and giving it the name **EHAT\_1**. The EViews command is

**series ehat\_1 = ehat(-1)**

Writing **ehat(-1)** has the effect of lagging the observations in **EHAT** by one period. To appreciate how lagged observations are stored, we create a group containing **EHAT** and **EHAT\_1** and examine the first few observations in the spreadsheet. They are illustrated on the following page. Notice what has happened. The 2nd observation for **EHAT\_1** is  $\hat{e}_1$ , the 3rd observation is  $\hat{e}_2$ , and so on. Because there is no observation  $\hat{e}_0$ , the first observation on **EHAT\_1** is “not available” and is recorded as **NA**. When asked to perform calculations that include this first observation, EViews will omit it.

Group: LAG_OF_EHAT W							
	View	Proc	Object	Print	Name	Freeze	Default
obs	EHAT		EHAT_1				
1	-0.303029						NA
2	0.254437						-0.303029
3	0.181515						0.254437
4	0.503053						0.181515
5	0.275078						0.503053

Now consider the first-order correlation  $r_1$  given in equation (9.18) on page 234 of the text

$$r_1 = \frac{\sum_{t=2}^T \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^T \hat{e}_t^2}$$

The numerator and denominator of this quantity can be computed using the following commands

```
series ee1 = ehat*ehat_1
scalar sum_ee1 = @sum(ee1)           (= \sum_{t=2}^T \hat{e}_t \hat{e}_{t-1})
series e1e1 = ehat_1*ehat_1
scalar sum_e1e1 = @sum(e1e1)       (= \sum_{t=2}^T \hat{e}_{t-1}^2)
```

We have created two new series,  $\hat{e}_t \hat{e}_{t-1}$  and  $\hat{e}_{t-1}^2$ , and then found their sum. The first observation in each of the series **EE1** and **E1E1** will be **NA**. The values obtained for their sums are  $\sum_{t=2}^T \hat{e}_t \hat{e}_{t-1} = 1.196874$  and  $\sum_{t=2}^T \hat{e}_{t-1}^2 = 2.997871$ , leading to a value for  $r_1$  of

$$r_1 = \frac{1.196874}{2.997871} = 0.399$$

This value differs slightly from that reported in the text which is  $r_1 = 0.404$ . The text value was obtained using the **EViews** command

```
scalar r1_text = @cor(ehat, ehat_1)
```

where **@cor(x1,x2)** is the **EViews** function for computing the correlation between two series **X1** and **X2**. The reason for the discrepancy is that, after omitting the first observation (or the last observation), the sample mean for  $\hat{e}_t$  is no longer zero. The formula used by the **@cor** function is

$$r_1 = \frac{\sum_{t=2}^T (\hat{e}_t - \bar{\hat{e}}_{[-1]}) (\hat{e}_{t-1} - \bar{\hat{e}}_{[-T]})}{\sum_{t=2}^T (\hat{e}_{t-1} - \bar{\hat{e}}_{[-T]})^2}$$

where  $\bar{\hat{e}}_{[-1]}$  is the sample mean of the  $\hat{e}_t$ , with the first observation excluded and  $\bar{\hat{e}}_{[-T]}$  is the sample mean of the  $\hat{e}_t$ , with the last observation excluded. In general the difference between the two alternative formulas will be slight and it disappears as the sample size gets larger.

If having two different formulas for  $r_1$  worries you, it may help to remember that they are simply two alternative estimators for the population correlation between the error and its lag

$$\rho = \frac{E(e_t e_{t-1})}{E(e_t^2)}$$

Having different estimators for the same population quantity is not unusual. The least squares and generalized least squares estimators in Chapter 8 are examples.

Now that you are comfortable with the idea of two estimators for the same population quantity, it is convenient to introduce one more. A 3rd estimator for  $\rho$  is relevant later in this chapter when we explain how EViews computes a sample **correlogram**. When sample size is large, the difference between it and the other estimators will be negligible. Its formula and value for the sugarcane residuals are

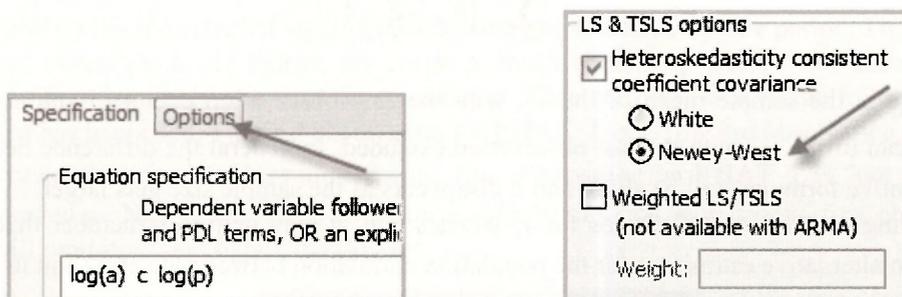
$$r_1 = \frac{\sum_{t=2}^T \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^T \hat{e}_t^2} = \frac{1.196874}{3.031571} = 0.395$$

Notice that the denominator includes all observations on  $\hat{e}$ . We can use EViews to compute this version of  $r_1$  as follows.

```
series ee = ehat*ehat
scalar sum_ee=@sum(ee)           (=  $\sum_{t=1}^T \hat{e}_t^2$ )
scalar r1_c = sum_ee1/sum_ee
```

## 9.2 NEWEY-WEST STANDARD ERRORS

In Chapter 8 when studying heteroskedasticity, we saw how least squares could be used instead of generalized least squares as long as we used White standard errors. A similar option exists for regression models with autocorrelated errors. In this case the standard errors are called Newey-West or HAC standard errors, with HAC being an acronym for heteroskedasticity-autocorrelation consistent. To compute the Newey-West standard errors for the sugarcane example, as reported on page 235 of the text, we choose **Options** in the **Equation Estimation** window. Then, in the Options window, go to the **LS & TSLS options** section, tick the **Heteroskedasticity consistent coefficient covariance** box, and select **Newey-West**. The Newey-West standard errors are consistent under both heteroskedasticity and autocorrelation.



The least-squares output with the corrected standard errors follows. Notice that Eviews has a note to tell you that it has calculated Newey-West standard errors.

Dependent Variable: LOG(A)				
Method: Least Squares				
Sample: 1 34				
Included observations: 34				
Newey-West HAC Standard Errors & Covariance (lag truncation=3)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.893256	0.062444	62.34761	0.0000
LOG(P)	0.776119	0.378207	2.052102	0.0484

### 9.3 ESTIMATING AN AR(1) ERROR MODEL

Continuing with the sugar cane example, we are interested in estimating the supply equation under the assumption that the errors follow an AR(1) model. These two components of the model can be written as

$$\ln(A_t) = \beta_1 + \beta_2 \ln(P_t) + e_t \quad e_t = \rho e_{t-1} + v_t$$

The main parameters to be estimated are  $\beta_1$ ,  $\beta_2$  and  $\rho$ . There are two error variances,  $\sigma_v^2$  and  $\sigma_e^2$ . The procedures we describe provide an estimate for  $\sigma_v^2$ . Once we have estimated  $\rho$  and  $\sigma_v^2$ , we can always estimate  $\sigma_e^2$  from the relationship  $\sigma_e^2 = \sigma_v^2 / (1 - \rho^2)$ . It is instructive to consider two ways of obtaining the nonlinear least squares estimates reported in equation (9.25) on page 237 of the text. The short way is to simply tell EViews to assume an AR(1) error in the **Equation specification** box. The long way is to write equation (9.24) in the **Equation specification** box.

#### 9.3.1 A short way

To estimate a model with an AR(1) error we begin, as usual, by selecting **Object/New Object/Equation**. After giving the equation object a name and clicking **OK**, the **Equation specification** box appears. Then, as before, you enter the names of the series that are in the equation, but this time you also add **AR(1)** to tell EViews the errors follow an AR(1) model.

Equation specification	
Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$ .	
log(a) c log(p) AR(1)	<b>Tells EViews to assume the error is an AR(1)</b>

You will notice several new features in the output that follows.

1. An estimate  $\hat{\rho} = 0.42214$  is provided next to the name **AR(1)**.
2. The **S.E of regression** is the estimate  $\hat{\sigma}_v = 0.2854$ .
3. The lagged variables in the equation lead to a loss of one observation. EViews automatically changes the **Sample** from **1 34** to **2 34**, and reports that 33 observations are

included. If why this happens is not clear to you, be patient. More will be said about how the lagged variables lead to one less observation when we move on to the “long way”.

- The note **Convergence achieved after 7 iterations** appears because of the nature of the nonlinear least squares estimator. This estimator is not a formula that calculates the required numbers. It is an iterative procedure that systematically tries different parameter values until it finds those that minimize the sum of squared residuals. The 7 iterations refer to the 7 different sets of parameters tried before it reached the minimum. If it fails to reach the minimum, the note will say **convergence not achieved**.

Dependent Variable: LOG(A) *lag leads to one less observation*  
 Method: Least Squares  
 Date: 12/06/07 Time: 13:00  
 Sample (adjusted): 2 34  
 Included observations: 33 after adjustments  
 Convergence achieved after 7 iterations *nonlinear estimator iterates*

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.898771	0.092165	42.30197	0.0000
LOG(P)	0.888370	0.259299	3.426048	0.0018
AR(1)	0.422140	0.166047	2.542284	0.0164

R-squared	0.277777	Mean dependent var	3.999309
Adjusted R-squared	0.229629	S.D. dependent var	0.325164
S.E. of regression	0.285399	Akaike info criterion	0.416650
Sum squared resid	2.443575	Schwarz criterion	0.552696

### 9.3.2 A long way

An alternative way of writing the AR(1) error model is

$$\ln(A_t) = \beta_1(1 - \rho) + \beta_2 \ln(P_t) + \rho \ln(A_{t-1}) - \rho\beta_2 \ln(P_{t-1}) + v_t$$

See page 236 of the text for a derivation of this result. We have made the substitutions  $y_t = \ln(A_t)$  and  $x_t = \ln(P_t)$ . Using  $C(1) = \beta_1$ ,  $C(2) = \beta_2$  and  $C(3) = \rho$ , this equation can be estimated by writing it directly into the **Equation specification** window.

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

$\log(a) = c(1)*(1-c(3)) + c(2)*\log(p) + c(3)*\log(a(-1)) - c(2)*c(3)*\log(p(-1))$

*A(-1) and P(-1) means A and P lagged one period*

Can you see what is different? Instead of writing in the name of the dependent variable followed by the explanatory variables, we have written out the whole equation. Also, the EVIEWS notation

for  $A_{t-1}$  and  $P_t$ ; is  $\mathbf{a}(-1)$  and  $\mathbf{p}(-1)$ , respectively. What would happen if we tried to estimate the equation using

$$\log(a) \quad c \quad \log(p) \quad \log(a(-1)) \quad \log(p(-1))$$

We would obtain estimates for 4 coefficients, one associated with each of the series  $\mathbf{c}$ ,  $\log(\mathbf{p})$ ,  $\log(\mathbf{a}(-1))$ , and  $\log(\mathbf{p}(-1))$ . In our case we only have 3 coefficients, and there is not an obvious way of associating them with each variable. This is the reason for writing out the equation in full. In general, equations which are nonlinear in the coefficients or that involve restrictions on the coefficients need to be written out in full.

It is useful to examine the lagged variables  $A_{t-1} = \mathbf{a}(-1)$  and  $P_{t-1} = \mathbf{p}(-1)$  in more detail. The following spreadsheet contains the first 5 observations on  $A_t$ ,  $A_{t-1}$ ,  $P_t$  and  $P_{t-1}$ . Notice that lagging has the effect of making the first observations for  $A_{t-1}$  and  $P_{t-1}$  not available. Accordingly, EViews omits it when carrying out estimation.

obs	A	A(-1)	P	P(-1)
1	28.96000	NA	0.749000	NA
2	67.81000	28.96000	1.093000	0.749000
3	55.15000	67.81000	0.920000	1.093000
4	78.62000	55.15000	0.960000	0.920000
5	60.15000	78.62000	0.912000	0.960000

The output appears below. The first thing you should notice in this output is that the results are identical to those obtained the "short way". The equation specifications for the short way and the long way are two different ways of telling EViews to do the same thing, namely, find values for  $\beta_1$ ,  $\beta_2$  and  $\rho$  that minimize

$$\sum_{t=2}^T v_t^2 = \sum_{t=2}^T (\ln(A_t) - \beta_1(1-\rho) - \beta_2 \ln(P_t) - \rho \ln(A_{t-1}) + \rho \beta_2 \ln(P_{t-1}))^2$$

Notice also that the sample has been adjusted to omit the first observation, convergence took 13 iterations in this case, and EViews writes out the equation that has been estimated so that you can readily see where each of the coefficients appears in the equation.

Dependent Variable: LOG(A)				
Method: Least Squares				
Sample (adjusted): 2 34				
Included observations: 33 after adjustments				
Convergence achieved after 13 iterations				
LOG(A) = C(1)*(1-C(3)) + C(2)*LOG(P) + C(3)*LOG(A(-1)) - C(2)*C(3)*LOG(P(-1))				
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	3.898771	0.092166	42.30155	0.0000
C(3)	0.422139	0.166047	2.542281	0.0164
C(2)	0.888372	0.259298	3.426062	0.0018
R-squared	0.277777	Mean dependent var	3.999309	
Adjusted R-squared	0.229629	S.D. dependent var	0.325164	
S.E. of regression	0.285399	Akaike info criterion	0.416650	
Sum squared resid	2.443575	Schwarz criterion	0.552696	

### 9.3.3 A more general model

In the previous section we asked what would happen if we used the following Equation specification

<p><b>Equation specification</b>          Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like <math>Y=c(1)+c(2)*X</math>.</p>
<p><math>\log(A) \ c \ \log(P) \ \log(P(-1)) \ \log(A(-1))</math></p>

In this case we are estimating the model

$$\ln(A_t) = \delta + \delta_0 \ln(P_t) + \delta_1 \ln(P_{t-1}) + \theta_1 \ln(A_{t-1}) + v_t$$

This model has the same variables as the AR(1) error model of the previous section, but it has 4 coefficients instead of 3. It is a more general model, known as an ARDL(1,1) model, that reduces to the AR(1) error model when  $\delta_1 = -\theta_1 \delta_0$ . ARDL models are discussed later in this chapter. The results below appear in equation (9.28) on page 239 of the text.

Dependent Variable: LOG(A)				
Method: Least Squares				
Sample (adjusted): 2 34				
Included observations: 33 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.366173	0.655701	3.608615	0.0011
LOG(P)	0.776629	0.279813	2.775530	0.0095
LOG(P(-1))	-0.610862	0.296644	-2.059245	0.0485
LOG(A(-1))	0.404284	0.166624	2.426323	0.0217

### 9.3.4 Testing the AR(1) error restriction

The restriction  $\delta_1 = -\theta_1 \delta_0$  can be tested using a Wald test with hypotheses  $H_0 : \delta_1 = -\theta_1 \delta_0$  and  $H_1 : \delta_1 \neq -\theta_1 \delta_0$ . It differs from the Wald tests we considered in Chapter 6 because the hypothesis is a nonlinear function of the coefficients. Nevertheless, EViews can perform the test using the same procedures described in Chapter 6. After estimating the equation, select **View/Coefficient Tests/Wald Coefficient Restrictions**. Recognizing that  $C(2) = \delta_0$ ,  $C(3) = \delta_1$  and  $C(4) = \theta_1$ , the null hypothesis is entered in the **Wald Test** dialog box as follows

<p><b>Coefficient restrictions</b> separated by commas</p> <p><math>C(3) = -C(2)*C(4)</math></p>
--

Clicking **OK**, yields the following test output.

Wald Test: Equation: GENERAL			
Test Statistic	Value	df	Probability
F-statistic	1.115006	(1, 29)	0.2997
Chi-square	1.115006	1	0.2910
Null Hypothesis Summary:			
Normalized Restriction (= 0)	Value	Std. Err.	
C(3) + C(2)*C(4)	-0.296884	0.281157	
Delta method computed using analytic derivatives.			

The test is performed in the same way as described in Chapter 6, although, because of the nonlinearity of the hypothesis, the formulas for the  $F$ - and  $\chi^2$ - statistics are different. These formulas as well as the **delta method** that is used to compute  $\text{se}(\hat{\delta}_1 + \hat{0}_1\hat{\delta}_0) = 0.281157$  are things that you will learn in a later stage of your econometric career. Since  $p\text{-value} = 0.29 > 0.05$ , we do not reject the restriction implied by the AR(1) error model. In this case the **normalized restriction** is  $\delta_1 + \theta_1\delta_0 = 0$  and its estimated left hand side is  $\hat{\delta}_1 + \hat{\theta}_1\hat{\delta}_0 = -0.296884$ .

## 9.4 TESTING FOR AUTOCORRELATION

### 9.4.1 Residual correlogram

Autocorrelation exists when the equation error  $e_t$  is correlated with any of its past values  $e_{t-1}, e_{t-2}, \dots$ . One way to investigate the possible existence of such correlation is to obtain the least squares residuals  $\hat{e}_t$  and to check whether the sample correlations between  $\hat{e}_t$  and  $\hat{e}_{t-1}, \hat{e}_{t-2}, \dots$  are significantly different from zero. The sequence of these correlations  $r_1, r_2, \dots$  is called the residual correlogram. Earlier in this chapter (Section 9.1.1) we saw that there are three slightly different formulas for computing  $r_1$ . Consider the correlation at a general lag  $k$ . The formula that EViews uses for computing  $r_k$  (the correlation between  $\hat{e}_t$  and  $\hat{e}_{t-k}$ ) is

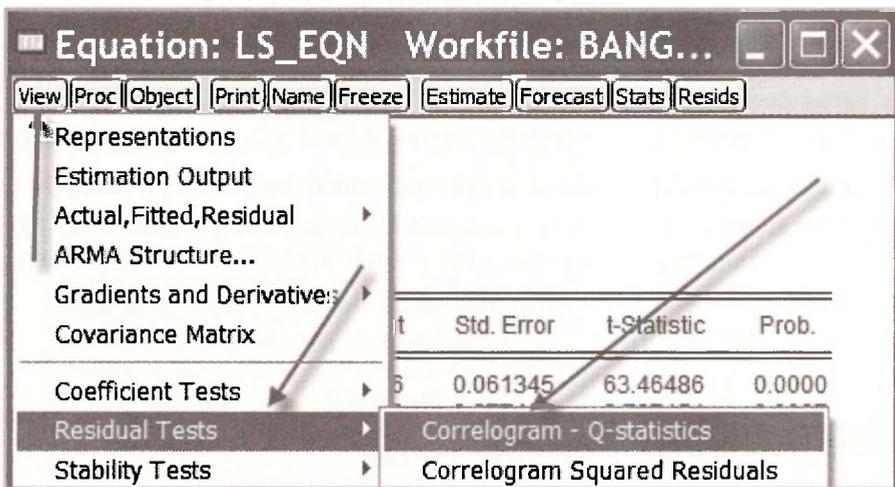
$$r_k = \frac{\sum_{t=k+1}^T \hat{e}_t \hat{e}_{t-k}}{\sum_{t=1}^T \hat{e}_t^2}$$

Another possible formula omits the last  $k$  terms in the summation in the denominator which then becomes  $\sum_{t=1}^{T-k} \hat{e}_t^2 = \sum_{t=k+1}^T \hat{e}_{t-k}^2$ . A third alternative is the EViews function **@cor( $\hat{e}_t, \hat{e}_{t-1}$ )**. It computes a mean-corrected version whose formula is

$$r_k = \frac{\sum_{t=k+1}^T (\hat{e}_t - \bar{\hat{e}}_{[last\ T-k]}) (\hat{e}_{t-k} - \bar{\hat{e}}_{[first\ T-k]})}{\sum_{t=k+1}^T (\hat{e}_{t-k} - \bar{\hat{e}}_{[first\ T-k]})^2}$$

where  $\bar{\hat{e}}_{[last\ T-k]}$  is the sample mean of the  $\hat{e}_t$  for the last  $T - k$  observations, and  $\bar{\hat{e}}_{[first\ T-k]}$  is the sample mean of the  $\hat{e}_t$  for the first  $T - k$  observations. In what follows we will report the EViews residual correlogram and describe how to obtain it. We will also explain any discrepancies with values in the text.

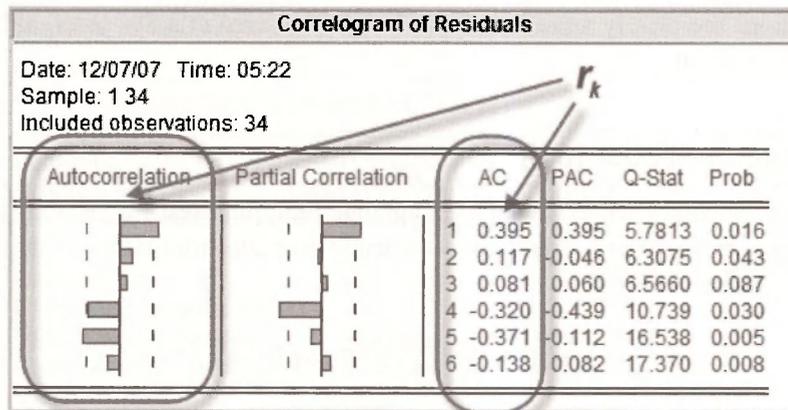
The EViews version of Figure 9.4 on page 241 of the text is obtained by first returning to the original least-squares estimated equation and selecting **View/Residual Tests/Correlogram - Q statistics**



You will then be faced with the following **Lag Specification** window. **Lags to include** is the number of correlations  $r_1, r_2, \dots, r_k$  that you would like EViews to calculate. In line with Figure 9.4, we choose 6. As we will see later in the chapter, larger numbers can be chosen when the sample size is larger.



Information on the  $r_k$  is presented in two ways. The numerical values appear in the column **AC**. A bar chart with each bar reflecting the magnitude and sign of each  $r_k$  is given in the column headed **Autocorrelation**. Bars long enough to obscure one of the dotted lines signify autocorrelations that are significantly different from zero at a 5% significance level.



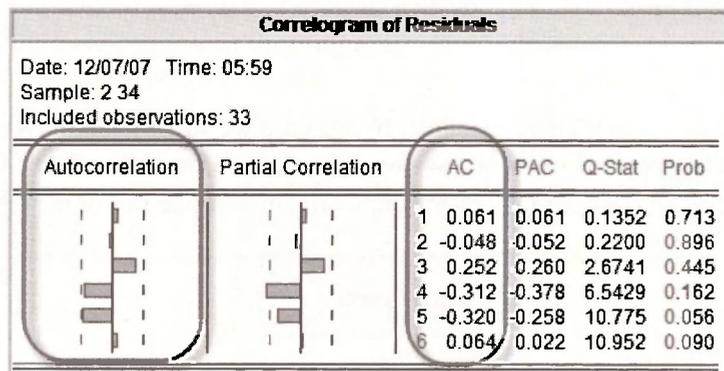
We will not be concerned with the remaining information. **Partial Correlation (PAC)**, **Q-Stat** and their *p*-values, **Prob**, are covered in specialist time-series courses.

The orientation of the EViews graph is different to that of Figure 9.4, and the values of the correlations are slightly different, but the message is the same. The residual correlations for lags 1 and 5 are significantly different from zero at a 5% significance level. That at lag 4 is marginal.

The correlations computed from the three alternative formulas are given in the table below. Those given on page 240 of the text correspond to those from the mean corrected formula.

Correlation	Not mean corrected	Mean corrected	<i>T</i> observations in denominator
$r_1$	0.399	0.404	0.395
$r_2$	0.120	0.122	0.117
$r_3$	0.083	0.084	0.081
$r_4$	-0.327	-0.353	-0.320
$r_5$	-0.381	-0.420	-0.371
$r_6$	-0.143	-0.161	-0.138

After estimating the model assuming that the errors follow an AR(1) model, one would hope that the new residuals, the  $\hat{v}_t$ , no longer exhibit autocorrelation. We can check them out by examining the residual correlogram from the estimated AR(1) error model. After opening the equation object for that model, and following the steps described above, we get the EViews version of Figure 9.5 on page 242 of the text.



No autocorrelations are clearly significantly different from zero at a 5% level, although those at lags 4 and 5 are marginal.

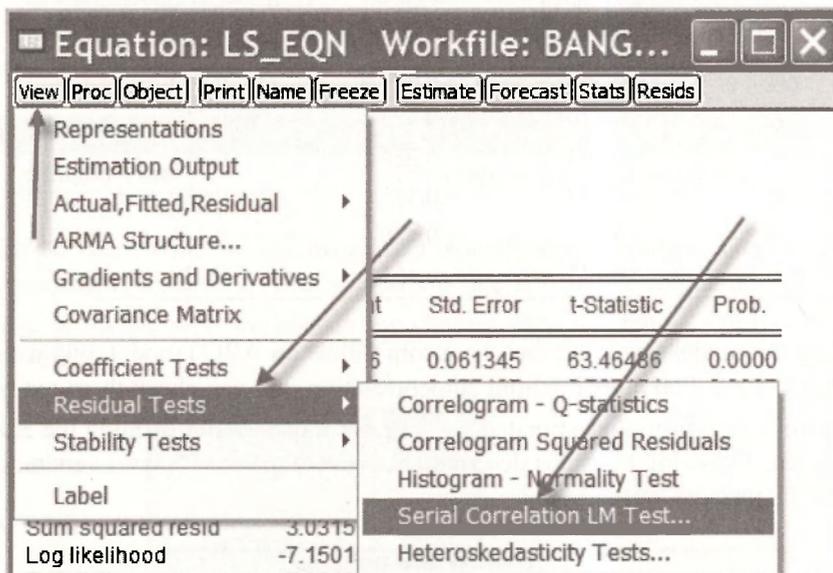
### 9.4.2 Lagrange multiplier (LM) test

In the context of the sugarcane example, the Lagrange multiplier test for an AR(1) error is a test of the significance of  $\hat{\rho}$ , where  $\hat{\rho}$  is the least squares estimate from either of the following two equations.

$$\log(A_t) = \beta_1 + \beta_2 \log(P_t) + \rho \hat{e}_{t-1} + v_t$$

$$\hat{e}_t = \gamma_1 + \gamma_2 \log(P_t) + \rho \hat{e}_{t-1} + v_t$$

In both cases the  $\hat{e}_t$  are the least squares residuals. We will focus on the second equation. Both equations yield identical results for  $F$ - and  $t$ -tests on the significance of  $\hat{\rho}$ . The second equation has the advantage of producing a further test value of the form  $LM = T \times R^2$ . To obtain these values re-open the least squares estimated equation and select **View/Residual Tests/Serial Correlation LM Test**.



You will be asked how many lags to include. In this case we specify just 1. We are interested in testing for an AR(1) error and we only have one lag of  $\hat{e}_t$  on the right side of the equation. The correlogram was used to consider the general autocorrelation properties of the residuals.

Lags to include:

The test results appear in the following output.

Breusch-Godfrey Serial Correlation LM Test				
F-statistic	5.949152	Prob. F(1,31)		0.0206
Obs*R-squared	5.474312	Prob. Chi-Square(1)		0.0193
Test Equation: Dependent Variable: RESID Method: Least Squares Date: 12/07/07 Time: 07:42 Sample: 1 34 Included observations: 34 Presample missing value lagged residuals set to zero.				
	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.008116	0.057186	-0.141927	0.8881
LOG(P)	0.091601	0.260934	0.351012	0.7279
RESID(-1)	0.407821	0.167202	2.439088	0.0206
R-squared	0.161009	Mean dependent var		2.45E-17

The two test values and their corresponding  $p$ -values are given at the top of the output. The value  $F = 5.949$  is a test of the significance of  $\hat{\rho}$ , the coefficient of **RESID(-1)** that appears in the bottom half of the output. Because  $F = 5.949 = t^2 = 2.439^2$ , the test can be performed as a  $t$ - or an  $F$ -test and the  $p$ -value of 0.0206 is the same in both cases. The other test is a  $\chi^2$ - test with the test value being given by  $LM = T \times R^2 = 34 \times 0.161 = 5.474$ , and a  $p$ -value of 0.0193. In both cases a null hypothesis of  $H_0 : \rho = 0$  is rejected at a 5% significance level. Make sure that you can locate these various values on the output. And check them against p.242-3 of the text.

### 9.4.3 Durbin-Watson test

You may have noticed a Durbin-Watson value that is automatically provided on the least squares output. The Durbin Watson test is a test for AR(1) errors. It considered in Appendix 9B of the text. Its critical values and  $p$ -values are less readily computed than those for other tests for AR(1) errors, and so its popularity as a test is declining. Although EViews computes the value of the test statistic, it does not have commands for computing corresponding critical or  $p$ -values. As a rough guide, values of the Durbin-Watson statistic of 1.3 or less could be suggestive of autocorrelation. The value from the least-squares estimated sugarcane equation is 1.169.

Log likelihood	-7.150159	Hannan-Quinn criter	0.568864
F-statistic	7.824072	Durbin-Watson stat	1.168987
Prob(F-statistic)	0.008653		

## 9.5 AUTOREGRESSIVE MODELS

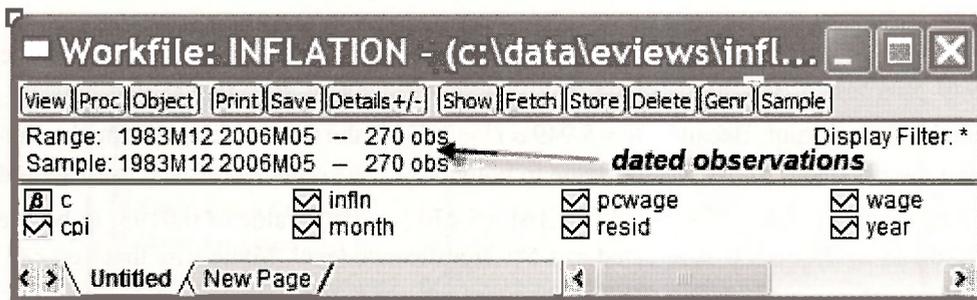
Autoregressive models can be specified not just for errors in an equation, but also for observable variables of interest. Furthermore, general models with more lags than the 1 assumed for the sugarcane example can be specified. In Section 9.5 of the text we are concerned with an AR(3) model for the inflation rate. It is given by

$$INFLN_t = \delta + \theta_1 INFLN_{t-1} + \theta_2 INFLN_{t-2} + \theta_3 INFLN_{t-3} + v_t$$

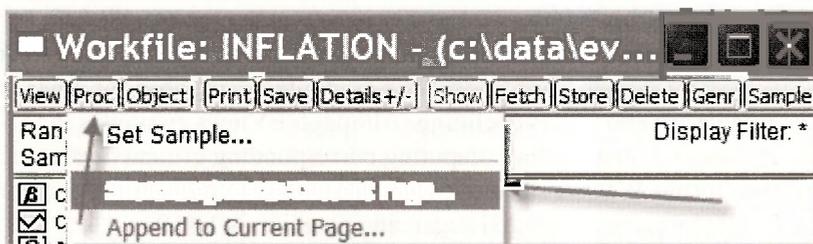
Data for inflation are stored in the file *inflation.wfl*, along with a number of other variables. Let us examine some special characteristics of this file and the observations on *CPI* and *INFLN*.

### 9.5.1 Workfile structure for time series data

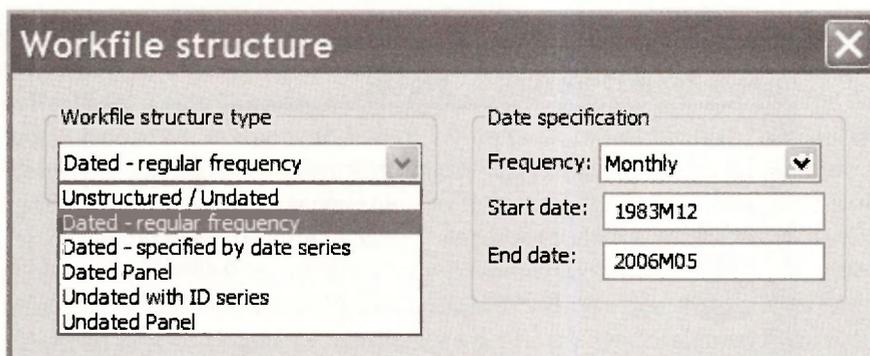
In the following screenshot the series in the file *inflation.wfl* are displayed. Notice how the **Range** and the **Sample** are specified. They contain dates. So far in the book we have mainly been concerned with cross-section observations which do not necessarily have a natural ordering and which were simply numbered from 1 to the number of the last observation. EViews calls workfiles with such observations **Unstructured/Undated**. We will check out the other alternatives.



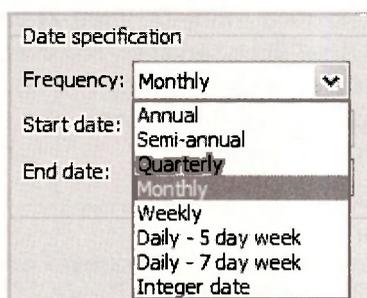
To examine the **workfile structure** of the workfile *inflation.wfl*, select **Proc/Structure/Resize Current Page**.



In the left panel of the Workfile structure window you can see a list of the **Workfile structure types**. When using cross-section data in earlier examples, the structure was **Unstructured / Undated**. Now we have moved on to time series data with specific dates for the observations, we use the **Dated – regular frequency** structure. In the **Date specification** panel on the right side the observations have been designated as **Monthly** in the **Frequency** box with **1983M12** (December, 1983) and **2006M05** (May, 2006) as the **Start date** and **End date**, respectively.



The other alternatives for a **Date specification** are illustrated below. The **integer date** option is used when specific dates have not been assigned to the observations. Such was the case with the Bangladesh sugarcane data that were simply allocated integer dates from 1 to 34.



It should be kept in mind that the dating of observations is generally a convenience factor, not one that has a bearing on your results from estimation. As long as the sequence of the observations is the same it does not matter how you label them. Panel data situations that are dealt with later in the book are an exception. In this case it is important to set up the labeling to distinguish between time periods and cross sections, but providing this is done, the labeling of the time series observations does not matter. How to set up your workfile structure when reading data from another source is covered in Chapter 17.

### 9.5.2 Estimating AR models

Now that we have finished our short digression on workfile structure for time series data, we return to the AR(3) model for inflation. The first few observations in a spreadsheet for a group containing *CPI* and *INFLN* are given below, superimposed on the workfile window. There are 270 observations on *CPI*. The series *INFLN* has been generated using the command

```
series infln = (log(CPI) - log(CPI(-1)))*100
```

Because **CPI(-1)** is needed to compute *INFLN*, no observation on *INFLN* is available for the first observation in December, 1983. EViews records it as NA. That leaves 269 observations for estimating the AR(3) model. The need for values of  $INFLN_{t-1}$ ,  $INFLN_{t-2}$  and  $INFLN_{t-3}$  in the estimation process reduces the sample size for estimation by a further 3 to 266. This will become more apparent as we consider the results from estimation.

Range: 1983M12 2006M05 - 270 obs      Dist

Sample: 1983M12 2006M05 - 270 obs

<input checked="" type="checkbox"/> c			
<input checked="" type="checkbox"/> cpi			
<input checked="" type="checkbox"/> cpi_inf			
<input checked="" type="checkbox"/> infln	obs	CPI	INFLN
<input checked="" type="checkbox"/> month	1983M12	101.4000	NA
<input checked="" type="checkbox"/> pcwage	1984M01	102.1000	0.687963
<input checked="" type="checkbox"/> resid	1984M02	102.6000	0.488521
<input checked="" type="checkbox"/> wage	1984M03	102.9000	0.291971
<input checked="" type="checkbox"/> year			

The AR model can be estimated using least squares and so estimation proceeds using the familiar EViews **Equation specification** window. The only new feature, but not something that is totally new, is how to specify the lagged variables  $INFLN_{t-1}$ ,  $INFLN_{t-2}$  and  $INFLN_{t-3}$  as explanatory variables. We do that using the notation **infln(-1)**, **infln(-2)** and **infln(-3)** as illustrated below. There is a shorter way of writing these three variables, however, one that is particularly useful if the number of lags is large. This short version is **infln(-1 to -3)**. The only other difference at this stage is the nature of the **Sample** setting. In line with the workfile structure the sample is described in terms of the dates of the observations.

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

infln c infln(-1) infln(-2) infln(-3)

**dated sample**

**short version**  
 infln c infln(-1 to -3)

Estimation settings

Method: LS / Least Squares (NLS and ARMA)

Sample: 1983m12 2006m05

The output that appears matches that in equation (9.37) on page 245 of the text. Note again the way in which the sample is expressed and that it has been adjusted to accommodate the observations lost through lagging, leaving a total of 266.

Dependent Variable: INFLN  
 Method: Least Squares  
 Sample (adjusted): 1984M04 2006M05  
 Included observations: 266 after adjustments

Variables	Coefficient	Std. Error	t-Statistic	Prob.
C	0.188335	0.025290	7.446877	0.0000
INFLN(-1)	0.373292	0.061481	6.071690	0.0000
INFLN(-2)	-0.217919	0.064472	-3.380029	0.0008
INFLN(-3)	0.101254	0.061268	1.652641	0.0996

Ideally, the residuals from the estimated AR model should not exhibit any autocorrelation. This fact can be checked by examining the residual correlogram. After opening the equation object, select **View/Residual Tests/Correlogram – Q statistics**. EViews will ask you for the number of lags to include. Choose **24**, in line with Figure 9.6 on page 246 of the text. The correlogram, with a host of information, will appear. We are primarily interested in the autocorrelations and whether they are significantly different from zero. In the following screenshots we have isolated the bar charts of the correlations for Figures 9.6, 9.7 and 9.8. Relative to the figures in the text, the EViews bar charts are rotated 90 degrees. They have the lag on the “y-axis” and the correlations on the “x-axis”. Check the EViews version of Figure 9.6. We see that all correlations are very small, with those at lags 6, 11 and 13 marginally significant. Discussion of Figures 9.7 and 9.8 is deferred until later sections.

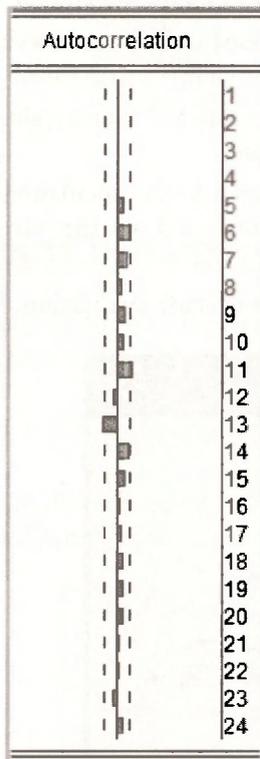


Figure 9.6

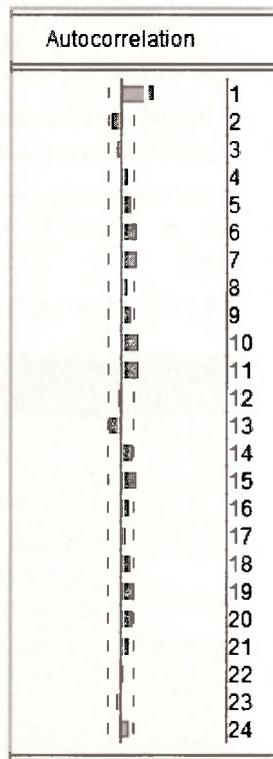


Figure 9.7

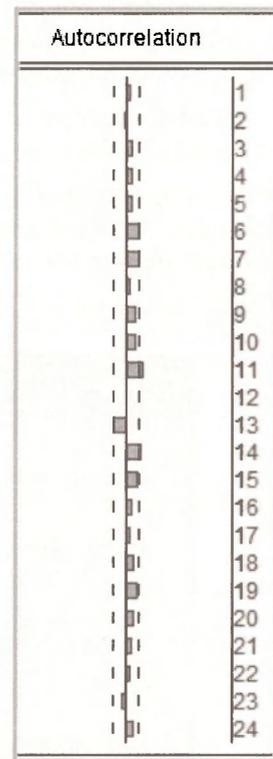
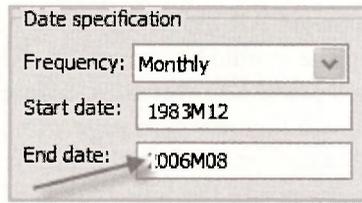


Figure 9.8

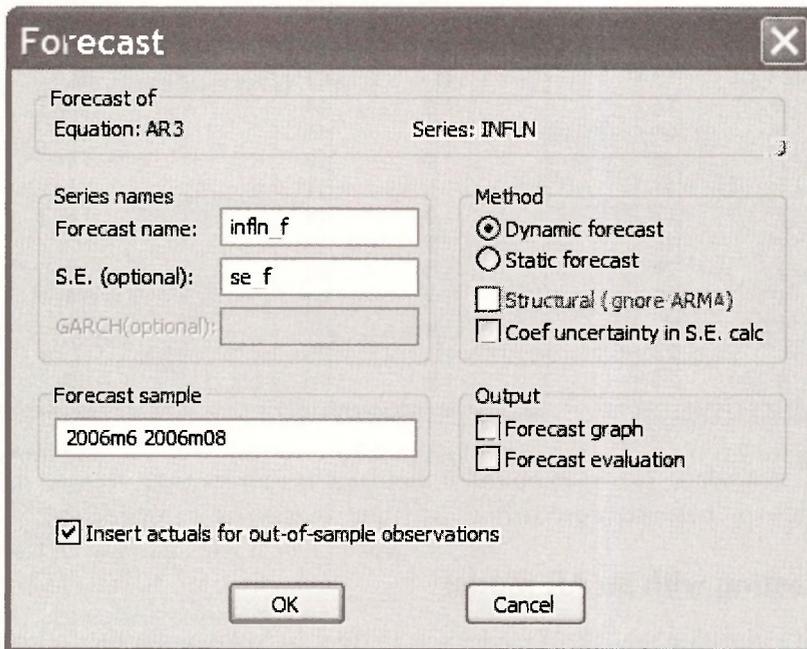
### 9.5.3 Forecasting with an AR model

The purpose of estimating the AR(3) model was to forecast inflation for the following 3 months, June, July and August of 2006. To make these forecasts we begin by extending the **range** of our workfile. Go back and have a quick re-read of Section 9.5.1. There you will see that we access the **Workfile structure** by selecting **Proc/Structure/Resize Current Page**. We change the **End date** of the **Date specification** to **2006M08** (August, 2006) and click **OK**. EViews will check whether you really want to make this change by asking **Resize involves inserting 3 observations. Continue?** Click **Yes**.



To compute the forecasts you fill in the Forecast dialog box that is obtained by opening the estimated equation, and then select Forecast. Several pieces of information are required.

1. We have assigned **INFLN\_F** as the series name for the forecasts (**Forecast name**) and **SE\_F** as the series name for the standard errors of the forecasts (**S.E. (optional)**).
2. The **Forecast sample** is June to August, 2006 that we specify as **2006m6 2006m8**.
3. Ticking the box **Insert actuals for out-of-sample observations** means actual values for *INFLN* will be inserted in the series **INFLN\_F** for the period 1983M12 to 2006M5.
4. **Dynamic forecast** is chosen for the **Method** because forecasts for future values will depend on earlier forecasts when actual values are not available.
5. Only error uncertainty, not coefficient uncertainty, is considered in the calculation of the forecast standard errors presented in Table 9.2 of the text and so the box **Coef uncertainty in S.E. calc** is not ticked.
6. We have not worried about **Output** for **Forecast graph** and **Forecast evaluation**.



Clicking **OK** creates the series **INFLN\_F** and **SE\_F**. The relevant values are given in the last 3 rows of their respective spreadsheets. To complete the information in Table 9.2 of the text we need the upper and lower values for the 95% forecast intervals. These values can be created using the commands below. Ask yourself where the 262 comes from.

```
scalar tc = @qtdist(0.975, 262)
series fint_low = infln_f - tc*se_f
```

series `fint_up = infln_f + tc*se_f`

After collecting the relevant information into a group

```
List of series, groups, and/or series expressions
infln_f se_f fint_low fint_up
```

the values in Table 9.2 can be presented as follows.

obs	INFLN F	SE F	FINT_LOW	FINT UP
2006M05	0.446762	NA	NA	NA
2006M06	0.260154	0.197247	-0.128237	0.648544
2006M07	0.248722	0.210542	-0.165847	0.663291
2006M08	0.269725	0.211111	-0.145965	0.685416

### 9.6 FINITE DISTRIBUTED LAGS

The finite distributed lag model in the text relates the inflation rate to current and past changes in the wage rate. The data are stored in the file `inflation.wf1`. The estimated model is

$$INFLN_t = \alpha + \beta_0 PCWAGE_t + \beta_1 PCWAGE_{t-1} + \beta_2 PCWAGE_{t-2} + \beta_3 PCWAGE_{t-3} + v_t$$

where `PCWAGE` denotes the percentage change in wages. The Equation specification for estimating this model is

Equation specification

Dependent variable, followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

`infln c pcwage(0 to -3)`

**includes lags 0, -1, -2 and -3**

**change end of sample back to May, 2006**

---

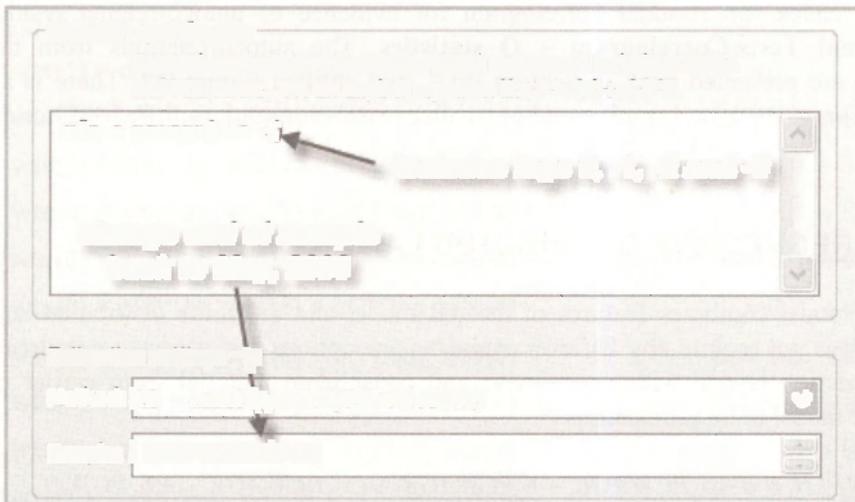
Estimation settings

Method: LS - Least Squares (NLS and ARMA)

Sample: 1983M12 2006M05

Note the shorthand notation for including all lags of `PCWAGE`. Also, remembering that we had extended the sample for forecasting, you made need to change the end of sample back to 2006M05 for estimating this equation.

The results are those given in Table 9.3 of the text.



Dependent Variable: INFLN				
Method: Least Squares				
Sample (adjusted): 1984M04 2006M05				
Included observations: 266 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.121873	0.048655	2.504862	0.0129
PCWAGE	0.156089	0.088502	1.763684	0.0790
PCWAGE(-1)	0.107498	0.085055	1.263861	0.2074
PCWAGE(-2)	0.049485	0.085258	0.580418	0.5621
PCWAGE(-3)	0.199014	0.087885	2.264475	0.0244

The delay multipliers in Table 9.4 are equal to the above coefficient estimates. The interim multipliers can be stored in a vector called **INTERIM** using the following commands.

```
vector(4) interim
interim(1) = c(2)
interim(2) = interim(1) + c(3)
interim(3) = interim(2) + c(4)
interim(4) = interim(3) + c(5)
```

INTERIM	
0.156089	
0.263587	
0.313072	
0.512086	

Finally, we check the residual correlogram for evidence of autocorrelated residuals. Select **View/Residual Tests/Correlogram – Q statistics**. The autocorrelations from the resulting correlogram are presented back in Section 9.5.2, and entitled Figure 9.7. There is a significant autocorrelation at lag 1, and significant but smaller correlations at lags 6, 7, 10, 11 and 15.

## 9.7 AUTOREGRESSIVE DISTRIBUTED LAG MODELS

The ARDL model combines features of the AR model and the finite distributed lag model. Its estimation does not require any EViews commands or options that we have not already covered. Thus, this section is one where we revise and consolidate material from earlier parts of the chapter. The model to be estimated is

$$\begin{aligned}
 INFLN_t = & \delta + \delta_0 PCWAGE_t + \delta_1 PCWAGE_{t-1} + \delta_2 PCWAGE_{t-2} + \delta_3 PCWAGE_{t-3} \\
 & + \theta_1 INFLN_{t-1} + \theta_2 INFLN_{t-2} + v_t
 \end{aligned}$$

The corresponding equation specification and results follow. See page 251 of the text.

Equation specification  
 Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .  
 infln c pcwage(0 to -3) infln(-1 to -2)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.098877	0.046807	2.112438	0.0356
PCWAGE	0.114903	0.083390	1.377901	0.1694
PCWAGE(-1)	0.037734	0.081245	0.464440	0.6427
PCWAGE(-2)	0.059275	0.081174	0.730221	0.4659
PCWAGE(-3)	0.236130	0.082944	2.846862	0.0048
INFLN(-1)	0.353640	0.060411	5.853884	0.0000
INFLN(-2)	-0.197561	0.060421	-3.269733	0.0012

The autocorrelations from the residual correlogram are presented back in Section 9.5.2, and entitled Figure 9.8. There are significant but very small autocorrelations at lags 6, 7, 11, 13 and 14.

### 9.7.1 Graphing the lag weights

The lag weights that are graphed in Figure 9.9 can be obtained recursively using the commands below. As you can see, calculating them individually as we have done is an unrewarding repetitive task. It can be avoided by programming a do loop, something that you will learn as you study more econometrics and start using the full version of EViews.

