

# Using Excel

For *Principles of Econometrics, Third Edition*

Version 1.0

# Using Excel

*For Principles of Econometrics, Third Edition*

**ASLI K. OGUNC**

*Texas A&M University-Commerce*

**R. CARTER HILL**

*Louisiana State University*



**JOHN WILEY & SONS, INC**

*New York / Chichester / Weinheim / Brisbane / Singapore / Toronto*

*Asli Ogunc dedicates this work to Patara*

*Carter Hill dedicates this work to Todd and Peter*

|                          |                 |
|--------------------------|-----------------|
| ACQUISITIONS EDITOR      | XXXXXX XXXXXXXX |
| MARKETING MANAGER        | XXXXXX XXXXXXXX |
| PRODUCTION EDITOR        | XXXXXX XXXXXXXX |
| PHOTO EDITOR             | XXXXXX XXXXXXXX |
| ILLUSTRATION COORDINATOR | XXXXXX XXXXXXXX |

This book was set in Times New Roman and printed and bound by .XXXXXX XXXXXXXXXXXX. The cover was printed by .XXXXXXXXX XXXXXXXXXXXXXXXX

This book is printed on acid-free paper. ∞

The paper in this book was manufactured by a mill whose forest management programs include sustained yield harvesting of its timberlands. Sustained yield harvesting principles ensure that the numbers of trees cut each year does not exceed the amount of new growth.

Copyright © John Wiley & Sons, Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (508) 750-8400, fax (508) 750-4470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

ISBN 0-471-xxxxx-x

Printed in the United States of America

**10 9 8 7 6 5 4 3 2 1**

# PREFACE

This book is a supplement 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 Excel 2003. This book will be useful to students taking econometrics, as well as their instructors, and others who wish to use Excel for econometric analysis.

In addition to this computer manual for Excel, there are similar manuals and support for the software packages EViews, Excel, Gretl, Shazam and Stata. In addition, all the data for *POE* in various formats, including Excel, are available at <http://www.wiley.com/college/hill>.

Individual Excel data files, errata for this manual and the textbook can be found at <http://www.bus.lsu.edu/hill/poe>. Templates for routine tasks can also be found at this web site.

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 Excel manual sections.

We welcome comments on this book, and suggestions for improvement. \*

Asli K. Ogunc  
Department of Accounting, Economics and Finance  
Texas A&M University-Commerce  
Commerce, TX 75429  
[Asli\\_Ogunc@tamu-commerce.edu](mailto:Asli_Ogunc@tamu-commerce.edu)

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

\* Microsoft product screen shot(s) reprinted with permission from Microsoft Corporation. Our use does not directly or indirectly imply Microsoft sponsorship, affiliation, or endorsement.

# **BRIEF CONTENTS**

|                                                                                           |            |
|-------------------------------------------------------------------------------------------|------------|
| 1. Introducing Excel                                                                      | <b>1</b>   |
| 2. Simple Linear Regression                                                               | <b>21</b>  |
| 3. Interval Estimation and Hypothesis Testing                                             | <b>39</b>  |
| 4. Prediction, Goodness of Fit and Modeling Issues                                        | <b>49</b>  |
| 5. Multiple Linear Regression                                                             | <b>68</b>  |
| 6. Further Inference in the Multiple Regression Model                                     | <b>80</b>  |
| 7. Nonlinear Relationships                                                                | <b>104</b> |
| 8. Heteroskedasticity                                                                     | <b>125</b> |
| 9. Dynamic Models, Autocorrelation, and Forecasting                                       | <b>142</b> |
| 10. Random Regressors and Moment Based Estimation                                         | <b>155</b> |
| 11. Simultaneous Equations Models                                                         | <b>169</b> |
| 12. Nonstationary Time Series Data and Cointegration                                      | <b>178</b> |
| 13. An Introduction to Macroeconometrics: VEC and VAR Models                              | <b>188</b> |
| 14. An Introduction to Financial Econometrics: Time-Varying Volatility<br>and ARCH Models | <b>200</b> |
| 15. Panel Data Models                                                                     | <b>204</b> |
| 16. Qualitative and Limited Dependent Variable Models                                     | <b>215</b> |
| 17. Importing Internet Data                                                               | <b>219</b> |
| Appendix B. Review of Probability Concepts                                                | <b>223</b> |