**series(13) lag\_wts**

**lag\_wts(1) = c(2)**

**lag\_wts(2) = c(6)\*lag\_wts(1) + c(3)**

**lag\_wts(3) = c(6)\*lag\_wts(2) + c(7)\*lag\_wts(1) + c(4)**

**lag\_wts(4) = c(6)\*lag\_wts(3) + c(7)\*lag\_wts(2) + c(5)**

**lag\_wts(5) = c(6)\*lag\_wts(4) + c(7)\*lag\_wts(3)**

**lag\_wts(6) = c(6)\*lag\_wts(5) + c(7)\*lag\_wts(4)**

**lag\_wts(7) = c(6)\*lag\_wts(6) + c(7)\*lag\_wts(5)**

**lag\_wts(8) = c(6)\*lag\_wts(7) + c(7)\*lag\_wts(6)**

**lag\_wts(9) = c(6)\*lag\_wts(8) + c(7)\*lag\_wts(7)**

**lag\_wts(10) = c(6)\*lag\_wts(9) + c(7)\*lag\_wts(8)**

**lag\_wts(11) = c(6)\*lag\_wts(10) + c(7)\*lag\_wts(9)**

**lag\_wts(12) = c(6)\*lag\_wts(11) + c(7)\*lag\_wts(10)**

**lag\_wts(13) = c(6)\*lag\_wts(12) + c(7)\*lag\_wts(11)**

$$\hat{\beta}_0 = \hat{\delta}_0$$

$$\hat{\beta}_1 = \hat{\theta}_1 \hat{\beta}_0 + \hat{\delta}_1$$

$$\hat{\beta}_2 = \hat{\theta}_1 \hat{\beta}_1 + \hat{\theta}_2 \hat{\beta}_0 + \hat{\delta}_2$$

$$\hat{\beta}_3 = \hat{\theta}_1 \hat{\beta}_2 + \hat{\theta}_2 \hat{\beta}_1 + \hat{\delta}_3$$

$$\hat{\beta}_4 = \hat{\theta}_1 \hat{\beta}_3 + \hat{\theta}_2 \hat{\beta}_2$$

$$\hat{\beta}_5 = \hat{\theta}_1 \hat{\beta}_4 + \hat{\theta}_2 \hat{\beta}_3$$

$$\hat{\beta}_6 = \hat{\theta}_1 \hat{\beta}_5 + \hat{\theta}_2 \hat{\beta}_4$$

$$\hat{\beta}_7 = \hat{\theta}_1 \hat{\beta}_6 + \hat{\theta}_2 \hat{\beta}_5$$

$$\hat{\beta}_8 = \hat{\theta}_1 \hat{\beta}_7 + \hat{\theta}_2 \hat{\beta}_6$$

$$\hat{\beta}_9 = \hat{\theta}_1 \hat{\beta}_8 + \hat{\theta}_2 \hat{\beta}_7$$

$$\hat{\beta}_{10} = \hat{\theta}_1 \hat{\beta}_9 + \hat{\theta}_2 \hat{\beta}_8$$

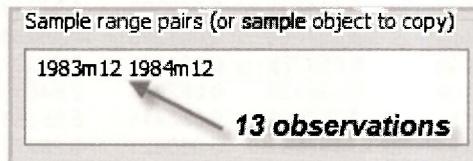
$$\hat{\beta}_{11} = \hat{\theta}_1 \hat{\beta}_{10} + \hat{\theta}_2 \hat{\beta}_9$$

$$\hat{\beta}_{12} = \hat{\theta}_1 \hat{\beta}_{11} + \hat{\theta}_2 \hat{\beta}_{10}$$

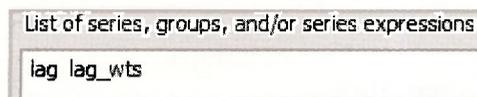
To create the graph in Figure 9.9 we also need a series called **LAG**. A command that produces this series is

**series lag = @trend**

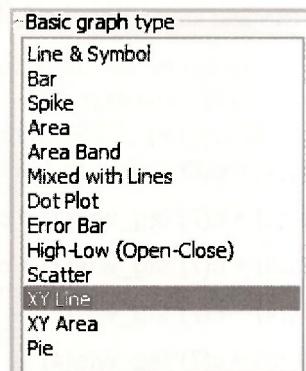
The EViews function **@trend** defines a trend series beginning at zero and with each observation incremented by 1, giving the values 0, 1, 2, ... Now, since there are only 13 values to graph, we need to restrict the sample to the first 13 observations. The monthly dates EViews has attached to its sample observations are not really relevant in this case, but we can trick EViews by cutting the sample back to the first 13 months. Select **Sample** from the workfile window and insert the following start and end dates.



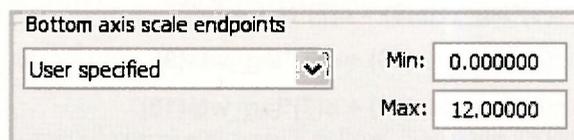
Then select **Object/New Object/Graph**. Name it **FIGURE9\_9**. Enter the two series for the graph, with the one on the *x*-axis first.



The graph **FIGURE9\_9** will appear in your workfile. However, it will need a bit of work before it is presentable. To ensure it is of the correct type, go to **Options/Type/XY Line**. Click **Apply**.



To make the *x*-axis compatible with Figure 9.9, go to **Axis/Scale/Edit Axis/Bottom Axis and Scale**. Then, for **Bottom axis scale endpoints**, choose **User specified** and specify **0** as the **Min** and **12** as the **Max**. Click **Apply**. Click **OK**.



To draw a horizontal line at 0, go to **Line/Shade**, select **Line** for **Type**. Choose **Orientation** and specify the **Data value** as follows.

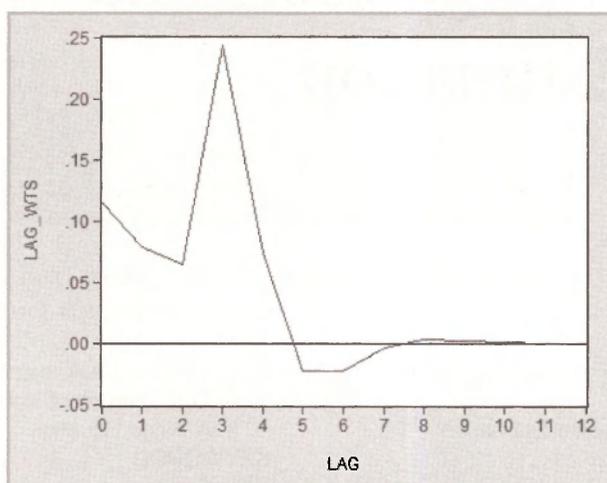
Orientation

Horizontal - Left axis

Position

Data value: 0

Voila! A respectable looking Figure 9.9 appears.



## Keywords

@cor  
 @sum  
 @trend  
 AC  
 AR(1)  
 AR(1) error  
 ARDL  
 autocorrelation  
 autoregressive models  
 axis/scale  
 basic graph  
 convergence  
 correlation  
 correlogram  
 date specification  
 delay multipliers  
 delta method  
 Durbin-Watson test  
 dynamic forecasting

edit axis  
 end date  
 finite distributed lags  
 forecast name  
 forecast standard errors  
 forecasting  
 freeze  
 frequency  
 graph  
 HAC standard errors  
 integer date  
 interim multipliers  
 lag specification  
 lag weights  
 lagging a series  
 Lagrange multiplier test  
 line & symbol  
 line/shade  
 LS & TSLS options

NA  
 Newey-West standard errors  
 nonlinear least squares  
 normalized restriction  
 orientation  
 range  
 resid  
 residual correlogram  
 residual graph  
 residual tests  
 residuals  
 sample (adjusted)  
 serial correlation LM test  
 start date  
 time series data  
 unstructured/undated  
 Wald test  
 workfile structure  
 XY line

# CHAPTER 10

## Random Regressors and Moment-Based Estimation

### CHAPTER OUTLINE

10.1 The Inconsistency of the Least Squares Estimator

10.2 Instrumental Variables Estimation

10.3 The Hausman Test

10.4 Test for Weak Instruments

10.5 Test Instrument Validity

10.6 A Wage Equation

KEYWORDS

Chapter 10 introduces the violation of assumption SR5 (and MR5) of the linear regression model, which states that the regressors ( $x_{it}$ ) are nonrandom. When this assumption is relaxed, the explanatory variables are sometimes said to be **stochastic**, which is another word meaning random. To begin our consideration of estimation in the simple linear regression framework in the presence of random regressors, open the workfile *ch10.wfl*. Save it under a new name, such as *chap10\_mc.wfl*.

### 10.1 THE INCONSISTENCY OF THE LEAST SQUARES ESTIMATOR

To reproduce Figure 10.2, showing the positive correlation between the  $x$  and  $e$  generated by the Monte Carlo experiment discussed in the text, first we must “create” the true errors  $e$ . In a Monte Carlo world we know the true parameters, and that  $y$  is created by

$$y = E(y) + e = \beta_1 + \beta_2 x + e = 1 + 1 \times x + e$$

Therefore we can create the series

$$e = y - \beta_1 - \beta_2 x = y - 1 - x$$

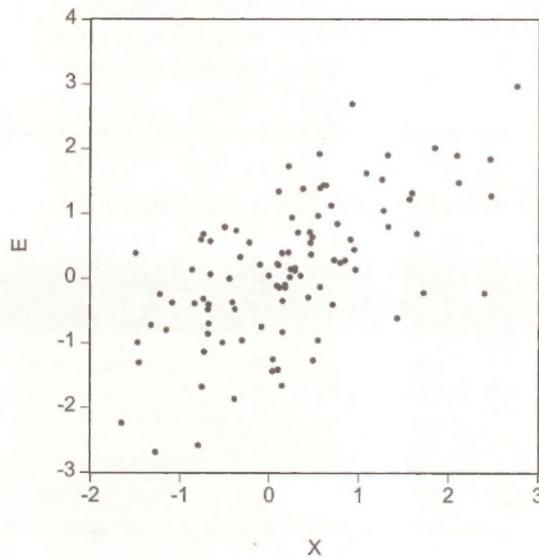
by entering the command

**series e = y - 1 - x**

From the main EViews menu select **Quick/Graph**. Enter into the **Series List** dialog box



In the **Graph Options** box choose a basic **Scatter** diagram. Copying (**Ctrl+C**) and pasting (**Ctrl+V**) the figure into our document



To generate Figure 10.3, we must first create

$$E(y) = \beta_1 + \beta_2 x = 1 + x$$

**series ey = 1 + x**

To create

$$\hat{y} = b_1 + b_2 x$$

we estimate the simple regression and then obtain the forecasted (predicted) values of  $y$ . To estimate the equation enter

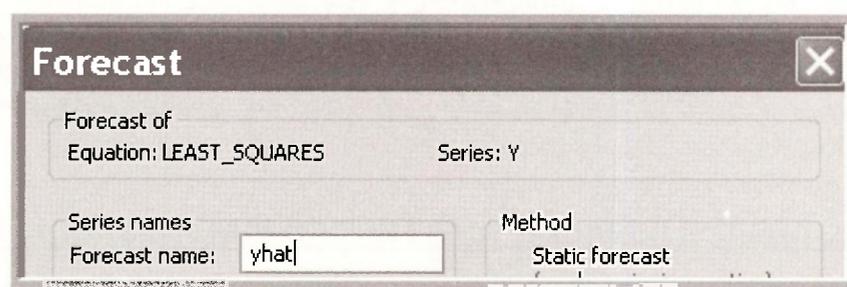
**ls y c x**

The result is

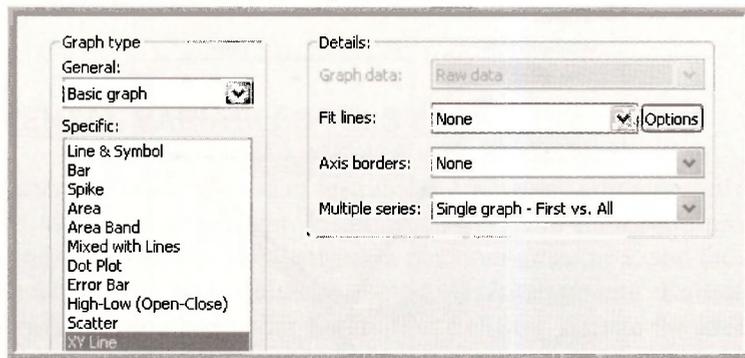
Dependent Variable: Y  
Method: Least Squares  
Sample: 1 100  
Included observations: 100

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.978893	0.088281	11.08838	0.0000
X	1.703431	0.089950	18.93754	0.0000
R-squared	0.785385	Mean dependent var		1.386287
Adjusted R-squared	0.783195	S.D. dependent var		1.838819
S.E. of regression	0.856198	Akaike info criterion		2.547166
Sum squared resid	71.84128	Schwarz criterion		2.599270
Log likelihood	-125.3583	Hannan-Quinn criter.		2.568253
F-statistic	358.6306	Durbin-Watson stat		2.103601
Prob(F-statistic)	0.000000			

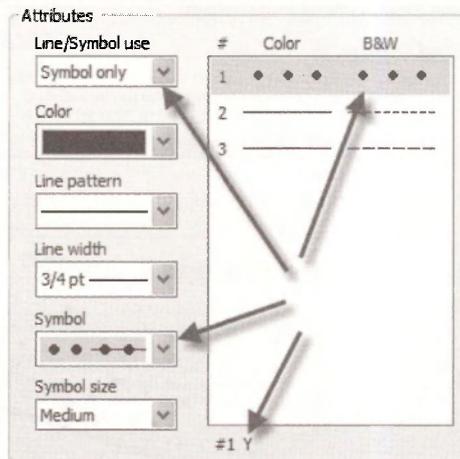
On the regression menu bar select **Forecast**. In the dialog box enter



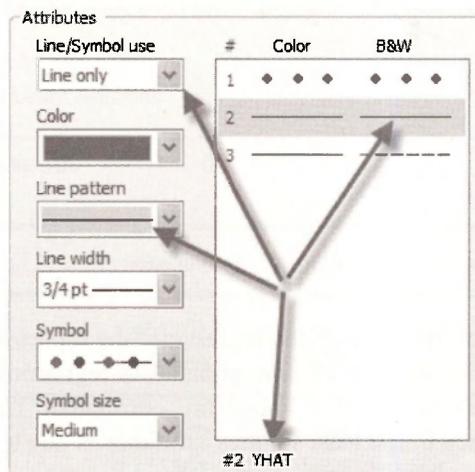
Select  $X$ ,  $Y$ , **YHAT**, and **EY** from the workfile, double-click on one of these variables, and select **Open Group**. After the group is open as a spreadsheet, select **View/Graph**. In the **Graph Options** dialog box enter



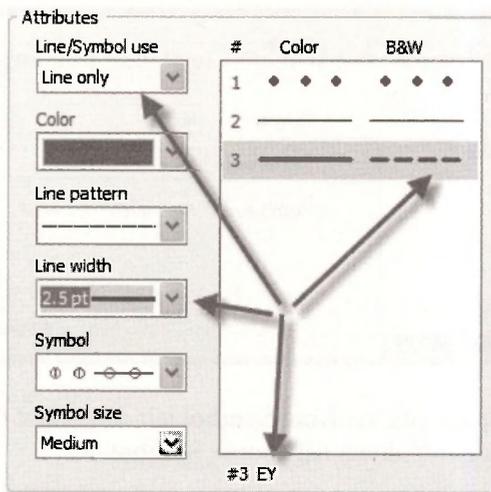
In the **Graph Options** dialog choose the **Line/Symbol** tab. Highlight the first series (#1) which is **Y**. From the **Line/Symbol use** drop down list choose **Symbol only**.



Select series #2, **YHAT**, and choose **Line & Symbol**.



Repeat this for series #3 which is **EY**.



The resulting figure, in black and white, is

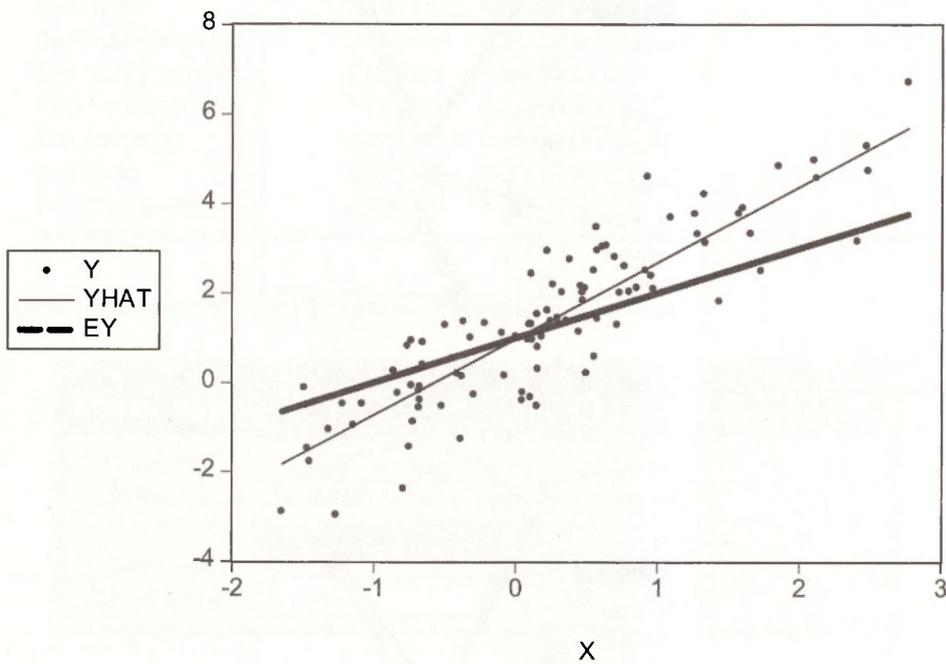
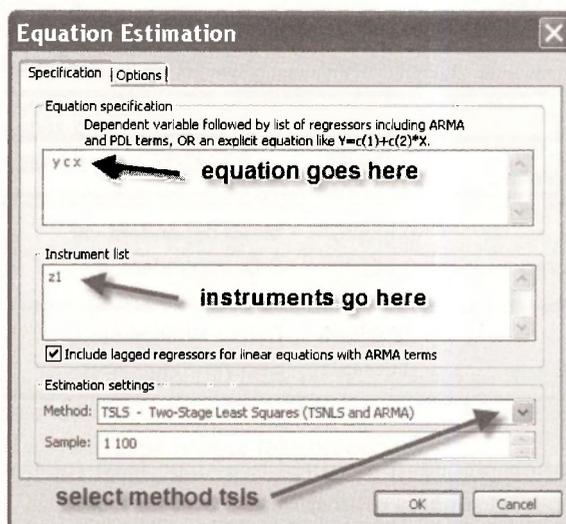


Figure 10.3 shows clearly that the slope of the regression line represented by the fitted dependent variable **YHAT** overstates the slope of the true population regression function. Hence, ordinary least squares is invalid in cases where  $x$  and  $e$  are correlated. The variable  $x$  is said to be **endogenous**. The inconsistency of the least squares estimator is due to an **endogeneity problem**.

## 10.2 INSTRUMENTAL VARIABLES/TSLS ESTIMATION

We now turn our attention to the method of instrumental variables estimation, also known as two-stage least squares, described in the text. Instrumental variables estimation produces consistent estimators in the presence of correlation between a random regressor,  $x$ , and the error term,  $e$ . To obtain the instrumental variables estimates, select **Quick/Estimate Equation**, and in the **Equation Specification** dialog box under **Estimation Settings**, change the **Method** to: **TSLS – Two-Stage Least Squares (TSNLS and ARMA)**. Next we enter the savings equation in list form  $Y C X$ , in the **Equation Specification** field. Finally, we list the instrument,  $Z1$ , in the **Instrument List** field.



Our results replicate those on page 280 of *POE*.

Dependent Variable: Y  
 Method: Two-Stage Least Squares  
 Sample: 1 100  
 Included observations: 100  
 Instrument list: Z1

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.101101	0.109128	10.08998	0.0000
X	1.192445	0.194518	6.130243	0.0000
R-squared	0.714712	Mean dependent var		1.386287
Adjusted R-squared	0.711801	S.D. dependent var		1.838819
S.E. of regression	0.987155	Sum squared resid		95.49855
F-statistic	37.57988	Durbin-Watson stat		1.997541
Prob(F-statistic)	0.000000	Second-Stage SSR		298.1235

Two-stage least squares, using instrumental variables Z1 and Z2, can be carried out by entering on the command line

**tsls y c x @ z1 z2**

The command is now **tsls** rather than just **ls**, and after the equation the instrumental variables are specified after the **@**-sign. The result is as shown in *POE* equation (10.27), page 284.

Dependent Variable: Y  
 Method: Two-Stage Least Squares  
 Sample: 1 100  
 Included observations: 100  
 Instrument list: Z1 Z2

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.137591	0.116444	9.769431	0.0000
X	1.039872	0.194223	5.354022	0.0000
R-squared	0.666207	Mean dependent var		1.386287
Adjusted R-squared	0.662801	S.D. dependent var		1.838819
S.E. of regression	1.067780	Sum squared resid		111.7351
F-statistic	28.66555	Durbin-Watson stat		1.967390
Prob(F-statistic)	0.000001	Second-Stage SSR		302.0610

As noted there the estimation procedure is called two-stage least squares because it can actually be implemented using two least squares estimations.

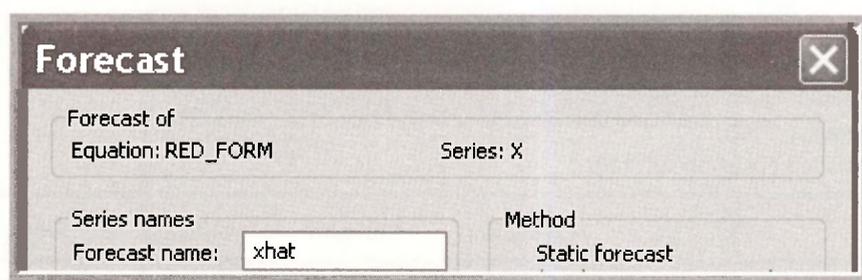
The first stage is a least squares regression of *X* on the instruments Z1 and Z2.

**ls x c z1 z2**

Dependent Variable: X  
 Method: Least Squares  
 Sample: 1 100  
 Included observations: 100

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.194732	0.079499	2.449486	0.0161
Z1	0.569978	0.088785	6.419747	0.0000
Z2	0.206786	0.077161	2.679940	0.0087

Save this estimated regression with the Name **RED\_FORM** which stands for **reduced form**.. Select **Forecast** on the regression menu bar.



Now estimate a least squares regression of  $Y$  on **XHAT** using

**ls y c xhat**

Dependent Variable: Y  
 Method: Least Squares  
 Sample: 1 100  
 Included observations: 100

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.137591	0.191456	5.941782	0.0000
XHAT	1.039872	0.319339	3.256324	0.0016

The estimated coefficients using this method are correct (compare them to *POE* (10.27)) but the standard errors **Std. Error** are incorrect. Thus two-stage least squares should not actually be implemented this way. Always use the proper **tsls** procedure.

### **10.3 THE HAUSMAN TEST**

Here we conduct the Hausman Test for correlation between an explanatory variable,  $x$ , and the error term. We continue with the simulated data example. We enter the following commands in the EViews command window:

```
equation hausman.ls x c z1 z2          estimate reduced form
series vhat = resid                   save residuals
equation endogtest.ls y c x vhat      artificial regression with vhat
```

The results are on the next page.

Dependent Variable: Y  
 Method: Least Squares  
 Sample: 1 100  
 Included observations: 100

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.137591	0.079746	14.26510	0.0000
X	1.039872	0.133013	7.817819	0.0000
VHAT	0.995728	0.162939	6.111053	0.0000

Note that the  $t$ -statistic for the coefficient on the residuals from the step one regression is 6.11 and the  $p$ -value of this test clearly shows that the  $t$ -statistic is statistically significant at the 1% level, so we reject the null hypothesis of no correlation between  $x$  and the error term  $e$  in favor of the alternative that  $x$  and  $e$  are correlated.

#### **10.4 TEST FOR WEAK INSTRUMENTS**

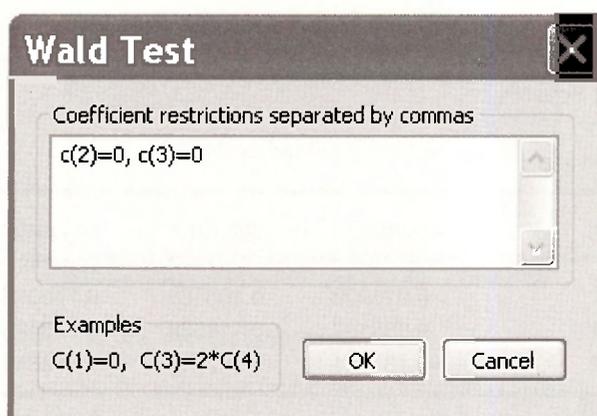
A requirement of good instrumental variables is that they be correlated with the right-hand side variable  $x$ , which is correlated with the error  $e$ . To test this we can examine the reduced form equation. Here consider the reduced form regression of  $x$  on the instruments Z1 and Z2.

Dependent Variable: X  
 Method: Least Squares  
 Sample: 1 100  
 Included observations: 100

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.194732	0.079499	2.449486	0.0161
Z1	0.569978	0.088785	6.419747	0.0000
Z2	0.206786	0.077161	2.679940	0.0087

The key is that the instruments are VERY significant, with  $t$ -values as a rule of thumb, greater than 3.3. In this case, we have two instruments but only require one to carry out two-stage least squares. Thus we can test the joint null hypothesis that the coefficients of the instruments are zero using an  $F$ -test. The alternative hypothesis is that at least one of the two reduced form parameters is not zero, which is exactly what we need.

In the regression window of the **RED\_FORM** select **View/Coefficient Tests/Wald – Coefficient Restrictions**. In the dialog box enter



The test result shows that we strongly reject the null hypothesis, and we can conclude that at least one of the reduced form parameters is not zero. Also the  $F$ -value is 24.28, which exceeds the rule of thumb guideline of  $F > 10$ .

Wald Test:			
Equation: RED_FORM			
Test Statistic	Value	df	Probability
F-statistic	24.27844	(2, 97)	0.0000

## 10.5 TEST INSTRUMENT VALIDITY

In addition to being strongly correlated to the variable  $x$ , the instruments must be uncorrelated with the error term  $e$ . Because we need one instrument to carry out two-stage least squares estimation, we can only check the validity of this condition for the surplus instruments. In the econometrics literature this is called a test of the **over-identifying restrictions**, and the test is often called the **Sargan test**. While there are several variants of this test, we will show a version that is based on the two-stage least squares residuals. We compute the TSLS residuals and regress them on all available instrumental variables. The test statistic is  $NR^2$  from this regression, where  $N$  is the sample size and  $R^2$  is the usual goodness-of-fit measure. If the surplus instruments are valid, the statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of surplus instruments. The validity of the surplus instruments is rejected if the test statistic value  $NR^2$  is greater than the critical value from the chi-square distribution.

The steps are

```

tsls y c x @ z1 z2
series ehat = resid
ls ehat c z1 z2

```

```

tsls estimation
tsls residuals
artificial regression

```

The results are on the following page.

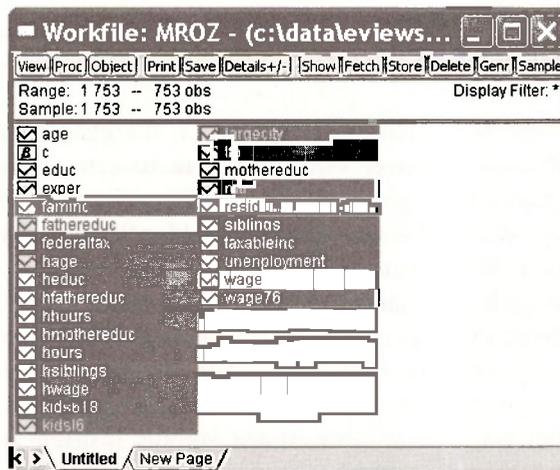
Dependent Variable: EHAT  
 Method: Least Squares  
 Sample: 1 100  
 Included observations: 100

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.018900	0.106168	0.178016	0.8591
Z1	0.088109	0.118568	0.743106	0.4592
Z2	-0.181754	0.103045	-1.763844	0.0809
R-squared	0.036276	Mean dependent var		4.00E-17

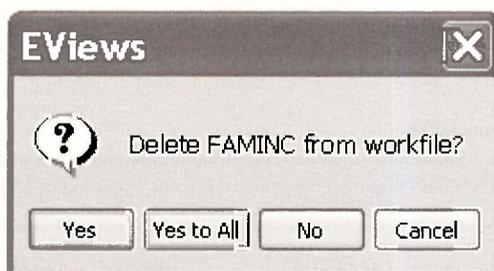
The  $R^2$  from this regression is .03628, and  $NR^2 = 3.628$ . The .05 critical value for the chi-square distribution with one degree of freedom is 3.84, thus we fail to reject the validity of the surplus instrumental variable.

## 10.6 A WAGE EQUATION

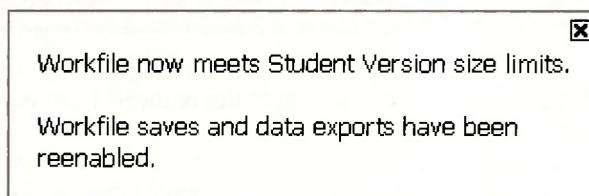
In Chapter 10.2 we introduced an important example, the estimation of the relationship between wages, specifically  $\log(WAGE)$ , and years of education ( $EDUC$ ). We will use the data on married women in the workfile *mroz.wfl* to examine this relationship. Open this workfile



If you are using the EViews 6 Student Version you will get a message saying the workfile is too large. Select all the variable shown above by clicking while holding down the **Ctrl**-key. Right-click in the blue area, and select **Delete**. A message like the following will appear, depending on which variable you selected first.



Click **Yes to All**. You will find the cheerful message

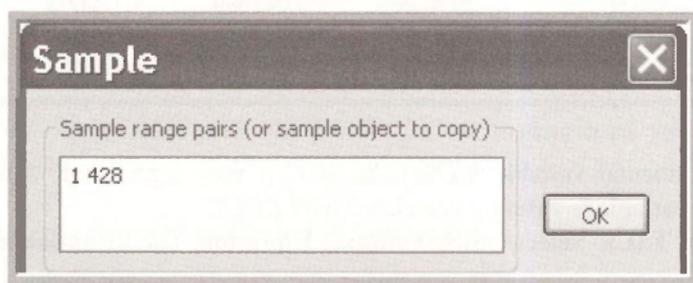


**Save** the workfile under a new name, such as *mroz\_chap10.wf1*.

A second problem is that in the data only the first 428 women have wage data. The remainder have  $WAGE = 0$  because they do not participate in the labor market. In the workfile window select the **Sample** button.



Fill in the dialog box to include in the estimation sample only the first 428 observations.



Now we can estimate the equation

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

The command is

**ls log(wage) c educ exper exper^2**

The result matches those on *POE* page 281. As noted in *POE* the concern is that the variable *EDUC* might be correlated with factors in the error term, such as ability. If that is the case, then the least squares estimator is biased and the bias will not disappear even if the sample size becomes very large.

Dependent Variable: LOG(WAGE)  
 Method: Least Squares  
 Sample: 1 428  
 Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.522041	0.198632	-2.628179	0.0089
EDUC	0.107490	0.014146	7.598332	0.0000
EXPER	0.041567	0.013175	3.154906	0.0017
EXPER^2	-0.000811	0.000393	-2.062834	0.0397

To implement two-stage least squares we first obtain the reduced form equation, adding mother's education *MOTHEREDUC* as an instrumental variable.

#### ls educ c exper exper^2 mothereduc

Dependent Variable: EDUC  
 Method: Least Squares  
 Sample: 1 428  
 Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	9.775103	0.423889	23.06055	0.0000
EXPER	0.048862	0.041669	1.172603	0.2416
EXPER^2	-0.001281	0.001245	-1.029046	0.3040
MOTHEREDUC	0.267691	0.031130	8.599183	0.0000

Note that the instrumental variable *MOTHEREDUC* is very significant, with a *t*-value of 8.6, indicating that this variable is strongly correlated with *EDUC*.

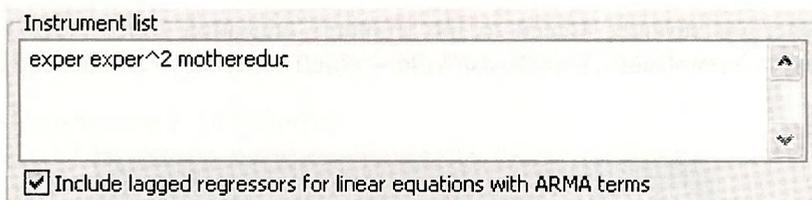
Now implement **TSLS**. Select **Quick/Estimate Equation**. The **Equation specification** is

Equation specification

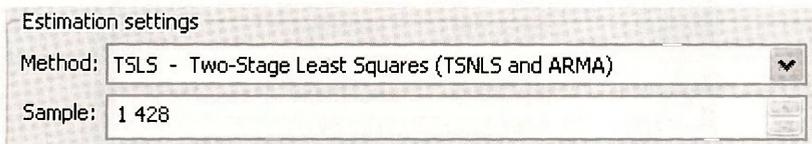
Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

log(wage) c educ exper exper^2

The **Instrument list** must include all the variables that are NOT correlated with the error term. These variables are said to be **exogenous**.



The estimation settings are



Click **OK**.

Dependent Variable: LOG(WAGE)  
 Method: Two-Stage Least Squares  
 Sample: 1 428  
 Included observations: 428  
 Instrument list: EXPER EXPER^2 MOTHEREDUC

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.198186	0.472877	0.419107	0.6754
EDUC	0.049263	0.037436	1.315924	0.1889
EXPER	0.044856	0.013577	3.303856	0.0010
EXPER^2	-0.000922	0.000406	-2.268993	0.0238

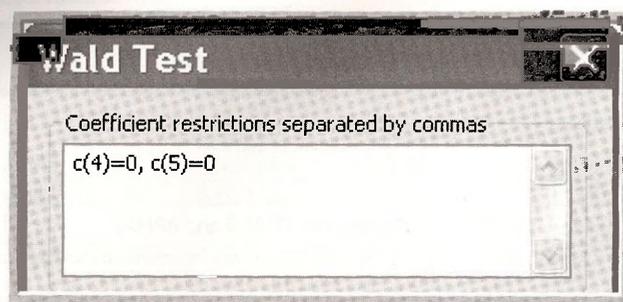
If we use both *MOTHEREDUC* and *FATHEREDUC* as instrumental variables the estimated reduced form is obtained using

**ls educ c exper exper^2 mothereduc fathereduc**

Dependent Variable: EDUC  
 Method: Least Squares  
 Sample: 1 428  
 Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	9.102640	0.426561	21.33958	0.0000
EXPER	0.045225	0.040251	1.123593	0.2618
EXPER^2	-0.001009	0.001203	-0.838572	0.4022
MOTHEREDUC	0.157597	0.035894	4.390609	0.0000
FATHEREDUC	0.189548	0.033756	5.615173	0.0000

Both instruments are strongly related to the woman's education *EDUC*. To test their joint significance select **View/Coefficient Tests/Wald – Coefficient Restrictions**. In the dialog box enter



The result shows an  $F$  value of 55.4, giving strong evidence that at least one of the instruments has a non-zero coefficient in the reduced form equation.

Wald Test:  
Equation: REDFORM\_MOM\_DAD

Test Statistic	Value	df	Probability
F-statistic	55.40030	(2, 423)	0.0000

To test the endogeneity of *EDUC* we obtain the reduced form residuals and then include them in the wage equation as an extra explanatory variable.

```
series vhat = resid
ls log(wage) c educ exper exper^2 vhat
```

The estimation results show that the variable **VHAT** has a  $p$ -value of 0.0954, which is not strong evidence that *EDUC* is endogenous.

Dependent Variable: LOG(WAGE)  
Method: Least Squares  
Sample: 1 428  
Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.048100	0.394575	0.121904	0.9030
EDUC	0.061397	0.030985	1.981499	0.0482
EXPER	0.044170	0.013239	3.336272	0.0009
EXPER^2	-0.000899	0.000396	-2.270623	0.0237
VHAT	0.058167	0.034807	1.671105	0.0954

Two-stage least squares estimates can be obtained with the command

**tsls log(wage) c educ exper exper^2 @ exper exper^2 mothereduc fathereduc**

Dependent Variable: LOG(WAGE)

Method: Two-Stage Least Squares

Sample: 1 428

Included observations: 428

Instrument list: EXPER EXPER^2 MOTHEREDUC FATHEREDUC

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.048100	0.400328	0.120152	0.9044
EDUC	0.061397	0.031437	1.953024	0.0515
EXPER	0.044170	0.013432	3.288329	0.0011
EXPER^2	-0.000899	0.000402	-2.237993	0.0257

To test the validity of the surplus instrumental variable we save the TOLS residuals, and regress them on all the instrumental variables.

**series ehat = resid****ls ehat c exper exper^2 mothereduc fathereduc**

Dependent Variable: EHAT

Method: Least Squares

Sample: 1 428

Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.010964	0.141257	0.077618	0.9382
EXPER	-1.83E-05	0.013329	-0.001376	0.9989
EXPER^2	7.34E-07	0.000398	0.001842	0.9985
MOTHEREDUC	-0.006607	0.011886	-0.555804	0.5786
FATHEREDUC	0.005782	0.011179	0.517263	0.6052
R-squared	0.000883	Mean dependent var		5.54E-16

For the artificial regression  $R^2 = .000883$ , and the test statistic value is

$$NR^2 = 428 \times .000883 = .3779$$

The .05 critical value for the chi-square distribution with one degree of freedom is 3.84, thus we fail to reject the surplus instrument as valid. With this result we are reassured that our instrumental variables estimator for the wage equation is consistent.

**Keywords**

chi-square test  
endogeneity problem  
endogenous regressor  
forecast  
graph options  
Hausman test  
identified  
inconsistent estimator  
instrument list

instrumental variables  
instruments  
least squares  
line & symbol  
Monte Carlo  
Mroz data  
over-identified  
random regressors  
reduced form

Sargan statistic  
scatter plot  
stochastic  
surplus instruments  
TSLS  
two-stage least squares  
validity of surplus instruments  
wage equation  
Wald test

# CHAPTER 11

## Simultaneous Equations Models

### CHAPTER OUTLINE

11.1 Examining the Data

11.2 Estimating the Reduced Form

11.3 TSLS Estimation of an Equation

11.4 TSLS Estimation of a System of Equations

11.5 Supply and Demand at Fulton Fish Market

KEYWORDS

Until now, we have considered estimation and hypothesis testing in a variety of single equation models. Here we introduce models for the joint estimation of two or more equations. While there are countless applications for simultaneous equations models in economics, some applications with which you will be familiar include market demand and supply models and the multi-equation Keynesian models that we analyze in macroeconomics.

### 11.1 EXAMINING THE DATA

In this section, the text introduces a two-equation demand and supply model for truffles, a French gourmet mushroom delicacy. To estimate the truffles model in EViews, open the workfile *truffles.wfl*. Open a **Group** containing the data. While holding down the **Ctrl**-key, select *P*, *Q*, *PS*, *DI* and *PF*. The first few observations look like

obs	P	Q	PS	DI	PF
1	29.64000	19.89000	19.97000	2.103000	10.52000
2	40.23000	13.04000	18.04000	2.043000	19.67000
3	34.71000	19.61000	22.36000	1.870000	13.74000
4	41.43000	17.13000	20.87000	1.525000	17.95000
5	53.37000	22.55000	19.79000	2.709000	13.71000

The summary statistics for the variables are obtained from the spreadsheet by selecting **View/Descriptive Stats/Common Sample**

Sample: 1 30

	P	Q	PS	DI	PF
Mean	62.72400	18.45833	22.02200	3.526967	22.75333
Median	63.07500	19.27000	22.68500	3.708000	24.14500
Maximum	105.4500	26.27000	28.98000	5.125000	34.01000
Minimum	29.64000	6.370000	15.21000	1.525000	10.52000
Std. Dev.	18.72346	4.613088	4.077237	1.040803	5.329654

## 11.2 ESTIMATING THE REDUCED FORM

We first estimate the reduced form equations of *POE* Section 11.6.2 by regressing each endogenous variable,  $Q$ , and  $P$ , on the exogenous variables,  $PS$ ,  $DI$ , and  $PF$ . We can quickly accomplish this task with the following statements typed in the EViews command window. The results match those in *POE* Table 11.2, page 313.

### equation redform\_q.ls q c ps di pf

Dependent Variable: Q  
Method: Least Squares  
Sample: 1 30

	Coefficient	Std. Error	t-Statistic	Prob.
C	7.895099	3.243422	2.434188	0.0221
PS	0.656402	0.142538	4.605115	0.0001
DI	2.167156	0.700474	3.093842	0.0047
PF	-0.506982	0.121262	-4.180896	0.0003
R-squared	0.697386	Mean dependent var		18.45833

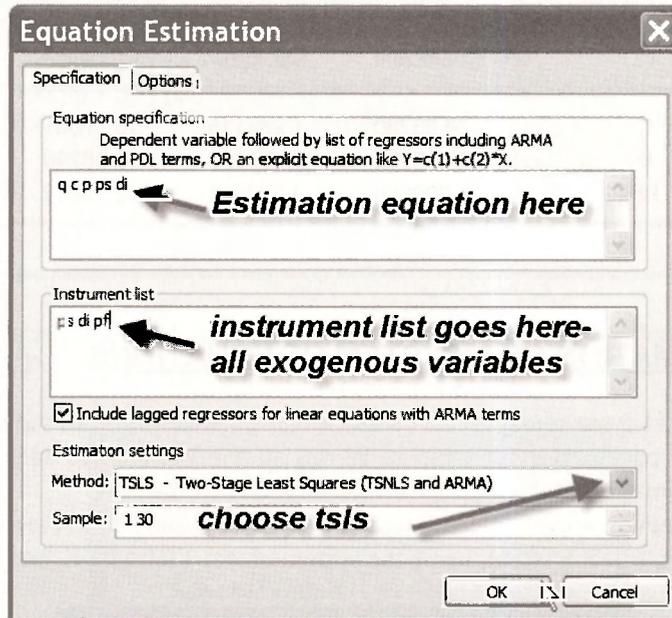
### equation redform\_p.ls p c ps di pf

Dependent Variable: P  
Method: Least Squares  
Sample: 1 30

	Coefficient	Std. Error	t-Statistic	Prob.
C	-32.51242	7.984235	-4.072077	0.0004
PS	1.708147	0.350881	4.868172	0.0000
DI	7.602491	1.724336	4.408939	0.0002
PF	1.353906	0.298506	4.535603	0.0001
R-squared	0.888683	Mean dependent var		62.72400

### 11.3 TSLS ESTIMATION OF AN EQUATION

Any identified equation within a system of simultaneous equations can be estimated by two-stage least squares (2SLS/TSLS). Click on **Quick/Estimate Equation**.



To estimate the demand equation by 2SLS select the method to be **TSLS**, and fill in the demand equation variables in **Equation specification**, the upper area of the dialog box, and list all the exogenous variables in the system in the **Instrument list**. Click **OK**. Name the resulting equation **DEMAND**.

Dependent Variable: Q  
 Method: Two-Stage Least Squares  
 Sample: 1 30  
 Instrument list: PS DI PF

	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.279471	5.543884	-0.771926	0.4471
P	-0.374459	0.164752	-2.272869	0.0315
PS	1.296033	0.355193	3.648812	0.0012
DI	5.013977	2.283556	2.195688	0.0372

To estimate the supply equation we illustrate the use of the command line.

**equation supply.tsls q c p pf @ ps di pf**

In this command we name the estimation **SUPPLY** and the estimation technique *TSLS* by **equation supply.tsls**. The specification of the equation is followed by the instrumental variables, which follow **@**.

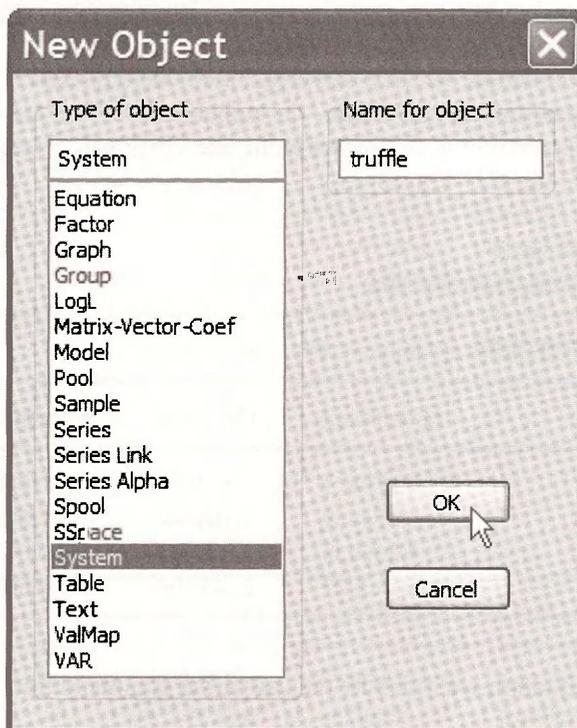
Dependent Variable: Q  
 Method: Two-Stage Least Squares  
 Sample: 1 30  
 Instrument list: PS DI PF

	Coefficient	Std. Error	t-Statistic	Prob.
C	20.03280	1.223115	16.37851	0.0000
P	0.337982	0.024920	13.56290	0.0000
PF	-1.000909	0.082528	-12.12813	0.0000

### 11.4 TSLS ESTIMATION OF A SYSTEM OF EQUATIONS

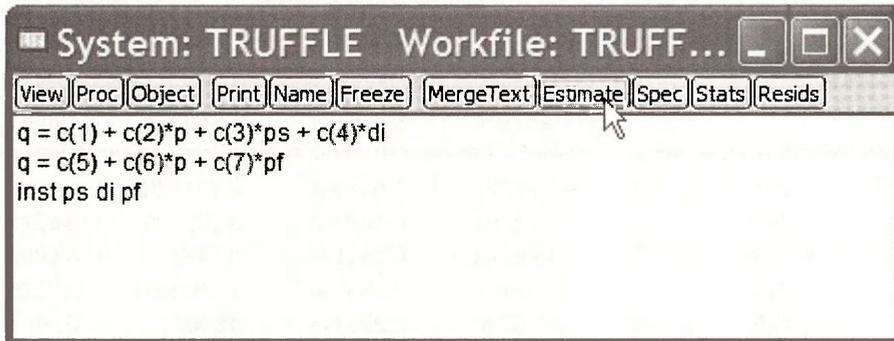
As noted in Section 11.2 we can apply *TSLS* equation by equation for all the identified equations within a system of equations. If all the equations in the system are identified, then all the equations can be estimated in one step.

We introduce a new EViews object here: the **SYSTEM**. From the EViews menubar, click on **Objects/New Object**, select **System**, name the system object **TRUFFLE**, and click **OK**.

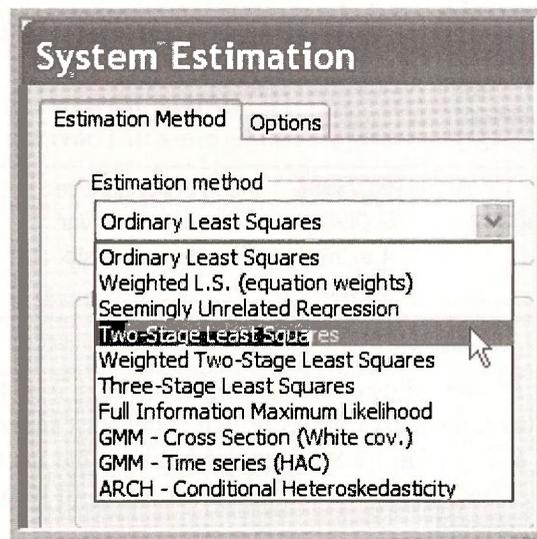


Next, enter the system equation specification given in *POE* equations (11.1) and (11.2), on page 304. Note that you must enter a line that contains the exogenous (determined outside the model) variables in the system, *PS*, *DI*, and *PF*. In the context of two-stage least squares estimation of

our truffles system, EViews refers to these exogenous variables as “instruments”. Enter the line **INST PS DI PF** directly below the supply equation, and click **Estimate** on the system’s toolbar.



To reproduce the results found in Tables 11.3a and 11.3b in your text, under **Estimation Method**, check the **Two-Stage Least Squares** checkbox, and click **OK**.



The results, on the following page, are identical to the equation by equation approach of estimating demand and then supply, but this system estimation approach opens the window to many advanced procedures that you may learn about in subsequent econometrics courses.

System: TRUFFLE  
 Estimation Method: Two-Stage Least Squares  
 Sample: 1 30  
 Total system (balanced) observations 60

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-4.279471	5.543884	-0.771926	0.4436
C(2)	-0.374459	0.164752	-2.272869	0.0271
C(3)	1.296033	0.355193	3.648812	0.0006
C(4)	5.013977	2.283556	2.195688	0.0325
C(5)	20.03280	1.223115	16.37851	0.0000
C(6)	0.337982	0.024920	13.56290	0.0000
C(7)	-1.000909	0.082528	-12.12813	0.0000

At the bottom of the **System Table** we find each equation represented and summarized.

Equation:  $Q = C(1) + C(2)*P + C(3)*PS + C(4)*DI$

Instruments: PS DI PF C

Observations: 30

R-squared	-0.023950	Mean dependent var	18.45833
Adjusted R-squared	-0.142098	S.D. dependent var	4.613088
S.E. of regression	4.929960	Sum squared resid	631.9171
Prob(F-statistic)	1.962370		

Note in these results that  $R^2$  and adjusted- $R^2$  are negative. This is not uncommon when using generalized least squares, instrumental variables or two-stage least squares. For any estimator but least squares, the identity  $SST = SSR + SSE$  does not hold, so the usual  $R^2 = 1 - SSE/SST$  can produce negative numbers. This just shows that the goodness-of-fit measure is not appropriate in this context, and should be ignored.

## 11.5 SUPPLY AND DEMAND AT FULTON FISH MARKET

A second example of a simultaneous equations model is given by the Fulton Fish Market discussed in *POE* Section 11.7, page 314. Open the workfile *fultonfish.wfl*. Let us specify the demand equation for this market as

$$\ln(QUAN_t) = \alpha_1 + \alpha_2 \ln(PRICE_t) + \alpha_3 MON_t + \alpha_4 TUE_t + \alpha_5 WED_t + \alpha_6 THU_t + e_t^d$$

Where  $QUAN_t$  is the quantity sold, in pounds, and  $PRICE_t$  the average daily price per pound. Note that we are using the subscript “ $t$ ” to index observations for this relationship because of the time series nature of the data. The remaining variables are dummy variables for the days of the week, with Friday being omitted. The coefficient  $\alpha_2$  is the price elasticity of demand, which we

expect to be negative. The daily dummy variables capture day to day shifts in demand. The supply equation is

$$\ln(QUAN_t) = \beta_1 + \beta_2 \ln(PRICE_t) + \beta_3 STORMY_t + e_t^s$$

The coefficient  $\beta_2$  is the price elasticity of supply. The variable *STORMY* is a dummy variable indicating stormy weather during the previous three days. This variable is important in the supply equation because stormy weather makes fishing more difficult, reducing the supply of fish brought to market.

The reduced form equations specify each endogenous variable as a function of all exogenous variables

$$\ln(QUAN_t) = \pi_{11} + \pi_{21} MON_t + \pi_{31} TUE_t + \pi_{41} WED_t + \pi_{51} THU_t + \pi_{61} STORMY_t + v_{1t}$$

$$\ln(PRICE_t) = \pi_{12} + \pi_{22} MON_t + \pi_{32} TUE_t + \pi_{42} WED_t + \pi_{52} THU_t + \pi_{62} STORMY_t + v_{2t}$$

The least squares estimates of the reduced forms are given by

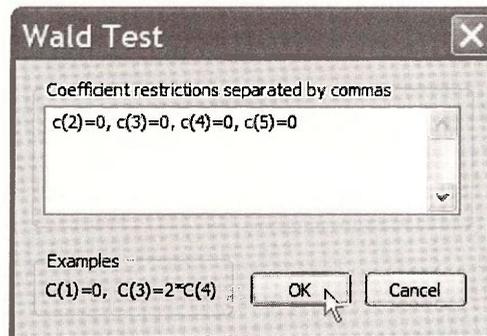
**ls lquan c mon tue wed thu stormy**  
**ls lprice c mon tue wed thu stormy**

The key reduced form equation is the second, for  $\ln(PRICE)$ .

Dependent Variable: LPRICE

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.271705	0.076389	-3.556867	0.0006
MON	-0.112922	0.107292	-1.052480	0.2950
TUE	-0.041149	0.104509	-0.393740	0.6946
WED	-0.011825	0.106930	-0.110587	0.9122
THU	0.049646	0.104458	0.475268	0.6356
STORMY	0.346406	0.074678	4.638681	0.0000

See *POE* pages 316-317 for a discussion of the importance of the strong significance of the variable *STORMY* and the lack of the significance of the day dummies, individually or jointly.



Wald Test:  
Equation: REDFORM\_PRICE

Test Statistic	Value	df	Probability
F-statistic	0.618762	(4, 105)	0.6501

The TSLS estimates of the demand equation in *POE* Table 11.5, page 317, are obtained using

**tsls lquan c lprice mon tue wed thu @ mon tue wed thu stormy**

Dependent Variable: LQUAN  
Method: Two-Stage Least Squares  
Sample: 1 111  
Instrument list: MON TUE WED THU STORMY

	Coefficient	Std. Error	t-Statistic	Prob.
C	8.505911	0.166167	51.18896	0.0000
LPRICE	-1.119417	0.428645	-2.611524	0.0103
MON	-0.025402	0.214774	-0.118274	0.9061
TUE	-0.530769	0.208000	-2.551775	0.0122
WED	-0.566351	0.212755	-2.661989	0.0090
THU	0.109267	0.208787	0.523345	0.6018

**Keywords**

demand equation  
endogenous variables  
exogenous variables

instrument list  
instrumental variables  
reduced form equation

supply equation  
system of equations  
Wald test

# CHAPTER 12

## Nonstationary Time-Series Data and Cointegration

### CHAPTER OUTLINE

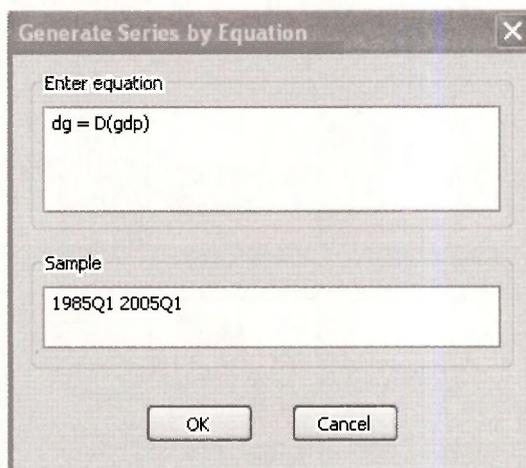
- 12.1 Stationary and Nonstationary Variables
- 12.2 Spurious Regressions

- 12.3 Unit Root Tests
  - 12.4 Cointegration
- KEYWORDS

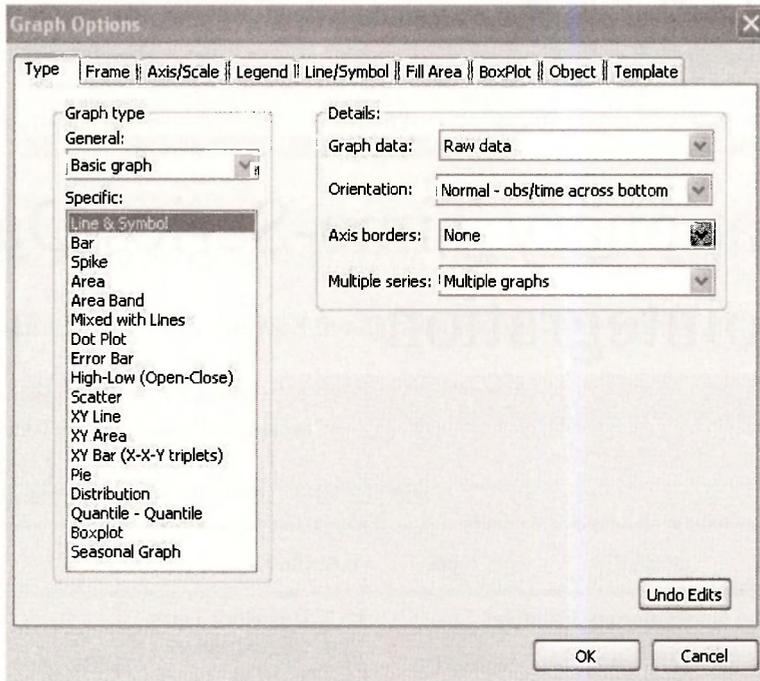
### 12.1 STATIONARY AND NONSTATIONARY VARIABLES

Time-series data display a variety of behavior. The data shown in Figure 12.1 is stored in the EViews workfile *usa.wf1*. They are real gross domestic product (*GDP*), inflation rate (*INF*), Federal funds rate (*F*) and the 3-year Bond rate (*B*). The changes for real gross domestic product (*DG*), inflation (*DI*), Federal funds rate (*DF*) and the 3-year Bond rate (*DB*) are computed using **Genr** and EViews first-difference operator **d**. For example,

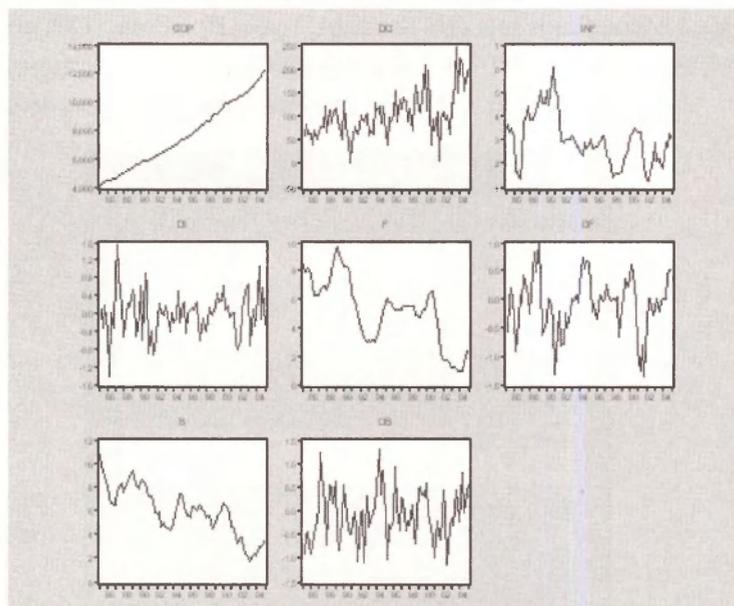
$$d(\text{dgp}) = \Delta GDP_t = GDP_t - GDP_{t-1}$$



To plot the graphs, select the 8 variables, open the **Group**, then select **View/ Graph/ Line & Symbol/ Multiple graphs** as shown below.



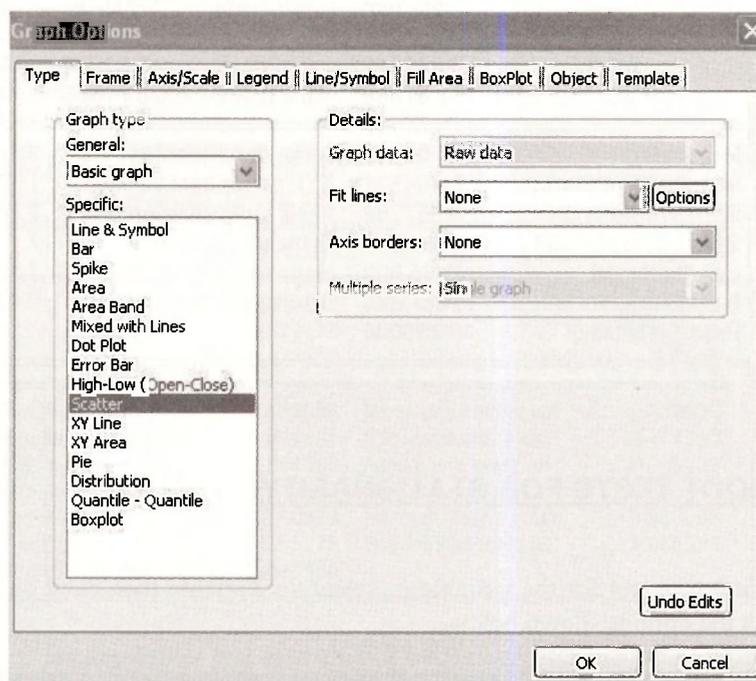
Clicking on **OK** produces the EViews output below. This set of graphs illustrate the variety of behavior observed with time series data, such as ‘trending’ (see *GDP*), ‘wandering around a trend’ (see *F*, and *B*), ‘wandering around a constant’ (see *INF*), ‘fluctuating around a trend’ (see *DG*) and fluctuating around a constant (see *DI*, *DF*, and *DB*). In general, nonstationary variables display wandering behavior (around constant and/or trend) while stationary data display fluctuating behavior (around constant and/or trend).



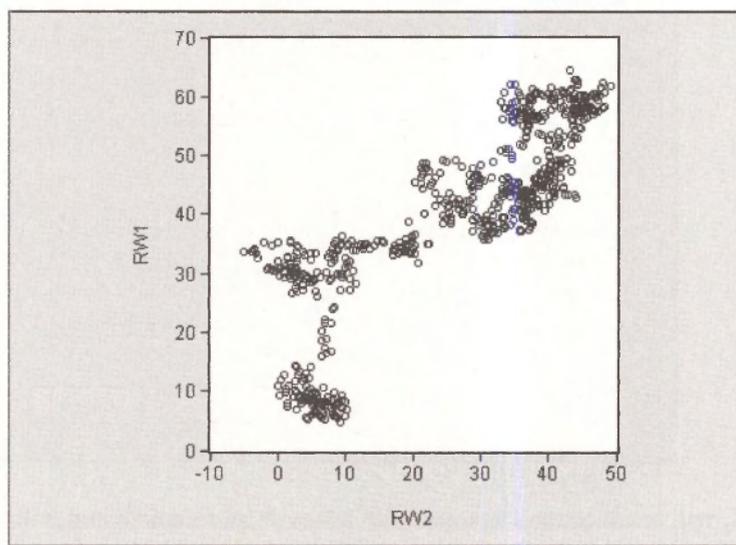
## 12.2 SPURIOUS REGRESSIONS

The main reason why it is important to know whether a time series is stationary or nonstationary before one embarks on a regression analysis is that there is a danger of obtaining apparently significant regression results from unrelated data when nonstationary series are used in regression analysis. Such regressions are said to be **spurious**.

The EViews workfile *spurious.wf1* contains the 2 random variables (*RW1* and *RW2*) shown in Figure 12.3(a). To plot the scatter graph, select the 2 variables, open **Group, View / Graph/** and select **Scatter** as shown below.



Clicking **OK** will produce the EViews output below.



Although the series (*RW1* and *RW2*) were generated independently and, in truth, have no relation to one another, the scatter plot suggests a positive relationship between them. The **spurious regression** of series one (*RW1*) on series two (*RW2*) is shown in the EViews output below.

Dependent Variable: RW1  
Method: Least Squares

Sample: 1 700  
Included observations: 700

	Coefficient	Std. Error	t-Statistic	Prob.
RW2	0.842041	0.020620	40.83684	0.0000
C	17.81804	0.620478	28.71665	0.0000
R-squared	0.704943	Mean dependent var		39.44163
Adjusted R-squared	0.704521	S.D. dependent var		15.74242
S.E. of regression	8.557268	Akaike info criterion		7.134292
Sum squared resid	51112.33	Schwarz criterion		7.147295
Log likelihood	-2495.002	Hannan-Quinn criter.		7.139319
F-statistic	1667.648	Durbin-Watson stat		0.022136
Prob(F-statistic)	0.000000			

## 12.3 UNIT ROOT TESTS FOR STATIONARITY

To obtain the unit root test for the variable  $F$ , select the variable then click on **View /Unit Root Test/** and select the options shown below.

Clicking on **OK**, will produce the Dickey-Fuller test with an intercept and with one lag term.

Null Hypothesis: F has a unit root  
 Exogenous: Constant  
 Lag Length: 1 (Fixed)

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-2.090304	0.2491
Test critical values:		
1% level	-3.515536	
5% level	-2.898623	
10% level	-2.586605	

\*Mackinnon (1996) one-sided p-values.

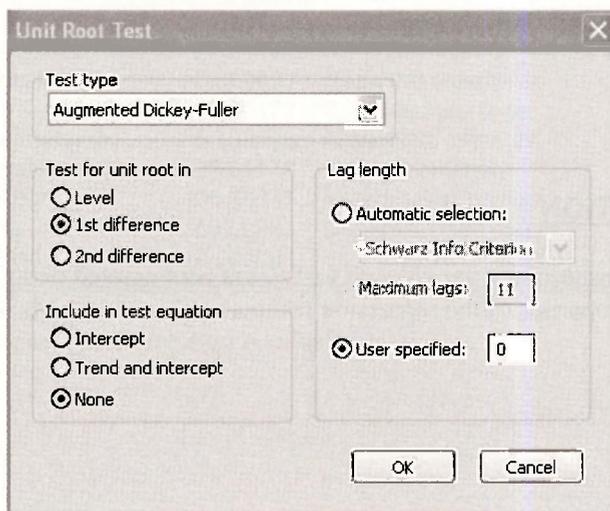
Augmented Dickey-Fuller Test Equation  
 Dependent Variable: D(F)  
 Method: Least Squares

Sample (adjusted): 1985Q3 2005Q1  
 Included observations: 79 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
F(-1)	-0.037067	0.017733	-2.090304	0.0399
D(F(-1))	0.672478	0.085366	7.877548	0.0000
C	0.177862	0.100751	1.765356	0.0815
R-squared	0.456006	Mean dependent var		-0.069029
Adjusted R-squared	0.441691	S.D. dependent var		0.474252
S.E. of regression	0.354362	Akaike info criterion		0.800238
Sum squared resid	9.543490	Schwarz criterion		0.890217
Log likelihood	-28.60939	Hannan-Quinn criter.		0.836286
F-statistic	31.85376	Durbin-Watson stat		1.855064
Prob(F-statistic)	0.000000			

Since the calculated Dickey-Fuller test statistic (-2.090) is greater than the 5% critical value of (-2.899), do not reject the null of nonstationarity. In other words, the variable  $F$  is a nonstationary series.

To perform the test for the first-difference of  $F$ , select the options shown below:



Clicking on **OK** gives the output below.

Null Hypothesis: D(F) has a unit root  
Exogenous: None  
Lag Length: 0 (Fixed)

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-4.007355	0.0001
Test critical values: 1% level	-2.594563	
5% level	-1.944969	
10% level	-1.614082	

\*Mackinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation  
Dependent Variable: D(F,2)  
Method: Least Squares

Sample (adjusted): 1985Q3 2005Q1  
Included observations: 79 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
D(F(-1))	-0.340465	0.084960	-4.007355	0.0001
R-squared	0.169739	Mean dependent var		0.013567
Adjusted R-squared	0.169739	S.D. dependent var		0.395101
S.E. of regression	0.360010	Akaike info criterion		0.807210
Sum squared resid	10.10938	Schwarz criterion		0.837203
Log likelihood	-30.88478	Hannan-Quinn criter.		0.819226
Durbin-Watson stat	1.798788			

Since the calculated Dickey-Fuller test statistic ( $-4.007$ ) is less than the 5% critical value of ( $-1.945$ ) we reject the null of nonstationarity. In other words, the variable  $d(f) = \Delta F$  is a stationary series.

It follows that since  $F$  has to be differenced once to obtain stationarity, it is integrated of order 1.

## 12.4 COINTEGRATION

To test whether the nonstationary variables,  $B$  and  $F$ , are cointegrated or spuriously related, we need to examine the properties of the regression residuals. The first step is to estimate the least squares regression:

Dependent Variable: B  
Method: Least Squares

Sample: 1985Q1 2005Q1  
Included observations: 81

	Coefficient	Std. Error	t-Statistic	Prob.
F	0.832505	0.034476	24.14743	0.0000
C	1.643733	0.194819	8.437233	0.0000
R-squared	0.880682	Mean dependent var		5.947408
Adjusted R-squared	0.879172	S.D. dependent var		2.037110
S.E. of regression	0.708106	Akaike info criterion		2.171935
Sum squared resid	39.61170	Schwarz criterion		2.231057
Log likelihood	-85.96337	Hannan-Quinn criter.		2.195656
F-statistic	583.0983	Durbin-Watson stat		0.413856
Prob(F-statistic)	0.000000			

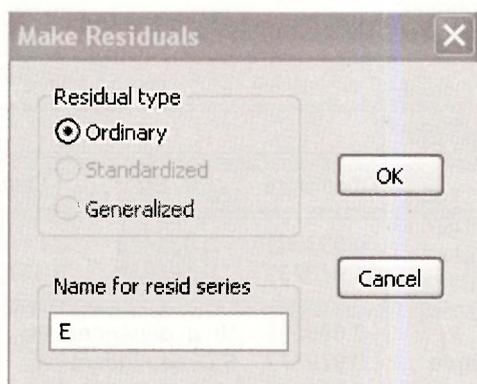
Next, to generate the residuals from the regression equation, click on **Proc** and select **Make Residual Series** from the drop-down menu.

The screenshot shows the EViews software interface. The 'Proc' menu is open, and 'Make Residual Series...' is highlighted. The background displays the regression results table from the previous block.

	Coefficient	Std. Error	t-Statistic	Prob.
F	0.832505	0.034476	24.14743	0.0000
C	1.643733	0.194819	8.437233	0.0000
R-squared	0.880682	Mean dependent var		5.947408
Adjusted R-squared	0.879172	S.D. dependent var		2.037110
S.E. of regression	0.708106	Akaike info criterion		2.171935
Sum squared resid	39.61170	Schwarz criterion		2.231057
Log likelihood	-85.96337	Hannan-Quinn criter.		2.195656
F-statistic	583.0983	Durbin-Watson stat		0.413856
Prob(F-statistic)	0.000000			

To conform to the text, call the regression residuals *E*.

Click on OK to create the residuals.



Next perform a Dickey-Fuller test by regressing the change of  $E$ , (namely,  $d(e)$ ) on lagged  $E$  (namely  $e(-1)$ ) and the lagged term  $d(e(-1))$ .

Dependent Variable: D(E)  
Method: Least Squares

Sample (adjusted): 1985Q3 2005Q1  
Included observations: 79 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
E(-1)	-0.314320	0.069191	-4.542819	0.0000
D(E(-1))	0.314748	0.102156	3.081034	0.0029
R-squared	0.240040	Mean dependent var		-0.020381
Adjusted R-squared	0.230171	S.D. dependent var		0.455110
S.E. of regression	0.399313	Akaike info criterion		1.026849
Sum squared resid	12.27773	Schwarz criterion		1.086835
Log likelihood	-38.56055	Hannan-Quinn criter.		1.050881
Durbin-Watson stat	2.002243			

Since the calculated Dickey-Fuller test statistic ( $-4.543$ ) is less than the 5% critical value of ( $-3.37$ ) we reject the null of no cointegration. Recall that the critical values are those from Table 12.3 and it is for the case where the regression model includes an intercept term.

### Keywords

cointegration  
d: difference operator  
Dickey-Fuller tests  
make residual series

multiple graphs  
nonstationary variables  
order of Integration  
spurious regression

stationary variables  
stationarity tests  
unit root test of residuals  
unit root tests of variables

# CHAPTER 13

## VEC and VAR Models: An Introduction to Macroeconometrics

### CHAPTER OUTLINE

13.1 VEC and VAR Models  
13.2 Estimating a VEC Model

13.3 Estimating a VAR Model  
13.4 Impulse Responses and Variance Decompositions  
KEYWORDS

### 13.1 VEC AND VAR MODELS

A VAR model describes a system of equations in which each variable is a function of its own lag and the lag of the other variables in the system. A VEC model is a special form of the VAR for I(1) variables which are cointegrated.

### 13.2 ESTIMATING A VEC MODEL

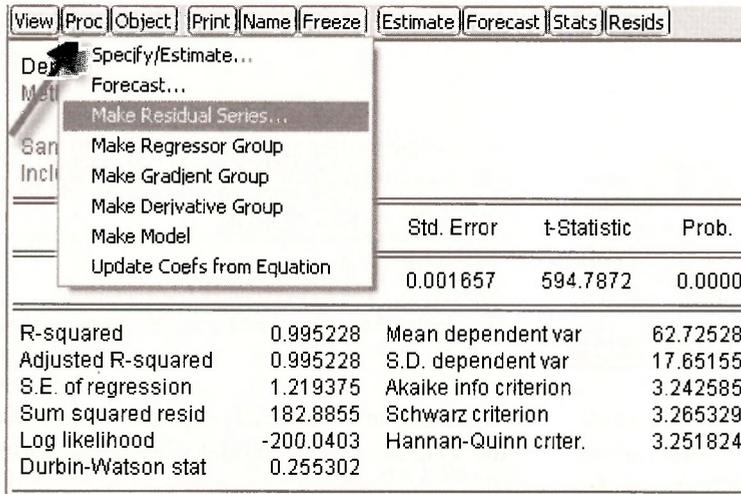
The results in the text were based on data contained in the EViews workfile *gdp.wf1*. The variables are *AUS* (real GDP for Australia) and *USA* (real GDP for US). To check whether the variables *AUS* and *USA* are cointegrated or spuriously related, we need to test the regression residuals for stationarity. To do this, first estimate the following least squares equation.

Dependent Variable: AUS  
 Method: Least Squares

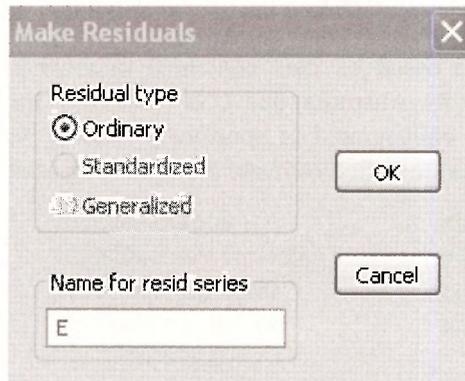
Sample: 1970Q1 2000Q4  
 Included observations: 124

	Coefficient	Std. Error	t-Statistic	Prob.
USA	0.985350	0.001657	594.7872	0.0000
R-squared	0.995228	Mean dependent var		62.72528
Adjusted R-squared	0.995228	S.D. dependent var		17.65155
S.E. of regression	1.219375	Akaike info criterion		3.242585
Sum squared resid	182.8855	Schwarz criterion		3.265329
Log likelihood	-200.0403	Hannan-Quinn criter.		3.251824
Durbin-Watson stat	0.255302			

Next click **Proc** and select **Make Residual Series** from the drop-down menu to generate the residuals.



Following notation in the text, call the residual  $E$ .



Next perform the unit root test by regressing the change of the residual  $d(e)$  on the lagged residual  $e(-1)$  as shown below.

Dependent Variable: D(E)  
Method: Least Squares

Sample (adjusted): 1970Q2 2000Q4  
Included observations: 123 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
E(-1)	-0.127937	0.044279	-2.889318	0.0046
R-squared	0.064044	Mean dependent var		-0.000498
Adjusted R-squared	0.064044	S.D. dependent var		0.618638
S.E. of regression	0.598500	Akaike info criterion		1.819315
Sum squared resid	43.70063	Schwarz criterion		1.842179
Log likelihood	-110.8879	Hannan-Quinn criter		1.828602
Durbin-Watson stat	1.978150			

Since the calculated unit root test value (-2.889) is less than the critical value (-2.76, see Table 12.3), the null of no cointegration is rejected.

As an aside, extra lags of the dependent variable (for example  $d(e(-1))$ ) were not introduced in the test equation above, as they were insignificant. See for example the case below.

Dependent Variable: D(E)  
Method: Least Squares

Sample (adjusted): 1970Q3 2000Q4  
Included observations: 122 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
E(-1)	-0.128207	0.046197	-2.775232	0.0064
D(E(-1))	-0.008532	0.093408	-0.091341	0.9274
R-squared	0.065661	Mean dependent var		-0.004108
Adjusted R-squared	0.057875	S.D. dependent var		0.619836
S.E. of regression	0.601632	Akaike info criterion		1.837917
Sum squared resid	43.43537	Schwarz criterion		1.883884
Log likelihood	-110.1129	Hannan-Quinn criter.		1.856588
Durbin-Watson stat	1.960758			

The estimated error-correction equations are shown below. The error correction coefficients are the parameters of the lagged residual term, namely  $e(-1)$  above.

Dependent Variable: D(AUS)  
Method: Least Squares

Sample (adjusted): 1970Q2 2000Q4  
Included observations: 123 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.491706	0.057909	8.490936	0.0000
E(-1)	-0.098703	0.047516	-2.077267	0.0399
R-squared	0.034434	Mean dependent var		0.499554
Adjusted R-squared	0.026454	S.D. dependent var		0.649528
S.E. of regression	0.640879	Akaike info criterion		1.964174
Sum squared resid	49.69782	Schwarz criterion		2.009901
Log likelihood	-118.7967	Hannan-Quinn criter.		1.982748
F-statistic	4.315037	Durbin-Watson stat		1.640143
Prob(F-statistic)	0.039893			

Dependent Variable: D(USA)  
Method: Least Squares

Sample (adjusted): 1970Q2 2000Q4  
Included observations: 123 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.509884	0.046677	10.92372	0.0000
E(-1)	0.030250	0.038299	0.789837	0.4312
R-squared	0.005129	Mean dependent var		0.507479
Adjusted R-squared	-0.003093	S.D. dependent var		0.515771
S.E. of regression	0.516568	Akaike info criterion		1.532907
Sum squared resid	32.28793	Schwarz criterion		1.578633
Log likelihood	-92.27376	Hannan-Quinn criter.		1.551481
F-statistic	0.623843	Durbin-Watson stat		1.367645
Prob(F-statistic)	0.431168			

### 13.3 ESTIMATING A VAR MODEL

The results in the text were based on data contained in the EViews workfile *growth.wf1*. The variables are  $G$  (log of GDP) and  $P$  (log of the CPI). To check whether the variables are cointegrated or spuriously related, we need to test the regression residuals for stationarity. To do this, first estimate the following least squares equation.

Dependent Variable: G  
Method: Least Squares

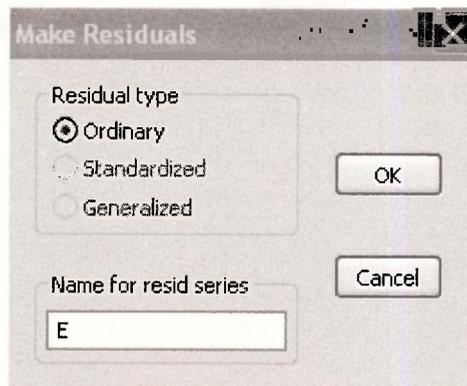
Sample: 1960Q1 2004Q4  
Included observations: 180

	Coefficient	Std. Error	t-Statistic	Prob.
P	0.623454	0.010140	61.48235	0.0000
C	1.631501	0.039323	41.48936	0.0000
R-squared	0.955029	Mean dependent var		4.015101
Adjusted R-squared	0.954776	S.D. dependent var		0.415151
S.E. of regression	0.088286	Akaike info criterion		-2.005428
Sum squared resid	1.387398	Schwarz criterion		-1.969950
Log likelihood	182.4885	Hannan-Quinn criter.		-1.991043
F-statistic	3780.079	Durbin-Watson stat		0.015086
Prob(F-statistic)	0.000000			

Then, click on **Proc** and select **Make Residual Series** from the drop down menu.

The screenshot shows the EViews software interface. The 'Proc' menu is open, and 'Make Residual Series...' is highlighted. The background window displays the same regression results table as shown in the previous block.

Following notation in the text, call the residuals  $E$ :



Next perform the unit root test by regressing the change of the residual  $d(e)$  on the lagged residual  $e(-1)$  as shown below.

Dependent Variable: D(E)  
Method: Least Squares

Sample (adjusted): 1960Q2 2004Q4  
Included observations: 179 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
E(-1)	-0.009037	0.009250	-0.976931	0.3299
R-squared	-0.018991	Mean dependent var		0.001674
Adjusted R-squared	-0.018991	S.D. dependent var		0.010713
S.E. of regression	0.010815	Akaike info criterion		-6.210247
Sum squared resid	0.020819	Schwarz criterion		-6.192441
Log likelihood	556.8171	Hannan-Quinn criter.		-6.203027
Durbin-Watson stat	0.929793			

Since the calculated tau statistic ( $-0.977$ ) is greater than the 5% critical value ( $-3.37$ , see Table 12.3), we accept the null of no cointegration. In other words, the variables are spuriously related.

The estimated VAR equations are estimated by least squares as shown below.

Dependent Variable: D(P)  
Method: Least Squares

Sample (adjusted): 1960Q3 2004Q4  
Included observations: 178 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001433	0.000710	2.016796	0.0452
D(P(-1))	0.826816	0.044707	18.49419	0.0000
D(G(-1))	0.046442	0.039858	1.165183	0.2455
R-squared	0.667250	Mean dependent var		0.010474
Adjusted R-squared	0.663447	S.D. dependent var		0.007684
S.E. of regression	0.004457	Akaike info criterion		-7.971758
Sum squared resid	0.003477	Schwarz criterion		-7.918133
Log likelihood	712.4865	Hannan-Quinn criter.		-7.950012
F-statistic	175.4599	Durbin-Watson stat		2.194622
Prob(F-statistic)	0.000000			

Dependent Variable: D(G)  
 Method: Least Squares

Sample (adjusted): 1960Q3 2004Q4  
 Included observations: 178 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.009814	0.001251	7.845799	0.0000
D(P(-1))	-0.326952	0.078719	-4.153419	0.0001
D(G(-1))	0.228505	0.070181	3.255925	0.0014
R-squared	0.168769	Mean dependent var		0.008260
Adjusted R-squared	0.159269	S.D. dependent var		0.008560
S.E. of regression	0.007849	Akaike info criterion		-6.840246
Sum squared resid	0.010780	Schwarz criterion		-6.786620
Log likelihood	611.7819	Hannan-Quinn criter.		-6.818499
F-statistic	17.76557	Durbin-Watson stat		2.110267
Prob(F-statistic)	0.000000			

### 13.4 IMPULSE RESPONSES AND VARIANCE DECOMPOSITION

In the text, we discuss the interpretation of impulse responses and variance decomposition for the special case where the shocks are uncorrelated. This is not usually the case, and EViews is set up for the more general cases when identification is an issue.

#### Keywords

error correction  
 identification

impulse responses  
 VAR

variance decomposition  
 VEC

# CHAPTER 14

## Time-Varying Volatility and ARCH Models: An Introduction to Financial Econometrics

### CHAPTER OUTLINE

- 14.1 Time-Varying Volatility
- 14.2 Testing for ARCH Effects
- 14.3 Estimating an ARCH Model

- 14.4 Generalized ARCH
- 14.5 Asymmetric ARCH
- 14.6 GARCH in Mean Model

KEYWORDS

### 14.1 TIME-VARYING VOLATILITY

In this chapter we are concerned with **variances that change over time**, i.e., time-varying variance processes. The model we focus on is called the **AutoRegressive Conditional Heteroskedastic (ARCH)** model.

$$y_t = \beta_0 + e_t$$

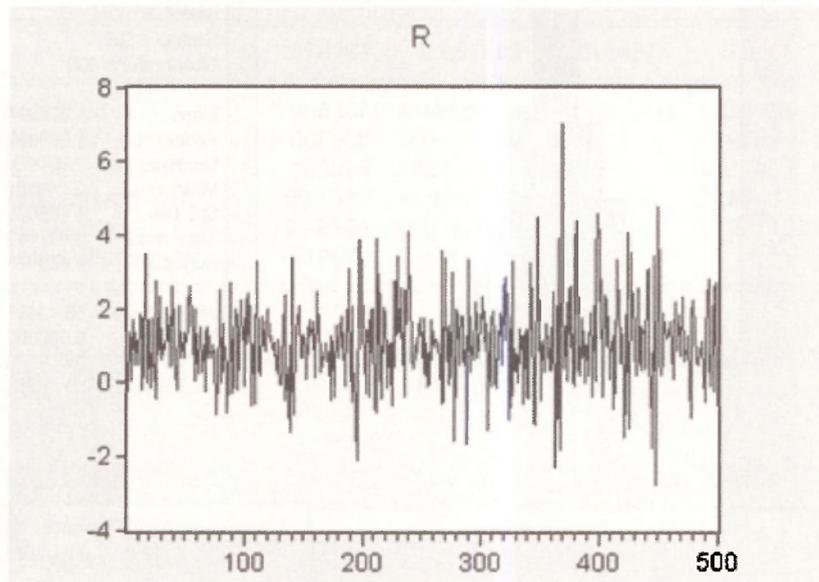
$$e_t | I_{t-1} \sim N(0, h_t)$$

$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2, \quad \alpha_0 > 0, \quad 0 \leq \alpha_1 < 1$$

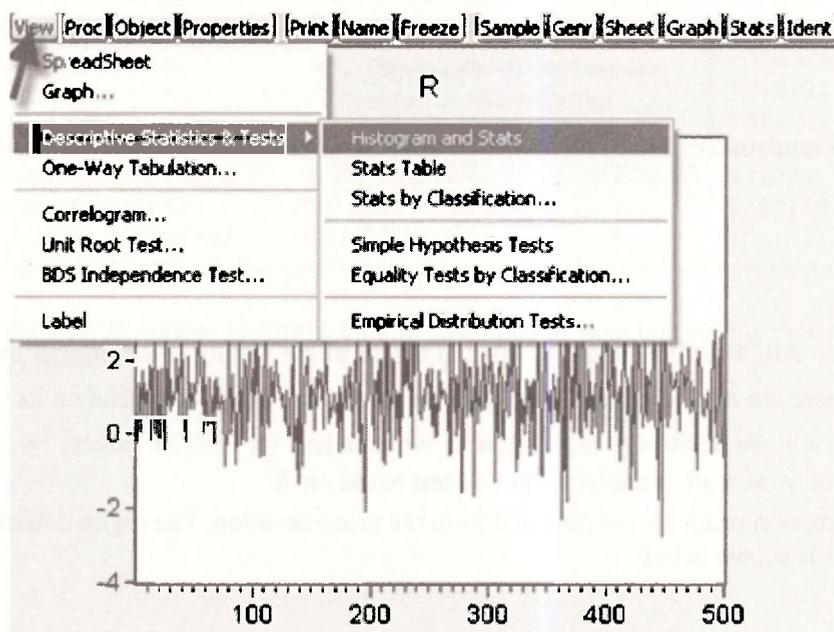
This is an example of an ARCH(1) model since the time varying variance  $h_t$  is a function of a constant term ( $\alpha_0$ ) plus a term lagged once, the square of the error in the previous period ( $\alpha_1 e_{t-1}^2$ ). The coefficients,  $\alpha_0$  and  $\alpha_1$ , have to be positive to ensure a positive variance. The coefficient  $\alpha_1$  must be less than 1, otherwise  $h_t$  will continue to increase over time, eventually exploding.

Conditional normality means that the distribution is a function of known information at time  $t - 1$  i.e., when  $t = 2$ ,  $(e_2 | I_1) \sim N(0, \alpha_0 + \alpha_1 e_1^2)$  and so on.

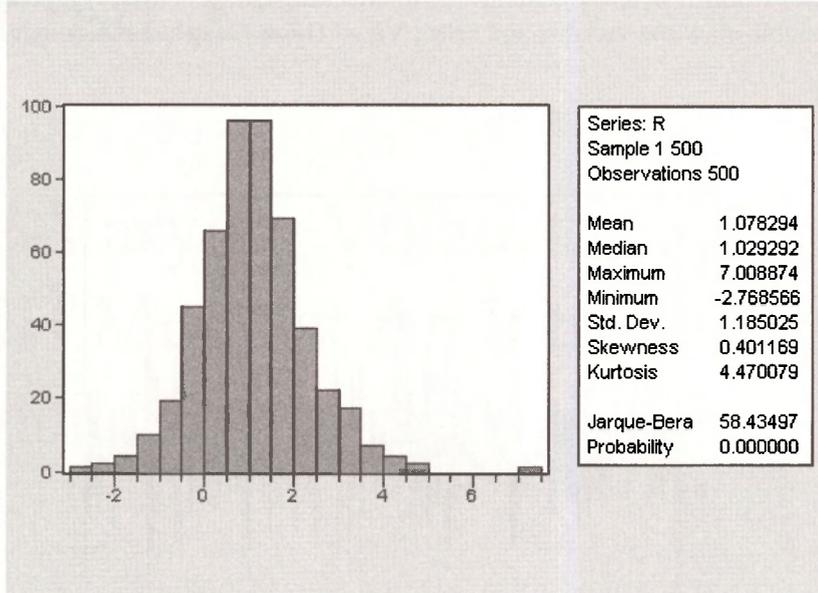
The EViews workfile *byd.wfl* contains the returns to BrightenYourDayLighting. To plot the times series, double-click the variable and select **View/ Open Graph/ Line & symbol** and click **OK**.



To generate the histogram, select **View/ Descriptive Statistics & Tests/ Histogram and Stats**.



Clicking this option gives the distribution below.



## 14.2 TESTING FOR ARCH EFFECTS

To test for first order ARCH, regress the squared regression residuals  $\hat{e}_t^2$  on their lags  $\hat{e}_{t-1}^2$ :

$$\hat{e}_t^2 = \gamma_0 + \gamma_1 \hat{e}_{t-1}^2 + v_t$$

where  $v_t$  is a random term. The null and alternative hypotheses are:

$$H_0 : \gamma_1 = 0$$

$$H_1 : \gamma_1 \neq 0$$

If there are no ARCH effects, then  $\gamma_1 = 0$  and the fit of the testing equation will be poor with a low  $R^2$ . If there are ARCH effects, we expect the magnitude of  $\hat{e}_t^2$  to depend on its lagged values and the  $R^2$  will be relatively high. Hence, we can test for ARCH effects by checking the significance of  $\gamma_1$  as well as applying the *LM* test based on  $R^2$ .

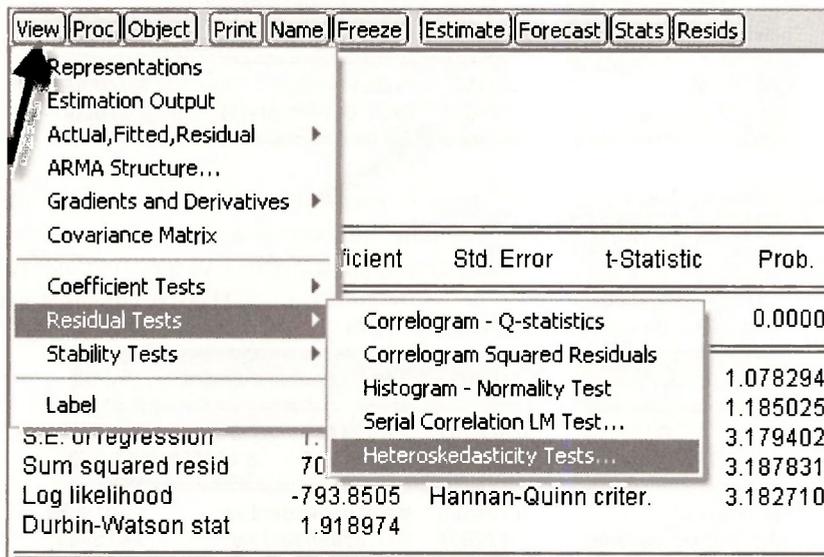
The regression residuals are obtained from the mean equation. The regression of returns on a constant term is shown below.

Dependent Variable: R  
Method: Least Squares

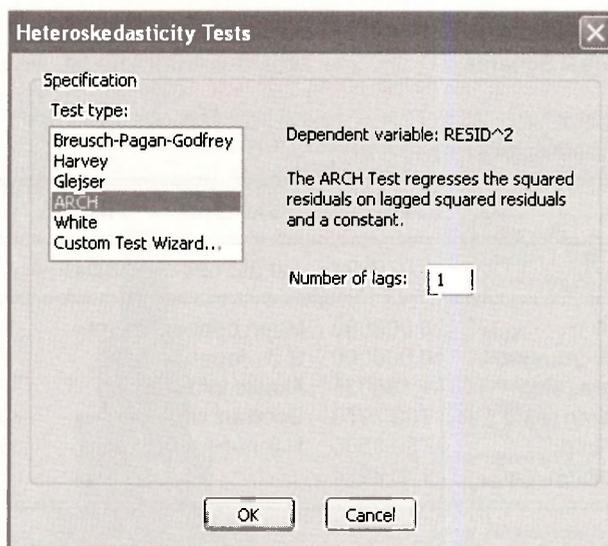
Sample: 1 500  
Included observations: 500

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.078294	0.052996	20.34674	0.0000
R-squared	0.000000	Mean dependent var		1.078294
Adjusted R-squared	0.000000	S.D. dependent var		1.185025
S.E. of regression	1.185025	Akaike info criterion		3.179402
Sum squared resid	700.7373	Schwarz criterion		3.187831
Log likelihood	-793.8505	Hannan-Quinn criter.		3.182710
Durbin-Watson stat	1.918974			

To generate the regression residuals, select **View**, select **Residual Tests/Heteroskedasticity Tests** from the drop down menus.



Then select the ARCH option. Inserting **1** in the **Number of lags** box means that we are testing for ARCH(1) effects.



Clicking on **OK** gives the ARCH test results below.

#### Heteroskedasticity Test: ARCH

F-statistic	70.71980	Prob. F(1,497)	0.0000
Obs*R-squared	62.15950	Prob. Chi-Square(1)	0.0000

Test Equation:  
Dependent Variable: RESID^2  
Method: Least Squares

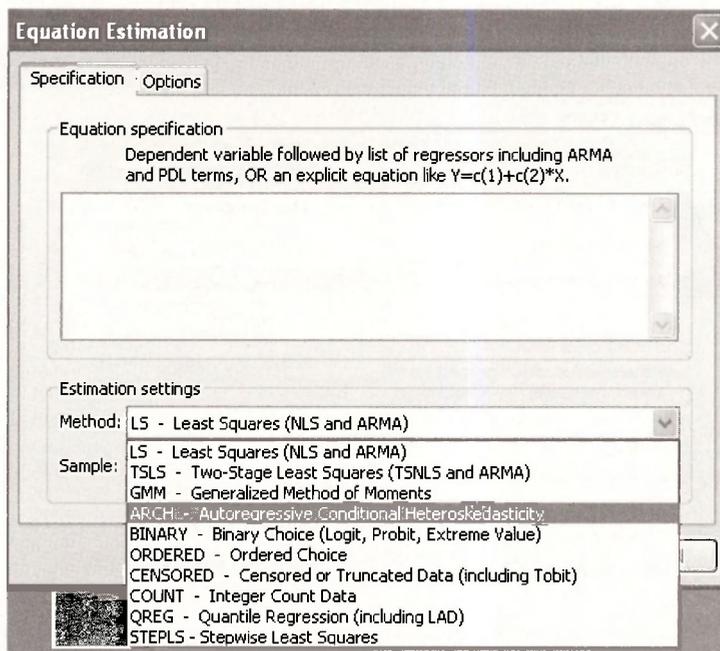
Sample (adjusted): 2 500  
Included observations: 499 after adjustments

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.908262	0.124401	7.301068	0.0000
RESID^2(-1)	0.353071	0.041985	8.409506	0.0000
R-squared	0.124568	Mean dependent var		1.401953
Adjusted R-squared	0.122807	S.D. dependent var		2.615903
S.E. of regression	2.450018	Akaike info criterion		4.634068
Sum squared resid	2983.286	Schwarz criterion		4.650952
Log likelihood	-1154.200	Hannan-Quinn criter.		4.640694
F-statistic	70.71980	Durbin-Watson stat		2.067032
Prob(F-statistic)	0.000000			

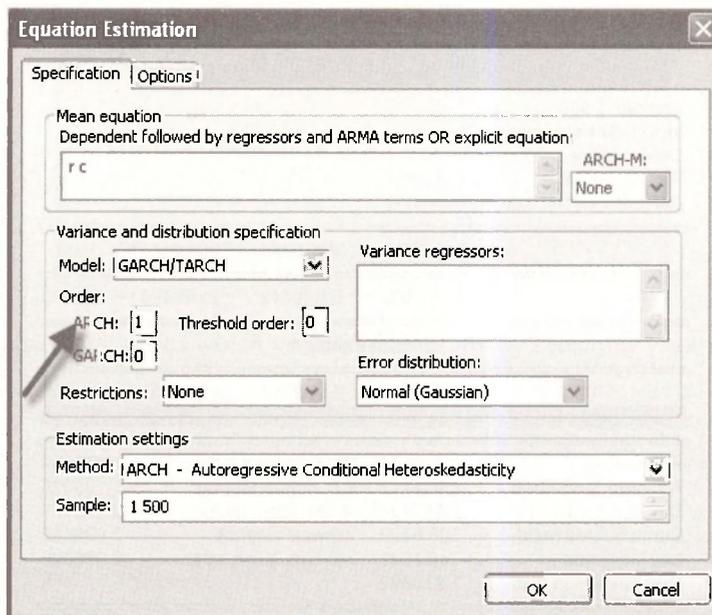
Since the *LM* statistic (62.159) is significant, we reject the null hypothesis that there is no first-order ARCH effects. Note that the *LM* statistic in EViews is calculated as  $LM = T \times R^2 = 499 \times 0.124568 = 62.16$ . Furthermore, the *F*- and *t*-statistics ( $62.16 = 8.4095^2$ ) corroborate the presence of first order ARCH effects.

### 14.3 ESTIMATING AN ARCH MODEL

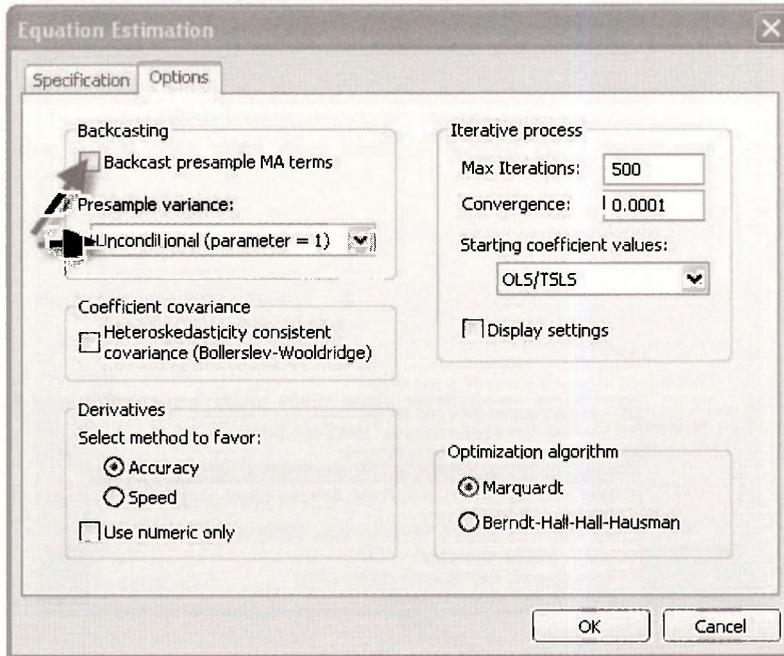
To estimate an ARCH model, click **Quick/ Estimate equation** and select the ARCH option from the drop-down menu in **Method**.



A screen with an upper **Mean equation** and a lower **Variance and distribution specification** section will open up. In the mean equation section, enter the regression of the returns,  $R$ , on a constant,  $C$ . In the variance and distribution specification section, to estimate an ARCH model of order 1, type a **1** against **ARCH**.



To obtain the standard errors reported in the text, click on **Options** (top left hand corner) and then pick the options noted below. As discussed in the text, time series models require an initial starting value, in this case the initial variance  $h_0$ . The options suggested here set the initial variance to the unconditional sample variance.



Clicking on **OK** will give the EViews output below. Note that we have used the default Marquardt algorithm to generate these results.

Dependent Variable: R  
 Method: ML - ARCH (Marquardt) - Normal distribution

Sample: 1 500  
 Included observations: 500  
 Convergence achieved after 10 iterations  
 Presample variance: unconditional  
 GARCH = C(2) + C(3)\*RESID(-1)^2

	Coefficient	Std. Error	z-Statistic	Prob.
C	1.063939	0.039442	26.97458	0.0000

Variance Equation				
	Coefficient	Std. Error	z-Statistic	Prob.
C	0.642140	0.063214	10.15827	0.0000
RESID(-1)^2	0.569343	0.102845	5.535932	0.0000

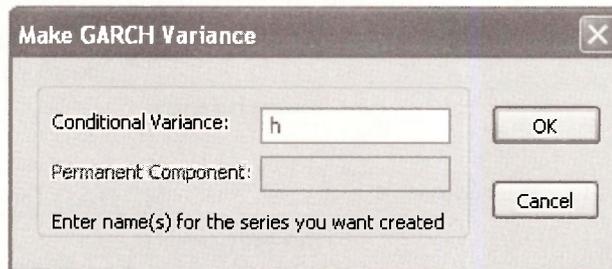
R-squared	-0.000147	Mean dependent var	1.078294
Adjusted R-squared	-0.004172	S.D. dependent var	1.185025
S.E. of regression	1.187494	Akaike info criterion	2.975173
Sum squared resid	700.8403	Schwarz criterion	3.000460
Log likelihood	-740.7932	Hannan-Quinn criter.	2.985096
Durbin-Watson stat	1.918692		

The top section is the mean equation. It shows that the average return is 1.063939. The lower section is the variance equation that gives the result of the ARCH model, namely, that the time varying volatility  $h_t$  includes a constant component (0.642140) plus a component which depends on past errors ( $0.569343e_{t-1}^2$ ). The shaded line highlights the significant ARCH effects.

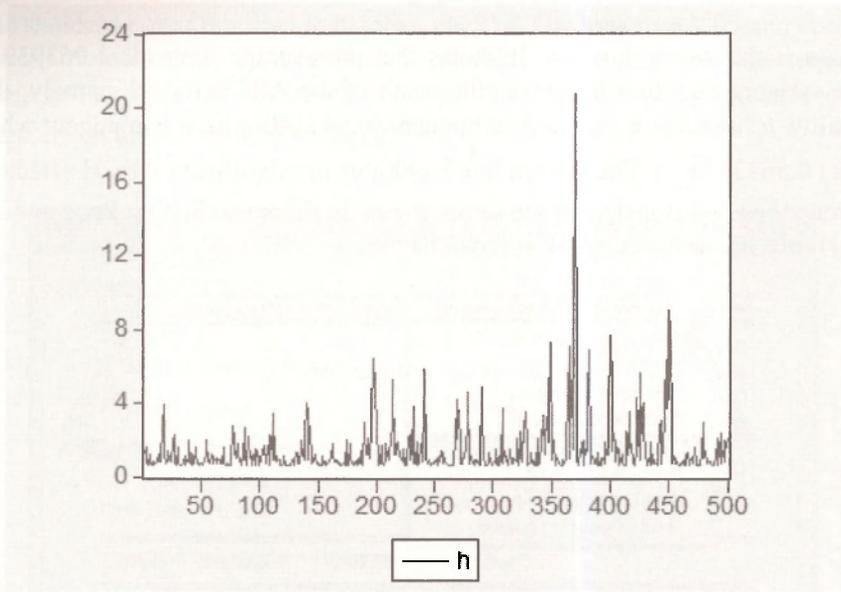
To generate the conditional variance series shown in the text, click on **Proc** and select **Make GARCH Variance Series** from the drop-down menu.