## CONTENTS

### CHAPTER 1 Introduction to Excel 1

- 1.1 Starting Excel 1
- 1.2 Entering Data 4
- 1.3 Using Excel for Calculations 6
  - 1.3.1 Arithmetic operations 8
  - 1.3.2 Mathematical functions 9
- 1.4 Excel Files for Principles of Econometrics 15
  - 1.4.1 John Wiley & Sons website 15
  - 1.4.2 Principles of Econometrics website 16
  - 1.4.3 Definition files 16
  - 1.4.4 The food expenditure data 16

### CHAPTER 2 The Simple Linear Regression Model 21

- 2.1 Plotting the Food Expenditure Data 21
- 2.2 Estimating a Simple Regression 28
- 2.3 Plotting a Simple Regression 31
- 2.4 Plotting the Least Squares Residuals 35
- 2.5 Prediction Using Excel 36

### CHAPTER 3 Interval Estimation and Hypothesis Testing 39

- 3.1 Interval Estimation 39
  - 3.1.1 Automatic interval estimates 39
  - 3.1.2 Constructing interval estimates 41
- 3.2 Hypothesis Testing 43
  - 3.2.1 Right-Tail tests 43
  - 3.2.2 Left-Tail tests 45
  - 3.2.3 Two-Tail tests 46

### CHAPTER 4 Prediction, Goodness-of-Fit and Modeling Issues 49

- 4.1 Prediction for the Food Expenditure Model 49
  - 4.1.1 Calculating the standard error of the forecast 49
  - 4.1.2 Prediction interval 52

- 4.2 Measuring Goodness-of-Fit 53
  - 4.2.1 Calculating  $R^2$  54
  - 4.2.2 Covariance and correlation analysis 54
- 4.3 Residual Diagnostics 57
  - 4.3.1 The Jarque-Bera test for normality 59
- 4.4 Modeling Issues 60
  - 4.4.1 Scaling the data 60
  - 4.4.2 The log-linear model 61
  - 4.4.3 The linear-log model 62
  - 4.4.4 The log-log model 62
- 4.5 More Examples 63
  - 4.5.1 Residual analysis with wheat data 63
  - 4.5.2 Log-linear model with wage data 65
  - 4.5.3 Generalized  $R^2$  67

### CHAPTER 5 Multiple Linear Regression 68

- 5.1 Big Andy's Burger Barn 68
- 5.2 Prediction 70
- 5.3 Sampling Precision 71
- 5.4 Confidence Intervals 76
- 5.5 Hypothesis Testing 77
- 5.6 Goodness-of-Fit 78

### CHAPTER 6 Further Inference in the Multiple Regression Model 80

- 6.1 The  $F$ -test 80
- 6.2 Testing the Overall Significance of the Model 84
- 6.3 An Extended Model 85
- 6.4 Testing Some Economic Hypotheses 86
  - 6.4.1 The Significance of advertising 86
  - 6.4.2 Optimal level of advertising 87
- 6.5 Nonsample Information 90
- 6.6 Model Specification 93
  - 6.6.1 Omitted variables 93
  - 6.6.2 Irrelevant variables 94
  - 6.6.3 Choosing the model 94
- 6.7 Poor Data, Collinearity and Insignificance 99

## **CHAPTER 7 Nonlinear Relationships 104**

- 7.1 Nonlinear Relationships 104
  - 7.1.1 Summarize data and estimate regression 104
  - 7.1.2 Calculating a marginal effect 106
- 7.2 Dummy Variables 107
  - 7.2.1 Creating dummy variables 107
  - 7.2.2 Estimating a dummy variable regression 107
  - 7.2.3 Testing the significance of the dummy variables 109
  - 7.2.4 Further calculations 109
- 7.3 Applying Dummy Variables 110
  - 7.3.1 Interactions between qualitative factors 110
  - 7.3.2 Adding regional dummy variables 114
  - 7.3.3 Testing the equivalence of two regressions 116
- 7.4 Interactions Between Continuous Variables 119
- 7.5 Dummy Variables in Log-linear Models 121