View	Proc	Object	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dep	Specify/Estimate...								
Me	Forecast...								Normal distribution
Dist	Make Residual Series...								
San	Make Regressor Group								
Incl	Make GARCH Variance Series...								
Cor	Make Gradient Group								
Pre	Make Derivative Group								
GAP	Update Coefs from Equation								
						Coefficient	Std. Error	z-Statistic	Prob.
	C	1.063939	0.039442	26.97458	0.0000				
Variance Equation									
	C	0.642140	0.063214	10.15827	0.0000				
	RESID(-1)^2	0.569343	0.102845	5.536932	0.0000				
R-squared		-0.000147	Mean dependent var		1.078294				
Adjusted R-squared		-0.004172	S.D. dependent var		1.185025				
S.E. of regression		1.187494	Akaike info criterion		2.975173				
Sum squared resid		700.8403	Schwarz criterion		3.000460				
Log likelihood		-740.7932	Hannan-Quinn criter.		2.985096				
Durbin-Watson stat		1.918692							

Clicking opens the window below. We have used **H** to label the conditional variance.

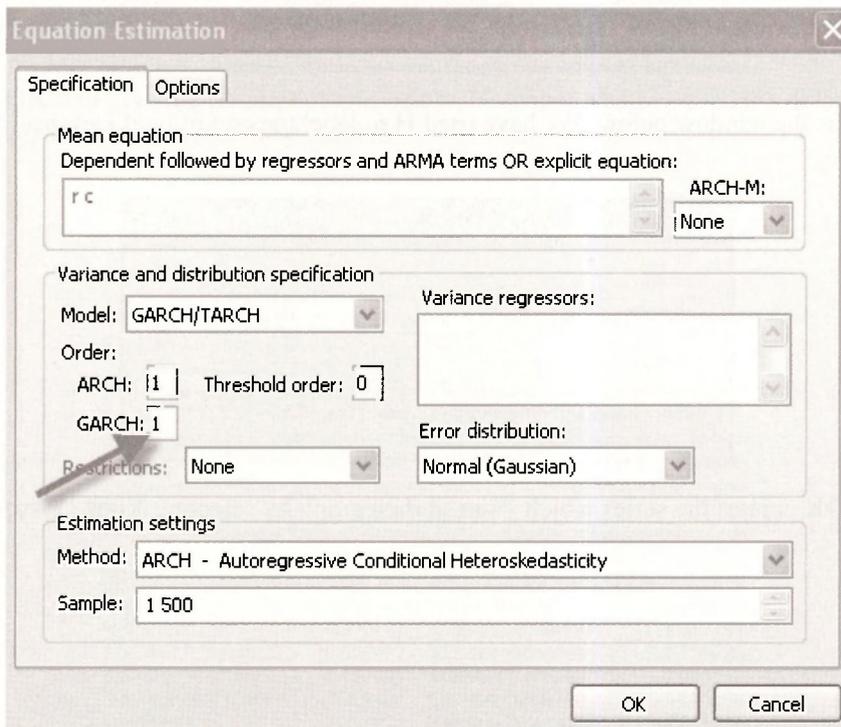


Clicking on **OK** creates the series which you can then graph by selecting **View / Graph/ Line & Symbol/**.



### 14.4 GENERALIZED ARCH

To estimate a GARCH(1,1) model, select the option shown below.



Clicking on **OK** produces the EViews results below.

Dependent Variable: R  
Method: ML - ARCH (Marquardt) - Normal distribution

Sample: 1 500  
Included observations: 500  
Convergence achieved after 17 iterations  
Presample variance: unconditional  
GARCH = C(2) + C(3)\*RESID(-1)^2 + C(4)\*GARCH(-1)

	Coefficient	Std. Error	z-Statistic	Prob.
C	1.049864	0.040465	25.94520	0.0000
Variance Equation				
C	0.401044	0.089940	4.459026	0.0000
RESID(-1)^2	0.491027	0.101570	4.834362	0.0000
GARCH(-1)	0.238007	0.111500	2.134585	0.0328
R-squared	-0.000577	Mean dependent var		1.078294
Adjusted R-squared	-0.006629	S.D. dependent var		1.185025
S.E. of regression	1.188946	Akaike info criterion		2.960112
Sum squared resid	701.1414	Schwarz criterion		2.993829
Log likelihood	-736.0281	Hannan-Quinn criter.		2.973343
Durbin-Watson stat	1.917868			

Recall that the generalized GARCH(1,1) model is of the form:

$$h_t = \delta + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1}$$

We also note that we need  $\alpha_1 + \beta_1 < 1$  for stationarity; if  $\alpha_1 + \beta_1 \geq 1$  we have a so-called “integrated GARCH” process, or IGARCH.

The shaded line in the EViews output shows the significance of the GARCH term. These results show that the volatility coefficients, the one in front of the ARCH effect (0.491027) and the one in front of the GARCH effect (0.238007) are both positive and their sum is between zero and one, as required by theory.

## 14.5 ASYMMETRIC GARCH

The threshold ARCH model, or T-ARCH, is one example where positive and negative news are treated asymmetrically. In the T-GARCH version of the model, the specification of the conditional variance is:

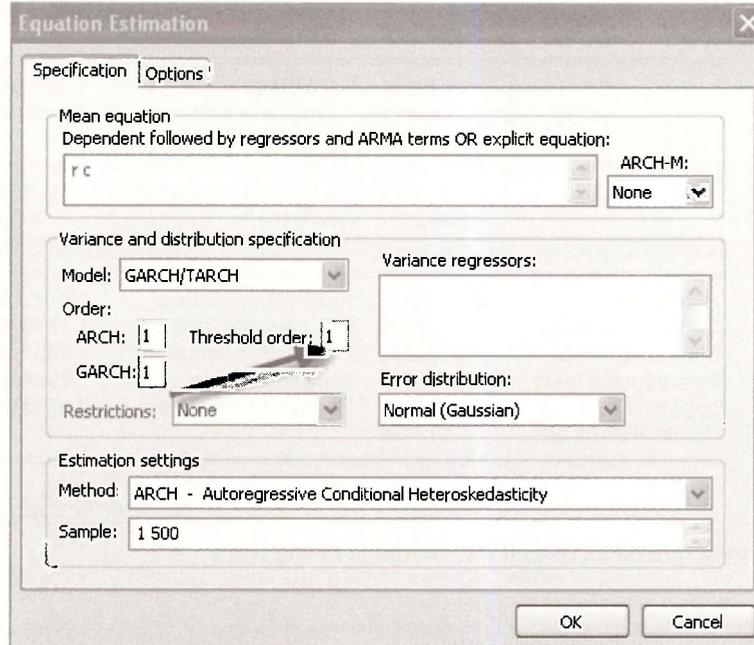
$$h_t = \delta + \alpha_1 e_{t-1}^2 + \gamma d_{t-1} e_{t-1}^2 + \beta_1 h_{t-1}$$

$$d_t = \begin{cases} 1 & e_t < 0 \text{ (bad news)} \\ 0 & e_t > 0 \text{ (good news)} \end{cases}$$

where  $\gamma$  is known as the asymmetry or leverage term.

When  $\gamma = 0$ , the model collapses to the standard GARCH form. Otherwise, when the shock is positive (i.e. good news) the effect on volatility is  $\alpha_1$  but when the news is negative (i.e. bad news) the effect on volatility is  $\alpha_1 + \gamma$ . Hence, so long as  $\gamma$  is significant and positive, negative shocks have a larger effect on  $h_t$  than positive shocks.

To estimate a threshold GARCH model, select the option shown below.



Clicking **OK** gives the EViews output

Dependent Variable: R  
Method: ML - ARCH (Marquardt) - Normal distribution

Sample: 1 500  
Included observations: 500  
Convergence achieved after 34 iterations  
Presample variance: unconditional

$$\text{GARCH} = C(2) + C(3) * \text{RESID}(-1)^2 + C(4) * \text{RESID}(-1)^2 * (\text{RESID}(-1) < 0) + C(5) * \text{GARCH}(-1)$$

	Coefficient	Std. Error	z-Statistic	Prob.
C	0.994809	0.042918	23.17943	0.0000
Variance Equation				
C	0.355662	0.090047	3.949717	0.0001
RESID(-1)^2	0.262576	0.080374	3.266924	0.0011
RESID(-1)^2 * (RESID(-1) < 0)	0.491902	0.204566	2.404609	0.0162
GARCH(-1)	0.287370	0.115485	2.488375	0.0128
R-squared	-0.004973	Mean dependent var		1.078294
Adjusted R-squared	-0.013094	S.D. dependent var		1.185025
S.E. of regression	1.192758	Akaike info criterion		2.942215
Sum squared resid	704.2222	Schwarz criterion		2.984361
Log likelihood	-730.5537	Hannan-Quinn criter.		2.958753
Durbin-Watson stat	1.909478			

Since the coefficient on the asymmetric term (0.492) is significant, we infer that there is evidence that positive and negative shocks have different effects. In particular, when the shock is positive, the estimate of the time-varying volatility is

$$h_t = 0.355662 + 0.262576e_{t-1}^2 + 0.287370h_{t-1}$$

and when the shock is negative, the estimate of the time-varying volatility is

$$h_t = 0.355662 + (0.262576 + 0.491902)e_{t-1}^2 + 0.287370h_{t-1}$$

## 14.6 GARCH-IN-MEAN

The equations of a GARCH-in-mean model are shown below:

$$y_t = \beta_0 + \theta h_t + e_t$$

$$e_t | I_{t-1} \sim N(0, h_t)$$

$$h_t = \delta + \alpha_1 e_{t-1}^2 + \beta_1 h_{t-1}, \quad \delta > 0, \quad 0 \leq \alpha_1 < 1, \quad 0 \leq \beta_1 < 1$$

The first equation is the mean equation; it now shows the effect of the conditional variance on the dependent variable. In particular, note that the model postulates that the conditional variance  $h_t$  affects  $y_t$  by a factor  $\theta$ . The other two equations are as before.

To estimate a GARCH in mean model, select the option shown below.

The screenshot shows the 'Equation Estimation' dialog box with the following settings:

- Specification:** Options
- Mean equation:** Dependent followed by regressors and ARMA terms OR explicit equation:  $r c$ . **ARCH-M:** Variance
- Variance and distribution specification:**
  - Model:** GARCH/TARCH
  - Order:** ARCH: 1, Threshold order: 1
  - GARCH:** 1
  - Restrictions:** None
  - Variance regressors:** (empty)
  - Error distribution:** Normal (Gaussian)
- Estimation settings:**
  - Method:** ARCH - Autoregressive Conditional Heteroskedasticity
  - Sample:** 1 500

Buttons: OK, Cancel

Clicking on **OK** produces the EViews output below.

Dependent Variable: R  
 Method: ML - ARCH (Marquardt) - Normal distribution

Sample: 1 500  
 Included observations: 500  
 Convergence achieved after 144 iterations  
 Presample variance: unconditional  
 GARCH = C(3) + C(4)\*RESID(-1)^2 + C(5)\*RESID(-1)^2\*(RESID(-1)<0) +  
 C(6)\*GARCH(-1)

	Coefficient	Std. Error	z-Statistic	Prob.
GARCH	0.195787	0.067159	2.915291	0.0036
C	0.818237	0.071160	11.49860	0.0000

Variance Equation				
C	0.370564	0.081875	4.525960	0.0000
RESID(-1)^2	0.294997	0.086104	3.426047	0.0006
RESID(-1)^2*(RESID(-1)<0)	0.321186	0.162257	1.979493	0.0478
GARCH(-1)	0.278281	0.103908	2.678133	0.0074

R-squared	0.043888	Mean dependent var	1.078294
Adjusted R-squared	0.034211	S.D. dependent var	1.185025
S.E. of regression	1.164578	Akaike info criterion	2.922629
Sum squared resid	669.9834	Schwarz criterion	2.973205
Log likelihood	-724.6573	Hannan-Quinn criter.	2.942475
F-statistic	4.535169	Durbin-Watson stat	1.822325
Prob(F-statistic)	0.000472		

Since the coefficient on the GARCH in mean term (0.196) is significant, we infer that there is evidence that volatility affects returns.

## Keywords

ARCH test  
 asymmetric GARCH

GARCH  
 GARCH-in-mean

time-varying volatility  
 threshold GARCH

# CHAPTER 15

## Panel Data Models

### CHAPTER OUTLINE

- 15.1 Grunfeld Data: Two Equations
    - 15.1.1 Separate least squares estimation
    - 15.1.2 Stacking the data
    - 15.1.3 Least squares estimation with dummy variables
    - 15.1.4 Introducing the pool object
    - 15.1.5 Seemingly unrelated regressions
    - 15.1.6 Testing contemporaneous correlation
    - 15.1.7 Testing cross-equation restrictions
  - 15.2 Grunfeld Data: Ten Firms
    - 15.2.1 Structuring the workfile
    - 15.2.2 Fixed effects using dummy variables
    - 15.2.3 Testing the effects
    - 15.2.4 Pooled least squares
    - 15.2.5 The fixed effects estimator
  - 15.3 NLS Panel Data
    - 15.3.1 Fixed effects estimation
    - 15.3.2 Random effects estimation
    - 15.3.3 The Hausman test
- KEYWORDS

### 15.1 GRUNFELD DATA: TWO EQUATIONS

Panel data are data with two dimensions, a time dimension and a cross-section dimension. They typically comprise observations on a number of economic units, such as individuals or firms, over a number of time periods. The use of panel data involves new models, new econometric techniques and new ways of handling the data. EViews has the capacity to estimate a vast array of models, using many different estimation techniques. Also, the user has various options for handling the data and proceeding to estimation. Some but not all of those options will be introduced as we lead you through the examples in Chapter 15 of the text. The first example involves  $T = 20$  time series observations on just  $N = 2$  cross sectional units, the firms General Electric and Westinghouse. The data can be found in the file *grunfeld2.dat*. We are interested in estimating the two equations

$$INV_{GE} = \beta_{1,GE} + \beta_{2,GE}V_{GE} + \beta_{3,GE}K_{GE} + e_{GE}$$

$$INV_{WE} = \hat{\rho}_{1,WE} + \beta_{2,WE}V_{WE} + \hat{\rho}_{3,WE}K_{WE} + e_{WE}$$

where  $INV$  denotes investment,  $V$  denotes market value of stock and  $K$  denotes capital stock, with the subscripts  $GE$  and  $WE$  referring to General Electric and Westinghouse, respectively. There are various ways of estimating these two equations depending on what further assumptions are made

about the coefficients and the error terms in each of the equations. We first consider separate least squares estimation of each equation.

### 15.1.1 Separate least squares estimation

In the following screenshot the two separate equation specifications for the *GE* and *WE* equations have been superimposed on the workfile. There is nothing new in these specifications. They are straightforward least squares estimations. With respect to the structure of the workfile, there are two things worth noting. First, the observations are dated with the **range** and **sample** specified as annual data from 1935 to 1954. Second, each of the variables has a “subscript”, *GE* or *WE* to signify whether the observations are for General Electric or Westinghouse. These “subscripts” are known as **cross section identifiers**. They will be important in subsequent sections of this chapter.

The screenshot shows the EViews software interface for a workfile named 'GRUNFELD2'. The range and sample are both set to 1935-1954 with 20 observations. On the left, a list of variables is shown with checkboxes: *c*, *inv\_ge*, *inv\_we*, *k\_ge*, *k\_we*, *resid*, *v\_ge*, and *v\_we*. Two equation specifications are displayed:

- General Electric:** The dependent variable is *inv\_ge*, followed by regressors *c*, *v\_ge*, and *k\_ge*.
- Westinghouse:** The dependent variable is *inv\_we*, followed by regressors *c*, *v\_we*, and *k\_we*.

The outputs from each of these regressions follow. Note that they confirm the results in Table 15.1 on page 386 of the text.

Dependent Variable: INV_GE				
Method: Least Squares				
Sample: 1935 1954				
Included observations: 20				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.956306	31.37425	-0.317340	0.7548
V_GE	0.026551	0.015566	1.705705	0.1063
K_GE	0.151694	0.025704	5.901548	0.0000
R-squared	0.705307	Sum squared resid	13216.59	

Dependent Variable: INV_WE				
Method: Least Squares				
Sample: 1935 1954				
Included observations: 20				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.509390	8.015289	-0.063552	0.9501
V WE	0.052894	0.015707	3.367658	0.0037
K_WE	0.092406	0.056099	1.647205	0.1179
R-squared	0.744446	Sum squared resid	1773.234	

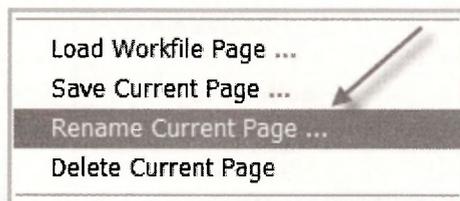
### 15.1.2 Stacking the data

In the previous section we estimated two regression equations with 20 observations for each. As noted in equation (15.6) of the text, the same least squares estimates can be obtained by pooling the observations into one sample of 40, and including intercept and slope dummy variables for each of the coefficients. The standard errors turn out differently, however. With separate least squares estimation we get separate estimates for  $\sigma_{GE}^2$  and  $\sigma_{WE}^2$ . When the observations are pooled into one sample, the implicit assumption is that  $\sigma_{GE}^2 = \sigma_{WE}^2$  and only one error variance estimate is obtained.

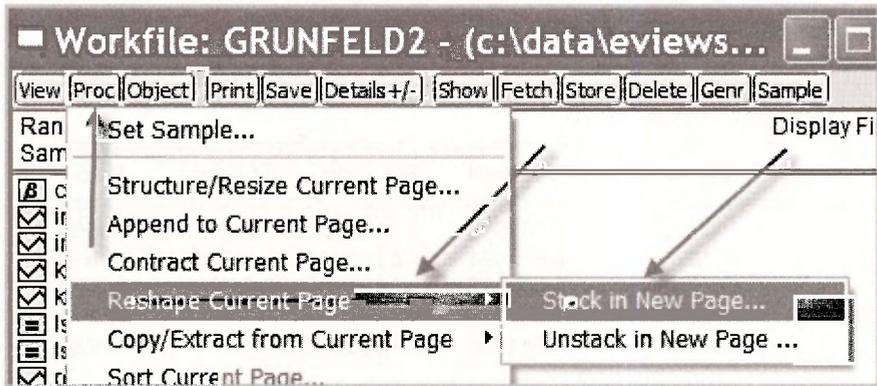
To obtain the pooled dummy variable estimates, it is convenient to stack the observations into one sample of size 40. In addition to stacking *INV*, *V* and *K*, we will create the required dummy variable by defining  $DUM\_WE = 1$  and  $DUM\_GE = 0$ , and also stacking these two series.

```
series dum_we = 1
series dum_ge = 0
```

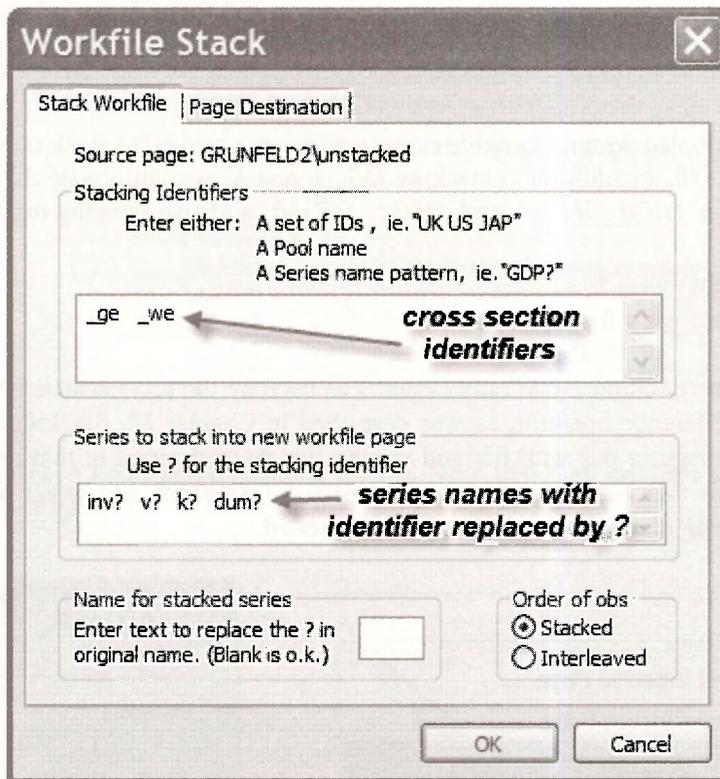
We have chosen the notation *DUM* rather than *D* as used by the text because EViews reserves *D* to be used as a difference operator, as was described in Chapter 12. Stacking is carried out by creating a second page in our workfile and storing the stacked series in that page. But, first we name the first page that contains the unstacked data. Go to **Proc**, and select **Rename Current Page**. In the resulting dialog box, call the page **unstacked**.



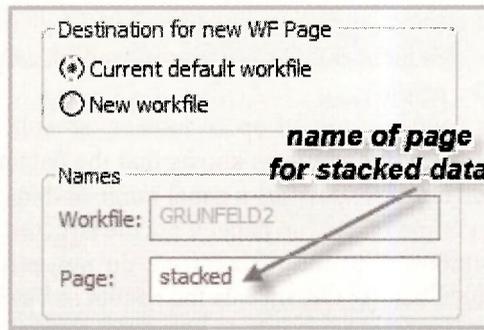
To create a new page with the stacked data, go to **Proc/Reshape Current Page/Stack in New Page**.



A **Workfile Stack** dialog box appears. The cross section identifiers *\_GE* and *\_WE* are inserted in the **Stacking Identifiers** box. In the box that says **Series to stack into new workfile page**, each of the series names is entered without the subscripts (identifiers), and with each identifier replaced by a question mark ?. Leaving the box below that blank will mean that the new series of length 40, with both the *GE* and *WE* observations, will be called *INV*, *V*, *K* and *DUM*.



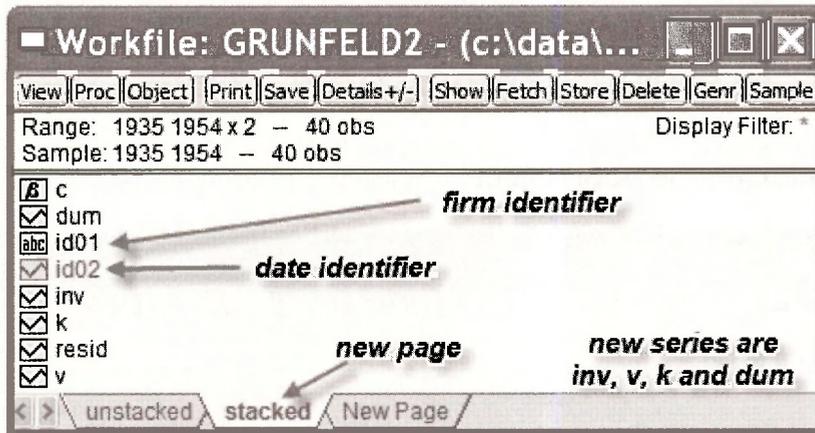
Notice the second tab in the **Workfile Stack** box called **Page Destination**. Click on that. We are keeping the current workfile and naming the new page **stacked**.



The new workfile page called **stacked** is illustrated below. Check out the following.

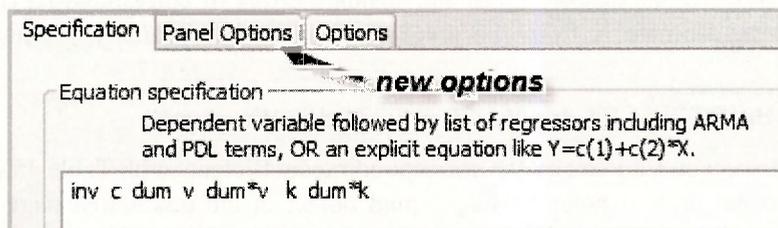
1. The **Range** is given as **1935 1954 × 2** implying we have two cross sections for the specified time period.
2. The names of the new series that include all 40 observations are *INV*, *V*, *K* and *DUM*.
3. There are two new series **ID01** and **ID02**. The first one indicates which observations are *GE* and which are *WE*. The second contains the date of each observation.

Open the various series and familiarize yourself with how EViews has set them up.



### 15.1.3 Least squares estimation with dummy variables

We are now in a position to obtain the estimates given in Table 15.2 on page 387 of the text. Follow the familiar routine of going to **Object/New Object/Equation**, name the equation object and fill in the **Equation specification**.



The variables specified are those that appear in equation (15.6) and Table 15.2 of the text. Notice that we are able to enter the products  $\text{dum} \times v$  and  $\text{dum} \times k$  without creating new series. EViews figures it out and gives you the results.

Something new that has suddenly turned up is another tab called **Panel Options**. Because you went through a stacking procedure, EViews knows that the data are panel data. Accordingly, it set up a panel workfile structure that specifies a panel range and includes objects describing the cross section and time series components. The panel workfile structure includes **Panel Options** in the **Equation Estimation** window. For the moment we do not need these options, but we do consider some of them shortly. Clicking OK reveals the results in Table 15.2 on page 387.

Dependent Variable: INV				
Method: Panel Least Squares				
Sample: 1935 1954				
Periods included: 20				
Cross-sections included: 2				
Total panel (balanced) observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.956306	23.62636	-0.421407	0.6761
DUM	9.446916	28.80535	0.327957	0.7450
V	0.026551	0.011722	2.265064	0.0300
DUM*V	0.026343	0.034353	0.766838	0.4485
K	0.151694	0.019356	7.836865	0.0000
DUM*K	-0.059287	0.116946	-0.506962	0.6155

*description of nature of panel observations*

### 15.1.4 Introducing the pool object

The dummy variable model estimated in the previous section is given by

$$INV = \beta_{1,GE} + \delta_1 DUM + \beta_{2,GE} V + \delta_2 (DUM \times V) + \beta_{3,GE} K + \delta_3 (DUM \times K) + e$$

Using the substitutions

$$\delta_1 = \beta_{1,WE} - \beta_{1,GE} \quad \delta_2 = \beta_{2,WE} - \beta_{2,GE} \quad \delta_3 = \beta_{3,WE} - \beta_{3,GE}$$

the dummy variable model can be written as

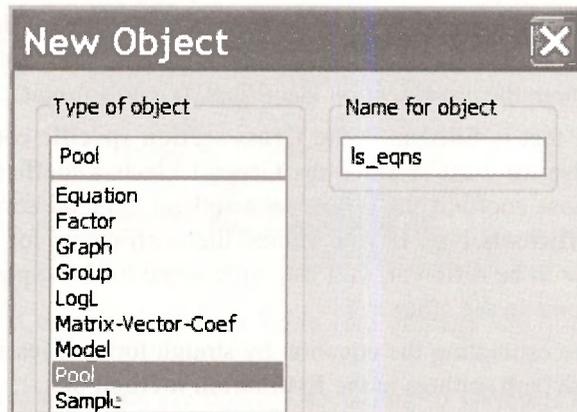
$$INV = \beta_{1,GE}(1 - DUM) + \beta_{1,WE} DUM + \beta_{2,GE}(1 - DUM) \times V + \beta_{2,WE} (DUM \times V) \\ + \beta_{3,GE}(1 - DUM) \times K + \beta_{3,WE} (DUM \times K) + e$$

Estimating this equation will give exactly the same results as those from the earlier dummy variable model in the sense that estimates and standard errors of corresponding coefficients will be equal. We can estimate it from the **stacked** page of *grunfeld2.wfl*, using the equation specification

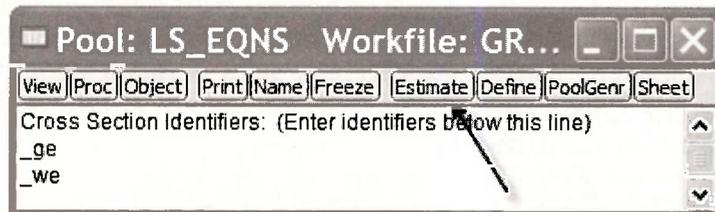
**inv (1-dum) dum (1-dum)\*v dum\*v (1-dum)\*k dum\*k**

Try it! See what you get. Can you match corresponding coefficients with Table 15.2?

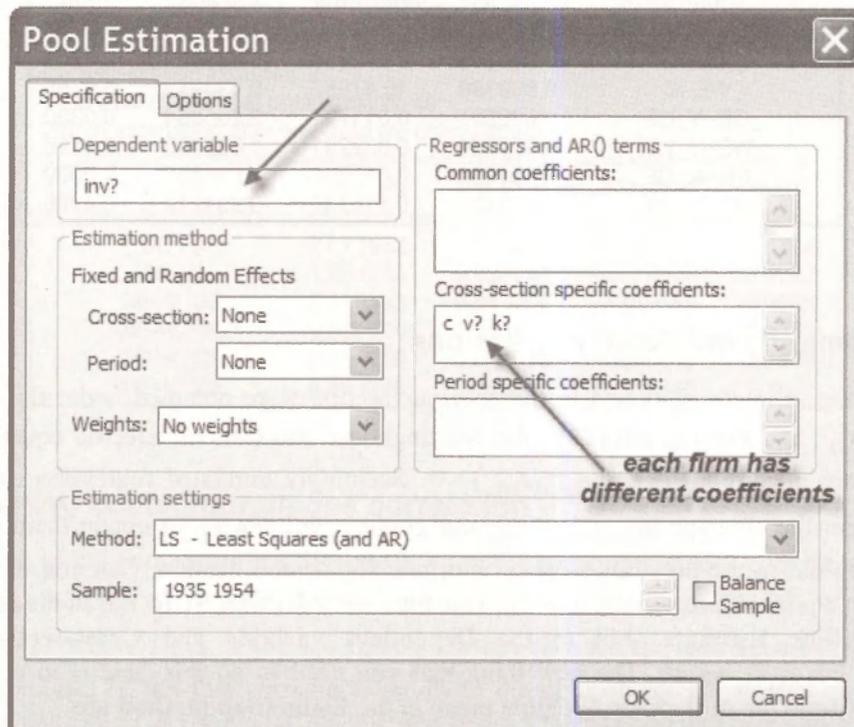
We can also get these estimates using a **pool** object in the **unstacked** page. Return to the **unstacked** page and select **Object/New Object/Pool**. We named the pool object **LS\_EQNS**.



EViews will then ask you for the cross section identifiers which in this case are *\_GE* and *\_WE*. For this procedure to work, series names should be expressed with a common component such as *INV*, *K* and *V* and with a cross section identifying component like *\_GE* and *\_WE*.



Then click on **Estimate**. Wow! Look at all the boxes you have to fill in. Don't be scared. At the moment we are only concerned with two of them.



1. For the **Dependent variable** we have written *INV?*. Writing it this way, with the question mark, tells EViews to consider all values on investment. Remember that you have already told EViews about the cross-section identifiers. It won't forget.
2. The other box that is filled in is the **Cross-section specific coefficients** box. We chose this one because we want to allow the General Electric coefficients to be different from the Westinghouse coefficients. If you wanted them to be the same, you would choose the **Common coefficients** box. If you wanted the coefficients for some variables to be the same and some to be different, you can write some of the explanatory variable names in one box and some in the other.
3. Because we are estimating the equation by straightforward least squares, we do not need to change the default settings in the **Estimation method** box.

The results follow. They are equivalent to those in Table 15.2, although at first glance you might not think so. We can see the equivalence by noting that

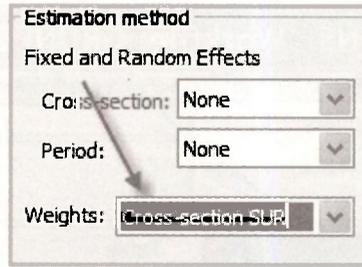
$$\hat{\delta}_1 = -0.5094 - (-9.9563) = 9.4469 \quad \hat{\delta}_2 = 0.052894 - 0.026551 = 0.026343$$

$$\hat{\delta}_3 = 0.09241 - 0.15169 = -0.05928$$

Dependent Variable: INV?				
Method: Pooled Least Squares				
Date: 12/10/07 Time: 01:46				
Sample: 1935 1954				
Included observations: 20				
Cross-sections included: 2				
Total pool (balanced) observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
_GE-C	-9.956306	23.62636	-0.421407	0.6761
_WE-C	-0.509390	16.47857	-0.030912	0.9755
_GE-V_GE	0.026551	0.011722	2.265064	0.0300
_WE-V_WE	0.052894	0.032291	1.638052	0.1106
_GE-K_GE	0.151694	0.019356	7.836865	0.0000
_WE-K_WE	0.092406	0.115333	0.801212	0.4286

### 15.1.5 Seemingly unrelated regressions

The coefficient estimates obtained in the previous section were obtained under the assumption that  $\sigma_{GE}^2 = \sigma_{WE}^2$ , and that the errors for the Westinghouse and General Electric equations, in the same year, are uncorrelated,  $\text{cov}(e_{GE,t}, e_{WE,t}) = 0$ . Seemingly unrelated regression estimates are obtained under the assumptions  $\sigma_{GE}^2 \neq \sigma_{WE}^2$  and  $\text{cov}(e_{GE,t}, e_{WE,t}) \neq 0$ . To obtain them we proceed exactly as we did in the previous section, with one slight modification. Can you remember the steps? Set up a pool object. Give it a name. This time we will call it **SUR**. Fill in the cross-section identifiers. Click **Estimate**. Fill in the **Dependent variable** and **Cross-section specific coefficients** boxes as before. The new thing that you need to do this time is to select **Cross-section SUR** from the drop-down **Weights** menu in the **Estimation method** box.



The results appear below. Compare them with Table 15.3 on page 388 of the text. The following points are worth noting.

1. In the output the coefficients are ordered according to variable. In Table 15.3 they are ordered according to equation.
2. EViews calls the estimation method **Pooled EGLS** (estimated generalized least squares). The SUR estimator is a particular kind of generalized least squares estimator.
3. Although the coefficient estimates are identical to those in Table 15.3, the standard errors are not. The difference arises because EViews uses  $T$  as the divisor when estimating the error variances and covariance, whereas the degrees of freedom corrected divisor  $T - K$  was used for the results in Table 15.3. Both are popular. To reconcile the two, consider the last standard error reported from both places and note that

$$0.0530 \times \sqrt{\frac{T-K}{T}} = 0.0530 \times \sqrt{\frac{17}{20}} = 0.0489$$

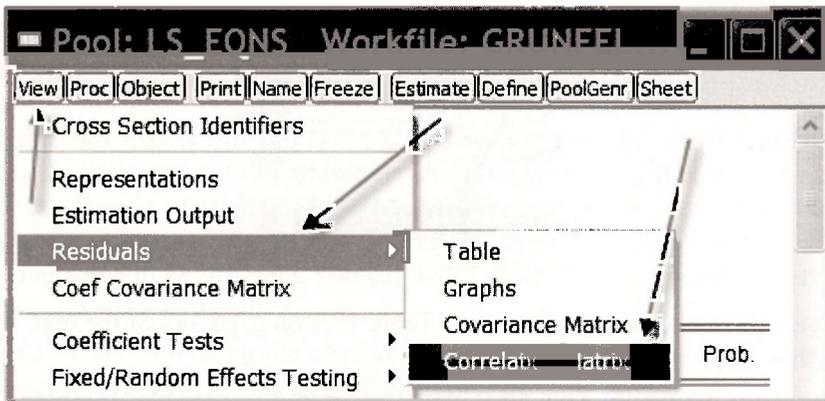
Dependent Variable: INV?  
 Method: Pooled EGLS (Cross-section SUR)  
 Date: 12/10/07 Time: 02:24  
 Sample: 1935 1954  
 Included observations: 20  
 Cross-sections included: 2  
 Total pool (balanced) observations: 40  
 Linear estimation after one-step weighting matrix

**generalized least squares method**

Variable	Coefficient	Std. Error	t-Statistic	Prob.
_GE-C	-27.71932	27.03283	-1.025395	0.3124
_WE-C	-1.251988	6.956347	-0.179978	0.8582
_GE-V GE	0.038310	0.013290	2.882609	0.0068
_WE-V WE	0.057630	0.013411	4.297200	0.0001
_GE-K GE	0.139036	0.023036	6.035716	0.0000
_WE-K WE	0.063978	0.048901	1.308318	0.1995

### 15.1.6 Testing contemporaneous correlation

In the context of the two-equation SUR model, a test for contemporaneous correlation is a test of  $H_0 : \text{cov}(e_{GE,t}, e_{WE,t}) = 0$ . The relevant test statistic, described on page 389 of the text, is  $LM = T \times r_{GE,WE}^2$  where  $r_{GE,WE}^2$  is the squared correlation between the least squares residuals from the two equations. To get this correlation return to the pool object **LS\_EQNS** (we want least squares residuals not SUR residuals), open it, and select **View/Residuals/Correlation Matrix**.



Residual Correlation Matrix		
	GE	WE
GE	1.000000	0.728965
WE	0.728965	1.000000

From the resulting matrix, we have  $r_{GE,WE}^2 = (0.728965)^2 = 0.53139$ , giving a test statistic value of  $LM = 20 \times 0.53139 = 10.628$ . The command

```
scalar pval = 1 - @cchisq(10.6278,1)
```

yields a  $p$ -value of 0.0011. We reject  $H_0$  and conclude that contemporaneous correlation between the equation errors exists.

### 15.1.7 Testing cross-equation restrictions

So far we have been assuming that General Electric and Westinghouse have different coefficients. Could they be the same? To answer this question we test the hypothesis

$$H_0: \beta_{1,GE} = \beta_{1,WE}, \beta_{2,GE} = \beta_{2,WE}, \beta_{3,GE} = \beta_{3,WE}$$

This hypothesis can be tested using the **Wald test** option from SUR estimation. For carrying out the test we can follow the same steps as described in Chapter 6, although in this case the formulas for the test statistics are more complicated than we have divulged. Also, special care must be exercised to make sure we are testing the coefficients that we want to test. Return to the SUR output. Note the order of the coefficients. This is the order in which EViews stores them in the **C** vector. Consequently, writing the null hypothesis in terms of EViews coefficients, we have

$$H_0: C(1)=C(2), C(3)=C(4), C(5)=C(6)$$

Select **View/Coefficient Tests/Wald - Coefficient Restrictions**. Enter the following restrictions in the **Wald test** box.

Coefficient restrictions separated by commas  
 c(1) = c(2), c(3) = c(4), c(5) = c(6)

Wald Test:			
Pool: SUR			
Test Statistic	Value	df	Probability
F-statistic	3.437931	(3, 34)	0.0275
Chi-square	10.31379	3	0.0161
Null Hypothesis Summary:			
Normalized Restriction (= 0)	Value	Std. Err.	
C(1) - C(2)	-26.46733	22.91810	
C(3) - C(4)	-0.019320	0.010793	
C(5) - C(6)	0.075058	0.041621	

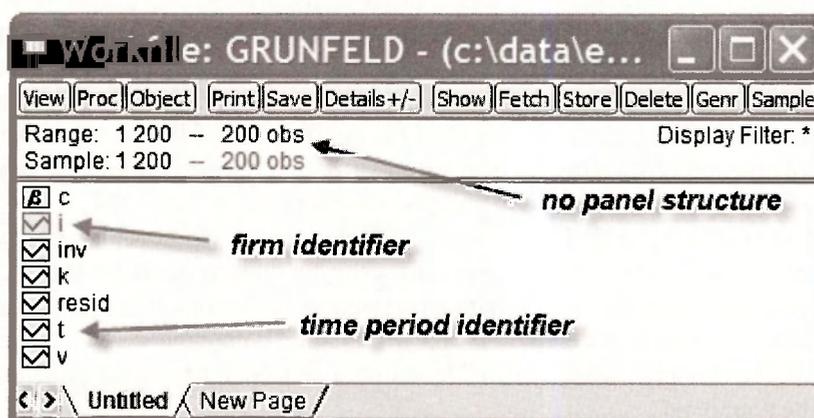
In the lower part of the output, the normalized restrictions are  $\beta_{1,GE} - \beta_{1,WE} = 0$ ,  $\beta_{2,GE} - \beta_{2,WE} = 0$ , and  $\beta_{3,GE} - \beta_{3,WE} = 0$ . Estimates of the left hand sides of the restrictions and their standard errors appear in the columns **Value** and **Std. Err.** The  $F$ - and  $\chi^2$ -values for the test are given in the upper part of the output, along with their corresponding  $p$ -values. The hypothesis of equal coefficients is rejected.

There is a discrepancy between the values in the text on pages 390-1 and those in the above output. Those in the text are  $F = 2.92$  and  $\chi^2 = 8.77$ . The difference is again attributable to the treatment of a degrees of freedom correction when estimating the error variances and covariance. To convert the EViews values to the text values, we multiply by  $(17/20)$ .

$$3.4379 \times \frac{17}{20} = 2.92 \qquad 10.3138 \times \frac{17}{20} = 8.77$$

## 15.2 GRUNFELD DATA: TEN FIRMS

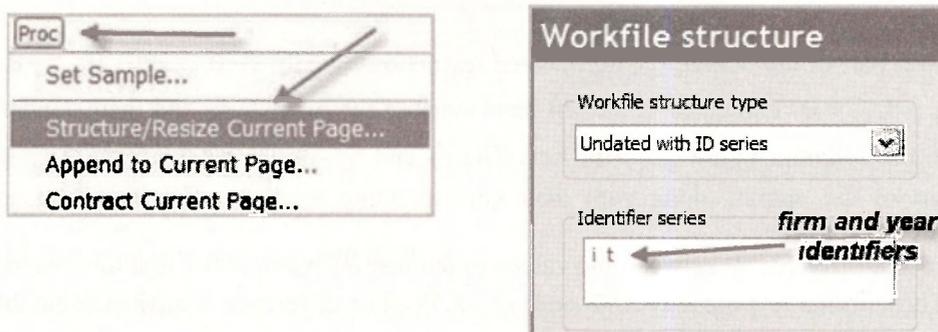
A more complete set of the Grunfeld data comprising  $T = 20$  observations on  $N = 10$  firms can be found in the workfile *grunfeld.wfl*. The contents of this workfile are displayed below.



This file contains the familiar series  $INV$ ,  $V$  and  $K$  and two new series  $I$  and  $T$ . The series  $I$  identifies observations for the  $i$ -th firm,  $i=1,2,\dots,10$ . The series  $T$  identifies observations for the  $t$ -th time period,  $t=1,2,\dots,20$ . The **Range** and **Sample** are both simply set at **1 200** without recognition of the panel structure of the data. So that EViews is fully informed, we begin this section by specifying the panel structure.

### 15.2.1 Structuring the workfile

Go to **Proc/Structure/Resize Current Page**. A **Workfile structure** dialog box appears. There are various options that we could choose from the drop-down menu **Workfile structure type**. Since we have the identifying series  $I$  and  $T$  in the workfile, we choose **undated with ID series** and insert the names of these series in the **Identifier series** box.



When you return to the workfile, you will see extra information under **Range** that says **Dim(10,20)**. In other words, the panel dimension is  $(10 \times 20)$ . EViews has figured out this dimension by checking the values in  $I$  and  $T$ .

Range: 1 200	Dim(10,20)	– 200 obs
Sample: 1 200	–	200 obs

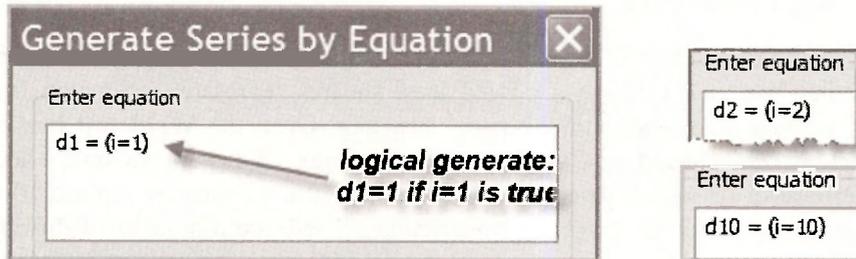
### 15.2.2 Fixed effects using dummy variables

Table 15.4 on page 392 of the text presents estimates of the investment functions for the 10 firms assuming (1) all firms have the same coefficients on  $V$  and  $K$ , (2) each firm has a different intercept, (3) the error variances are the same for all firms, and (4) there is no contemporaneous correlation between errors of different firms. Taken together, these assumptions comprise those of a standard fixed effects model. The different intercept terms are known as fixed effects. The fixed effects model can be estimated in one of two ways. Dummy variables can be included for each of the firms and the constant omitted. In this case the coefficients of the dummy variables are the intercepts (fixed effects). Alternatively, the data can be expressed in terms of deviations from firm means and estimated without any intercepts, as described on page 394 of the text. We will first estimate the model by including dummy variables. Later we consider EViews automatic fixed effects option, and relate it back to our results for the dummy variable specification. We do not explicitly consider estimation using data expressed as deviations from firm means, although that is undoubtedly the approach taken by EViews automatic command.

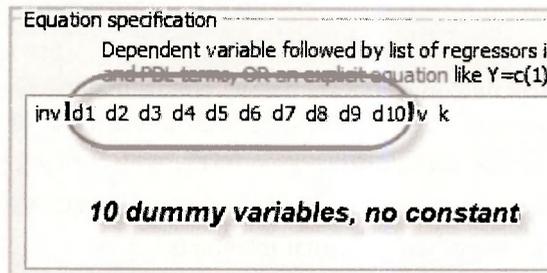
We generate the dummy variable series by using a sequence of logical generate commands. For example, the command

series  $d1 = (i=1)$

generates a series **D1** that is equal to one when  $(i=1)$  is true and equal to zero when  $(i=1)$  is false. Ten such commands are needed, one for each dummy variable.



To estimate the model, proceed to the **Equation specification** in the usual way. Enter the dependent variable *INV*, followed by each of the dummy variables, *V* and *K*. You will have noticed the **Panel options** tab at the top of the **Equation estimation** window. It is not needed. You might be tempted to select fixed effects. That would be wrong. Including the dummy variables means the fixed effects are already included. If you try to do it twice, EViews will get upset and send you a nasty **singular matrix** message.



Dependent Variable: INV  
 Method: Panel Least Squares  
 Sample: 1 200  
 Periods included: 20  
 Cross-sections included: 10  
 Total panel (balanced) observations: 200

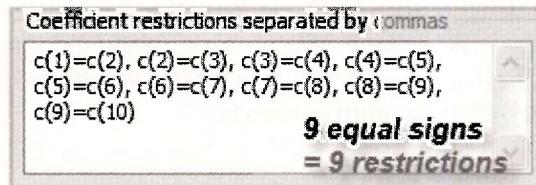
*panel description*

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D1	-69.14348	49.68547	-1.391624	0.1657
D2	100.8624	24.91366	4.048478	0.0001
D3	-235.1187	24.41825	-9.628812	0.0000
D4	-27.63498	14.06983	-1.964130	0.0510
D5	-115.3169	14.16199	-8.142703	0.0000
D6	-23.07357	12.66121	-1.822382	0.0700
D7	-66.68293	12.83763	-5.194332	0.0000
D8	-57.35860	13.98559	-4.101265	0.0001
D9	-87.27701	12.88512	-6.773473	0.0000
D10	-6.546269	11.81987	-0.553836	0.5803
V	0.109771	0.011855	9.259556	0.0000
K	0.310644	0.017370	17.88354	0.0000
Sum squared resid	522855.2			

Compare the above output with that in Table 15.4 of the text. Note the way EViews describes the panel structure in the top portion of the output.

### 15.2.3 Testing the effects

A test likely to be of interest is one that checks whether the intercepts for all firms could be identical. If they are, one can use a pooled least squares regression estimated from the 200 observations without any regard for the panel structure. Open the **Wald test** box by going to **View/Coefficient Tests/Wald – Coefficient Restrictions**. We want to test whether the 10 intercept coefficients are equal. Another way of putting it is we want to replace 10 coefficients with one coefficient. To do so involves 9 restrictions. There are a number of different ways of writing these restrictions. One way appears in the box below. Note that the intercepts represent the first 10 coefficients and so they will be numbered  $C(1)$ ,  $C(2)$ , ...,  $C(10)$ . Another alternative is to set  $C(1) = C(2)$ ,  $C(1) = C(3)$ , ...,  $C(1) = C(10)$ . You will be able to think of other ways.

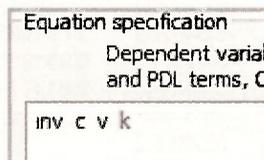


The upper panel of the test outcome appears below. Notice that the  $F$ -value is the same as that on page 393 of the text, obtained using restricted and unrestricted sums of squared errors. The relationship between the two test values is  $\chi^2 = 9 \times F$ , with 9 being the degrees of freedom for the  $\chi^2$ -test and the numerator degrees of freedom for the  $F$ -test. With  $p$ -values of 0.0000, both tests clearly reject the null hypothesis of equal intercepts.

Wald Test:			
Equation: TABLE15_4			
Test Statistic	Value	df	Probability
F-statistic	48.99152	(9, 188)	0.0000
Chi-square	440.9237	9	0.0000

### 15.2.4 Pooled least squares

The pooled least squares estimates that make no special assumptions to accommodate the panel structure are given in Table 15.7 on page 395 of the text. No special commands are required to produce these estimates. Following the usual steps, leads to the **Equation specification** and results that appear below.



Dependent Variable: INV  
 Method: Panel Least Squares  
 Sample: 1 200  
 Periods included: 20  
 Cross-sections included: 10  
 Total panel (balanced) observations: 200

*panel description*

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-43.02448	9.497896	-4.529896	0.0000
V	0.115374	0.005830	19.78916	0.0000
K	0.231931	0.025465	9.107895	0.0000
Sum squared resid	1749128.			

### 15.2.5 The fixed effects estimator

Now consider EViews automatic command for estimating a fixed effects (dummy variable) model. You fill in the same **Equation specification** as you did for the pooled least squares estimator in the previous section, but this time you need to click on the **Panel options** tab and choose **Fixed** for the **Cross-section Effects specification**.

**Equation specification**

Dependent variable and PDL terms, C

inv c v k

Specification Panel Options Options

Effects specification

Cross-section: Fixed

Period: None

*fixed effects for firms*

The upper part of the output appears below. You should notice that the estimates for the coefficients of *V* and *K* are identical to those obtained when we explicitly included the dummy variables in Section 15.2.2. Also, if you took time to do the arithmetic, you would discover that the new intercept  $-58.729$  is equal to the average of the dummy variable coefficients obtained earlier.

Dependent Variable: INV  
 Method: Panel Least Squares  
 Sample: 1 200  
 Periods included: 20  
 Cross-sections included: 10  
 Total panel (balanced) observations: 200

*average of fixed effects*

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-58.72901	12.44628	-4.718601	0.0000
V	0.109771	0.011855	9.259556	0.0000
K	0.310644	0.017370	17.88354	0.0000

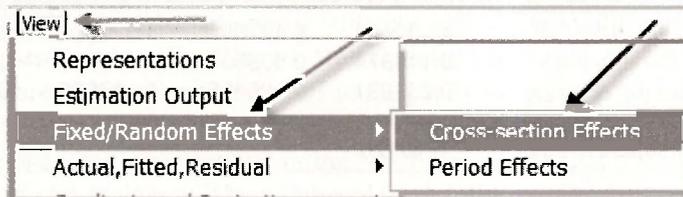
Effects Specification

---

Cross-section fixed (dummy variables)

**15.2.5a Retrieving the fixed effects**

Sometimes the fixed effects (intercept estimates) are of special interest. They can be used to analyze the extent of firm heterogeneity and to examine any particular firms that may be of interest. For many examples the number of fixed effects is enormous and so rather than print them on the output, EViews puts them in a special spreadsheet. To locate this spreadsheet go to **View/Fixed/Random Effects/Cross-section Effects**.

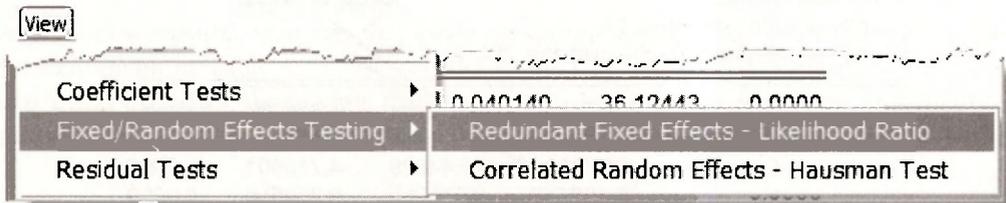


The spreadsheet for the fixed effects for each of the 10 firms is given in the left-hand side of the panel below. A comparison with the dummy variable coefficients from Table 15.4 reveals that they are not the same. The difference is that EViews has expressed them in terms of deviations from the mean of  $-58.729$  that was reported on the output. To get the original fixed effects you add the mean as is done on the right-hand side of the below panel.

Cross-section Fixed Effects		<i>fixed effects: not mean corrected</i>
I	Effect	
1.000000	-10.41447	$-10.41447 - 59.729 = -69.14$
2.000000	159.5914	$159.5914 - 59.729 = 100.86$
3.000000	-176.3897	$-176.3897 - 59.729 = -235.11$
4.000000	31.09403	$31.09403 - 59.729 = -27.64$
5.000000	-56.58790	$-56.58790 - 59.729 = -115.32$
6.000000	35.65544	$35.65544 - 59.729 = -23.07$
7.000000	-7.953918	$-7.953918 - 59.729 = -66.68$
8.000000	1.370412	$1.370412 - 59.729 = -57.36$
9.000000	-28.54801	$-28.54801 - 59.729 = -87.28$
10.000000	52.18274	$52.18274 - 59.729 = -6.55$

**15.2.5b Testing the fixed effects**

Can we use the fixed effects output to test for equality of the fixed effects (dummy variable coefficients) like we did earlier using the dummy variable specification? The answer is yes. Go to **View/Fixed/Random Effects Testing/Redundant Fixed Effects – Likelihood Ratio**.



Two versions of the likelihood ratio test appear in the output, an  $F$ -test and a  $\chi^2$ -test. The  $F$ -test is identical to the one we considered earlier, and gives the same test results. The  $\chi^2$ -test has a

different origin, and so leads to a different test value. The details are beyond the level of our current description, but you can get a feel for where it comes from by checking equation (C.25) on page 537 of the text.

Redundant Fixed Effects Tests			
Equation: TABLE15_7			
Test cross-section fixed effects			
Effects Test	Statistic	d.f.	Prob.
Cross-section F	48.991522	(9,188)	0.0000
Cross-section Chi-square	241.513596	9	0.0000

### 15.3 THE NLS PANEL DATA

The data in the file *nls\_panel.wfl* is from the National Longitudinal Surveys conducted by the U.S. Department of Labor. This file is a large one that, in its current form, cannot be saved by the Student Version of EViews. We can nevertheless use the Student Version to analyze the data. After you have finished estimation, if you wish to save your results, you will need to reduce the range of the workfile structure and delete some of the series until the file is small enough to be saved by EViews Student Version. When saving it, name it differently, say *nls\_results.wfl*. You will then have two files, the original one with the data and another one with your results. This is an inconvenient state of affairs, but not an impossible scenario. The other alternative is to pay more for convenience and buy the EViews full version.

Opening the file reveals 3580 observations with a panel structure comprising 5 time series observations (1982, 1983, 1987, 1988) on 716 individuals.

Range: 1982 1988 x 716 -- 3580 obs
Sample: 1982 1988 -- 3580 obs

We can check the data against that in Table 15.8 by collecting those variables into a group and examining the following spreadsheet.

ID	YEAR	LWAGE	EDUC	COLLGRAD	BLACK	UNION	EXPER	TENURE
1.000000	82.00000	1.808289	12.00000	0.000000	1.000000	1.000000	7.666667	7.666667
1.000000	83.00000	1.863417	12.00000	0.000000	1.000000	1.000000	8.583333	8.583333
1.000000	85.00000	1.789367	12.00000	0.000000	1.000000	1.000000	10.17949	1.833333
1.000000	87.00000	1.846530	12.00000	0.000000	1.000000	1.000000	12.17949	3.750000
1.000000	88.00000	1.856449	12.00000	0.000000	1.000000	1.000000	13.62179	5.250000
2.000000	82.00000	1.280933	17.00000	1.000000	0.000000	0.000000	7.576923	2.416667
2.000000	83.00000	1.515855	17.00000	1.000000	0.000000	0.000000	8.384615	3.416667
2.000000	85.00000	1.930170	17.00000	1.000000	0.000000	0.000000	10.38461	5.416667
2.000000	87.00000	1.919034	17.00000	1.000000	0.000000	1.000000	12.03846	0.333333
2.000000	88.00000	2.200974	17.00000	1.000000	0.000000	1.000000	13.21154	1.750000
3.000000	82.00000	1.814825	12.00000	0.000000	0.000000	0.000000	11.41667	11.41667
3.000000	83.00000	1.919913	12.00000	0.000000	0.000000	1.000000	12.41667	12.41667
3.000000	85.00000	1.958377	12.00000	0.000000	0.000000	0.000000	14.41667	14.41667
3.000000	87.00000	2.007068	12.00000	0.000000	0.000000	0.000000	16.41667	16.41667
3.000000	88.00000	2.089854	12.00000	0.000000	0.000000	0.000000	17.82051	17.75000

### 15.3.1 Fixed effects estimation

The first model estimated using the NLS panel is a fixed effects model with  $\ln(WAGE)$  as the dependent variable and explanatory variables *EXPER*, *EXPER2*, *TENURE*, *TENURE2*, *SOUTH* and *UNION*. It is also suggested that we try a fixed effects model with *EDUC* and *BLACK* included to see what happens. Estimation should fail because *EDUC* and *BLACK* are constant over time for each individual. Their effects will be captured by the individual fixed effects. The **Equation specification** and **Effects specification** (selected from the Panel options) for this model are

Equation specification

Dependent variable followed by list of regressors including ARMA and PDL terms, OR an explicit equation like  $Y=c(1)+c(2)*X$ .

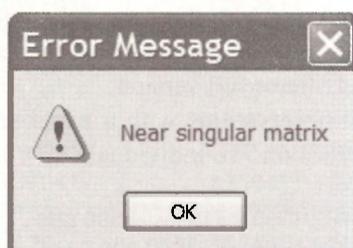
lwage c exper exper2 tenure tenure2 south union educ black

Effects specification

Cross-section: Fixed

Period: None

EViews' response is



This is a message that you will see if you try to estimate a model with perfect collinearity among the explanatory variables. In fact, EViews is being kind. The relevant matrix is singular not just “nearly” singular. We have not been specific about the matrix to which EViews refers. At this stage of your career it is sufficient to know that the singularity is caused by collinear explanatory variables.

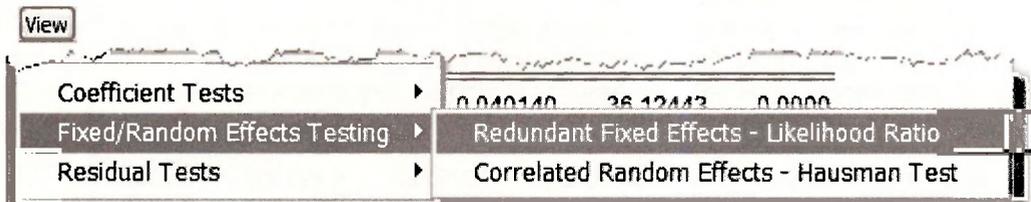
After dropping the offending variables *EDUC* and *BLACK*, the specification is

<p>Equation specification</p> <p>Dependent variable followed by list of regressors in and PDL terms, OR an explicit equation like <math>Y=c(1)</math></p> <p>lwage c exper exper2 tenure tenure2 south union</p>	<p>Effects specification</p> <p>Cross-section: Fixed</p> <p>Period: None</p>
--	--

The output follows. Note the correspondence with the results in Table 15.9 on page 397 of the text.

Dependent Variable: LWAGE				
Method: Panel Least Squares				
Sample: 1982 1988				
Periods included: 5				
Cross-sections included: 716				
Total panel (balanced) observations: 3580				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.450034	0.040140	36.12443	0.0000
EXPER	0.041083	0.006620	6.205904	0.0000
EXPER2	-0.000409	0.000273	-1.496532	0.1346
TENURE	0.013909	0.003278	4.243324	0.0000
TENURE2	-0.000896	0.000206	-4.353571	0.0000
SOUTH	-0.016322	0.036149	-0.451531	0.6516
UNION	0.063697	0.014254	4.468790	0.0000
Effects Specification				
Cross-section fixed (dummy variables)				

To test for the presence of individual differences we test the equality of the fixed effects as described in Section 15.2.5b. Go to **View/Fixed/Random Effects Testing/Redundant Fixed Effects – Likelihood Ratio**.



The resulting test details confirm the  $F$ -value of 19.66 reported on page 398 of the text.

Redundant Fixed Effects Tests			
Equation: TABLE15_9			
Test cross-section fixed effects			
Effects Test	Statistic	d.f.	Prob.
Cross-section F	19.658186	(715,2858)	0.0000

### 15.3.2 Random effects estimation

Individual effects that were modeled by fixed coefficients in the fixed effects model are treated as random draws from a larger population in the random effects model. For estimation purposes they become part of the error term. Also, estimation of the random effects model takes into account variation between individuals as well as variation within individuals. For our data set, this means it is possible to include *EDUC* and *BLACK* in the model. Doing so leads to the following **Equation** and **effects specifications**.

Equation specification
Dependent variable followed by list of regressors including and PDL terms, OR an explicit equation like $Y=c(1)+c(2)*X$ .
lwage c educ exper exper2 tenure tenure2 black south union

Effects specification	
Cross-section:	Random
Period:	None

The output that follows yields the results in Table 15.10 on page 402 of the text. In the lower part of the output, **cross section random** refers to the estimate  $\hat{\sigma}_u = 0.3291$  and **idiosyncratic random** refers to the estimate  $\hat{\sigma}_e = 0.1951$ . The values in the column **rho** are the proportions of total error variance attributable to each of the components. Thus,

$$\hat{\rho}_u = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{(0.3291)^2}{(0.3291)^2 + (0.1951)^2} = 0.7399$$

and

$$\hat{\rho}_e = 1 - \hat{\rho}_u = 0.2601$$

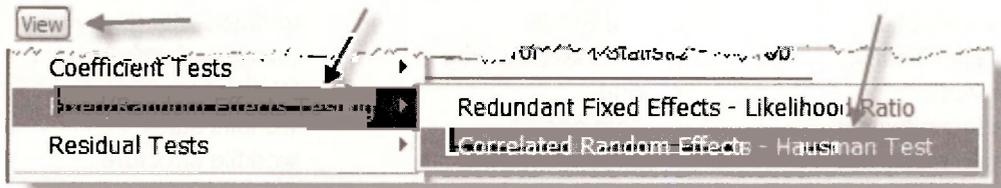
Dependent Variable: LWAGE				
Method: Panel EGLS (Cross-section random effects)				
Sample: 1982 1988				
Periods included: 5				
Cross-sections included: 716				
Total panel (balanced) observations: 3580				
Swamy and Arora estimator of component variances				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.533929	0.079722	6.697408	0.0000
EDUC	0.073254	0.005320	13.76943	0.0000
EXPER	0.043617	0.006345	6.874478	0.0000
EXPER2	-0.000561	0.000262	-2.140427	0.0324
TENURE	0.014154	0.003160	4.478901	0.0000
TENURE2	-0.000755	0.000194	-3.886833	0.0001
BLACK	-0.116737	0.030148	-3.872140	0.0001
SOUTH	-0.081812	0.022366	-3.657897	0.0003
UNION	0.080235	0.013187	6.084619	0.0000
Effects Specification			S.D.	Rho
Cross-section random			0.329050	0.7399
Idiosyncratic random			0.195110	0.2601

### 15.3.3 The Hausman test

The ability of the random effects model to take into account variation between individuals as well as variation within individuals makes it an attractive alternative to fixed effects estimation. However, for the random effects estimator to be unbiased in large samples the effects must be uncorrelated with the explanatory variables, an assumption that is often unrealistic. This

assumption can be tested using a Hausman test. The Hausman test is a test of the significance of the difference between the fixed effects estimates and the random effects estimates. Correlation between the random effects and the explanatory variables will cause these estimates to diverge; their difference will be significant. If the difference is not significant, there is no evidence of the offending correlation. The differences between the two sets of estimates can be tested separately using *t*-tests, or as a block using a  $\chi^2$ -test.

You can ask EViews to perform a Hausman test by opening the random-effects estimated equation and going to **View/Fixed/Random Effects Testing/Correlated Random Effects – Hausman Test**.



For the wage equation we get the following results. The value of the  $\chi^2_{(6)}$ -statistic for testing differences between all coefficients is  $\chi^2 = 20.437$ . Its corresponding *p*-value of 0.0023 suggests the null hypothesis of no correlation between the explanatory variables and the random effects should be rejected. The *p*-values for separate tests on the differences between each pair of coefficients are given in the column **Prob**. The results here are mixed. At a 5% significance level, the null hypothesis is rejected for *TENURE2*, *SOUTH* and *UNION*, but not for *EXPER*, *EXPER2* and *TENURE*. These results are slightly different to those reported on pages 205-206 of the text, but not enough to suggest anything is wrong. Differences may have occurred because of different covariance matrix estimators.

Correlated Random Effects - Hausman Test				
Equation: TABLE15_10				
Test cross-section random effects				
Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.	
Cross-section random	20.437076	6	0.0023	
<b><i>p-values for separate tests on each coefficient</i></b>				
Cross-section random effects test comparisons:				
Variable	Fixed	Random	Var(Diff.)	Prob.
EXPER	0.041083	0.043617	0.000004	0.1798
EXPER2	-0.000409	-0.000561	0.000000	0.0504
TENURE	0.013909	0.014154	0.000001	0.7782
TENURE2	-0.000896	-0.000755	0.000000	0.0380
SOUTH	-0.016322	-0.081812	0.000807	0.0211
UNION	0.063697	0.080235	0.000029	0.0022

## Keywords

- common coefficients
- contemporaneous correlation
- correlated random effects
- cross section random
- cross-equation restrictions
- cross-section coefficients
- cross-section identifiers
- cross-section SUR
- dim
- dummy variables
- effects specification
- fixed effects
- fixed effects: testing
- Hausman test
- identifier series
- idiosyncratic random
- logical generate
- new page
- NLS panel
- page destination
- panel options
- panel structure
- pool object
- pooled EGLS
- pooled least squares
- pooling
- random effects
- redundant fixed effects
- rename page
- reshape page
- residual correlation matrix
- seemingly unrelated regression
- singular matrix
- stack in new page
- stacked data
- stacking identifiers
- SUR
- undated with ID series
- unstacked data
- Wald test
- workfile stack
- workfile structure

# CHAPTER 16

## Qualitative and Limited Dependent Variables

### CHAPTER OUTLINE

- 16.1 Models with Binary Dependent Variables
  - 16.1.1 Examine the data
  - 16.1.2 The linear probability model
  - 16.1.3 The probit model
  - 16.1.4 Predicting probabilities
  - 16.1.5 Marginal effects in the probit model
- 16.2 Ordered Choice Models
  - 16.2.1 Ordered probit predictions
  - 16.2.2 Ordered probit marginal effects
- 16.3 Models for Count Data
  - 16.3.1 Examine the data
  - 16.3.2 Estimating a Poisson model
  - 16.3.3 Predicting with a Poisson model
  - 16.3.4 Poisson model marginal effects
- 16.4 Limited Dependent Variables
  - 16.4.1 Least squares estimation
  - 16.4.2 Tobit estimation and interpretation
  - 16.4.3 The Heckit selection bias model

### KEYWORDS

Microeconomics is a general theory of choice, and many of the choices that individuals and firms make cannot be measured by a continuous outcome variable. In this chapter we examine some fascinating models that are used to describe choice behavior, and which do not have the usual continuous dependent variable. Our descriptions will be brief, since we will not go into all the theory, but we will reveal to you a rich area of economic applications.

We also introduce a class of models with dependent variables that are *limited*. By that, we mean that they are continuous, but their range of values is constrained in some way and their values are not completely observable. Alternatives to least squares estimation must be considered for such cases, since the least squares estimator is both biased and inconsistent.

### 16.1 MODELS WITH BINARY DEPENDENT VARIABLES

We will illustrate **binary choice models** using an important problem from transportation economics. How can we explain an individual's choice between driving (private transportation) and taking the bus (public transportation) when commuting to work, assuming, for simplicity,

that these are the only two alternatives? We represent an individual's choice by the dummy variable

$$y = \begin{cases} 1 & \text{individual drives to work} \\ 0 & \text{individual takes bus to work} \end{cases}$$

If we collect a random sample of workers who commute to work, then the outcome  $y$  will be unknown to us until the sample is drawn. Thus,  $y$  is a random variable. If the probability that an individual drives to work is  $p$ , then  $P[y=1]=p$ . It follows that the probability that a person uses public transportation is  $P[y=0]=1-p$ . The probability function for such a binary random variable is

$$f(y) = p^y(1-p)^{1-y}, \quad y = 0, 1$$

where  $p$  is the probability that  $y$  takes the value 1. This discrete random variable has expected value  $E(y) = p$  and variance  $\text{var}(y) = p(1-p)$ .

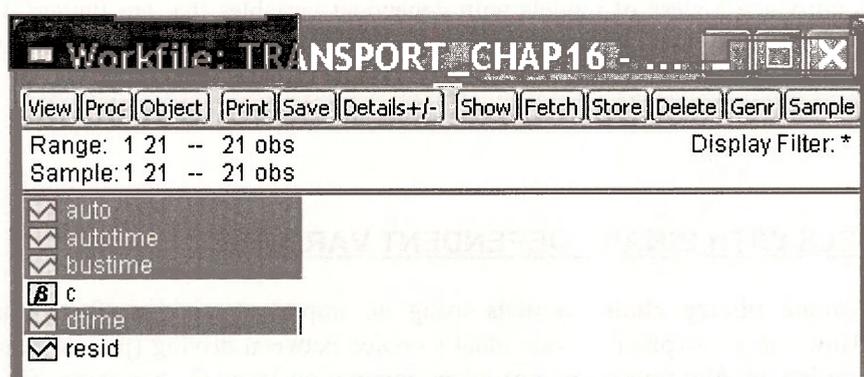
What factors might affect the probability that an individual chooses one transportation mode over the other? One factor will certainly be how long it takes to get to work one way or the other. Define the explanatory variable

$$x = (\text{commuting time by bus} - \text{commuting time by car})$$

There are other factors that affect the decision, but let us focus on this single explanatory variable. *A priori* we expect that as  $x$  increases, and commuting time by bus increases relative to commuting time by car, an individual would be more inclined to drive. That is, we expect a positive relationship between  $x$  and  $p$ , the probability that an individual will drive to work.

### 16.1.1 Examine the data

Open the workfile *transport.wfl*. Save the workfile with a new name to *transport\_chap16.wfl* so that the original workfile will not be changed. Highlight the series *AUTOTIME*, *BUSTIME*, *DTIME* and *AUTO* in order. Double-click in the blue to open the **Group**. The data are shown on the next page.



The key point is that *AUTO*, which is to be the dependent variable in the model, only takes the values 0 and 1.

obs	AUTOTIME	BUSTIME	DTIME	AUTO
1	52.90000	4.400000	-48.50000	0.000000
2	4.100000	28.50000	24.40000	0.000000
3	4.100000	86.90000	82.80000	1.000000
4	56.20000	31.60000	-24.60000	0.000000
5	51.80000	20.20000	-31.60000	0.000000

Obtain the descriptive statistics from the spreadsheet view: Select **View/Descriptive Stats/Common Sample**.

Group Members	BUSTIME	DTIME	AUTO
Spreadsheet	00000	-48.50000	0.000000
Dated Data Table	50000	24.40000	0.000000
Graph...	80000	82.80000	1.000000
	60000	-24.60000	0.000000
	00000	-31.60000	0.000000
			1.000000

The summary statistics will be useful later, but for now notice that the **SUM** of the *AUTO* series is 10, meaning that of the 21 individuals in the sample, 10 take their automobile to work and 11 take public transportation (the bus.)

Sample: 1 21

	AUTOTIME	BUSTIME	DTIME	AUTO
Mean	49.34762	48.12381	-1.223810	0.476190
Median	51.40000	38.00000	-7.000000	0.000000
Maximum	99.10000	91.50000	91.00000	1.000000
Minimum	0.200000	1.600000	-90.70000	0.000000
Std. Dev.	32.43491	34.63082	56.91037	0.511766
Sum	1036.300	1010.600	-25.70000	10.00000

### 16.1.2 The linear probability model

Our objective is to estimate a model explaining why some choose *AUTO* and some choose *BUS* transportation. Because the outcome variable is binary, its expected value is the probability of observing  $AUTO = 1$ ,

$$E(y) = p = \beta_1 + \beta_2 x$$

The model

$$y = E(y) + e = \beta_1 + \beta_2 x + e$$

is called the **linear probability model**. It looks like a regression, but as noted in *POE*, page 420, there are some problems. Nevertheless, apply least squares using  $y = AUTO$  and  $x = DTIME$ .

### Is auto c dtime

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.484795	0.071449	6.785151	0.0000
DTIME	0.007031	0.001286	5.466635	0.0000

The problems with this estimation procedure can be observed by examining the predicted values, which we call **PHAT**. In the regression window select the **Forecast** button

Forecast

Fill in the dialog box with a **Forecast name**.

An object **PHAT** appears in the workfile. Double-click to open. Examining just a few observations shows the unfortunate outcome that the linear probability model has predicted some probabilities to be greater than 1 or less than 0.

Series: PHAT Workfile: TRANSPORT	
PHAT	
Last updated: 12/08/07 - 10:23	
Modified: 1 21 # linear_prob.fit(=actual) phat	
1	0.143792
2	0.656351
3	1.066961
4	0.311833
5	0.262616
6	1.124615
7	0.851110
8	-0.131823

Now, examine the summary statistics for **PHAT** from the spreadsheet view, by selecting **View/Descriptive Statistics & Tests/Stats Table**.

Series: PHAT Workfile: TRANSPORT	
View Proc Object Properties Print Name Freeze Sample Genr St	
SpreadSheet	
Graph...	
Descriptive Statistics & Tests	Histogram and Stats
One-Way Tabulation...	Stats Table

Note that the average value of the predicted probability is .476, which is exactly equal to the fraction (10/21) of riders who choose *AUTO* in the sample. But also note that the minimum and maximum values are outside the feasible range.

Series: PHAT Wo	
View Proc Object Properties Print	
PHAT	
Mean	0.476190
Median	0.435578
Maximum	1.124615
Minimum	-0.152916

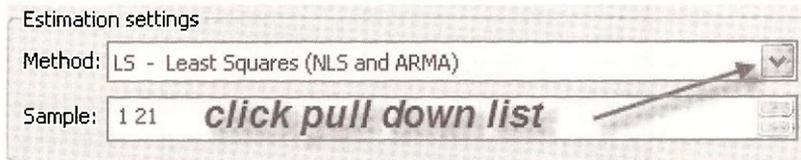
### 16.1.3 The probit model

The probit statistical model expresses the probability  $p$  that  $y$  takes the value 1 to be

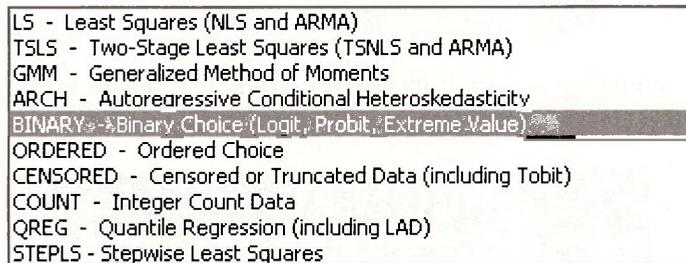
$$p = P[Z \leq \beta_1 + \beta_2 x] = \Phi(\beta_1 + \beta_2 x)$$

where  $\Phi(z)$  is the probit function, which is the standard normal cumulative distribution function (CDF). This is a **nonlinear model** because the parameters  $\beta_1$  and  $\beta_2$  are inside the very nonlinear function  $\Phi(\cdot)$ . Using numerical optimization procedures, that are outside the scope of this book, we can obtain **maximum likelihood** estimates. From the EViews menubar select **Quick/Estimate**

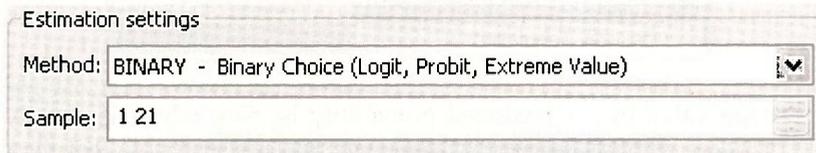
**Equation.** In the resulting dialog box, click the pull down list in the **Method** section of **Estimation settings**



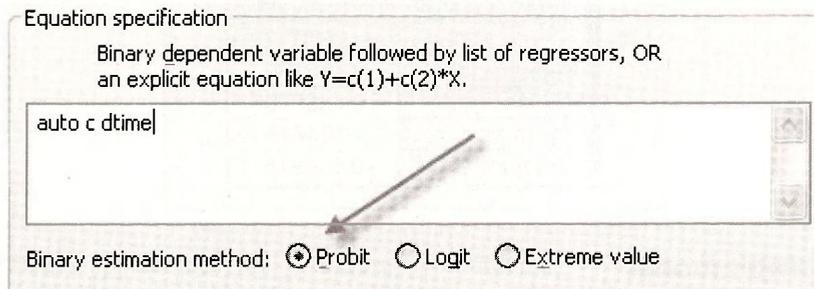
A long list of options appears. Choose **BINARY**.



The estimation settings should look like



In the **Equation specification** box enter the equation as usual, but select the radio button for **Probit**.



Click **OK**. The estimation results appear on the next page. In most ways the output looks similar to the regression output we have seen many times. The **Coefficients**, **Std. Error** and **Prob.** columns are familiar. There are many items included in the output you will not understand, and we are just omitting. However, we note the following:

- The **Method: ML** means that the model was estimated by maximum likelihood.
- The usual t-Statistic has been replaced by **z-Statistic**. The reason for this change is that the standard errors given are only valid in **large samples**. As we know the *t*-distribution

converges to the standard normal distribution in large samples, so using “z” rather than “t” recognizes this fact. The  $p$ -values **Prob.** are calculated using the  $N(0,1)$  distribution rather than the  $t$ -distribution.

Dependent Variable: AUTO  
Method: ML - Binary Probit  
Sample: 1 21  
Included observations: 21

	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.064434	0.399244	-0.161390	0.8718
DTIME	0.029999	0.010287	2.916279	0.0035
McFadden R-squared	0.575761	Mean dependent var		0.476190
LR statistic	16.73423			
Prob(LR statistic)	0.000043			
Obs with Dep=0	11	Total obs		21

- In the bottom portion of the output we see an  $R^2$  value called **McFadden R-squared**. This is not a typical  $R^2$  and cannot be interpreted like an  $R^2$ . As a child your mother pointed to a pan of boiling water on the stove and said **Hot! Don't touch!** We have a similar attitude about this value. We don't want you to “get burned,” so please disregard this number until you know much more.
- The **LR statistic** is comparable to the overall  $F$ -test of model significance in regression. It is a test statistic for the null hypothesis that all the model coefficients are zero except the intercept, against the alternative that at least one of the coefficients is not zero. The **LR** statistic has a chi-square distribution if the null hypothesis is true, with degrees of freedom equal to the number of explanatory variables, here 1. **Prob(LR statistic)** is the  $p$ -value for this test, and it is used in the standard way. If  $p \leq \alpha$  then we reject the null hypothesis at the  $\alpha$  level of significance.

### 16.1.4 Predicting probabilities

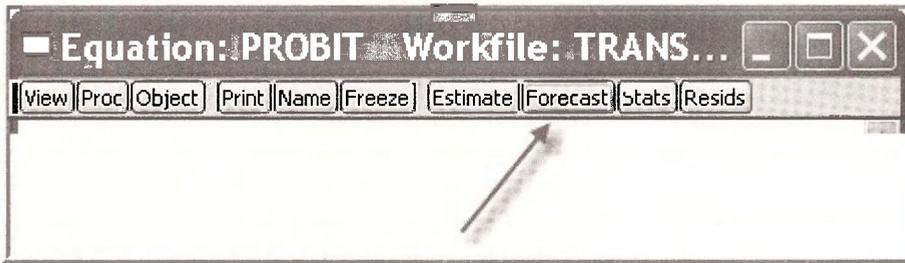
The “prediction” problem in probit is to predict the choice by an individual. We can predict the probability that individuals in the sample choose *AUTO*. In order to predict the probability that an individual chooses the alternative *AUTO* ( $y = 1$ ) we can use the probability model  $p = \Phi(\beta_1 + \beta_2 x)$  using estimates  $\tilde{\beta}_1 = -0.0644$  and  $\tilde{\beta}_2 = 0.02999$  of the unknown parameters obtained in the previous section.. Using these we estimate the probability  $p$  to be

$$\hat{p} = \Phi(\tilde{\beta}_1 + \tilde{\beta}_2 x)$$

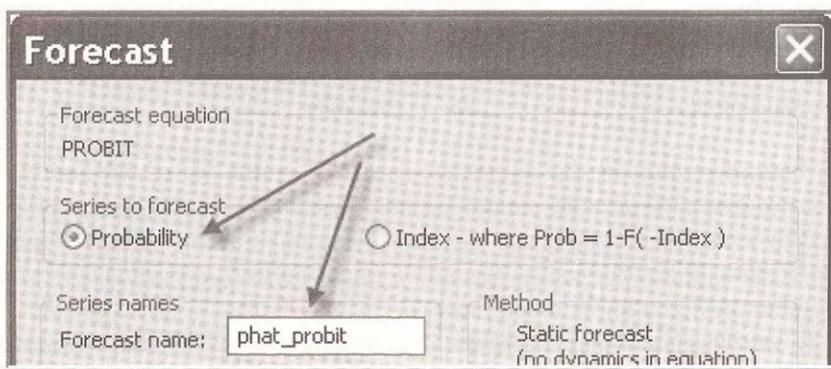
By comparing to a threshold value, like 0.5, we can predict choice using the rule

$$\hat{y} = \begin{cases} 1 & \hat{p} > 0.5 \\ 0 & \hat{p} \leq 0.5 \end{cases}$$

The predicted probabilities are easily obtained in EViews. Within the probit estimation window select **Forecast**.



In the resulting **Forecast** dialog box choose the **Series to forecast** to be the **Probability**, and assign the **Forecast name** **PHAT\_PROBIT**. Click **OK**.



Open the series **PHAT\_PROBIT** by double-clicking the series icon in the workfile window. The values of the predicted probabilities are given for each individual in the sample, based on their actual *DTIME*.

PHAT_PROBIT	
Last updated: 12/08/07 - 10:57	
Modified: 1 21 // probit.fit(f=actual,d) phat_probit	
1	0.064333
2	0.747787
3	0.992229
4	0.211158
5	0.155673
6	0.996156

It is useful to see that these predicted probabilities can be computed directly using the EViews function **@cnorm** which is the CDF of a standard normal random variable, what we have called “ $\Phi$ ”. EViews places the estimates of the unknown parameters in the coefficient vector  $C$ ,

$$\boxed{B} \ c$$

$C(1) = \tilde{\beta}_1 = -0.0644$  and  $C(2) = \tilde{\beta}_2 = 0.02999$ . We can create a series of predicted probabilities using the command

```
series phat_probit_calc = @cnorm(c(1)+c(2)*dtime)
```

The predicted probabilities from the two methods are the same

obs	DTIME	PHAT_PROBIT	PHAT_PROBIT_CALC
1	-48.50000	0.064333	0.064333
2	24.40000	0.747787	0.747787
3	82.80000	0.992229	0.992229
4	-24.60000	0.211158	0.211158
5	-31.60000	0.155673	0.155673

### 16.1.5 Marginal effects in the probit model

In this model we can examine the effect of a one unit change in  $x$  on the probability that  $y = 1$  by considering the derivative, which is often called **marginal effect** by economists.

$$\frac{dp}{dx} = \phi(\beta_1 + \beta_2 x) \beta_2$$

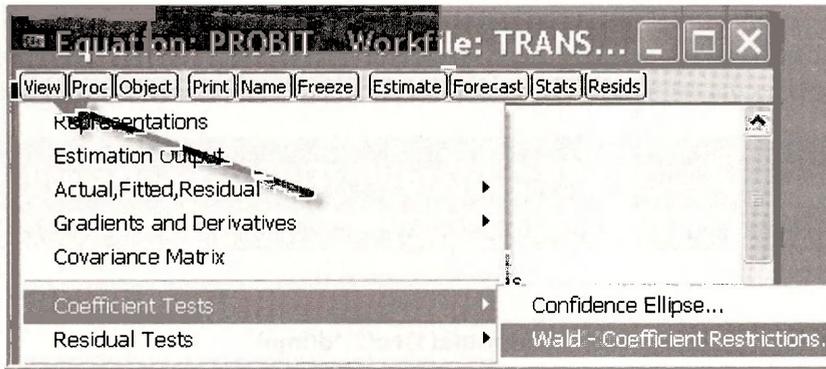
This quantity can be computed using the EViews function **@dnorm**, which gives the standard normal density function value, that we have represented by  $\phi$ . To generate the series of marginal effects for each individual in the sample, enter the command

```
series mfx_probit = @dnorm(c(1)+c(2)*dtime)*c(2)
```

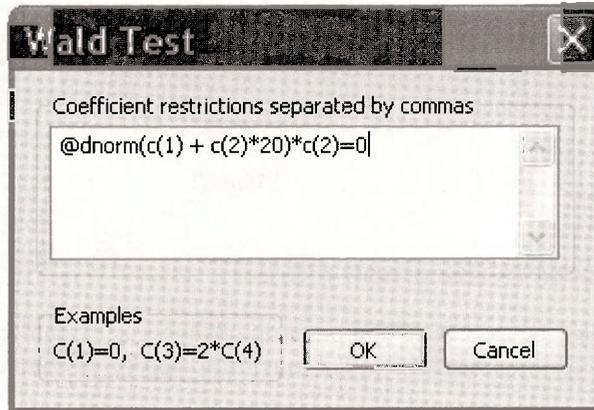
The marginal effect at a particular point uses the same calculation for a particular value of *DTIME*, such as 20.

```
scalar mfx_probit_20 = @dnorm(c(1)+c(2)*20)*c(2)
```

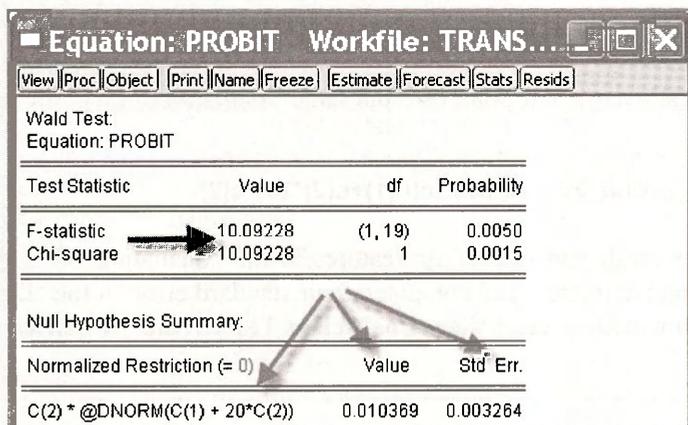
EViews is very powerful, and one of its features is the calculation of complicated nonlinear expressions involving parameters and computed their standard error by the “Delta” method. In the **PROBIT** estimation window, select **View/Coefficient Tests/Wald – Coefficient Restrictions**.



In the dialog window enter the expression for the marginal effect, assuming  $DTIME = 20$ , setting it equal to zero as if it were a hypothesis test.

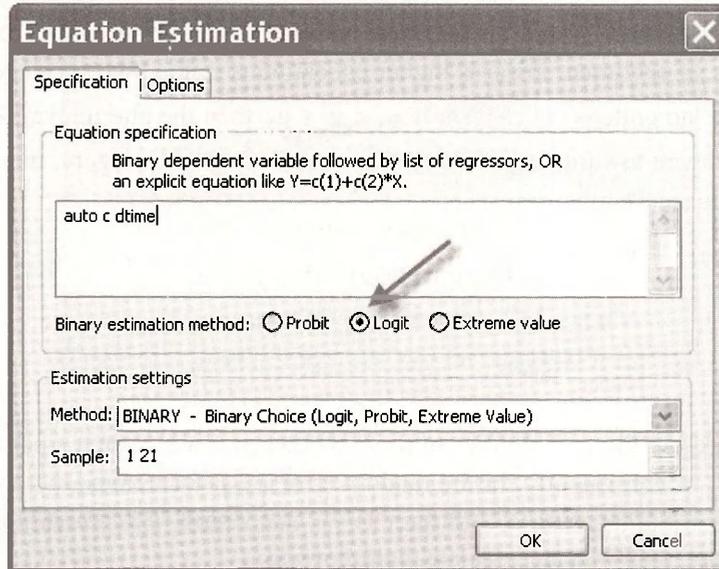


This returns the  $F$ -test statistic for the null hypothesis that the marginal effect is zero. The  $p$ -value is 0.005 leading us to reject the null hypothesis that additional *BUS* time has no effect on the probability of *AUTO* travel when  $DTIME = 20$ . Furthermore, the **Value** and the **Std. Err.** are computed. The value matches the scalar `MFX_PROBIT_20` computed earlier, and we now have a standard error that can be used to construct a confidence interval. Very cool.



`Scalar MFX_PROBIT_20 = 0.0103689956214`

To make the estimations using the **Logit** model simply change the **Equation Estimation** entries to



## 16.2 ORDERED CHOICE MODELS

In *POE* Chapter 16.3 we considered the problem of choosing what type of college to attend after graduating from high school as an illustration of a choice among unordered alternatives. However, in this particular case there may in fact be natural ordering. We might rank the possibilities as

$$y = \begin{cases} 3 & \text{4-year college (the full college experience)} \\ 2 & \text{2-year college (a partial college experience)} \\ 1 & \text{no college} \end{cases}$$

The usual linear regression model is not appropriate for such data, because in regression we would treat the  $y$  values as having some numerical meaning when they do not. When faced with a ranking problem, we develop a “sentiment” about how we feel concerning the alternative choices, and the higher the sentiment the more likely a higher ranked alternative will be chosen. This sentiment is, of course, unobservable to the econometrician. Unobservable variables that enter decisions are called **latent variables**, and we will denote our sentiment towards the ranked alternatives by  $y_i^*$ , with the “star” reminding us that this variable is unobserved.

As a concrete example, let us think about what factors might lead a high school graduate to choose among the alternatives “no college,” “2-year college” and “4-year college” as described by the ordered choices above. For simplicity, let us focus on the single explanatory variable *GRADES*. The model is then

$$y_i^* = \beta \times \text{GRADES}_i + e_i$$

This model is not a regression model because the dependent variable is unobservable. Consequently it is sometimes called an **index model**.

Because there are  $M = 3$  alternatives there are  $M - 1 = 2$  thresholds  $\mu_1$  and  $\mu_2$ , with  $\mu_1 < \mu_2$ . The index model does not contain an intercept because it would be exactly collinear with the threshold variables. If sentiment towards higher education is in the lowest category, then  $y^* \leq \mu_1$  and the alternative “no college” is chosen, if  $\mu_1 < y^* \leq \mu_2$  then the alternative “2-year college” is chosen, and if sentiment towards higher education is in the highest category, then  $y^* > \mu_2$  and “4-year college” is chosen. That is,

$$y = \begin{cases} 3 \text{ (4-year college)} & \text{if } y^* > \mu_2 \\ 2 \text{ (2-year college)} & \text{if } \mu_1 < y^* \leq \mu_2 \\ 1 \text{ (no college)} & \text{if } y^* \leq \mu_1 \end{cases}$$

We are able to represent the probabilities of these outcomes if we assume a particular probability distribution for  $y_i^*$ , or equivalently for the random error  $e_i$ . If we assume that the errors have the standard normal distribution,  $N(0,1)$ , and the CDF is denoted  $\Phi$ , an assumption that defines the ordered probit model, then we can calculate the following:

$$P[y = 1] = \Phi(\mu_1 - \beta \text{GRADES}_i)$$

$$P[y = 2] = \Phi(\mu_2 - \beta \text{GRADES}_i) - \Phi(\mu_1 - \beta \text{GRADES}_i)$$

and the probability that  $y = 3$  is

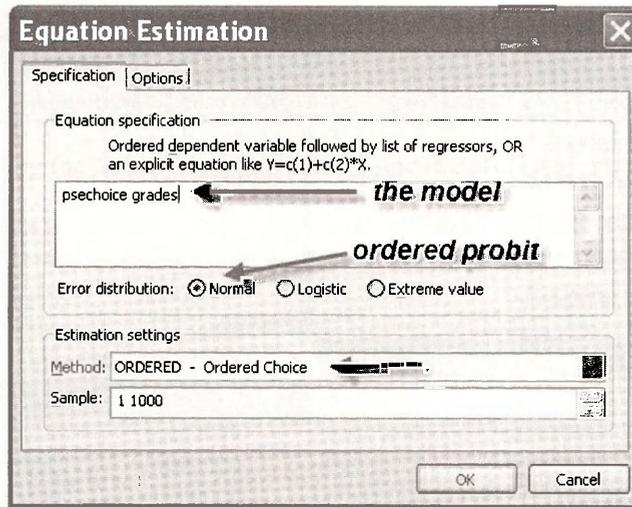
$$P[y = 3] = 1 - \Phi(\mu_2 - \beta \text{GRADES}_i)$$

In this model we wish to estimate the parameter  $\beta$ , and the two threshold values  $\mu_1$  and  $\mu_2$ . These parameters are estimated by maximum likelihood.

In EViews open the workfile *nels\_small.wfl*. Save it under the name *nels\_small\_oprobit.wfl*. The dependent variable of interest is *PSECHOICE* and the explanatory variable is *GRADES*. Select **Quick/Estimate Equation**. In the drop down menu of estimation methods choose **Ordered Choice**.

LS - Least Squares (NLS and ARMA)
TOLS - Two-Stage Least Squares (TSNLS and ARMA)
GMM - Generalized Method of Moments
ARCH - Autoregressive Conditional Heteroskedasticity
BINARY - Binary Choice (Logit, Probit, Extreme Value)
<input checked="" type="radio"/> ORDERED - Ordered Choice
CENSORED - Censored or Truncated Data (including Tobit)
COUNT - Integer Count Data
QREG - Quantile Regression (including LAD)
STEPLS - Stepwise Least Squares

Enter the estimation equation with NO INTERCEPT. Make sure the **Normal** radio button is selected so that the model is **Ordered Probit**.



The results, edited to remove things that are not of interest, are

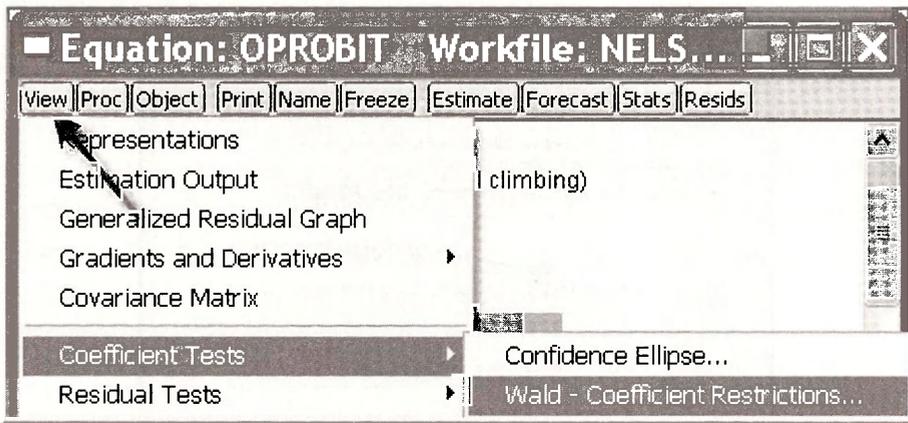
Dependent Variable: PSECHOICE  
 Method: ML - Ordered Probit  
 Sample: 1 1000  
 Number of ordered indicator values: 3

	Coefficient	Std. Error	z-Statistic	Prob.
GRADES	-0.306625	0.019173	-15.99217	0.0000
Limit Points				
LIMIT_2:C(2)	-2.945600	0.146828	-20.06154	0.0000
LIMIT_3:C(3)	-2.089993	0.135768	-15.39385	0.0000

The coefficient of *GRADES* is the maximum likelihood estimate  $\tilde{\beta}$ . The values labeled **LIMIT\_2:C(2)** and **LIMIT\_3:C(3)** are the maximum likelihood estimates of  $\mu_1$  and  $\mu_2$ . The notation points out that these parameter estimates are saved into the coefficient vector as C(2) and C(3). C(1) contains  $\tilde{\beta}$ . Name this equation **OPROBIT**.

### 16.2.1 Ordered probit predictions

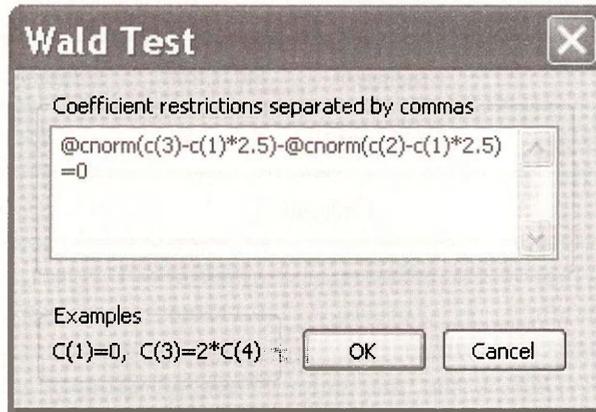
To predict the probabilities of various outcomes, as shown on page 436 of *POE*, we can again use the computing abilities of EViews. In the **OPROBIT** estimation window select **View/Coefficient Tests/Wald – Coefficient Restrictions**.



To compute the probability that a student with  $GRADES = 2.5$  will attend a 2-year college we calculate

$$P[y = 2 | GRADES = 2.5] = \Phi(\tilde{\mu}_2 - \tilde{\beta} \times 2.5) - \Phi(\tilde{\mu}_1 - \tilde{\beta} \times 2.5)$$

Enter into the **Wald Test** dialog box



The predicted probability is the relatively low 0.078, which makes sense because  $GRADES = 2.5$  is very high on the 13 points scale..

Normalized Restriction (= 0)	Value	Std. Err.
-@CNORM(-2.5*C(1) + C(2)) + @CNORM(-2.5*C(1) + C(3))	0.078182	0.011972

We can use the same general approach to compute the probabilities for each option for all the individuals in the sample. Recall that the maximum likelihood estimates of  $\mu_1$  and  $\mu_2$  are saved into the coefficient vector as C(2) and C(3). C(1) contains  $\tilde{\beta}$ .

```

series phat_y1 = @cnorm(c(2) - c(1)*grades)
series phat_y2 = @cnorm(c(3) - c(1)*grades) - @cnorm(c(2) - c(1)*grades)
series phat_y3 = 1 - phat_y1 - phat_y2

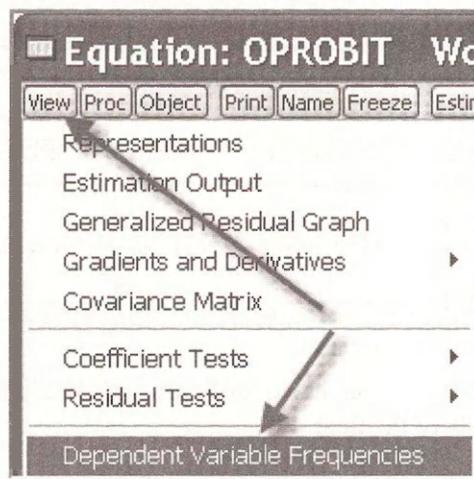
```

Open a **Group** showing the *GRADES*, *PSECHOICE* and the predicted probabilities.

obs	GRADES	PSECHOICE	PHAT_Y1	PHAT_Y2	PHAT_Y3
1	9.080000	2.000000	0.435872	0.320338	0.243790
2	8.310000	2.000000	0.345483	0.331063	0.323454
3	7.420000	3.000000	0.251288	0.322162	0.426550
4	7.420000	3.000000	0.251288	0.322162	0.426550
5	7.420000	3.000000	0.251288	0.322162	0.426550
6	7.460000	3.000000	0.255213	0.323042	0.421745
7	9.670000	2.000000	0.507765	0.301467	0.190767
8	11.770000	1.000000	0.746456	0.189161	0.064383
9	8.810000	3.000000	0.403526	0.325999	0.270476
10	6.440000	3.000000	0.165791	0.288302	0.545907

A standard procedure is to predict the actual choice using the highest probability. Thus we would predict that person 1 would attend no college, and the same with person 2. Both of these predictions are in fact incorrect because they choose a 2-year college. Individual 3 we predict will attend a 4-year college, and they did.

In the EViews window containing the estimated model **OPROBIT**, select **View/Dependent Variable Frequencies**

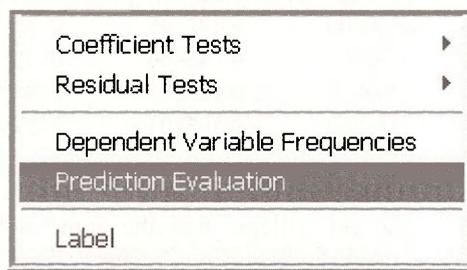


We see the choices made in the data

Dependent Variable Frequencies  
Equation: OPROBIT

Dep. Value	Count	Percent	Cumulative	
			Count	Percent
1	222	22.00	222	22.20
2	251	25.00	473	47.30
3	527	52.00	1000	100.00

Now select **View/Prediction Evaluation**.



Using the “highest probability” prediction rule, EViews calculates

Prediction Evaluation for Ordered Specification  
Equation: OPROBIT

Estimated Equation					
Dep. Value	Obs.	Correct	Incorrect	% Correct	% Incorrect
1	222	116	106	52.252	47.748
2	251	0	251	0.000	100.000
3	527	471	56	89.374	10.626
Total	1000	587	413	58.700	41.300

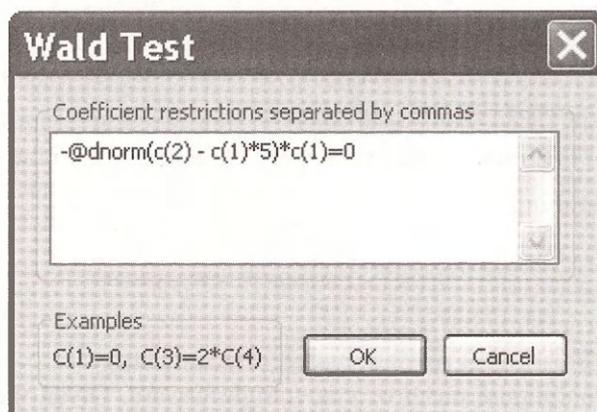
This model, being a very simple one, has a difficult time predicting who will attend 2-year colleges, being incorrect 100% of the time.