## **CHAPTER 8 Heteroskedasticity 125**

- 8.1 The Nature of Heteroskedasticity 125
- 8.2 Using the Least Squares Estimator 126
- 8.3 The Generalized Least Squares Estimator 128
  - 8.3.1 Transforming the model 128
  - 8.3.2 Estimating the variance function 130
  - 8.3.3 A heteroskedastic partition 131
- 8.4 Detecting Heteroskedasticity 134
  - 8.4.1 Residual plots 134
  - 8.4.2 The Goldfeld-Quandt test 135
  - 8.4.3 Testing the variance function 139

## **CHAPTER 9 Dynamic Models, Autocorrelation, and Forecasting 142**

- 9.1 Lags in the Error Term 142
- 9.2 Area Response for Sugar 143

- 9.3 Estimating an AR(1) Model 145
  - 9.3.1 Least squares 145
- 9.4 Detecting Autocorrelation 148
  - 9.4.1 The Durbin-Watson test 148
  - 9.4.2 An LM test 149
- 9.5 Autoregressive Models 150
- 9.6 Finite Distributed Lags 151
- 9.7 Autoregressive Distributed Lag (ARDL) Model 153

## **CHAPTER 10 Random Regressors and Moment Based Estimation 155**

- 10.1 Least Squares with Simulated Data 155
- 10.2 Instrumental Variables Estimation with Simulated Data 157
  - 10.2.1 Correction of IV standard errors 159
  - 10.2.2 Corrected standard errors for simulated data 160
- 10.3 The Hausman Test: Simulated Data 162
- 10.4 Testing for Weak Instruments: Simulated Data 163
- 10.5 Testing for Validity of Surplus Instruments: Simulated Data 164
- 10.6 Estimation using Mroz Data 164
  - 10.6.1 Least squares regression 164
  - 10.6.2 Two-stage least squares 165
- 10.7 Testing the Endogeneity of Education 167
- 10.8 Testing for Weak Instruments 167
- 10.9 Testing the Validity of Surplus Instruments 168

## **CHAPTER 11 Simultaneous Equations Models 169**

- 11.1 Truffle Supply and Demand 169
- 11.2 Estimating the Reduced Form Equations 170
- 11.3 2SLS Estimates of Truffle Demand and Supply 171
  - 11.3.1 Correction of 2SLS standard errors 173
  - 11.3.2 Corrected standard errors in truffle demand

- and supply 174
- 11.4 Supply and Demand of Fish 175
- 11.5 Reduced Forms for Fish Price  
and Quantity 176

## **CHAPTER 12 Nonstationary Time-Series Data and Cointegration 178**

- 12.1 Stationary and Nonstationary  
Data 178
- 12.2 Spurious Regression 179
- 12.3 Unit Root Test for  
Stationarity 181
- 12.4 Integration and  
Cointegration 185
- 12.5 Engle-Granger Test 185

## **CHAPTER 13 An Introduction to Macroeconometrics: VEC and VAR Models 188**

- 13.1 VEC and VAR Models 188
- 13.2 Estimating a VEC Model 188
- 13.3 Estimating VAR 197

## **CHAPTER 14 An Introduction to Financial Econometrics: Time-Varying Volatility and ARCH Models 200**

- 14.1 ARCH Model and Time Varying  
Volatility 200
- 14.2 Testing, Estimating and  
Forecasting 202

## **CHAPTER 15 Panel Data Models 205**

- 15.1 Sets of Regression Equations 205
- 15.2 Seemingly Unrelated Regressions 209
  - 15.2.1 Breusch-Pagan test of  
independence 209
- 15.3 The Fixed Effects Model 210
  - 15.3.1 A dummy variable model 211
- 15.4 Random Effects Estimation 214