### 16.2.2 Ordered probit marginal effects

The marginal effects in the ordered probit model measure the changed in probability of choosing a particular category given a 1-unit change in an explanatory variable. The calculations are different by category. The calculations involve the standard normal probability density function, denoted  $\phi$  and calculated in EViews by **@dnorm**. For example the marginal effect of *GRADES* on the probability that a student attends no college is

$$\frac{\partial P[y=1]}{\partial \text{GRADES}} = -\phi(\mu_1 - \beta \text{GRADES}) \times \beta$$

In the **OPROBIT** window select **View/Coefficient Tests/Wald – Coefficient Restrictions**. In the dialog box enter



Recalling that a higher value of *GRADES* is a poorer academic performance, we see that the probability of attending no college increases by 0.045 for a student with *GRADES* = 5.

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
-C(1) * @DNORM(-5*C(1) + C(2))	0.045112	0.003058

The marginal effect calculation can be carried out for each person in the sample using the command

```
series mfx_y1 = - @dnorm(c(2) - c(1)*grades)*c(1)
```

Open a **Group** showing *GRADES* and this marginal effect. Note that increasing *GRADES* by 1-point (worse grades) increases the probabilities of attending no college, but for students with better grades (*GRADES* lower) the effect is smaller.

obs	GRADES	MFX_Y1
1	9.080000	0.120742
2	8.310000	0.113032
3	7.420000	0.097704
4	7.420000	0.097704
5	7.420000	0.097704
6	7.460000	0.098503
7	9.670000	0.122303
8	11.77000	0.098165

### 16.3 MODELS FOR COUNT DATA

If  $Y$  is a Poisson random variable, then its probability function is

$$f(y) = P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, 2, \dots$$

The factorial (!) term  $y! = y \times (y-1) \times (y-2) \times \dots \times 1$ . This probability function has one parameter,  $\lambda$ , which is the mean (and variance) of  $Y$ . In a regression model we try to explain the behavior of  $E(Y)$  as a function of some explanatory variables. We do the same here, keeping the value of  $E(Y) \geq 0$  by defining

$$E(Y) = \lambda = \exp(\beta_1 + \beta_2 x)$$

This choice defines the **Poisson regression model** for count data.

Prediction of the conditional mean of  $y$  is straightforward. Given the maximum likelihood estimates  $\tilde{\beta}_1$  and  $\tilde{\beta}_2$ , and given a value of the explanatory variable  $x_0$ , then

$$\widehat{E(y_0)} = \tilde{\lambda}_0 = \exp(\tilde{\beta}_1 + \tilde{\beta}_2 x_0)$$

This value is an estimate of the expected number of occurrences observed, if  $x$  takes the value  $x_0$ . The probability of a particular number of occurrences can be estimated by inserting the estimated conditional mean into the probability function, as

$$\widehat{\Pr(Y = y)} = \frac{\exp(-\tilde{\lambda}_0) \tilde{\lambda}_0^y}{y!}, \quad y = 0, 1, 2, \dots$$

The marginal effect of a change in a continuous variable  $x$  in the Poisson regression model is not simply given by the parameter, because the conditional mean model is a nonlinear function of the parameters. Using our specification that the conditional mean is given by

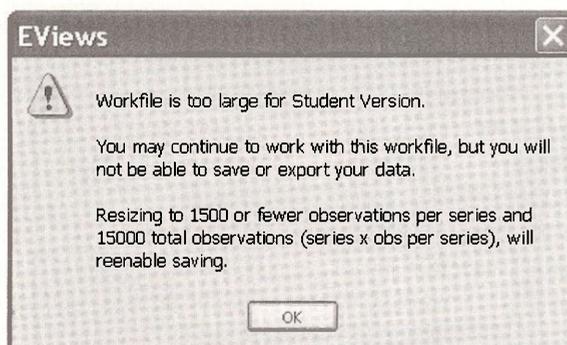
$$E(y_i) = \lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

and using rules for derivatives of exponential functions, we obtain the marginal effect

$$\frac{\partial E(y_i)}{\partial x_i} = \lambda_i \beta_2$$

To estimate this marginal effect, replace the parameters by their maximum likelihood estimates, and select a value for  $x$ . The marginal effect is different depending on the value of  $x$  chosen.

To illustrate open the workfile *olympics.wfl*. You will find a very rude message.



This workfile is too large because there are too many observations. The definition file *olympics.def* shows that there are 1610 observations.

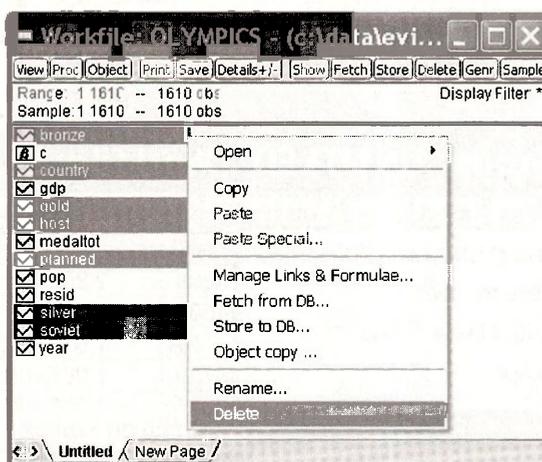
```
olympics.def

country year gdp pop gold silver bronze medaltot host planned soviet

Obs: 1610

country      country code
year         olympics year
```

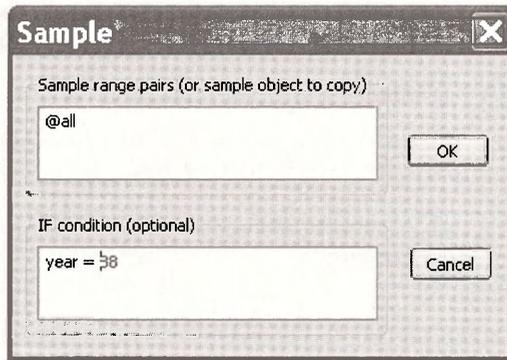
We can still operate with the workfile, but we cannot save it even if we delete some variables. Give this a try, deleting the variables that are needed in this example.



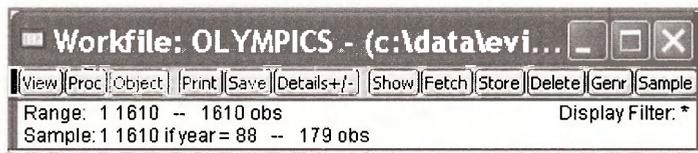
The example in the book uses only data from 1988. To modify the sample, click the **Sample** button on the EViews main menu.



In the **Sample** dialog box add the **IF condition**



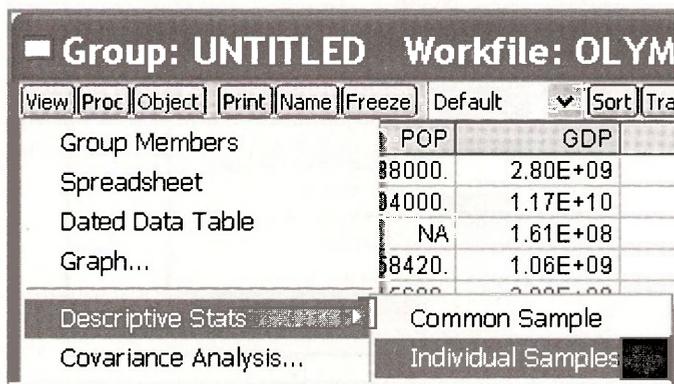
The workfile window now shows that the estimation sample is 179 observations from the year 1988.



Despite these changes the file still cannot be saved with the Student Version of EViews 6. Your options are to switch to the full version, or to make sure you print out all intermediate results as you go along.

### 16.3.1 Examine the data

Open a group consisting of *MEDALTOT*, *POP* and *GDP*. Obtain summary statistics for the individual samples



Finding the summary statistics for individual samples is important when some observations are missing, or NA.

Note that there are 152 observations for *MEDALTOT*, 176 for *POP* and 179 for *GDP*.

	MEDALTOT	POP	GDP
Mean	4.855263	28866147	1.38E+11
Median	0.000000	5921270.	5.51E+09
Maximum	132.0000	1.10E+09	6.07E+12
Minimum	0.000000	20000.00	41700000
Std. Dev.	16.57630	1.08E+08	5.92E+11
Skewness	5.543308	8.023542	7.908054
Kurtosis	36.84647	72.99436	71.93663
Jarque-Bera	8033.810	37815.94	37309.63
Probability	0.000000	0.000000	0.000000
Sum	738.0000	5.08E+09	2.46E+13
Sum Sq. Dev.	41490.82	2.03E+18	6.23E+25
Observations	152	176	179

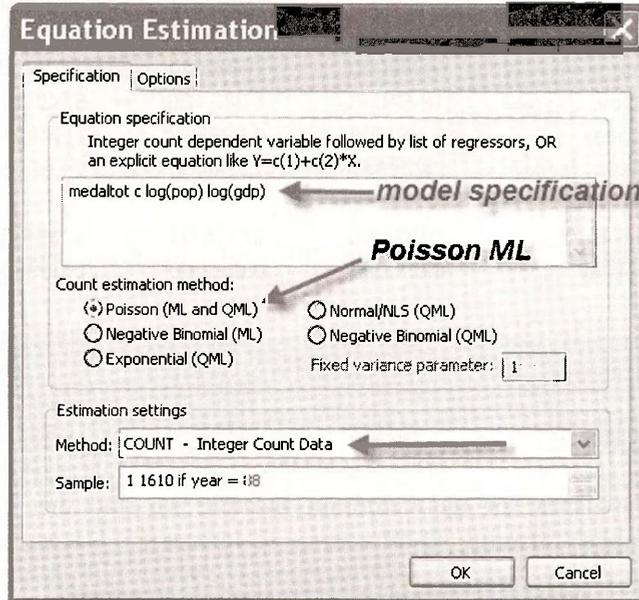
Obtaining summary statistics for the **Common Sample** we find that 151 observations are available on all 3 variables.

Object	POP	GDP
Group Members	38000.	2.80E+09
Spreadsheet	34000.	1.17E+10
Dated Data Table	NA	1.61E+08
Graph...	38420.	1.06E+09
Descriptive Stats	Common Sample	

	MEDALTOT	POP	GDP
Mean	4.887417	32337758	1.62E+11
Median	0.000000	6812400.	8.13E+09
Maximum	132.0000	1.10E+09	6.07E+12
Minimum	0.000000	20000.00	59700000
Std. Dev.	16.62670	1.16E+08	6.42E+11
Skewness	5.524132	7.437776	7.253078
Kurtosis	36.60302	62.78908	60.71151
Jarque-Bera	7872.302	23883.27	22279.09
Probability	0.000000	0.000000	0.000000
Sum	738.0000	4.88E+09	2.44E+13
Sum Sq. Dev.	41467.09	2.01E+18	6.17E+25
Observations	151	151	151

### 16.3.2 Estimating a Poisson model

To estimate the model by maximum likelihood choose **Quick/Estimate Equation**. In the dialog box make the choices shown below.



The estimated model is

Dependent Variable: MEDALTOT  
 Method: ML/QML = Poisson Count  
 Sample: 1 1610 IF YEAR = 88  
 Included observations: 151

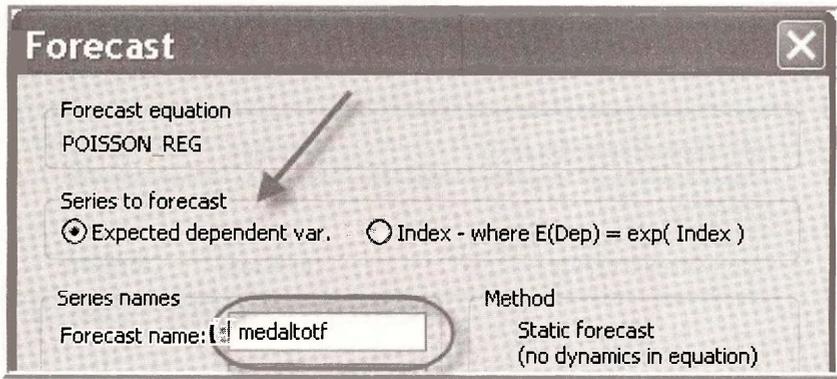
	Coefficient	Std. Error	z-Statistic	Prob.
C	-15.88746	0.511805	-31.04203	0.0000
LOG(POP)	0.180038	0.032280	5.577348	0.0000
LOG(GDP)	0.576603	0.024722	23.32376	0.0000

Note that the number of observations used in the estimation is only 151, which is the number of observations common to all variables.

Despite the fact that the workfile cannot be saved, we save these estimation results as an object named **POISSON\_REG**.

### 16.3.3 Predicting with a Poisson model

In the estimation window click **Forecast**. Choose the **Series to forecast** as **Expected dependent var.** and assign a name



Recall that the expected value of the dependent variable, in a simple model, is given by

$$E(y_i) = \lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

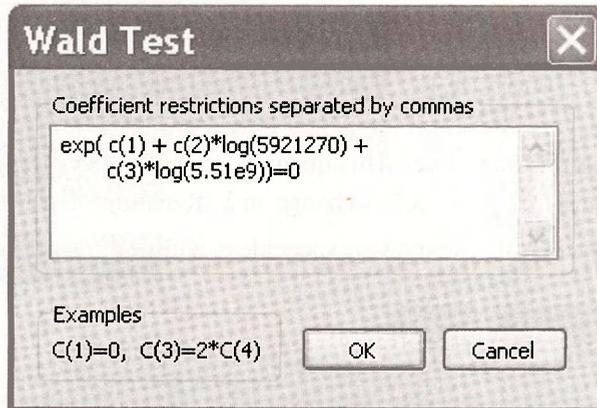
The forecast can be replicated using the following command

**series lam = exp(c(1) + c(2)\*log(pop) + c(3)\*log(gdp))**

obs	POP	GDP	MEDALTOT	MEDALTOTF	LAM
5	3138000.	2.80E+09	NA	0.521277	0.521277
20	7004000.	1.17E+10	NA	1.373831	1.373831
69	NA	1.61E+08	0.000000	NA	NA
89	5158420.	1.06E+09	NA	0.325605	0.325605
91	8115690.	2.09E+09	NA	0.522552	0.522552
99	327970.0	3.33E+08	NA	0.101693	0.101693

We have shown a few values.

To compute the predicted mean for specific values of the explanatory variables we again use the trick of applying the “Wald test.” Select **View/Coefficient Tests/Wald – Coefficient Restrictions**. We must choose some values for *POP* and *GDP* at which to evaluate the prediction. Enter the **median values from the individual samples** for *POP* and *GDP*.



The result shows that for these population and *GDP* values we predict that 0.8634 medals will be won.

Normalized Restriction (= 0)	Value	Std. Err.
EXP(C(1) + 22.429830460111234*C(3) + C(2)*LOG(5921270))	0.863443	0.075976

### 16.3.4 Poisson model marginal effects

As shown in *POE* equation (16.29) the marginal effects in the simple shown in the simple Poisson model

$$E(y_i) = \lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

are

$$\frac{\partial E(y_i)}{\partial x_i} = \exp(\beta_1 + \beta_2 x_i) \beta_2 = \lambda_i \beta_2$$

This marginal effect is correct if the values of the explanatory variable  $x$  is not transformed. In the Olympics medal example the explanatory variables are in logarithms, so the model is

$$E(y_i) = \lambda_i = \exp(\beta_1 + \beta_2 \ln(x_i))$$

and the marginal effect is, using the chain rule of differentiation,

$$\frac{\partial E(y_i)}{\partial x_i} = \exp(\beta_1 + \beta_2 \ln(x_i)) \frac{\beta_2}{x_i} = \lambda_i \frac{\beta_2}{x_i}$$

While this does not necessarily look very pretty, it has a rather nice interpretation. Rearrange it as

$$\frac{\partial E(y_i)}{100(\partial x_i/x_i)} = \exp(\beta_1 + \beta_2 \ln(x_i)) \frac{\beta_2}{100} = \lambda_i \frac{\beta_2}{100}$$

Are you still not finding this attractive? This quantity can be called a **semi-elasticity**, because it expresses the change in  $E(y)$  given a 1% change in  $x$ . Recalling that  $E(y_i) = \lambda_i$  we can make one further enhancement that will leave you speechless with joy. Divide both sides by  $E(y)$  to obtain

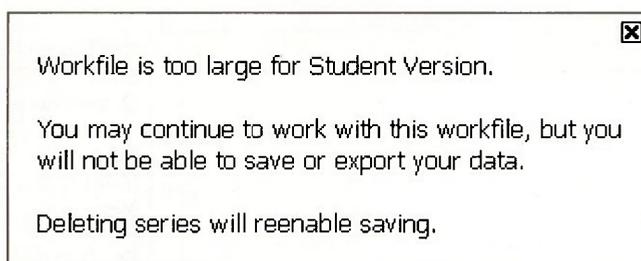
$$\frac{\partial E(y_i)/E(y_i)}{(\partial x_i/x_i)} = \beta_2 = \varepsilon$$

The parameter  $\beta_2$  is the **elasticity** of the output  $y$  with respect to  $x$ . A 1% change in  $x$  is estimated to change  $E(y)$  by  $\beta_2\%$ .

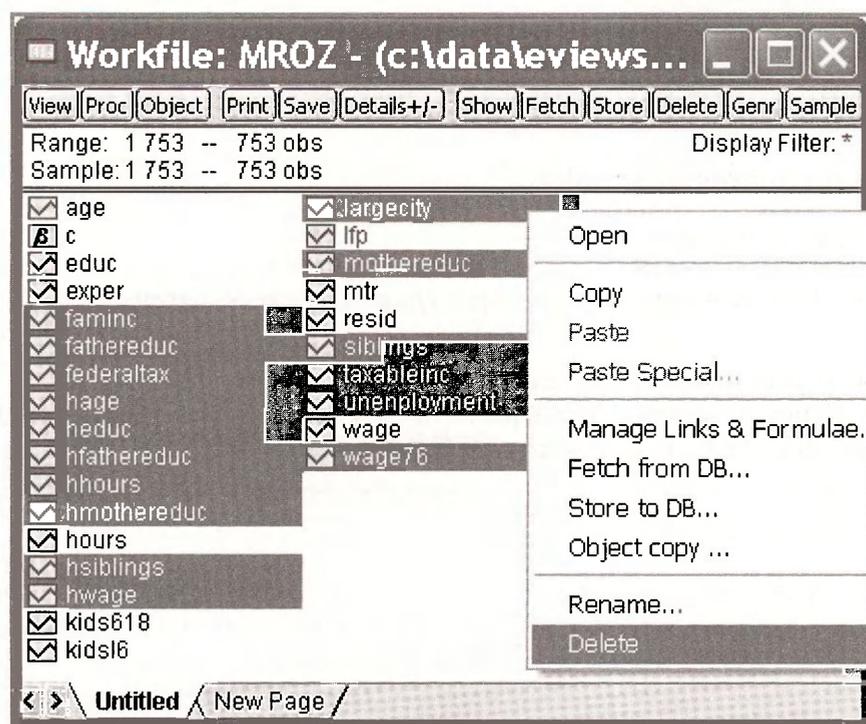
In the Olympics example, based on the estimation results, we conclude that a 1% increase population increases the expected medal count by 0.18%, and a 1% in GDP increases expected medal count by 0.5766%.

## 16.4 LIMITED DEPENDENT VARIABLES

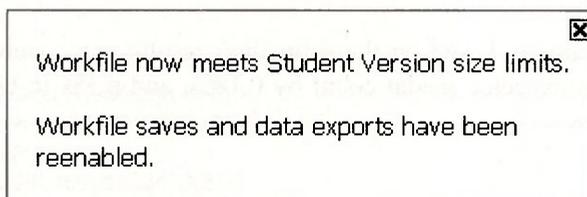
The idea of censored data is well illustrated by the Mroz data on labor force participation of married women. Open the workfile *mroz.wfl*. You will receive an unpleasant warning when using the Student Version of EViews 6:



However, this problem can be fixed by deleting some variables. Delete the variables indicated below.

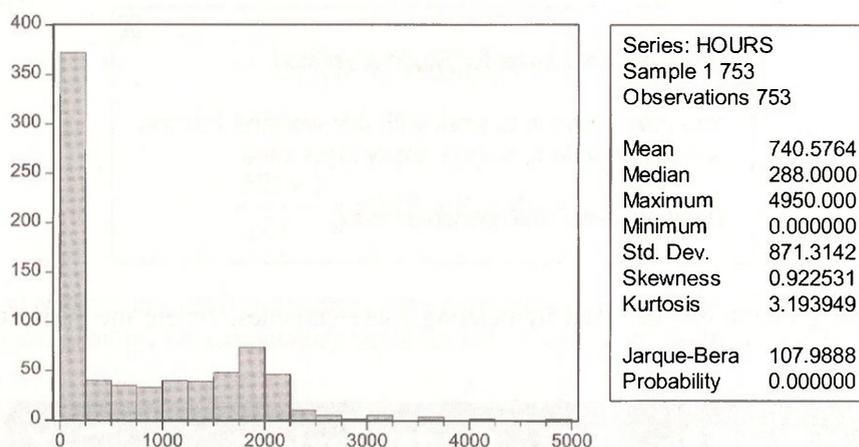


EViews now tells us we are OK, can save the workfile



Save the workfile as *mroz\_tobit.wfl* to as to keep the original file intact.

A **Histogram** of the variable *HOURS* shows the problem with the full sample. There are 753 observations on the wages of married women but 325 of these women did not engage in market work, and thus their *HOURS* = 0, leaving 428 observations with positive *HOURS*.



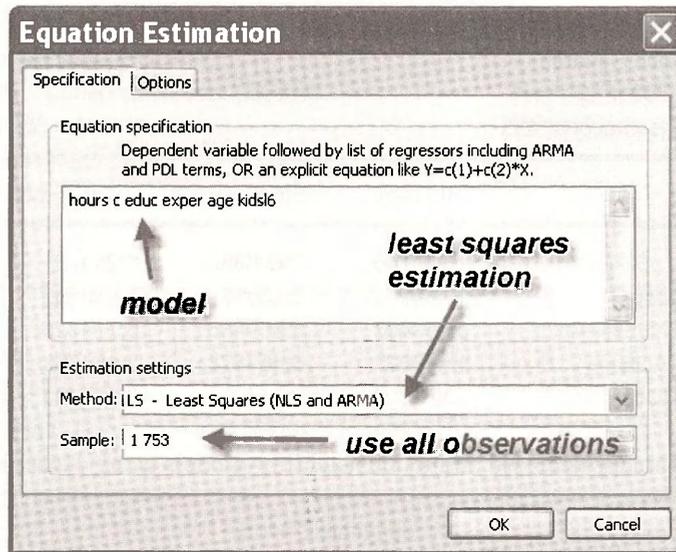
### 16.4.1 Least squares estimation

We are interested in the equation

$$HOURS = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 AGE + \beta_5 KIDSL6 + e$$

The question is “How shall we treat the observations with *HOURS* = 0”?

A first solution is to apply least squares to all the observations. Select **Quick/Estimate Equation** and fill in the **Equation Estimation** dialog box as follows:

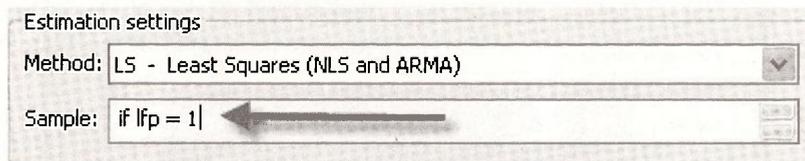


The estimation results are

Dependent Variable: HOURS  
 Method: Least Squares  
 Sample: 1 753  
 Included observations: 753

	Coefficient	Std. Error	t-Statistic	Prob.
C	1335.306	235.6487	5.666511	0.0000
EDUC	27.08568	12.23989	2.212902	0.0272
EXPER	48.03981	3.641804	13.19121	0.0000
AGE	-31.30782	3.960990	-7.904040	0.0000
KIDSL6	-447.8547	58.41252	-7.667100	0.0000

Repeat the estimation using only those women who “participated in the labor force.” Those women who worked are indicated by a dummy variable *LFP* which is 1 for working women, but zero otherwise.



The estimation results are shown below. Note that the included observations are 428. The estimation results now show the effect of education (*EDUC*) to have a negative, but insignificant, affect on *HOURS*. In the estimation using all the observations *EDUC* had a positive and significant effect on *HOURS*.

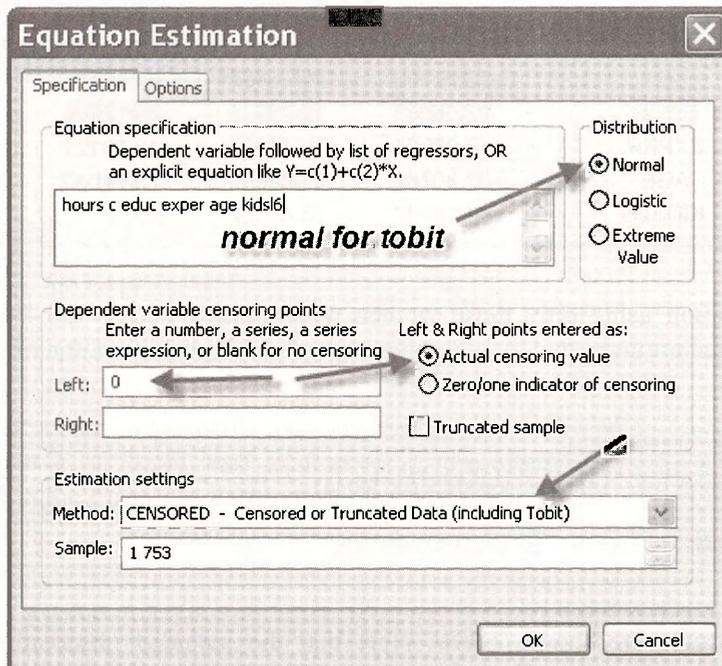
Dependent Variable: HOURS  
 Method: Least Squares  
 Sample: 1 753 IF LFP = 1  
 Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	1829.746	292.5356	6.254781	0.0000
EDUC	-16.46211	15.58083	-1.056562	0.2913
EXPER	33.93637	5.009185	6.774829	0.0000
AGE	-17.10821	5.457674	-3.134708	0.0018
KIDSL6	-305.3090	96.44904	-3.165495	0.0017

The least squares estimator is biased and inconsistent for models using censored data.

### 16.4.2 Tobit estimation and interpretation

An appropriate estimation procedure is **Tobit**, which uses maximum likelihood principles. Select **Quick/Estimate Equation**. In the **Equation Estimation** window fill in the options as shown below. Tobit estimation is predicated upon the regression errors being **Normal**, so tick that radio button. In our cases the observations that are “censored” take the actual value 0, and the dependent variable is said to be **Left censored** because 0 is a minimum value and all relevant values of *HOURS* are positive. The **Estimation settings** show the method to include **Tobit**.



Dependent Variable: HOURS  
 Method: ML - Censored Normal (TOBIT)  
 Sample: 1 753  
 Included observations: 753  
 Left censoring (value) at zero

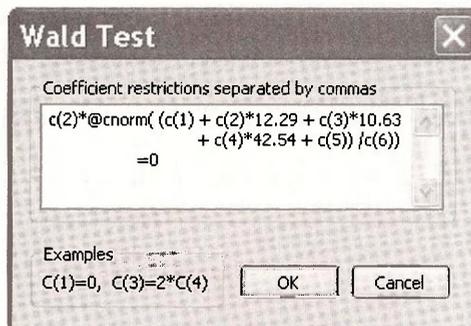
	Coefficient	Std. Error	z-Statistic	Prob.
C	1349.876	386.2991	3.494381	0.0005
EDUC	73.29099	20.47459	3.579607	0.0003
EXPER	80.53527	6.287808	12.80816	0.0000
AGE	-60.76780	6.888194	-8.822022	0.0000
KIDSL6	-918.9181	111.6607	-8.229557	0.0000
Error Distribution				
SCALE:C(6)	1133.697	42.06239	26.95274	0.0000
Left censored obs	325	Right censored obs	0	
Uncensored obs	428	Total obs	753	

The estimation output shows the usual **Coefficient** and **Std. Error** columns. Instead of a **t-Statistic** EViews reports a **z-Statistic** because the standard errors are only valid in large samples, making the test statistic only valid in large samples, and in large samples a *t*-statistic converges to the standard normal distribution. The *p*-value **Prob.** is based on the standard normal distribution.

The parameter called **SCALE:C(6)** is the estimate of  $\sigma$ , the square root of the error variance. This value is an important ingredient in Tobit model interpretation. As noted in *POE*, equation (16.35), the marginal effect of an explanatory variable in a simple model, is

$$\frac{\partial E(y|x)}{\partial x} = \beta_2 \Phi\left(\frac{\beta_1 + \beta_2 x}{\sigma}\right)$$

where as usual  $\Phi$  is the CDF of a standard normal variable. To evaluate the marginal effect of *EDUC* on *HOURS*, given that *HOURS* > 0, we can use **Wald** test dialog box. Select **View/Coefficient Tests/Wald – Coefficient Restrictions**. Enter in the expression for the marginal effect of *EDUC* at the sample means, as shown on page 447 of *POE*.



The obtained value is slightly different than the value in the text. Slight differences in results are inevitable when carrying out complicated nonlinear estimations and calculations. The maximum likelihood routines are all slightly different, and stop when “convergence” is achieved. These stopping rules are different from one software package to another.\*

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
$C(2) * @CNORM((C(1) + 12.29*C(2) + 10.63*C(3) + 42.54*C(4) + C(5)) / C(6))$	26.60555	7.548908

### 16.4.3 The Heckit selection bias model

If you consult an econometrician concerning an estimation problem, the first question you will usually hear is, “How were the data obtained?” If the data are obtained by random sampling, then classic regression methods, such as least squares, work well. However, if the data are obtained by a sampling procedure that is not random, then standard procedures do not work well. Economists regularly face such data problems. A famous illustration comes from labor economics. If we wish to study the determinants of the wages of married women we face a **sample selection** problem. If we collect data on married women, and ask them what wage rate they earn, many will respond that the question is not relevant since they are homemakers. We only observe data on market wages when the woman chooses to enter the workforce. One strategy is to ignore the women who are homemakers, omit them from the sample, then use least squares to estimate a wage equation for those who work. This strategy fails, the reason for the failure being that our sample is not a random sample. The data we observe are “selected” by a systematic process for which we do not account.

A solution to this problem is a technique called **Heckit**, named after its developer, Nobel Prize winning econometrician James Heckman. This simple procedure uses two estimation steps. In the context of the problem of estimating the wage equation for married women, a probit model is first estimated explaining why a woman is in the labor force or not. In the second stage, a least squares regression is estimated relating the wage of a working woman to education, experience, etc., and a variable called the “Inverse Mills Ratio,” or IMR. The IMR is created from the first step probit estimation, and accounts for the fact that the observed sample of working women is not random.

The econometric model describing the situation is composed of two equations. The first, is the **selection equation** that determines whether the variable of interest is observed. The sample consists of  $N$  observations, however the variable of interest is observed only for  $n < N$  of these. The selection equation is expressed in terms of a latent variable  $z_i^*$  which depends on one or more explanatory variables  $w_i$ , and is given by

$$z_i^* = \gamma_1 + \gamma_2 w_i + u_i \quad i = 1, \dots, N$$

\* The text book calculations were carried out using Stata.

For simplicity we will include only one explanatory variable in the selection equation. The latent variable is not observed, but we do observe the binary variable

$$z_i = \begin{cases} 1 & z_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

The second equation is the linear model of interest. It is

$$y_i = \beta_1 + \beta_2 x_i + e_i \quad i = 1, \dots, n \quad N > n$$

A **selectivity problem** arises when  $y_i$  is observed only when  $z_i = 1$ , and if the errors of the two equations are correlated. In such a situation the usual least squares estimators of  $\beta_1$  and  $\beta_2$  are biased and inconsistent.

Consistent estimators are based on the conditional regression function

$$E[y_i | z_i^* > 0] = \beta_1 + \beta_2 x_i + \beta_\lambda \lambda_i \quad i = 1, \dots, n$$

where the additional variable  $\lambda_i$  is “Inverse Mills Ratio.” It is equal to

$$\lambda_i = \frac{\phi(\gamma_1 + \gamma_2 w_i)}{\Phi(\gamma_1 + \gamma_2 w_i)}$$

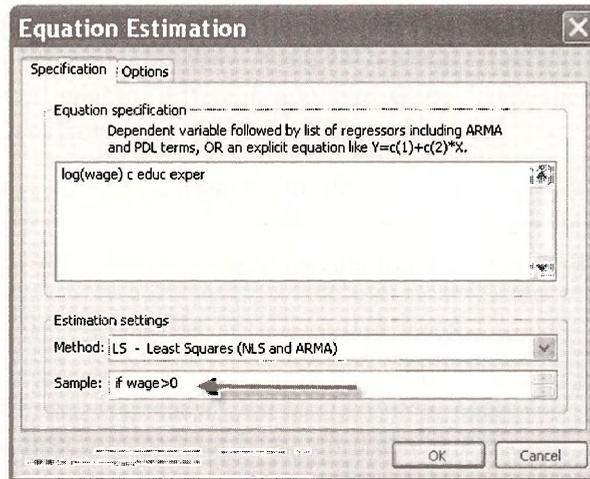
where, as usual,  $\phi(\cdot)$  denotes the standard normal probability density function, and  $\Phi(\cdot)$  denotes the cumulative distribution function for a standard normal random variable. While the value of  $\lambda_i$  is not known, the parameters  $\gamma_1$  and  $\gamma_2$  can be estimated using a probit model, based on the observed binary outcome  $z_i$ . Then the estimated IMR,

$$\tilde{\lambda}_i = \frac{\phi(\tilde{\gamma}_1 + \tilde{\gamma}_2 w_i)}{\Phi(\tilde{\gamma}_1 + \tilde{\gamma}_2 w_i)}$$

is inserted into the regression equation as an extra explanatory variable, yielding the estimating equation

$$y_i = \beta_1 + \beta_2 x_i + \beta_\lambda \tilde{\lambda}_i + v_i \quad i = 1, \dots, n$$

First, let us estimate a simple wage equation, explaining  $\ln(WAGE)$  as a function of the woman's education,  $EDUC$ , and years of market work experience ( $EXPER$ ), using the 428 women who have positive wages. Select **Quick/Estimate Equation**. Fill the dialog box as



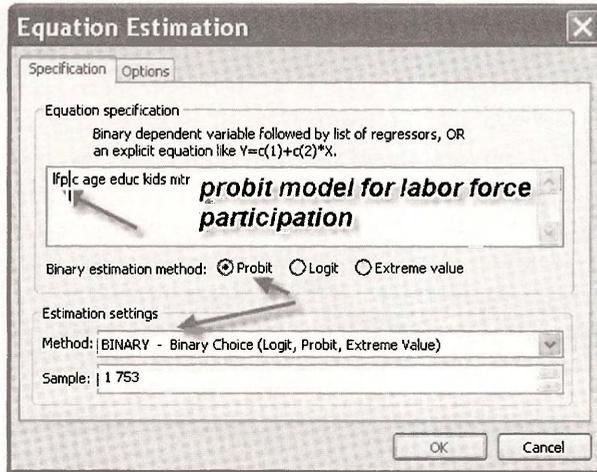
Dependent Variable: LOG(WAGE)  
 Method: Least Squares  
 Sample: 1 753 IF WAGE>0  
 Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.400174	0.190368	-2.102107	0.0361
EDUC	0.109489	0.014167	7.728334	0.0000
EXPER	0.015674	0.004019	3.899798	0.0001

Heckit estimation begins with a probit model estimation of the “participation equation,” in which *LFP* is taken to be a function of *AGE*, *EDUC*, a dummy variable for whether or not the woman as children (*KIDS*) and her marginal tax rate *MTR*. Create the dummy variable **KIDS** using

**series kids = (kids16 + kids618 > 0)**

Select **Quick/Estimate equation** and fill in the dialog box as shown.

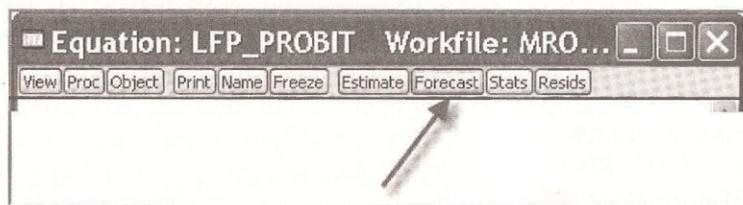


Using all the sample data we obtain

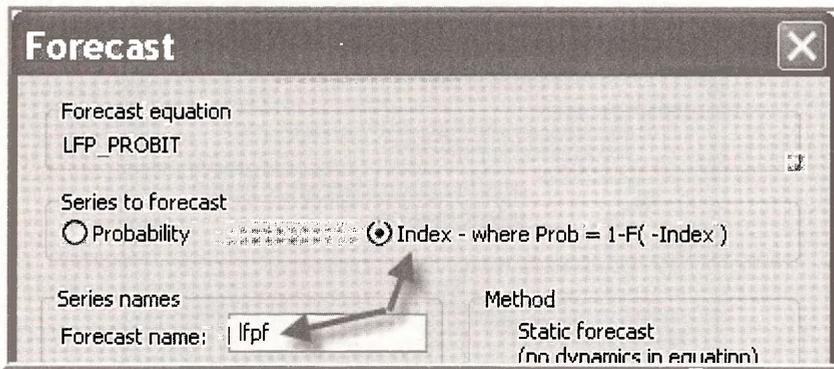
Dependent Variable: LFP  
 Method: ML - Binary Probit  
 Sample: 1.753  
 Included observations: 753

	Coefficient	Std. Error	z-Statistic	Prob.
C	1.192296	0.720544	1.654717	0.0980
AGE	-0.020616	0.007045	-2.926390	0.0034
EDUC	0.083775	0.023205	3.610225	0.0003
KIDS	-0.313885	0.123711	-2.537248	0.0112
MTR	-1.393853	0.616575	-2.260638	0.0238

The inverse Mills ratio *IMR* requires computation of the fitted **index model**. In the probit estimation window, select **Forecast**.



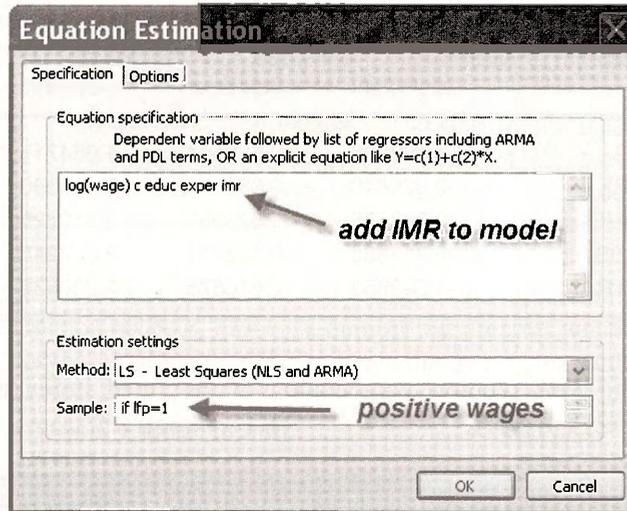
In the **Forecast** dialog box choose the radio button for **Index** and give this variable a name.



The inverse Mills ratio is then calculated using the EViews functions for the standard normal pdf **@dnorm** and the standard normal CDF **@cnorm**.

$$\text{series imr} = \text{@dnorm(lfpf)}/\text{@cnorm(lfpf)}$$

Include the **IMR** into the wage equation as an explanatory variable, using only those women who were in the labor force and had positive wages.



Dependent Variable: LOG(WAGE)

Method: Least Squares

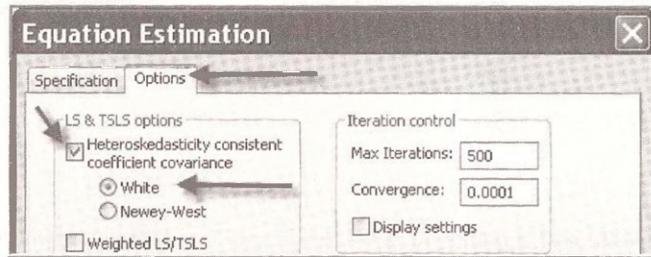
Sample: 1 753 IF LFP=1

Included observations: 428

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.810542	0.494472	1.639206	0.1019
EDUC	0.058458	0.023849	2.451122	0.0146
EXPER	0.016320	0.003998	4.081732	0.0001
IMR	-0.866439	0.326985	-2.649777	0.0084

This two-step estimation process is a consistent estimator, however the standard errors **Std. Error** do not account for the fact the **IMR** is in fact estimated. If the errors are homoskedastic, we however can carry out a test of the significance of the **IMR** variable based on the *t*-statistic that is reported by EViews. This is because under the null hypothesis that there is no selection bias the coefficient of **IMR** is zero, and thus under the null hypothesis the usual *t*-test is valid. Here we reject the null hypothesis of no selection bias and conclude that using the two-step Heckit estimation process is needed.

If the regression errors may be heteroskedastic, as they might be for this microeconomic example, a robust standard error can be used. On the **Options** tab of the **Equation Estimation** dialog check the box for **Heteroskedasticity consistent coefficient covariance** and click the **White** radio button.



The resulting *t*-statistic is still significant at the .05 level.

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.810542	0.498851	1.624816	0.1049
EDUC	0.058458	0.024143	2.421369	0.0159
EXPER	0.016320	0.004101	3.979199	0.0001
IMR	-0.866439	0.340561	-2.544155	0.0113

Correct standard errors for the two step estimation procedure are difficult to obtain without specially designed software. Such options, and maximum likelihood estimation of the Heckit model, are available in Limdep and Stata software packages.

### Keywords

@cnorm	individual samples	ordered choice models
@dnorm	inverse Mills ratio	ordered probit
binary choice models	latent variables	Poisson regression
censored data	limit values	prediction evaluation
common sample	linear probability model	probability forecast
count data models	logit	probit
elasticity	LR statistic	sample
EViews size limitations	marginal effect	semi-elasticity
heckit	maximum likelihood	threshold values
IMR	McFadden R-squared	tobit
index model	NA	

# CHAPTER 17

## Importing and Exporting Data

### CHAPTER OUTLINE

17.1 Obtaining Data from the Internet  
17.2 Importing An Excel Data File  
17.3 Date Conventions

17.4 Importing a Text (Ascii) Data File  
17.5 Entering Data Manually  
17.6 Exporting Data from EViews

KEYWORDS

### 17.1 OBTAINING DATA FROM THE INTERNET

Up to now we have taken you through various econometric methodologies and applications using already prepared EViews workfiles. In this chapter, we show you how to create a workfile and how to import data from an Excel spreadsheet. The first step is to create the Excel data file.

Getting data for economic research is much easier today than it was years ago. Before the Internet, hours would be spent in libraries, looking for and copying data by hand. Now we have access to rich data sources which are a few clicks away.

Suppose you are interested in analyzing the GDP of the United States. As suggested in Chapter 17, the website **Resources for Economists** contains a wide variety of data, and in particular the macro data we seek.

Websites are continually updated and improved. We shall guide you through an example, but be prepared for differences from what we show here.

First, open up the website: [www.rfe.org](http://www.rfe.org) :

**AEA**WEB

AEA Web  
RFE Home  
Econ Search Engine  
RFE Search  
Conferences  
Complete Contents  
Abridged Contents

WWW Virtual Library  
Economics

**RFE**  
ISSN 1081-4248  
Vol. 11, No. 8  
August, 2007

Editor: Bill Goffe  
Dept. of Economics, SUNY Oswego  
Editorial Assistant: Marian Aboud

- [Introduction](#)
- [Data](#)
- [Dictionaries, Glossaries, & Encyclopedias](#)
- [Economists, Departments, & Universities](#)
- [Forecasting & Consulting](#)
- [Jobs, Grants, Grad School, & Advice](#)
- [Mailing Lists, Forums, & Usenet](#)
- [Meetings & Conferences](#)
- [News Media](#)
- [Organizations & Associations](#)
- [Other Internet Guides](#)
- [Scholarly Communication](#)
- [Software](#)
- [Teaching Resources](#)
- [Blogs, Commentaries, and Podcasts](#)
- [Neat Stuff](#)

Select the **Data** option and then select **U.S. Macro and Regional Data**.

**AEA**WEB

AEA Web  
RFE Home  
Econ Search Engine  
RFE Search  
Conferences  
Complete Contents  
Abridged Contents

[Title Page / Data](#)  
[Table of Contents: Abridged](#) | [Complete Contents](#)  
[Search Economic Web Sites](#)

**Data**

- [U.S. Macro and Regional Data](#)
- [Other U.S. Data](#)
- [World and Non-U.S. Data](#)
- [Finance and Financial Markets](#)
- [Journal Data and Program Archives](#)

This will open up a range of sub-data categories. For the example discussed here, select the **National Income and Produce Accounts** to get data on GDP.

**AEA WEB**

AEA Web  
RFE Home  
Econ Search Engine  
RFE Search  
Conferences  
Complete Contents  
Abridged Contents

Title Page / Data / U.S. Macro and Regional Data  
Table of Contents; Abridged | Complete Contents  
Search Economic Web Sites

## U.S. Macro and Regional Data

Macro summary statistics (major data series for recent years)

Economic Statistics Briefing Room (ESBR) - handful of most important data series with graphs (White House site) | [details...](#)

"Primary" macro and regional sites that generate data (many long series)

- [Bureau of Economic Analysis \(BEA\)](#) - National Income and Produce Accounts (GDP, etc.), international and regional data | [details...](#)
- [Bureau of Labor Statistics \(BLS\)](#) - more than 250,000 long series; unemp. and price series most prominent | [details...](#)
- [Conference Board](#) - "Leading Economic Indicators", and "Consumer Confidence" and other non-govt. data | [details...](#)
- [Congressional Budget Office \(CBO\)](#) - current federal spending and revenue, macro forecasts | [details...](#)
- [Federal Budget for the Fiscal Years 1997 to 2007](#) - summary and very detailed federal budget info | [details...](#)
- [Survey of Consumers from the Univ. of Michigan](#) - well-known survey of consumer attitudes | [details...](#)
- [What Was the Exchange Rate Then?](#) - historical exchange rates | [details...](#)
- [What Was the GDP Then?](#) - U.S. GDP estimates starting in 1789 | [details...](#)

From the screen below, select the **Gross Domestic Product (GDP)** option.

### U.S. Economic Accounts

<p><b>National</b></p> <hr/> <p>Access National Economic Accounts Data</p> <ul style="list-style-type: none"> <li>▶ <a href="#">Gross Domestic Product (GDP)</a></li> <li>▶ <a href="#">Personal Income and Outlays</a></li> <li>▶ <a href="#">Corporate Profits</a></li> <li>▶ <a href="#">Fixed Assets</a></li> <li>▶ <a href="#">Satellite Account</a> <ul style="list-style-type: none"> <li>• <a href="#">Research and Development</a></li> </ul> </li> <li>▶ View all <a href="#">National Accounts</a> Information...</li> </ul>	<p><b>International</b></p> <hr/> <p>Access International Economic Accounts Data</p> <ul style="list-style-type: none"> <li>▶ <a href="#">Balance of Payments</a></li> <li>▶ <a href="#">Trade in Goods and Services</a></li> <li>▶ <a href="#">International Services</a></li> <li>▶ <a href="#">International Investment Position</a></li> <li>▶ <a href="#">Operations of Multinational Companies</a></li> <li>▶ <a href="#">Survey Forms and Related Materials</a></li> <li>▶ View all <a href="#">International Accounts</a> Information...</li> </ul>
<p><b>Regional</b></p> <hr/> <p>Access Regional Economic Accounts Data</p> <ul style="list-style-type: none"> <li>▶ <a href="#">GDP by State (formerly GSP)</a></li> <li>▶ <a href="#">State and Local Area Personal Income</a></li> <li>▶ <a href="#">RIMS II Regional Input-Output Multipliers</a></li> <li>▶ <a href="#">BEA's Regional FACT Sheets (BEARFACTS)</a></li> <li>▶ <a href="#">BEA Economic Areas</a></li> <li>▶ View all <a href="#">Regional Accounts</a> Information...</li> </ul>	<p><b>Industry</b></p> <hr/> <p>Access Industry Economic Accounts Data</p> <ul style="list-style-type: none"> <li>▶ <a href="#">Annual Industry Accounts</a> <ul style="list-style-type: none"> <li>• <a href="#">GDP by Industry</a></li> <li>• <a href="#">Input-Output Accounts</a></li> </ul> </li> <li>▶ <a href="#">Benchmark Input-Output Accounts</a></li> <li>▶ <a href="#">Satellite Accounts</a> <ul style="list-style-type: none"> <li>• <a href="#">Research and Development</a></li> <li>• <a href="#">Travel and Tourism</a></li> </ul> </li> <li>▶ <a href="#">Supplemental Estimates</a></li> <li>▶ View all <a href="#">Industry Accounts</a> Information...</li> </ul>

Most websites allow you to download data conveniently in an Excel format.

## National Economic Accounts

### Gross Domestic Product (GDP)

- ▶ News Release: [Gross Domestic Product](#)
  - ↳ includes highlights, technical note, and associated tables
- ▶ **Now Available** : [NIPA annual and comprehensive revision plans](#)
- ▶ [Current-dollar and "real" GDP \(Excel • 35KB\)](#)
- ▶ [Percent change from preceding period \(Excel • 35KB\)](#)
- ▶  [Interactive Tables: National Income and Product Accounts Tables](#)
- ▶ Selected NIPA Tables:
  - [Text format \(Text • 1,788KB\)](#)
  - [Self-extracting format \(EXE • 870KB\)](#)
  - [Compressed format \(ZIP • 802KB\)](#)
  - [Comma-delimited format \(CSV • 1,071KB\)](#)
  - [Portable document format \(PDF • 6,189KB\)](#)

Be sure to save the file which is called *gdplev.xls*.

Home > National Economic Accounts

### National Economic Accounts

#### Gross Domestic Product (GDP)

- ▶ News Release: [Gross Domestic Product](#)
  - ↳ includes highlights, technical note, and associated tables
- ▶ **Now Available** : [NIPA annual and comprehensive revision plans](#)
- ▶ [Current-dollar and "real" GDP \(Excel • 35KB\)](#)
- ▶ [Percent change from preceding period \(Excel • 35KB\)](#)
- ▶  [Interactive Tables: National Income and Product Accounts Tables](#)
- ▶ Selected NIPA Tables:
  - [Text format \(Text • 1,788KB\)](#)
  - [Self-extracting format \(EXE • 870KB\)](#)
  - [Compressed format \(ZIP • 802KB\)](#)
  - [Comma-delimited format \(CSV • 1,071KB\)](#)
  - [Portable document format \(PDF • 6,189KB\)](#)

**File Download**

Do you want to open or save this file?

 Name: *gdplev.xls*  
Type: Microsoft Excel Worksheet, 34.5 KB  
From: [www.bea.gov](http://www.bea.gov)

Always ask before opening this type of file

 While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. [What's the risk?](#)

### National Economic Accounts

#### Gross Domestic Product (GDP)

- ▶ News Release: [Gross Domestic Product](#)  
includes highlights, technical note, and a
- ▶ **Now Available** : [NIPA annual and comprehensive](#)
- ▶ [Current-dollar and "real" GDP](#) (Excel • 35KB)
- ▶ [Percent change from preceding period](#) (Excel • 35KB)
- ▶ Interactive Tables: [National Income and Product Accounts](#)
- ▶ Selected NIPA Tables:
  - [Text format](#) (Text • 1,788KB)
  - [Self-extracting format](#) (EXE • 370KB)
  - [Compressed format](#) (ZIP • 802KB)
  - [Comma-delimited format](#) (CSV • 1,071KB)
  - [Portable document format](#) (PDF • 6,189KB)

Download complete

Download Complete

Saved: [gdplev.xls](#) from [www.bea.gov](#)

---

Downloaded: 34.5 KB in 1 sec  
 Download to: C:\gdplev.xls  
 Transfer rate: 34.5 KB/Sec

Close this dialog box when download completes

Open Open Folder Close

Once the file has been downloaded (in this example, to *C:\gdplev.xls*), we can open the file and a sample of the data in Excel format is shown below.

	A	B	C	D	E	F	G	
1		<b>Current-Dollar and "Real" Gross Domestic Product</b>						
2								
3		Annual				Quarterly		
4						(Seasonally adjusted annual rates)		
5								
6		GDP in	GDP in			GDP in	GDP in	
7		billions of	billions of			billions of	billions of	
		current	chained			current	chained	
		dollars	2000 dollars			dollars	2000 dollars	
8	1929	103.6	865.2		1947q1	237.2	1,570.5	
9	1930	91.2	790.7		1947q2	240.5	1,568.7	
10	1931	76.5	739.9		1947q3	244.6	1,568.0	
11	1932	58.7	643.7		1947q4	254.4	1,590.9	
12	1933	56.4	635.5		1948q1	260.4	1,616.1	
13	1934	66.0	704.2		1948q2	267.3	1,644.6	
14	1935	73.3	766.9		1948q3	273.9	1,654.1	
15	1936	83.8	866.6		1948q4	275.2	1,658.0	
16	1937	91.9	911.1		1949q1	270.0	1,633.2	
17	1938	86.1	879.7		1949q2	266.2	1,628.4	

For illustrative purposes, let us now import the annual data (1929-2006) for nominal GDP (column B, first observation in cell B8) and real GDP (column C, first observation in cell C8) into an EViews workfile.

## 17.2 IMPORTING AN EXCEL DATA FILE

To create an EViews workfile, double click on your EViews icon to open the software, then select **File/New/Workfile**. The following screen will open up.

**Workfile Create**

Workfile structure type  
 Dated - regular frequency

Irregular Dated and Panel workfiles may be made from Unstructured workfiles by later specifying date and/or other identifier series.

Date specification  
 Frequency: Annual  
 Start date: 1929  
 End date: 2006

Names (optional)  
 WF:   
 Page:

OK Cancel

To create the workfile for annual data covering sample period **1929** to **2006**, select **Annual** from the drop-down menu in **Frequency** and type in the **Start** and **End** dates. Clicking on **OK** will create the **UNTITLED** workfile below.

**Workfile: UNTITLED**

View | Proc | Object | Print | Save | Details+/- | Show | Fetch | Store | Delete | Genr | Sample

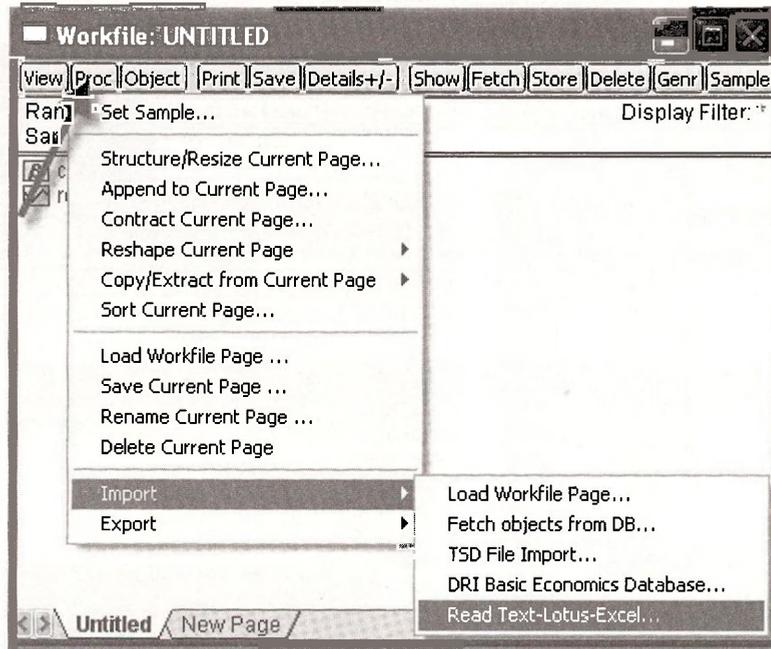
Range: 1929 2006 -- 78 obs      Display Filter: \*

Sample: 1929 2006 -- 78 obs

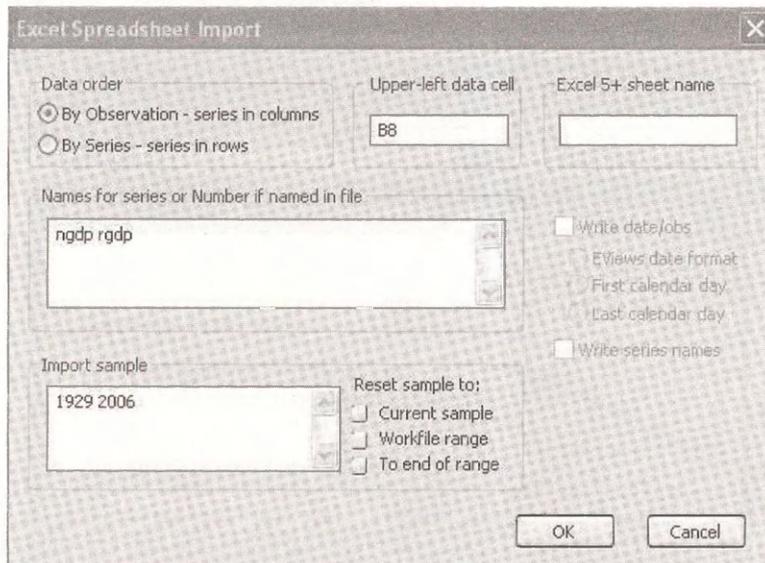
c  
 resid

Untitled / New Page

To import data select **Proc/ Import/ Read Text-Lotus-Excel**.



EViews will then ask you for the location of the Excel file. Open the *C:\gdpdplev.xls* file we have created and the following screen will open.



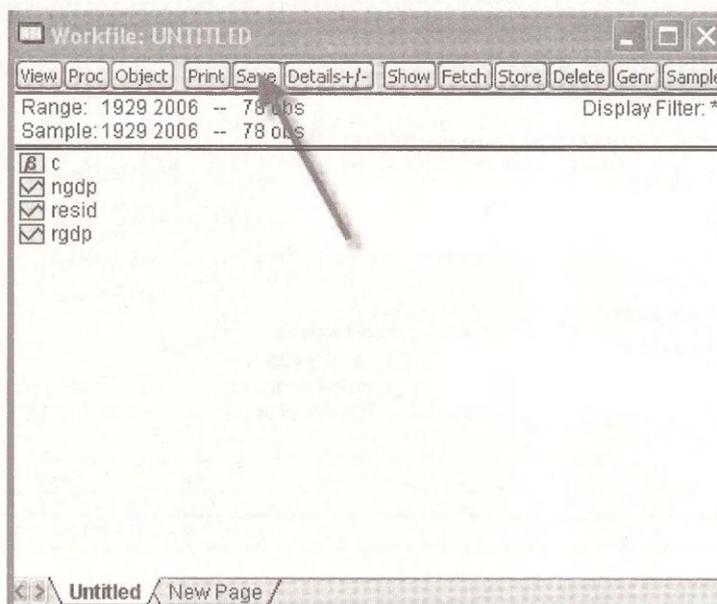
Be sure to pick the **By observation – series in columns** option, enter the correct location of the first observation (B8) and type in the names of the variables – in this case *NGDP* and *RGDP*.

Clicking on **OK** will import the data from the Excel datafile to the EViews workfile. As a check open the group *NGDP* and *RGDP* and you can see that we have successfully imported the data (do check this against the Excel spreadsheet shown above).

The screenshot shows the EViews 'Workfile: UNTITLED' window. The 'Group: UNTITLED' window is open, displaying a table of data. The table has columns for 'obs', 'NGDP', and 'RGDP'. The data rows range from 1929 to 1948, with values for NGDP and RGDP. The 'Save' button in the top menu bar is highlighted with a mouse cursor.

obs	NGDP	RGDP
1929	103.6000	865.2000
1930	91.20000	790.7000
1931	76.50000	739.9000
1932	58.70000	643.7000
1933	56.40000	635.5000
1934	66.00000	704.2000
1935	73.30000	766.9000
1936	83.80000	866.6000
1937	91.90000	911.1000
1938	86.10000	879.7000
1939	92.20000	950.7000
1940	101.4000	1034.100
1941	126.7000	1211.100
1942	161.9000	1435.400
1943	198.6000	1670.900
1944	219.8000	1806.500
1945	223.1000	1786.300
1946	222.3000	1589.400
1947	244.2000	1574.500
1948		

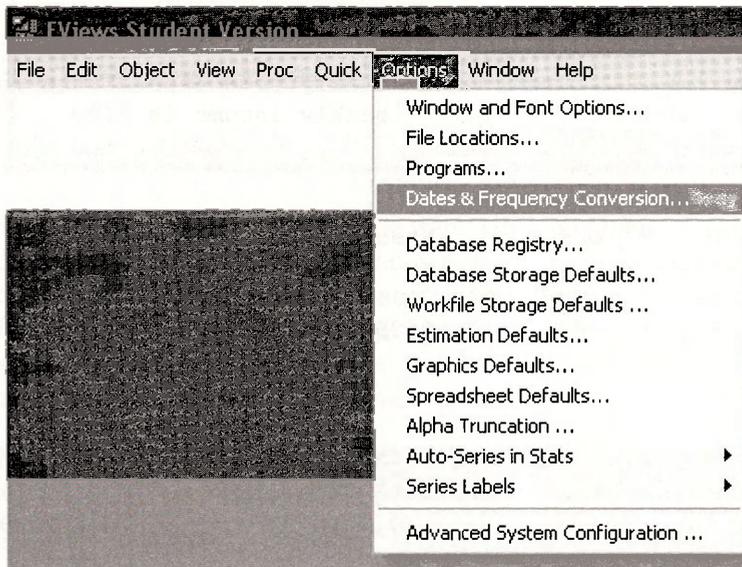
The final step is to save your workfile.



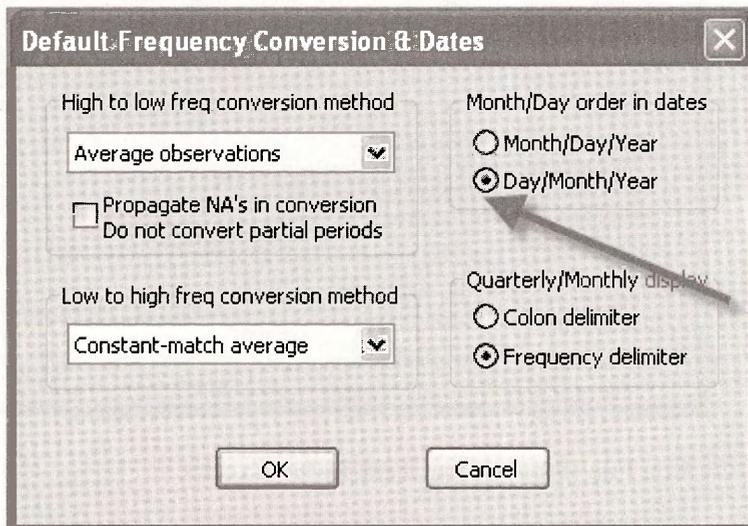
### 17.3 DATE CONVENTIONS

The rules for describing calendar or ordered data are:

- Annual: specify the year; for example, 1981, or 2007.
- Quarterly: the year, followed by a colon or period, and the quarter number. Examples: 1992:1, 65:4, 2007:3.
- Monthly: the year, followed by a colon or period, and the month number. Examples: 1956:1, 2007:11.
- Weekly and daily: by default, you should specify these dates as Month/Day/Year. Thus August 15, 2007 is 8/15/2007. However, you can easily change this to day/month/year using **Options/Dates & Frequency Conversion**.



Clicking on Day/Month/Year will give you 15/8/2007.



## 17.4 IMPORTING A TEXT (ASCII) DATA FILE

Excel data files are the most common way of handling data. However, some data also come in text form and so for completeness, we shall consider the case of importing a text data file. As an illustration we will import an ASCII file called *food.dat*. Before trying to import the data in *food.dat* examine the contents of the definition file *food.def*. It is an ASCII file that can be opened with NOTEPAD. The *\*.def* files contain variable names and descriptions. For example, open *food.def*.

```
food.def
```

```
food_exp income
```

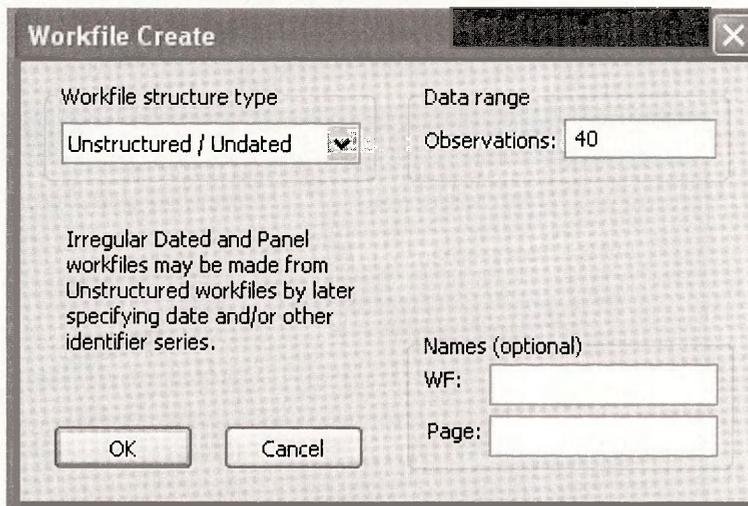
```
Obs: 40
```

```
1. food_exp (y)          weekly food expenditure in $
2. income (x)           weekly income in $100
```

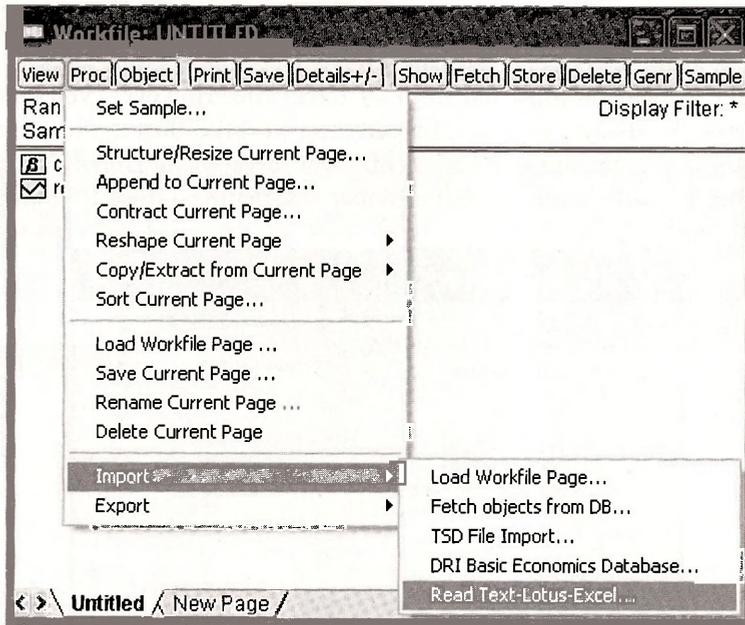
Variable	Obs	Mean	Std. Dev.	Min	Max
food_exp	40	283.5735	112.6752	109.71	587.66
income	40	19.60475	6.847773	3.69	33.4

This definition file shows that there are 40 observations on two variables, *Y* and *X*, in that order, and they are weekly food expenditure and weekly income, respectively.

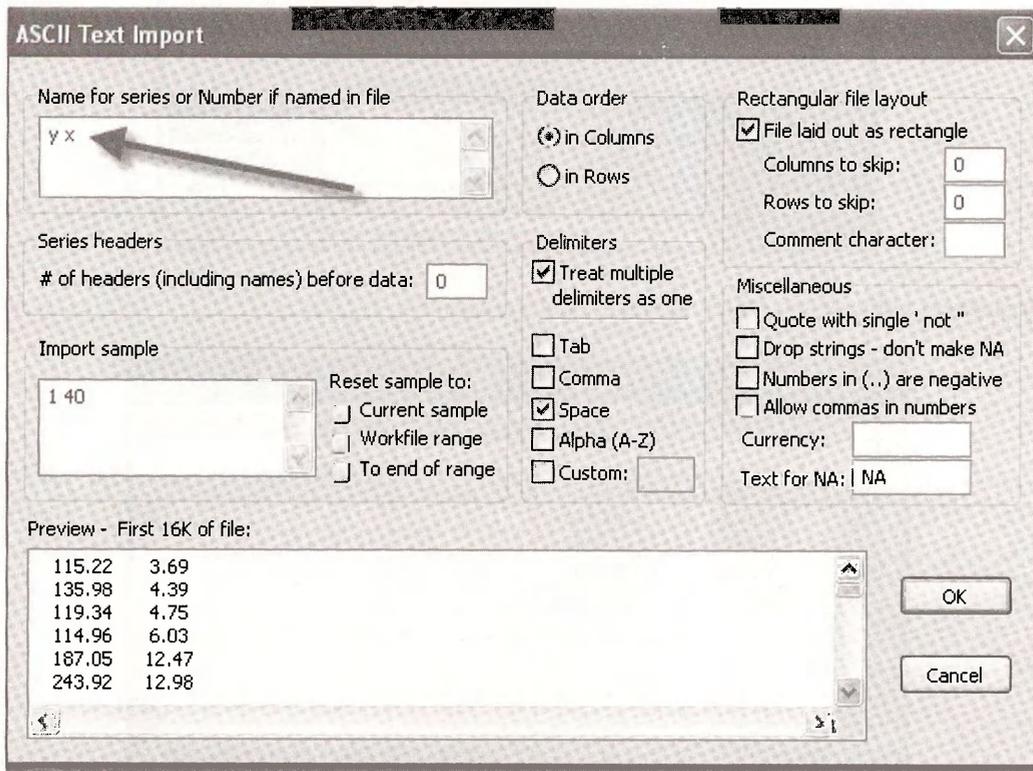
To import this data, create a workfile for 40 undated observations and click **OK**.



To import data, click on **File/Import/Read Text-Lotus-Excel**.

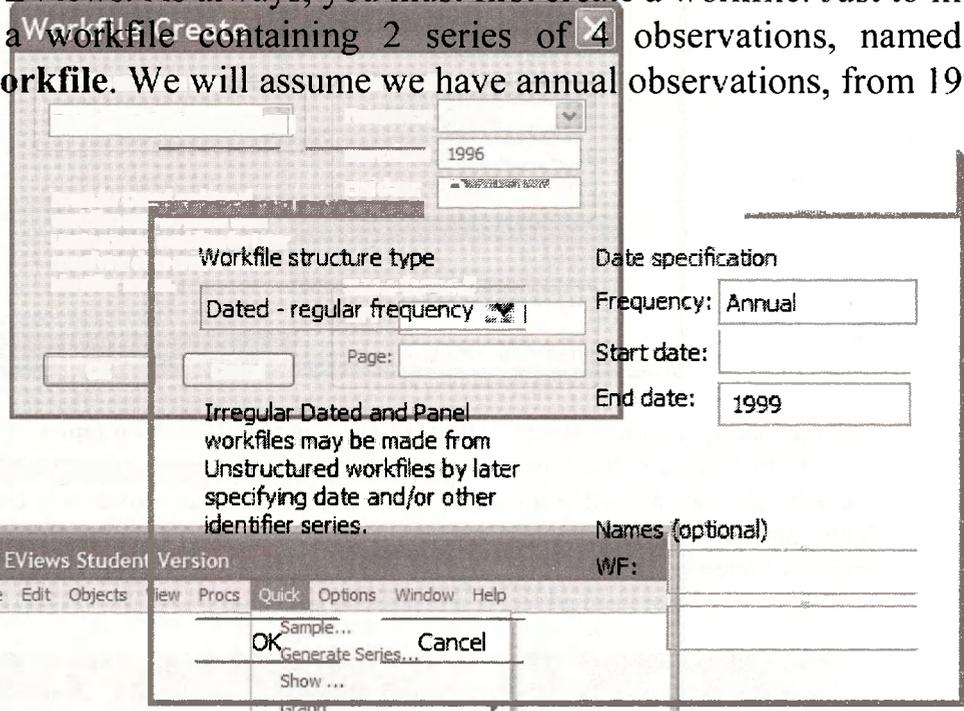


Use the dialog box to locate and select the file you want. Click on **Open**. A dialog box will open. Note at the bottom of the dialog box the first few observations in the data file are shown. Because the data file does not contain variable names, enter them as shown, and click **OK**. If there are a large number of long variable names, it is convenient to cut and paste them from the \*.def file into the EViews window using **Ctrl+C** followed by **Ctrl+V**. The workfile will then show that two new series have been added, *X* and *Y*. Save your file

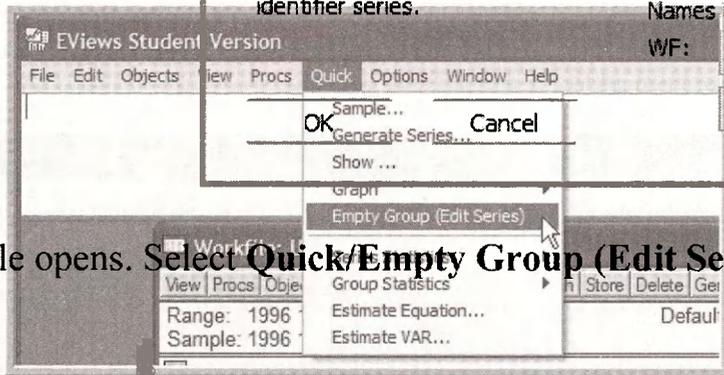


## 17.5 ENTERING DATA MANUALLY

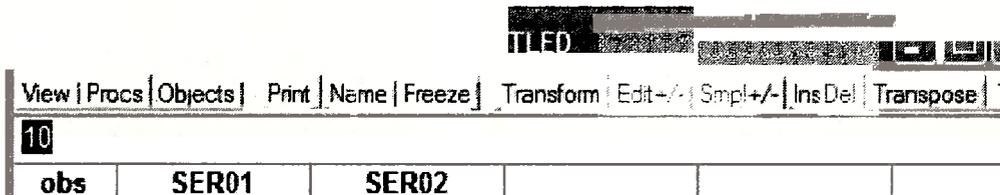
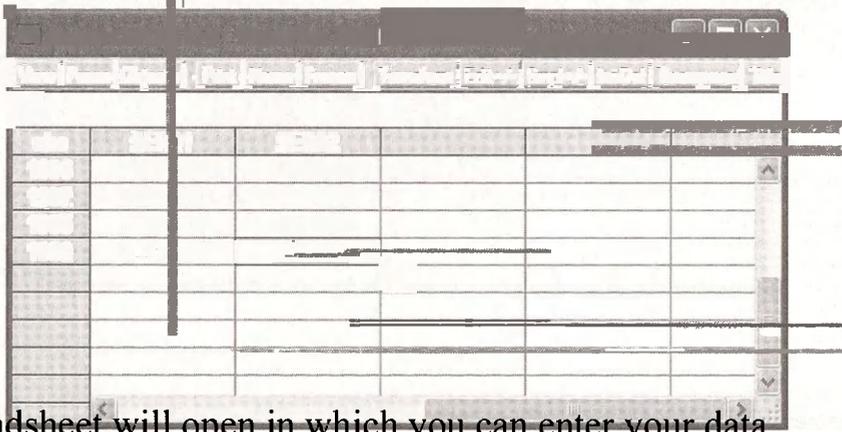
Most of the time, data will be imported from an Excel file. However, you can enter data directly into EViews. As always, you must first create a workfile. Just to illustrate, we will create a workfile containing 2 series of 4 observations, named X and Y. **File/New/Workfile.** We will assume we have annual observations, from 1996 to 1999.



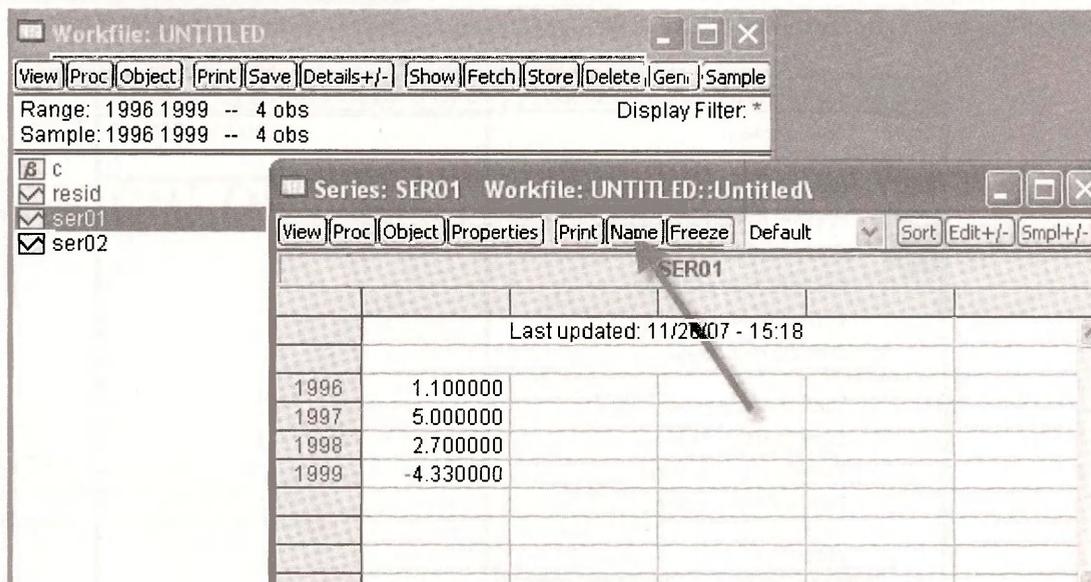
A workfile opens. Select **Quick/Empty Group (Edit Series)**.



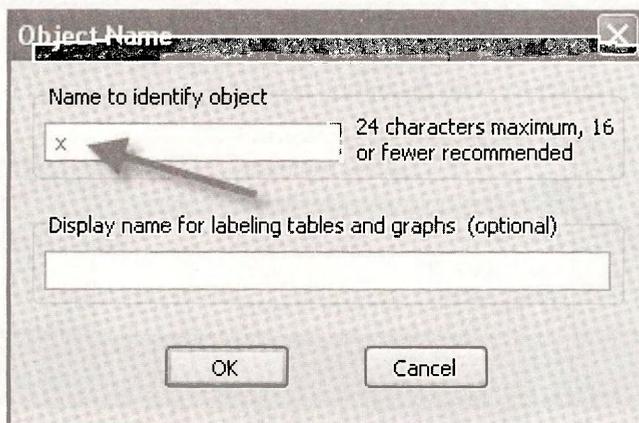
A spreadsheet will open in which you can enter your data.



As you fill in the data, EViews will assign temporary names, **SER01** and **SER02**, to the variables. To change those names, for example, to change **SER01** to **X**, click open **SER01**, and select **name**:



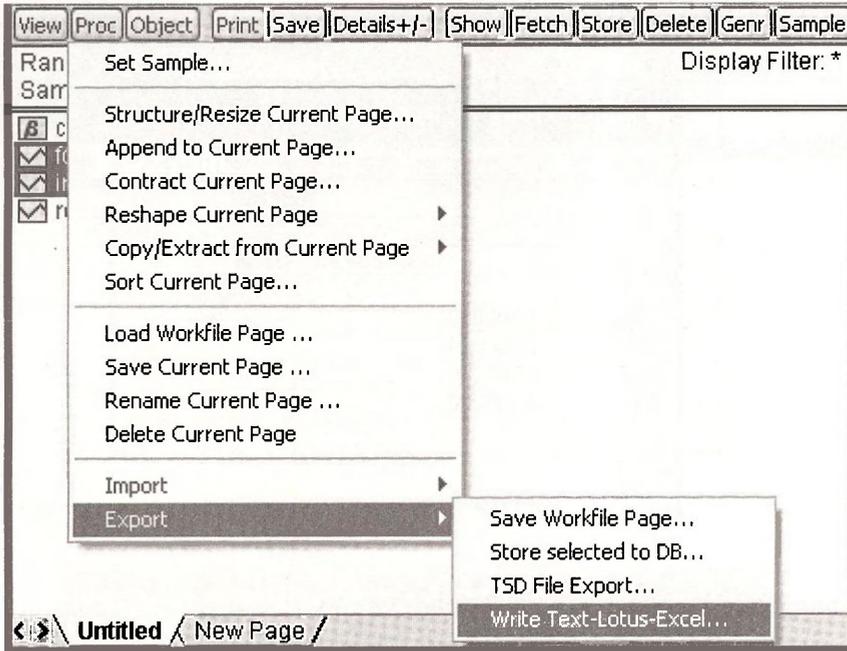
This will open the following box, and you can then type in **X**.



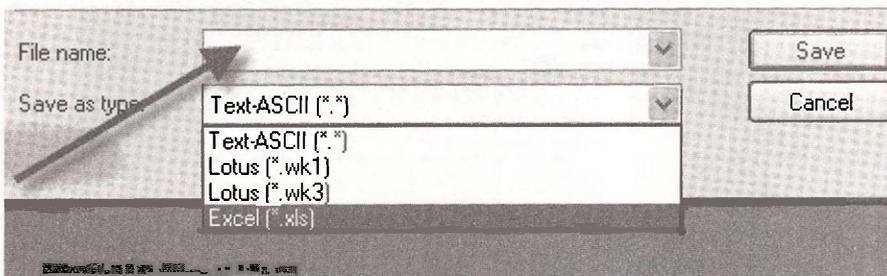
Repeat the process to change **SER02** to **Y**. You should now find the series **X** and **Y** in your workfile.

### 17.6 EXPORTING DATA FROM EViews

There are times when you would like to export data from an EViews workfile. To illustrate, let us work with *food.wf1* and export the two series. To do so, highlight the two series then click on **Proc/Export/Write Text-Lotus-Excel**.



This will then open up a directory with the option to save as a text or Excel file.



#### Keywords

entering data

importing data

exporting data

# APPENDIX **A**

## Review of Math Essentials

### CHAPTER OUTLINE

A.1 Mathematical Operations  
A.2 Logarithms and Exponentials

A.3 Graphing Functions  
KEYWORDS

### A.1 MATHEMATICAL OPERATIONS

EViews has many mathematical functions. Click on **Help/Help Reference/Function Reference**. Choose **Basic Mathematical Functions**. Some of these are:

<b>Name</b>	<b>Function</b>
@abs(x), abs(x)	absolute value
@ceiling(x)	smallest integer not less than
@exp(x), exp(x)	exponential
@fact(x)	factorial
@floor(x)	largest integer not greater than
@inv(x)	reciprocal
@log(x), log(x)	natural logarithm
@round(x)	round to the nearest integer
@sqrt(x), sqr(x)	square root

Some of these functions require the “@” sign in front, and some do not. Also recall the basic mathematical operators:

<b>Expression</b>	<b>Operator</b>	<b>Description</b>
+	add	$x+y$ adds the contents of X and Y.
-	subtract	$x-y$ subtracts the contents of Y from X.
*	multiply	$x*y$ multiplies the contents of X by Y.
/	divide	$x/y$ divides the contents of X by Y.
^	raise to the power	$x^y$ raises X to the power of Y.

To illustrate these operations create a workfile, with 101 undated observations. Name it *appendix\_a.wfl*. Create a scalar  $a = 3$ , and carry out some basic operations on this scalar by entering the following lines in the command window:

```
scalar a = 3
scalar acube = a^3
scalar roota = sqr(a)
scalar lna = log(a)
scalar expa = exp(a)
```

All of the results are scalars because we have defined  $a$  to be a scalar and declared the outcome a scalar as well. EViews uses **Scientific Notation** when reporting extremely large numbers. On the command line enter

```
scalar b = 510000/.00000034
```

EViews reports the value of  $B$  as

Scalar B = 1.5e+012

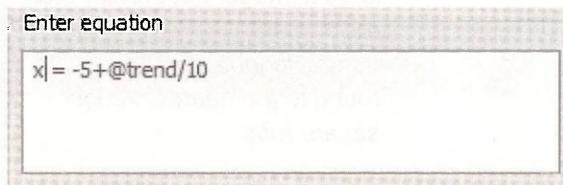
This means  $B = 1.5 \times 10^{12}$

## A.2 LOGARITHMS AND EXPONENTIALS

For practice purposes, create a variable  $X$  that ranges from  $-5$  to  $+5$  in increments of  $0.1$ . To do this, we create a **Trend** variable. The command **@trend** creates a variable starting at  $0$  and increasing by one for each observation.

```
series x = -5 + @trend/10
```

Alternatively, we could have clicked on the **Genr** tab on the EViews menubar and entered



Enter equation

x| = -5+@trend/10

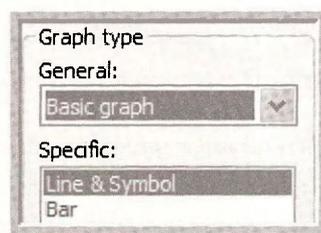
Now we can create new variables using the mathematical functions. Note that since we are using the command **series** we are in fact creating new series containing 101 observations.

```
series absx = @abs(x)
series expx = @exp(x)
```

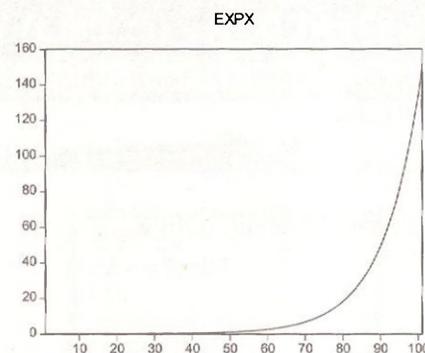
Click on each series to open it. For example the first few values of **EXPX** are

EXPX	
Last updated: 11/16/07 - 13:20	
Modified: 1 101 // expx = @exp(x)	
1	0.006738
2	0.007447
3	0.008230
4	0.009095
5	0.010052
6	0.011109
7	

In this spreadsheet window click on **View/Graph**. Select the default **Line & Symbol** plot

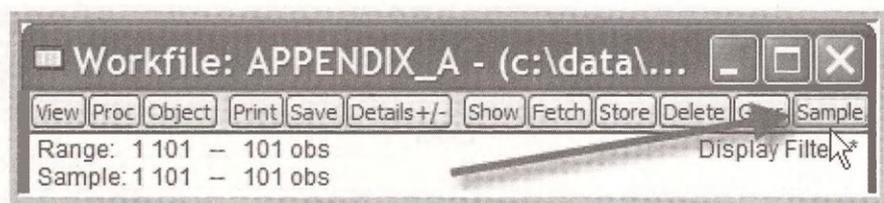


The resulting graph is

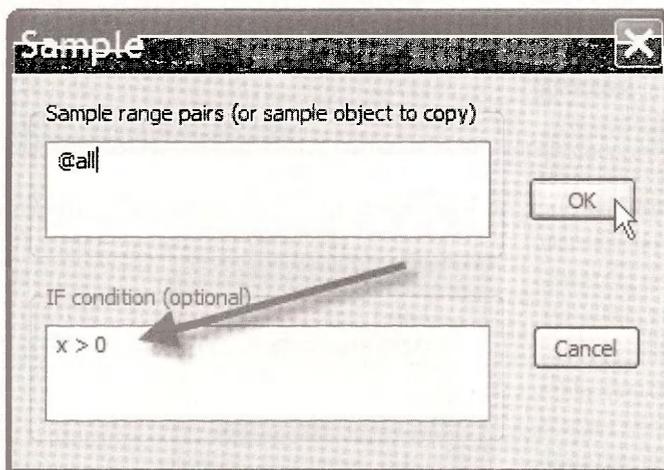


On the horizontal axis is the observation number. Observation 51 corresponds to  $X = 0$  and  $\exp(0) = 1$ . As  $X$  increases in value the value of  $\exp(x)$  becomes very large, and thus on the graph the value below observation 51, which are less than one, actually do not show up on the graph.

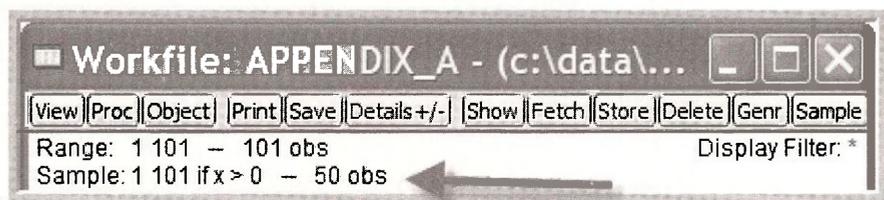
Logarithms are only defined for positive numbers. Thus to illustrate its use we will first change the sample to include only positive values of  $X$ . Click the **Sample** button on the EViews menubar.



In the dialog box that appears add an **IF condition** that ensures that  $X > 0$ . Click **OK**. All operations now will only take place on the positive values of  $X$ . This is important not only for the logarithm, but also for square roots.



Note that in the EViews workfile the header now indicates that the sample has a condition.



Now we may safely generate the natural logarithms of  $X$ .

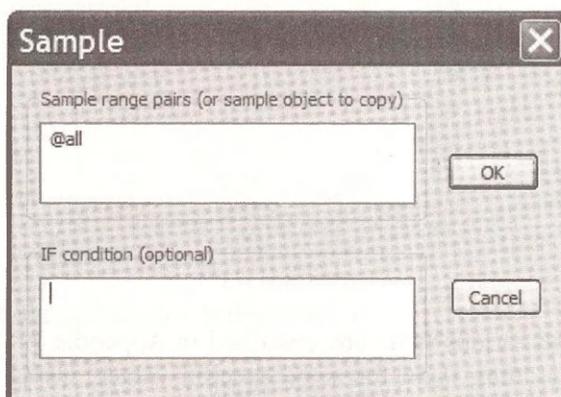
**series lnx = log(x)**

Examine the values of **LN $X$** . You will find “NA” for the nonpositive values of  $X$ . If you graph **LN $X$**  it will not show values for observations 51 (where  $X = 0$ ) and below.

### **A.3 GRAPHING FUNCTIONS**

Let us use the  $X$  variable we created in Section A.2 (**series x = -5 + @trend/10**) to explore the shapes of some functional forms. First, change the sample back to the full 101 observations, and

not just the positive values. Click on **Sample** in the main menubar and remove the condition that  $X > 0$ . It should look like

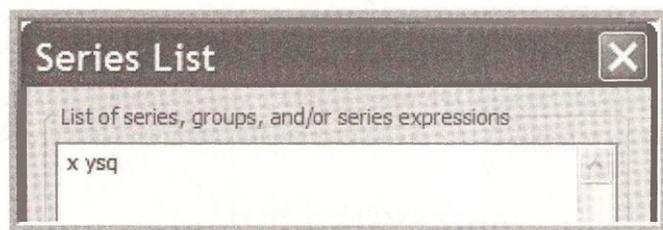


Then click **OK**.

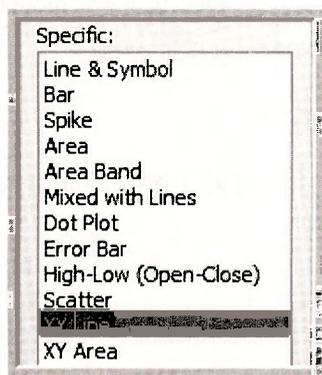
We will create the values of the quadratic function  $ysq = 1 + 2x + x^2$  by typing into the command line

**series ysq = 1 + 2\*x + x^2**

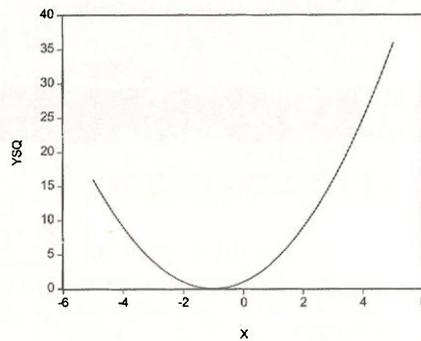
Plot the resulting series against **X** by selecting **Quick/Graph** from the EViews menu. In the dialog box enter the  $x$ -axis variable first



In the **Graph Options** dialog box select **XY Line** then **OK**.



The resulting graph shows the parabolic shape that we expected, with **YSQ** taking the value 1 when  $X = 0$ .



As you are examining the various functions described in Appendix A there is no better way to grasp their nature than plotting them for some specific values. Using EViews makes it easy. Because you are less familiar with the **log-linear** and **log-log** functions, let us plot some examples.

We will be working with logs, so for convenience let us again work with sample values for which  $X > 0$ . Click on **Sample** on the main menubar, and add the **IF** condition  $X > 0$ .

In Section A.4.4 the **log-log** function is described. The equation

$$\ln(y) = \beta_1 + \beta_2 \ln(x)$$

can be solved for  $y$  as

$$y = \exp[\beta_1 + \beta_2 \ln(x)]$$

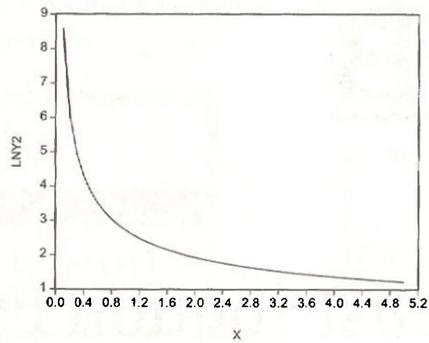
We must use this expression because we want to plot  $y$  values against  $x$  values. As an illustration let us select the values  $\beta_1 = 1$  and  $\beta_2 = -.5$ , so that the function we wish to graph is

$$y = \exp[1 - .5 \ln(x)]$$

The negative value of  $\beta_2$  will create an inverse relationship. This is a **constant elasticity** relationship and the elasticity is  $\beta_2 = -.5$ . That is, a 1% increase in  $X$  leads to a ½% reduction in  $Y$ . Into the EViews command line, type

```
series lny2 = exp(1 - .5*log(x))
```

Using **Quick/Graph/XY Line** plot the values **LNY2** against **X**, to produce



In Section A.4.5 the **log-linear** function is described. It is

$$\ln(y) = \beta_1 + \beta_2 x$$

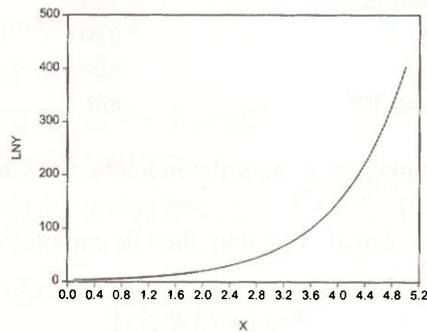
To plot this function we make use of the fact that taking antilogarithms we can express the dependent variable

$$y = \exp(\beta_1 + \beta_2 x)$$

For illustration let us plot  $y = \exp(1 + 1x)$ . Into the EViews command line type

```
series lny = exp(1 + 1* x)
```

Using **Quick/Graph/XY Line** plot the values **LNY** against **X**, to produce



### Keywords

@abs  
@exp  
@trend  
exp

exponential  
logarithm  
sample: if condition

scientific notation  
sqr  
XY line graph

# APPENDIX **B**

## Statistical Distribution Functions

### CHAPTER OUTLINE

- |   |   |
|---|---|
| B.1 Cumulative Normal Probabilities           | B.6 Plotting the Chi-square Distribution                      |
| B.2 Using Vectors                             | B.7 Plotting the $F$ Distribution                             |
| B.3 Computing Normal Distribution Percentiles | B.8 Probability Calculations for the $t$ , $F$ and Chi-square |
| B.4 Plotting Some Normal Distributions        | KEYWORDS  |
| B.5 Plotting the $t$ -Distribution            |   |

EViews has several types of functions for working with probability distributions. These are:

Function Type	Beginning of Name
Cumulative distribution (CDF)	@c
Density or probability	@d
Quantile (inverse CDF)	@q
Random number generator	@r

Of these four functions we are interested in, at the moment, the **Cumulative distribution (CDF)** and the **Quantile (inverse CDF)**.

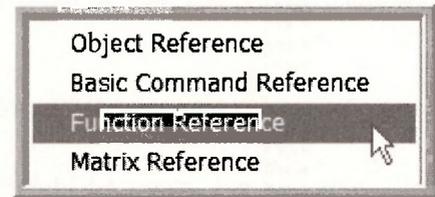
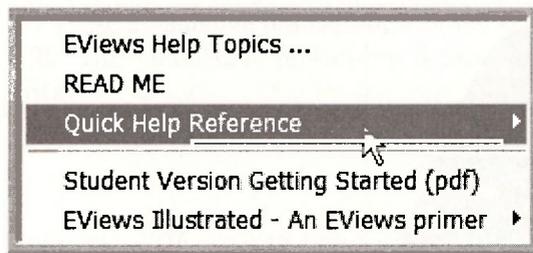
If  $f(x)$  is some probability density function, then its cumulative distribution function is

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

That is, the **CDF gives the probability that the random variable  $X$  takes a value less than, or equal to, the specified value,  $x$ .**

The quantile function works just the reverse. You provide a probability, say .10, and the **quantile function tells you the value of  $x$  such that  $F(x) = .10$ .** This answer is exact for continuous distributions, For discrete random variables there of course may not be an exact  $x$  value corresponding to any probability value that you select. The EViews help file says “*The quantile functions will return the smallest value where the CDF evaluated at the value equals or exceeds the probability of interest.*”

The list of probability distributions that EViews can work with is extensive. Click **Help/Quick Help Reference/Function Reference**



The links to different types of functions are listed.

## Operator and Function Reference

This material is divided into several topics:

- Operators.
- Basic mathematical functions.
- Time series functions.
- Financial functions.
- Descriptive statistics.
- Cumulative statistics functions.
- Moving statistics functions.
- Group row functions.
- By-group statistics.
- Additional and special functions.
- Trigonometric functions.
- Statistical distribution functions.

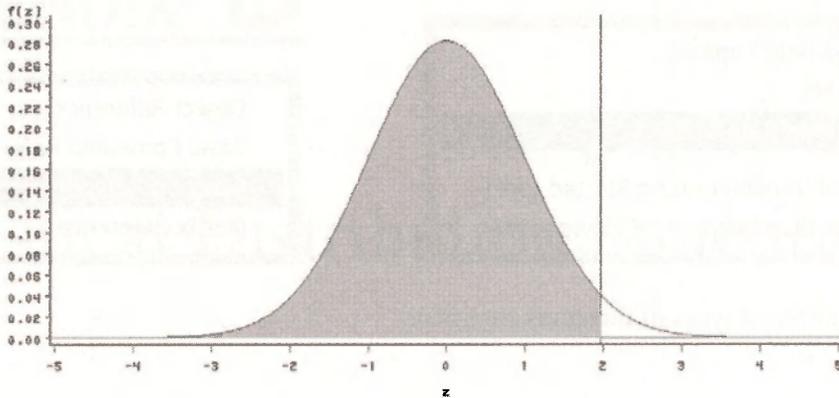


Select **Statistical distribution functions.**

### B.1 CUMULATIVE NORMAL PROBABILITIES

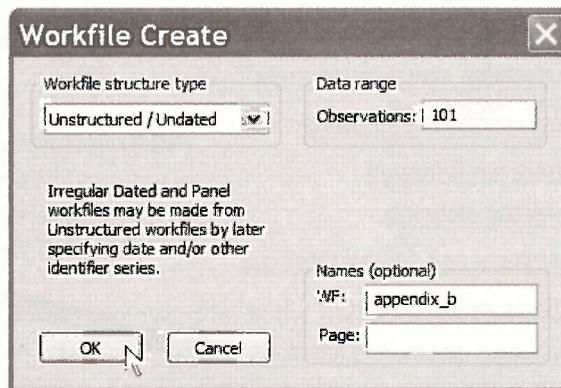
The EViews function **@cnorm(z)** returns the cumulative probability that a standard normal random variable falls to the left of the given value **z**, as shown below.

A Cumulative Probability  
Shaded Area is  $P(Z \leq 1.96)$



We will illustrate how to use EViews to compute cumulative probabilities. However, as always, you must begin by creating a workfile.

- Click on **File/New/Workfile**
- Click on **Unstructured/Undated**. Enter some number for observations, say 101, though it doesn't matter here since we will not be entering data. The reason for the odd choice will become clear later. Click **OK**.



Entering the **Name** for the workfile is optional, but doing it now, before we forget, saves time later. Name this workfile *appendix\_b.wfl*.

To compute the probability that a standard normal random variable takes a value less than or equal to 1.96, type into the command line

```
scalar p1 = @cnorm(1.96)
```

The scalar **P1** is added to the workfile. Scalars are indicated by # symbols. Highlight **P1** and double-click. The value of the scalar appears in the lower left corner of the EViews screen

Scalar P1 = 0.975002104852

That is, the probability that a standard normal random variable falls to the left of, or equals, 1.96 is .975. For a continuous random variable the probability of any one point is zero. We can also say that the probability that the standard normal random variable falls to the right of 1.96 is .025.

These cumulative probabilities for the standard normal are provided in Table 1 at the back of *POE*. This cumulative probability is used so often it is given its own symbol,  $\Phi$ . Thus we can write

$$P[Z < 1.96] = \Phi(1.96) = .9750$$

Using the CDF we can compute any probability we might be given. For example, suppose  $X \sim N(3,9)$ , that is,  $X$  is normally distributed with mean  $\mu = 3$  and variance  $\sigma^2 = 9$ . We can compute the probability that  $X$  falls in the interval  $[4,6]$  as

$$P[4 < X < 6] = P\left[\frac{4-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{6-\mu}{\sigma}\right] = P[.33 < Z < 1] = \Phi(1) - \Phi(.33)$$

To compute this in EViews we will create each of the cumulative probabilities, and then subtract.

```
scalar phi1=@cnorm(1)
scalar phi2=@cnorm(.33)
scalar prob = phi1-phi2
```

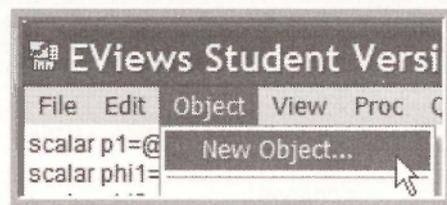
The calculated value of the probability, which we have called **PROB** = .2120.

Scalar PROB = 0.212044726913

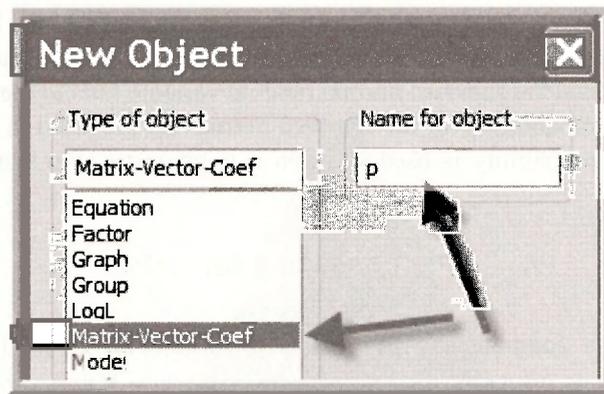
**Remark:** When writing EViews commands, try to make the names somewhat relevant to the algebraic form or the context of the problem. This will help you recall what you were doing when looking back at it later.

## B.2 USING VECTORS

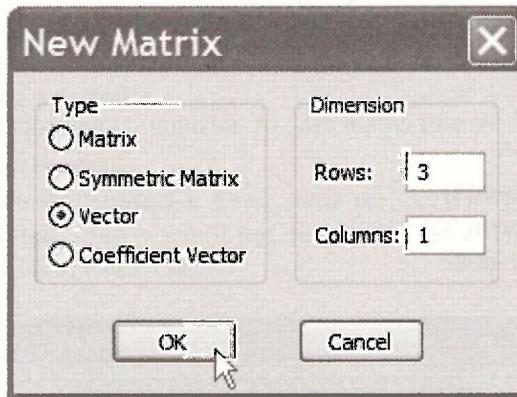
The approach above obviously works, but if you wish to have the results in a convenient form for a paper, using **coefficient vectors** is an option. Click on **Object/New Object** on the EViews menubar.



Click on **Matrix-Vector-Coeff**, giving the new object the name **P**.



Make this object a vector, which is just a single column of numbers, with 3 rows. Click **OK**.



Instead of pointing and clicking, you can simply type **vector(3) p** into the command line and press **Enter**.

Type into the command line the following series of statements

```
p(1)=(6-3)/3
p(2)=(4-3)/3
p(3)=@cnorm(p(1))-@cnorm(p(2))
```

With each command you will see an entry appear in the vector **P**, with the final entry being the probability you seek.

P	
	C1
R1	1.000000
R2	0.333333
R3	0.210786

Last updated: 11/16/07 - 14:57

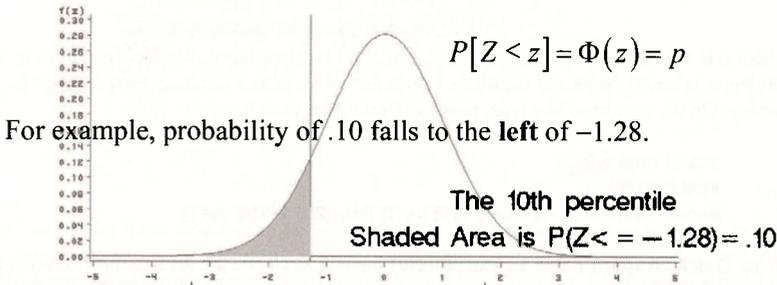
The advantage of this approach is that now we can **Freeze** this screen, and then **Name** it (**P\_TABLE**) and it will appear in our workbook. Furthermore, the contents of this table can be copied (highlight then **Ctrl+C**) and pasted (**Ctrl+V**) into a document

R1	1.000000
R2	0.333333
R3	0.210786

The result is a table that can be formatted, edited, etc.

### **B.3 COMPUTING NORMAL DISTRIBUTION PERCENTILES**

The function **@qnorm(p)** returns the percentile value **z** from a standard normal probability density, such that



Percentiles that identify regions containing a certain probability are often called **critical values**. To illustrate, type in the command

```
scalar z10 = @qnorm(.10)
```

That is, we are asking what is the value **z** from a standard normal distribution such that

$$P[Z < z] = \Phi(z) = .10$$

The value of the scalar **Z10** is

```
 Scalar Z10 = -1.28155156554
```

EViews will compute similar probabilities for many types of random variables. Click on **Help/Function Reference**. Scroll down to the section entitled *Statistical Distribution Functions*. There you will find a long list of probability distributions for which EViews can compute cumulative probabilities and percentiles. You have heard of some, like the **Binomial**. Many others will not be familiar. We will use several of these distributions in later chapters, such as the **Chi-square**, the **F-distribution** and **Student's t-distribution**.

#### **B.4 PLOTTING SOME NORMAL DISTRIBUTIONS**

It is useful to plot some normal distributions to see their shapes and locations. Recall that we set the sample size of the workfile to 101. Create a variable **X** that covers the  $[-5, 5]$  interval in increments of  $1/10^{\text{th}}$ . Type into the command line (or use the **Genr** button)

```
series x = -5 + @trend/10
```

Double-click on the generated series to verify.

The formula for the normal probability density function is given in Equation (B.26) of *POE*. It is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

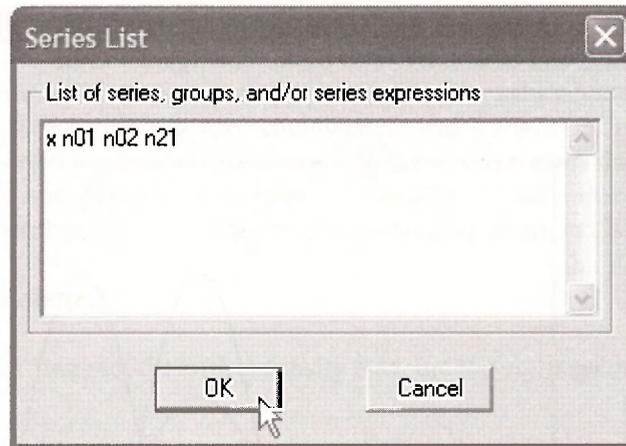
where  $\mu$  is the mean of the random variable and  $\sigma^2$  is its variance. We are simply going to type this formula into EViews and substitute values for the mean and variance. First we plot the  $N(0,1)$  density. On the command line type, pressing **Enter** after each line

```
scalar mu0 = 0
scalar var1 = 1
series n01 = exp( -(x-mu0)^2/(2*var1) )/sqrt(2*3.14159*var1)
```

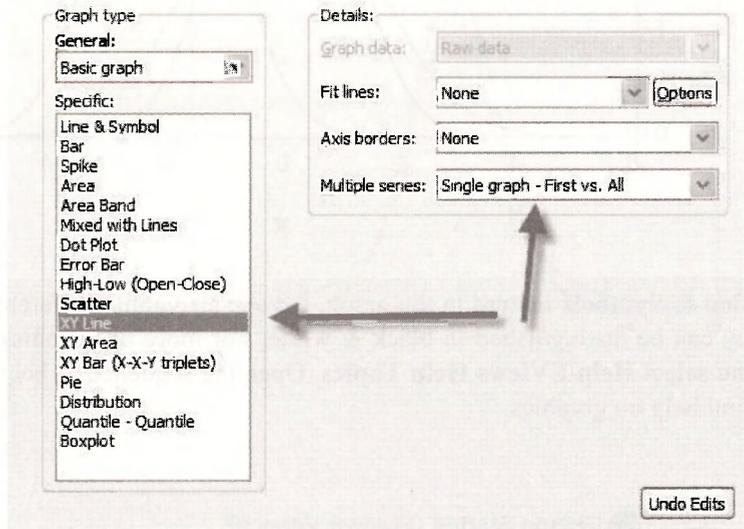
Using **Quick/Graph/XY line** we can plot **N01** against **X**. Create the density functions of a  $N(0,2)$  and a  $N(2,1)$ . To do so we can simply edit the items on the command line and press **Enter**. The commands are

```
scalar mu2 = 2
scalar var2 = 2
series n02 = exp( -(x-mu0)^2/(2*var2) )/sqrt(2*3.14159*var2)
series n21 = exp( -(x-mu2)^2/(2*var1) )/sqrt(2*3.14159*var1)
```

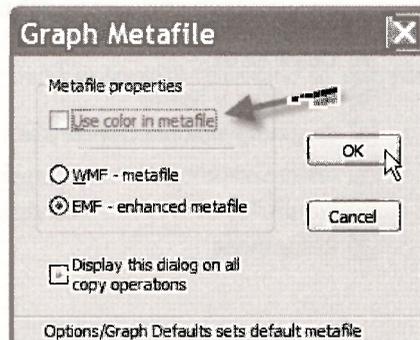
Using **Quick/Graph** we can plot all the graphs in the same picture. In the dialog box enter



In the **Graph Options** dialog box choose

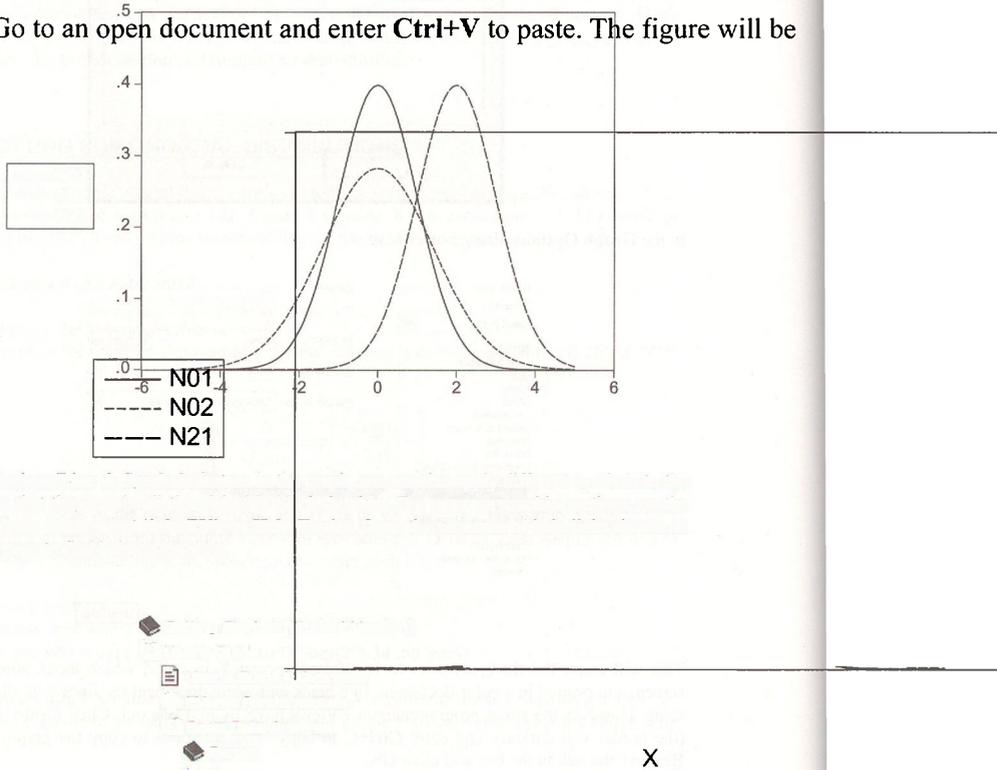


This will cause the three densities to be plotted against **X** in color, which looks good on the screen, and printed in a color document. In a black and white document the lines will all look the same. However the smart programmers at EViews have figured this out. Click inside the figure (the border will darken), and enter **Ctrl+C** to copy. The default is to copy the graph in color. **Remove** the tick in the box and click **OK**.



## 334 Appendix

Go to an open document and enter **Ctrl+V** to paste. The figure will be



Using **Options/Line & Symbols** feature in the graph, we can also apply different symbols to the curves so that they can be distinguished in black & white. For more on graphics options, on the main EViews menu select **Help/EViews Help Topics**. Open the sequence of help links shown on the next page to find help on graphics.

Getting Started (Student Version)

📖 User's Guide

User's Guide I Overview

📖 EViews Fundamentals

📖 Basic Data Analysis

Series

📖 Groups

📖 Graphing Data

Quick Start

Graphing a Series

Graphing Multiple Series (Groups)

Basic Customization

Graph Types

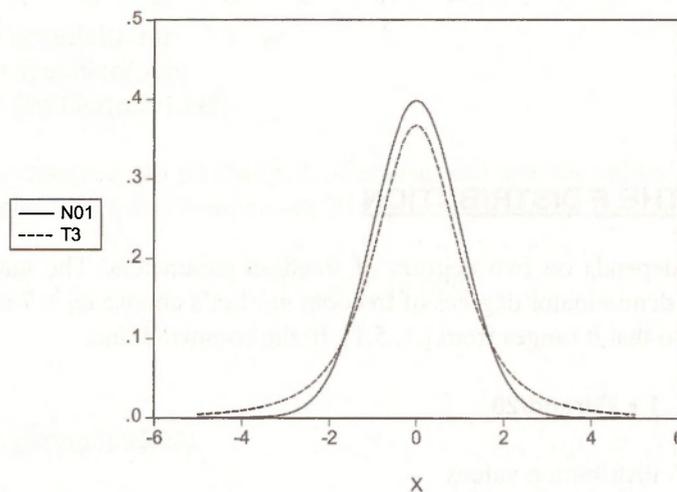
References

## B.5 PLOTTING THE T-DISTRIBUTION

The distribution that you will be using perhaps the most is the  $t$ -distribution. It is discussed in Section B.5.3 in *POE*. The formula is very complicated, and we will not report it, but EViews allows us to plot the distribution easily. The function `@dtdist(x,v)` returns the value of the density function of a  $t$ -random variable with  $v$  degrees of freedom for the value  $X$ . To generate the density values for a  $t$ -distribution with 3 degrees of freedom, the command is

```
series t3 = @dtdist(x,3)
```

Using **Quick/Graph/XY line plot** the  $N(0,1)$  density **N01** and the  $t_{(3)}$  density on the same graph.



## B.6 PLOTTING THE CHI-SQUARE DISTRIBUTION

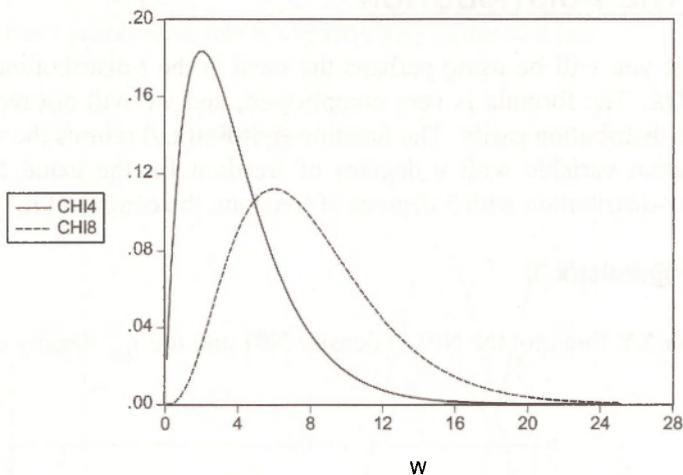
The  $F$  and Chi-square distributions are only defined for positive values. Let's create a new variable **W** that takes values in the interval  $[.1, 25.1]$  in increments of 0.25. These are simply positive values that can be used to construct the graphs.

```
series w = .1 + @trend/4
```

For the chi-square density we need to specify the value of one parameter, its degrees of freedom,  $m$ , which is also its mean.

```
series chi4 = @dchisq(w,4)
series chi8 = @dchisq(w,8)
```

Plot these values against **W**



### B.7 PLOTTING THE F DISTRIBUTION

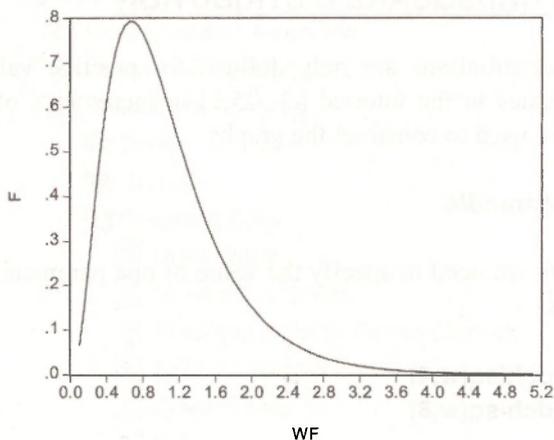
The *F*-distribution depends on two degrees of freedom parameters: The numerator degrees of freedom  $m_1$  and the denominator degrees of freedom  $m_2$ . Let's choose  $m_1 = 7$  and  $m_2 = 30$ . Define **WF** for these plots so that it ranges from [.1, 5.1]. In the command line,

```
series wf = .1 + @trend/20
```

Then generate the *F*- distribution values

```
series F = @dfdist(wf,7,30)
```

and plot.



## **B.8 PROBABILITY CALCULATIONS FOR THE T, F AND CHI-SQUARE**

Probability calculations for distributions other than the normal are set up the same way using functions beginning with **@c**. Thus we can compute, for some specified value  $x$ ,

The probability that  $t_{(m)} \leq x$  use the command scalar **pt = @ctdist(c,m)**

The probability that  $\chi_{(m)}^2 \leq x$  use the command scalar **pc = @cchisq(x,m)**

The probability that  $F_{(m_1, m_2)} \leq x$  use the command scalar **pf = @cfdist(x,m1,m2)**

We can compute the critical values such that the probability **p** falls to the left using for the three distributions the commands using quantile functions that begin with **@q**:

**scalar tc = @qtdist(p,m)**

**scalar cc = @qchisq(p,m)**

**scalar fc = @qfdist(p,m1,m2)**

The names we have assigned can be changed, of course, and specific values must be used for **c**, **m** and **p**. For example, for a  $t$ -distribution with 20 degrees of freedom, the value  $t_c$  such that

$$P[t_{(20)} \leq t_c] = 0.95$$

use the command

**scalar tc = @qtdist(.95,20)**

The resulting value is 1.7247, which you may compare with the value in Table 2 of *POE*.

### **Keywords**

@cchisq

@cfdist

@cnorm

@ctdist

@dchisq

@dfdist

@dtdist

@qchisq

@qfdist

@qnorm

@trend

chi-square distribution

cumulative distribution

F distribution

normal distribution

probability density function

quantile functions

scalar

t distribution

vector

XY Line graph

# APPENDIX C

## Review of Statistical Inference

### CHAPTER OUTLINE

C.1 A Histogram

C.2 Summary Statistics

C.2.1 The sample mean

C.2.2 Estimating higher moments

C.2.3 Create a table

C.2.4 Using the estimates

C.3 Interval Estimation

C.4 Hypothesis Tests About the Population Mean

C.4.1 One-tail test using the hip data

C.4.2 Two-tail test using the hip data

C.4.3 Testing the normality of the population

KEYWORDS

### C.1 A HISTOGRAM

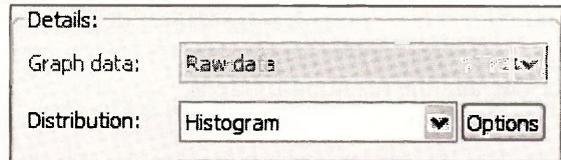
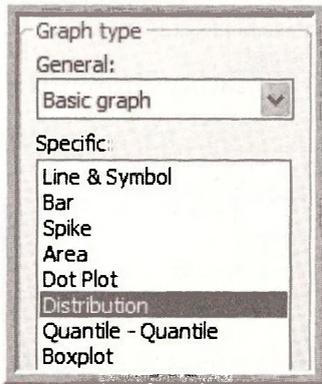
Open the EViews workfile *hip.wfl*. Double click on *Y*, which is hip width, to open the series into a spreadsheet view. Select **View/Descriptive Statistics & Tests/Stats Table**

Sample: 1 50

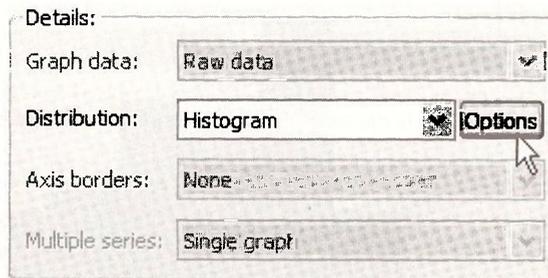
---

Mean	17.15820
Median	17.08500
Maximum	20.40000
Minimum	13.53000
Std. Dev.	1.807013
Skewness	-0.013825
Kurtosis	2.331534
Jarque-Bera	0.932523
Probability	0.627343
Sum	857.9100
Sum Sq. Dev.	159.9995
Observations	50

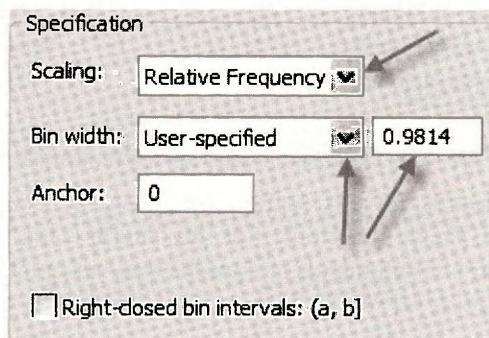
To construct a histogram, from the spreadsheet select **View/Graph**. Under the specific graph choose **Distribution**.



To have the figure closely resemble Figure C.1, select **Options**

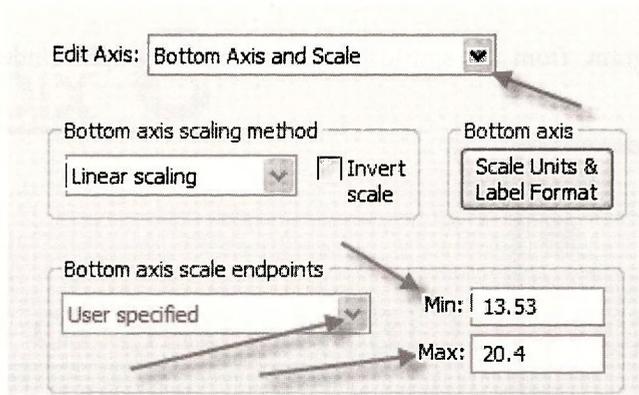


Choose **Relative Frequency** for **Scaling** and choose the **User-specified Bin width** of 0.9814. This seemingly odd choice is the maximum value (20.4) minus the minimum value (13.53) divided by 7, which will be the number of bins (or figure bars).

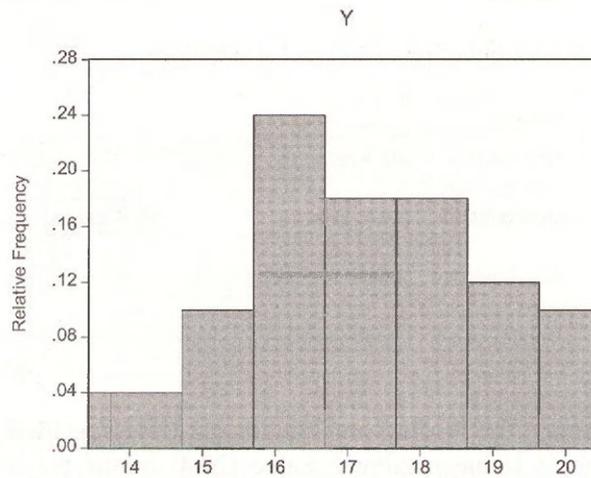


Click **OK**.

On the **Axis/Scale** tab select the **Bottom Axis and Scale**. For the **Bottom axis scale endpoints** select **User specified** and use the data minimum and maximum values



Select OK and the resulting figure is



## C.2 SUMMARY STATISTICS

The sample mean  $\bar{y} = \sum y_i / N = 17.1582$  is shown in the summary statistics table, along with several other summary measures. In this section we illustrate how to create a wide range of summary of statistics with EViews functions. For any single statistic there are perhaps several ways to do the calculation, so our commands are not the only acceptable ones.

### C.2.1 The sample mean

The sample mean is discussed in *POE* Section C.3. On page 505 we find the calculation for the “hip” data. Enter the commands

```
scalar ysum = @sum(y)
scalar n = @obs(y)
scalar ybar = @mean(y)
```

*sum of y values*  
*number of observations*  
*sample mean*

The scalar values created are:

Scalar YSUM = 857.91

Scalar N = 50

Scalar YBAR = 17.1582

### C.2.2 Estimating higher moments

In *POE* Section C.4.3 the hip data higher moments are estimated. The sample variance for the hip data is

$$\hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2}{N-1} = \frac{\sum (y_i - 17.1582)^2}{49} = \frac{159.9995}{49} = 3.2653$$

We can obtain this value as follows

**scalar ssy = @sumsq(y-@mean(y))**

*sum of squared deviations*

**scalar sig2 = @vars(y)**

*sample variance*

Scalar SSY = 159.999538

Scalar SIG2 = 3.26529669388

This means that the estimated variance of the sample mean is

$$\widehat{\text{var}}(\bar{Y}) = \frac{\hat{\sigma}^2}{N} = \frac{3.2653}{50} = .0653$$

and the standard error of the mean is

$$\text{se}(\bar{Y}) = \hat{\sigma} / \sqrt{N} = .2556$$

**scalar var = sig2/n**

*estimated variance of sample mean*

**scalar sig = @stdev(y)**

*standard deviation of y*

**scalar se = sig/sqr(n)**

*standard error of sample mean*

Scalar VAR = 0.0653059338776

Scalar SIG = 1.80701319693

Scalar SE = 0.255550257048

The estimated skewness is  $S = -.0138$  and the estimated kurtosis is  $K = 2.3315$  using

$$\hat{\sigma} = \sqrt{\sum (Y_i - \bar{Y})^2 / N} = \sqrt{159.9995/50} = 1.7889$$

$$\hat{\mu}_3 = \sum (Y_i - \bar{Y})^3 / N = -.0791$$

$$\hat{\mu}_4 = \sum (Y_i - \bar{Y})^4 / N = 23.8748$$

We can compute the skewness and kurtosis directly using built-in EViews functions

scalar sk = @skew(y)  
 scalar k = @kurt(y)

skewness  
 kurtosis

Scalar SK = -0.0138249736168

Scalar K = 2.33153416027

The intermediate calculations, shown on *POE* page 512, are

scalar sigtilde = @stdevp(y)  
 series y3 = (y-ybar)^3  
 scalar mu3 = @sum(y3)/n  
 series y4 = (y-ybar)^4  
 scalar mu4 = @sum(y4)/n

square root of sss divided by n  
 series deviations about mean cubed  
 3<sup>rd</sup> moment about mean  
 deviations about mean to fourth pwr  
 4<sup>th</sup> moment about mean

Scalar SIGTILDE = 1.78885179934

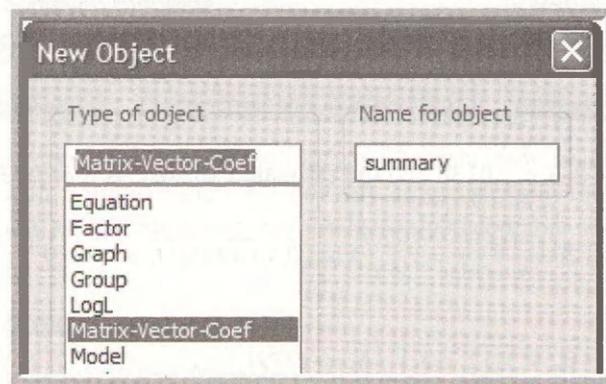
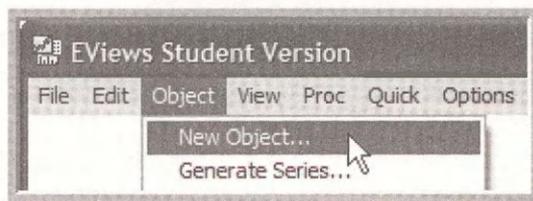
Scalar MU3 = -0.079138424064

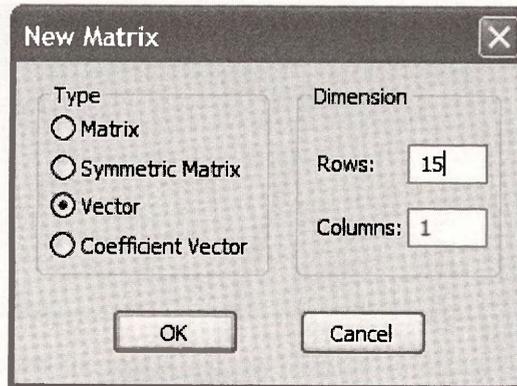
Scalar MU4 = 23.8747719237

Thus the hip data is slightly negatively skewed, and is slightly less peaked than would be expected for a normal distribution.

### C.2.3 Create a table

Rather than printing all these scalars it would be preferable in a report to create a **vector object** and store the results, creating a table. Create a new called **summary**.





Select **OK**. Enter the following series of commands to assign quantities to positions in **summary**.

```
summary(1)=ybar
summary(2)=ssy
summary(3)=n
summary(4)=sig2
summary(5)=sig
summary(6)=var
summary(7)=se
summary(8)=sigtilde
summary(9)=mu3
summary(10)=mu4
summary(11)=sk
summary(12)=k
```

To create a table, **Freeze** the results

	C1
R1	17.15820
R2	159.9995
R3	50.00000
R4	3.265297
R5	1.807013
R6	0.065306
R7	0.255550
R8	1.788852
R9	-0.079138
R10	23.87477
R11	-0.013825
R12	2.331534
R13	0.000000
R14	0.000000
R15	0.000000

Name the table.

	A	B	C	D	E
1		C1			
2		Last updated: 11/25/07 - 10:53			
3					
4	R1	17.15820			
5	R2	159.9995			
6	R3	50.00000			
7	R4	3.265297			
8	R5	1.807013			
9	R6	0.065306			
10	R7	0.255550			
11	R8	1.788852			
12	R9	-0.079138			
13	R10	23.87477			
14	R11	-0.013825			
15	R12	2.331534			
16	R13	0.000000			
17					

**Object Name**

Name to identify object

summary\_stats\_table 24 characters maximum, 16 or fewer recommended

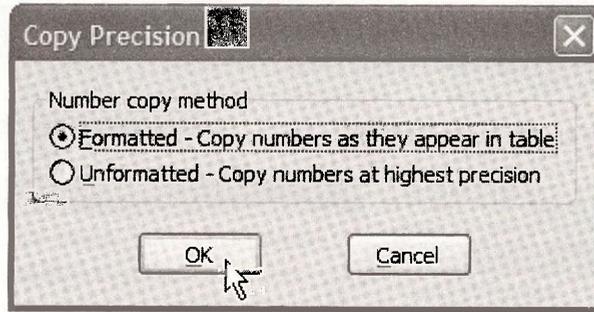
Display name for labeling tables and graphs (optional)

OK Cancel

Highlight table contents and enter **Ctrl+C** to copy to Windows Clipboard.

R1	17.15820		
R2	159.9995		
R3	50.00000		
R4	3.265297		
R5	1.807013		
R6	0.065306		
R7	0.255550		
R8	1.788852		
R9	-0.079138		
R10	23.87477		
R11	-0.013825		
R12	2.331534		
R13	0.000000		
R14	0.000000		
R15	0.000000		

Choose **Formatted** numbers.



Paste the contents into an open document, and the table will look like

R1	17.15820
R2	159.9995
R3	50.00000
R4	3.265297
R5	1.807013
R6	0.065306
R7	0.255550
R8	1.788852
R9	-0.079138
R10	23.87477
R11	-0.013825
R12	2.331534

Then edit the table contents with informative labels

Sample mean	17.15820
Sum of squares	159.9995
Sample size	50.00000
Sample variance	3.265297
Sample Std. Dev.	1.807013
Variance of mean	0.065306
Std. Err. of mean	0.255550
Root 2 <sup>nd</sup> moment	1.788852
Mu3	-0.079138
Mu4	23.87477
Skewness	-0.013825
Kurtosis	2.331534

We will be using this workfile more in this chapter, so **Save** the workfile as *hip\_appx\_c.wfl*.

### C.2.4 Using the estimates

In *POE* Section C.4.4 some calculations are carried out based on the estimates. First

$$\widehat{P}(Y > 18) \cong P\left(\frac{Y - \bar{y}}{\hat{\sigma}} > \frac{18 - 17.158}{1.8070}\right) = P(Z > .4659) = .3207$$

Use `@cnorm` to compute the cumulative normal probability.

`scalar pygt18 = 1 - @cnorm(.4659)`

`Scalar PYGT18 = 0.320643537685`

How wide would an airplane seat have to be to fit 95% of the population? If we let  $y^*$  denote the required seat size, then

$$\widehat{P}(Y \leq y^*) \cong P\left(\frac{Y - \bar{y}}{\hat{\sigma}} \leq \frac{y^* - 17.1582}{1.8070}\right) = P\left(Z \leq \frac{y^* - 17.1582}{1.8070}\right) = .95$$

The value of  $Z$  such that  $P(Z \leq z^*) = .95$  is  $z^* = 1.645$  which is computed using `@qnorm`.

`scalar z95 = @qnorm(.95)`

`Scalar Z95 = 1.64485362695`

Then the calculation of  $y^*$  is

$$\frac{y^* - 17.1582}{1.8070} = 1.645 \Rightarrow y^* = 20.1305$$

`scalar ystar = ybar + sig*z95`

`Scalar YSTAR = 20.1304722109`

### C.3 INTERVAL ESTIMATION

We have introduced the empirical problem faced by an airplane seat design engineer. Given a random sample of size  $N = 50$  we estimated the mean U.S. hip width to be

$$\bar{y} = 17.158 \text{ inches} = \mathbf{YBAR}$$

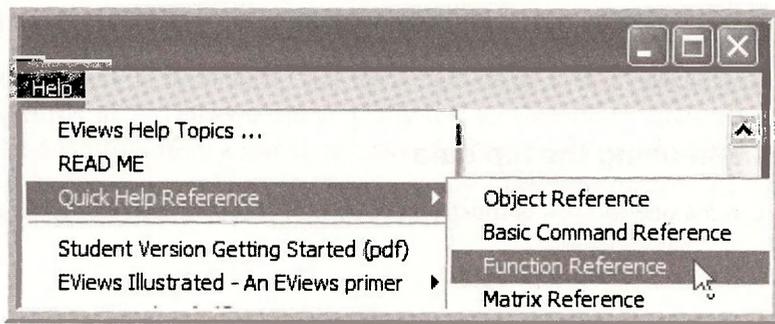
Furthermore we estimated the population variance to be  $\hat{\sigma}^2 = 3.265$ , thus the estimated standard deviation is  $\hat{\sigma} = 1.807$ . The standard error of the mean is

$$\hat{\sigma}/\sqrt{N} = 1.807/\sqrt{50} = .2556 = \mathbf{SE}$$

The critical value for interval estimation comes from a  $t$ -distribution with  $N-1 = 49$  degrees of freedom. While this value is not in Table 2, the correct value is

$$t_c = t_{(.975, 49)} = 2.0095752$$

which we round to  $t_c = 2.01$ . This value is found using the EViews command `@qtdist(p,v)` which calculates the **p-quantile**, or **percentile**, of the  $t$ -distribution with  $v$  degrees of freedom. Click **Help** on the EViews main menu, then **Quick Help Reference/Function Reference**



The **Statistical Distribution Functions** are described. For the *t*-distribution these are shown to be

Student's <i>t</i> -distribution	@ctdist(x,v), @dtdist(x,v), @qtdist(p,v), @rtdist(v)
----------------------------------	---

The names represent:

Function Type	Beginning of Name
Cumulative distribution (CDF)	@c
Density or probability	@d
Quantile (inverse CDF)	@q
Random number generator	@r

Thus the 97.5 percentile of the *t*-distribution with 49 degrees of freedom is obtained using

**scalar t975 = @qtdist(.975,49)**       Scalar T975 = 2.00957523713

The 100(1- $\alpha$ )% interval estimator for  $\mu$  is

$$\bar{Y} \pm t_c \frac{\hat{\sigma}}{\sqrt{N}} \text{ or } \bar{Y} \pm t_c \text{se}(\bar{Y}) \tag{C.15}$$

For the Hip data, use (C.15), replacing estimates for the estimators, to give

$$\bar{y} \pm t_c \frac{\hat{\sigma}}{\sqrt{N}} = 17.1582 \pm 2.01 \frac{1.807}{\sqrt{50}} = [16.6447, 17.6717]$$

**scalar lb = ybar - t975\*se**       Scalar LB = 16.6446525316  
**scalar ub = ybar + t975\*se**       Scalar UB = 17.6717474684

## C.4 HYPOTHESIS TESTS ABOUT THE POPULATION MEAN

### C.4.1 One-tail test using the hip data

In *POE* Section C.6.5 a one-tail test is illustrated. The null and alternative hypotheses are

$$H_0 : \mu = 16.5 \qquad H_1 : \mu > 16.5$$

The test statistic is

$$t = \frac{\bar{Y} - 16.5}{\hat{\sigma} / \sqrt{N}} \sim t_{(N-1)}$$

The value of the test statistic is calculated using

`scalar t1 = (ybar - 16.5)/se`

`Scalar T1 = 2.57561861844`

The right tail  $\alpha = 0.05$  critical value of the  $t$ -distribution with 49 degrees of freedom is

`scalar t95 = @qtdist(.95,49)`

`Scalar T95 = 1.67655089262`

The  $p$ -value is the probability that a  $t$ -statistic with 49 degrees of freedom is greater than **T1 = 2.5756**. Recalling the definition of a **cumulative distribution function**, this value is given by

`scalar p1 = 1 - @ctdist(t1,49)`

`Scalar P1 = 0.00653694452567`

### C.4.2 Two-tail test using the hip data

The null hypothesis is  $H_0 : \mu = 17$ . The alternative hypothesis is  $H_1 : \mu \neq 17$ . The test statistic is

$$t = \frac{\bar{Y} - 17}{\hat{\sigma} / \sqrt{N}} \sim t_{(N-1)}$$

For a test at the  $\alpha = 0.05$  level, the critical value is the 97.5 percentile of the  $t_{(49)}$  distribution which we used in Section C.3 above in the interval estimate calculation and named **t975**.

`scalar t2 = (ybar - 17)/se`

`Scalar T2 = 0.619056313334`

`scalar p2 = 2*(1 - @ctdist(abs(t2),49) )`

`Scalar P2 = 0.53874691505`

The complicated formula for the  $p$ -value is useful because its general setup will work for any two-tail test. It is twice the area in the right tail of the  $t$ -distribution beyond the **absolute value** of the  $t$ -statistic.

### C.4.3 Testing the normality of the population

The normal distribution is symmetric, and has a bell-shape with a peakedness and tail-thickness leading to a kurtosis of 3. Thus we can certainly test for departures from normality by checking the skewness and kurtosis from a sample of data. If skewness is not close to zero, and if kurtosis is not close to 3, then we would reject the normality of the population. In Appendix C.4.2 we developed sample measures of skewness and kurtosis

$$\widehat{\text{skewness}} = S = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3}$$

$$\widehat{\text{kurtosis}} = K = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$$

The **Jarque-Bera** test statistic allows a joint test of these two characteristics,

$$JB = \frac{N}{6} \left( S^2 + \frac{(K-3)^2}{4} \right)$$

If the true distribution is symmetric and has kurtosis 3, which includes the normal distribution, then the *JB* test statistic has a chi-square distribution with 2 degrees of freedom if the sample size is sufficiently large. If  $\alpha = .05$  then the critical value of the  $\chi_{(2)}^2$  distribution is 5.99. We reject the null hypothesis and conclude that the data are non-normal if  $JB \geq 5.99$ . To compute the critical value and *p*-value for the Chi-square distribution refer to help on **Statistical Distribution Functions**.

Chi-square	@cchisq(x, v), @dchisq(x, v), @qchisq(p, v), @rchisq(v)
------------	--

Using the cumulative and quantile functions we obtain

```
scalar jb = (n/6)*(sk^2 + (1/4)*(k-3)^2)
```

```
scalar pjb = 1 - @cchisq(jb,2)
```

```
scalar chic = @qchisq(.95,2)
```

```
 Scalar JB = 0.93252312182
```

```
 Scalar PJB = 0.627343174134
```

```
 Scalar CHIC = 5.99146454711
```

The **Jarque-Bera** statistic is calculated automatically by EViews. Open the series *Y*, select **View/Descriptive Statistics & Tests/Histogram and Stats**.

Series: Y	
Sample 1 50	
Observations 50	
Mean	17.15820
Median	17.08500
Maximum	20.40000
Minimum	13.53000
Std. Dev.	1.807013
Skewness	-0.013825
Kurtosis	2.331534
Jarque-Bera	0.932523
Probability	0.627343

## Keywords

@cchisq  
 @cnorm  
 @ctdist  
 @kurt  
 @mean  
 @obs  
 @qchisq  
 @qnorm  
 @qtdist  
 @skew  
 @stdev  
 @sum

@sumsq  
 @vars  
 chi-square distribution  
 confidence interval  
 critical value  
 descriptive statistics  
 freeze  
 histogram  
 hypothesis test  
 interval estimates  
 Jarque-Bera test  
 kurtosis

one-tail test  
 p-value  
 sample mean  
 sample variance  
 skewness  
 standard deviation  
 standard error of mean  
 Std. Dev.  
 Std. Err.  
 t-distribution  
 two-tail test  
 vector

# INDEX

- @abs, 319
- @cchisq, 114, 337, 349
- @cfdist, 114, 337
- @cnorm, 277, 327, 346
- @coefs, 62, 103
- @cor, 76, 173
- @cov, 120
- @ctdist, 61, 105, 337, 348
- @dchisq, 335
- @dfdists, 336
- @dnorm, 277
- @dtdist, 335
- @exp, 319
- @kurt, 342
- @mean, 115, 340
- @obs, 340
- @qchisq, 81, 114, 337, 349
- @qfdist, 114, 166, 337
- @qnorm, 331, 346
- @qtdist, 62, 72, 103, 346
- @regobs, 72
- @se, 72, 161
- @skew, 342
- @sqrt, 120, 158
- @ssdep, 75
- @ssr, 75, 113
- @stderrs, 62, 103
- @stdev, 341
- @sum, 173, 340
- @sumsq, 115, 341
- @trend, 133, 192, 320
- @vars, 341
- abs, 68
- absolute value, 67
- AC, 181
- actual, 148
- add text, 153
- apply, 152
- AR(1) error, 175
- AR(1), 175
- ARCH test, 236
- ARDL, 190
- arithmetic operators, 33
- asymmetric GARCH, 243
- autocorrelation, 180
- autoregressive models, 184
- axis/scale, 192
- basic graph, 11, 172
- binary choice models, 269
- Breusch-Pagan test, 167
- c object, 91
- censored data, 296
- chi-square distribution, 335, 349
- chi-square statistic, 110
- chi-square test, 110, 204
- Chow test, 140
- close series, 14
- coefficient tests, 104
- coefficient uncertainty in S.E., 101
- coefficient vector, 47, 62, 132
- coefficient, 103
- cointegration, 224
- collinearity, 128
- commands: saving, 108
- common coefficients, 254
- common sample, 271
- confidence interval, 346
- contemporaneous correlation, 255
- convergence, 176
- copy precision, 17
- copying a table, 17
- copying graph, 13
- correlated random effects, 267
- correlation matrix, 127
- correlation, 26, 172
- correlogram, 179
- count data models, 286
- covariance analysis, 127
- covariance matrix, 55, 102, 120,
- critical value, 62, 348
- cross section random, 266
- cross-equation restrictions, 256
- cross-section coefficients, 254
- cross-section identifiers, 253
- cross-section SUR, 254
- Ctrl+C, 13, 17
- Ctrl+V, 14, 17
- cubic equation, 84
- cumulative distribution, 326
- d: difference operator, 219
- data definition files, 2
- data range, 9, 82
- data sample, 82
- data scaling, 76
- date conventions, 313
- date conversion, 313
- date specification, 185
- delay multipliers, 190
- delta method, 145, 179
- demand equation, 213
- descriptive statistics, 10, 15, 39, 93, 122, 338
- df, 113
- Dickey-Fuller tests, 223
- dim, 258
- dummy variables, 134, 251, 259
- Durbin-Watson test, 183
- dynamic forecasting, 188
- edit +/-, 57, 100
- edit axis, 192
- effects specification, 261
- elasticity, 48, 79, 293
- end date, 184
- endogeneity problem, 198
- endogenous regressor, 198
- endogenous variables, 212
- entering data, 316
- equation name.ls, 83, 138
- equation representations, 47
- equation save, 46
- equation specification, 95
- error correction, 230
- error variance, 54

- estimate equation, 46, 95, 134
- estimation settings, 160
- EViews functions, 27
- EViews size limitations, 287
- exogenous variables, 212
- exp, 319
- exponential function, 87, 145
- exponential, 320
- exporting data, 318
- F distribution, 336
- finite distributed lags, 189
- fitted terms, 125
- fitted, 148
- fixed effects, 258
- fixed effects: testing, 260
- forecast name, 158, 188
- forecast sample, 101
- forecast stand error, 73, 101, 188
- forecast, 58, 73, 97, 126, 158, 196
- forecasting, 188
- freeze, 16, 172, 343
- frequency, 185, 310
- F-statistic, 115
- F-test, 108, 110, 163
- function reference, 5, 33
- F-value, 108, 118
- GARCH, 242
- Garch-in-mean, 245
- generalized least squares, 159
- generalized R-squared, 89
- generate series, 20, 53
- genr 21, 53
- Goldfeld-Quandt test, 163
- goodness-of-fit, 75
- graph axes/scale, 42
- graph copy to document, 44
- graph metafile, 14
- graph object, 149
- graph options, 11, 41, 150, 195
- graph regression line, 49
- graph save, 44
- graph symbol pattern, 43
- graph title, 42
- group, 122
- group: empty, 23
- group: naming, 93
- group: open, 38, 92
- HAC standard errors, 174
- Hausman test, 201, 266
- heckit, 298
- help, 4
- heteroskedastic partition, 159
- heteroskedasticity tests, 166
- histogram, 11, 81, 339
- hypothesis test, 64, 104, 348
- hypothesis test: left-tail, 65
- hypothesis test: one-tail, 66
- hypothesis test: right-tail, 64
- hypothesis test: two-tail, 67
- identification, 233
- identified, 203
- identifier series, 251
- idiosyncratic random, 266
- importing data: Excel, 310
- importing data: text, 314
- impulse responses, 233
- IMR, 299
- inconsistent estimator, 194
- index model, 280
- individual samples, 292
- instrument list, 199, 213
- instrumental variables, 199, 213
- instruments, 199
- integer date, 185
- interactions, 136, 143
- interim multipliers, 190
- internet data, 305
- interval estimates, 61, 102, 346
- inverse Mills ratio, 299
- Jarque-Bera test, 81, 349
- kurtosis, 341
- lag specification, 180
- lag weights, 191
- lagging a series, 172
- Lagrange multiplier test, 182
- latent variables, 279
- least squares line: plot, 152
- least squares, 95, 200
- legend, 153
- limit values, 281
- line/shade, 151, 193
- line/symbol, 153, 172, 197
- linear probability model, 271
- linear-log model, 79
- log function, 78, 144
- logarithm, 322
- logical generate, 259
- logit, 279
- log-linear model, 78
- log-log model, 79
- LR statistic, 275
- LS & TSLS options, 154, 159, 174
- ls, 77
- make GARCH variance, 241
- make residual series, 225, 228
- marginal effect, 131, 277, 284
- math functions, 34
- maximum likelihood, 273
- McFadden R-squared, 275
- mean dependent variable, 54
- mean specification, 239
- Monte Carlo, 194
- Mroz data, 204
- multiple graphs, 17, 220
- NA, 100, 172, 288
- name, 19
- new page, 249
- Newey-West standard errors, 174
- NLS panel, 263
- nonlinear hypothesis, 145
- nonlinear least squares, 176
- nonsample information, 122
- nonstationary variables, 220
- normal distribution, 327
- normality test, 80
- normalized restriction, 112, 179
- normalized restriction, 179
- null hypothesis: joint, 117
- null hypothesis: single, 112
- object name, 16, 40, 93
- object: creating, 92
- object: equation, 96
- object: group, 92
- object: text, 109
- one-tail test, 348

- open group, 15
- open series, 10
- operators, 136
- order of Integration, 224
- ordered choice models, 279
- ordered probit, 280
- orientation, 193
- outliers, 148
- over-identified, 203
- page destination, 250
- page: naming, 91
- page: resize, 99
- panel options, 252
- panel structure, 258
- path, 7
- plots, 148
- Poisson regression, 286
- polynomials, 130
- pool object, 252
- pooled EGLS, 255
- pooled least squares, 260
- pooling, 253
- prediction evaluation, 284
- prediction interval, 72
- prediction, 71
- prediction: corrected, 87
- prediction: log-linear model, 86
- prediction: natural, 87
- Prob(F-statistic), 115
- Prob., 67
- probability density function, 326
- probability forecast, 275
- probit, 273
- proc, 99
- p-value (Prob.), 104, 113
- p-value, 64, 348
- quantile functions, 326
- quick help reference, 27
- quick/empty group, 23
- quick/estimate equation, 45, 94
- quick/generate series, 20
- quick/graph, 21, 40
- quick/group statistics, 26
- quick/sample, 20
- quick/series statistics, 25
- quick/show, 25
- random effects, 265
- random regressors, 184
- range, 91
- range: change, 99
- reduced form equation, 212
- reduced form, 200
- redundant fixed effects, 262
- regression output, 95
- rename page, 249
- representations, 137
- RESET test, 124
- reshape page, 250
- resid, 47, 149, 171
- residual correlation matrix, 255
- residual correlogram, 179
- residual graph, 148, 171
- residual plot, 84
- residual table, 53
- residual tests, 166, 180
- residuals, 52, 147
- restricted least squares, 123
- R-squared, 75
- S.D. dependent variable, 54, 96
- S.E. of regression, 54, 96
- sample (adjusted), 178
- sample mean, 340
- sample range, 9, 56, 142
- sample range: change, 12, 20, 99
- sample variance, 341
- sample, 91, 160, 287
- sample: if condition, 322
- Sargan statistic, 203
- scalar, 27, 48, 62, 98
- scatter diagram/plot, 18, 41, 195
- scatter, 150
- scientific notation, 320
- seemingly unrelated regr, 254
- semi-elasticity, 292
- serial correlation LM test, 182
- series, 9, 133
- series: delete, 24
- series: rename, 24
- significance test, 64, 104, 111
- singular matrix, 264
- skewness, 341
- sort, 164
- spreadsheet view, 10
- spreadsheet, 38, 93
- spurious regression, 221
- sqr, 319
- SSE: restricted, 111
- SSE: unrestricted, 111
- stability tests, 125
- stack in new page, 249
- stacked data, 249
- stacking identifiers, 250
- standard deviation, 341
- standard error of mean, 341
- standard errors, 54, 102
- standard errors: White, 154
- standardized residual graph, 148
- start date, 184
- stationarity tests, 222
- stationary variables, 220
- Std. Dev., 338
- Std. Err., 345
- Std. Error, 55, 102
- stochastic, 194
- sum squared resid, 54, 113
- supply equation, 214
- SUR, 254
- surplus instruments, 203
- sym, 120
- symmetric matrix, 120
- system of equations, 214
- t-distribution, 335, 346
- t-distribution CDF, 61
- t-distribution critical value, 64
- test of significance, 64, 104, 111
- test: nonzero value, 106
- test: one-tail, 105
- test: two-tails, 104
- threshold GARCH, 243
- threshold values, 280
- time series data, 170
- time-varying volatility, 234
- tobit, 296
- transformed variables, 156
- TSLS, 199

## 354 Index

t-statistic, 67  
t-test, 104  
t-value (t-Statistic), 104  
two-stage least squares, 199  
two-tail test, 348  
type: graph, 150  
undated with ID series, 258  
unit root test of residuals, 226  
unit root tests of variables, 222  
unstacked data, 248  
unstructured/undated, 184  
validity of surplus instruments, 203  
VAR, 230  
variance decomposition, 233  
variance function, 157  
variance function: testing, 166  
variance specification, 239  
VEC, 227  
vector, 30, 329, 342  
wage equation, 204  
Wald coefficient restrictions, 112  
Wald test, 108, 112, 132, 179, 203  
weight, 155  
weighted least squares, 155  
weighted LS/TLS, 155  
White cross terms, 168  
White test, 168  
workfile stack, 250  
workfile structure, 82, 185, 258  
workfile, 7, 91  
workfile: open, 8, 36  
workfile: save, 22, 45  
XY line, 154, 192, 323