## **CHAPTER 16 Qualitative and Limited Dependent Variable Models 215**

- 16.1 Models with Binary Dependent  
Variables 215
  - 16.1.1 The linear probability  
model 216
  - 16.1.2 Least squares estimation of the  
linear probability model 216

## **CHAPTER 17 Importing Internet Data 219**

## **Appendix B Review of Probability Concepts 223**

- B.1 Binomial Probabilities 223
  - B.1.1 Computing binomial  
probabilities directly 223
  - B.1.2 Computing binomial  
probabilities using  
BINOMDIST 225
- B.2 The Normal Distribution 227



# CHAPTER 1

## Introduction to Excel

### CHAPTER OUTLINE

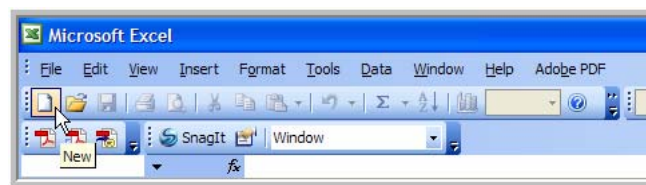
- 1.1 Starting Excel
- 1.2 Entering Data
- 1.3 Using Excel for Calculations
  - 1.3.1 Arithmetic operations
  - 1.3.2 Mathematical functions
- 1.4 Excel Files for Principles of Econometrics
  - 1.4.1 John Wiley & Sons website
  - 1.4.2 Principles of Econometrics website
  - 1.4.3 Definition files
  - 1.4.4 The food expenditure data

### 1.1 STARTING EXCEL

Start Excel by clicking the Start menu and locating the program, or by clicking a shortcut, such as,

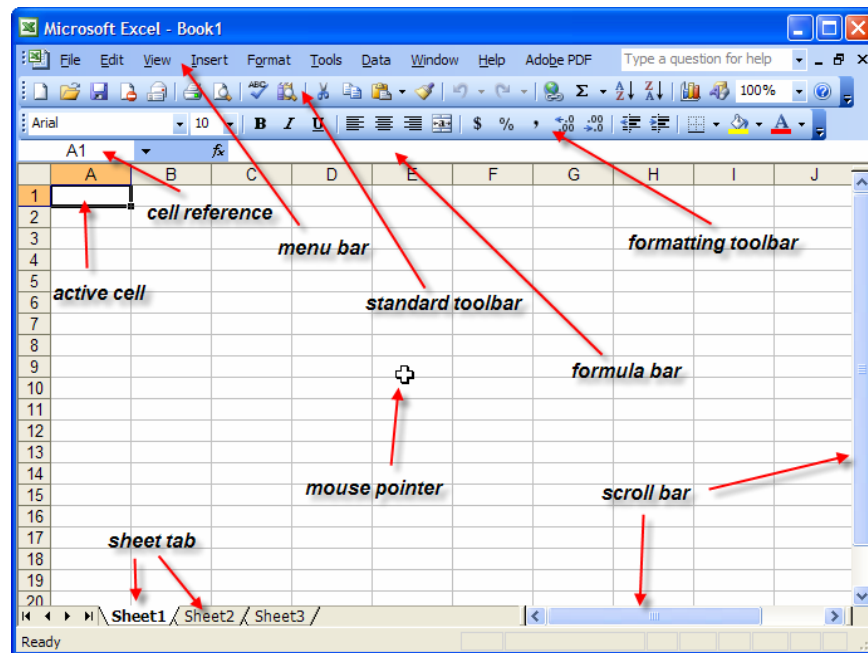


Excel opens. Click on the **New Workbook** icon.



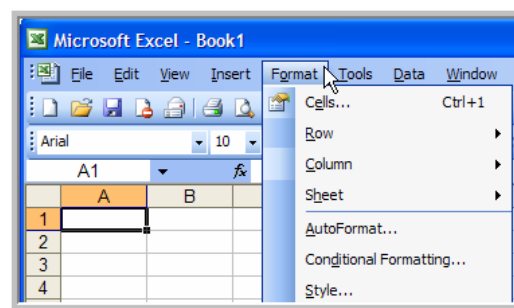
The worksheet looks like this

## 2 Chapter 1

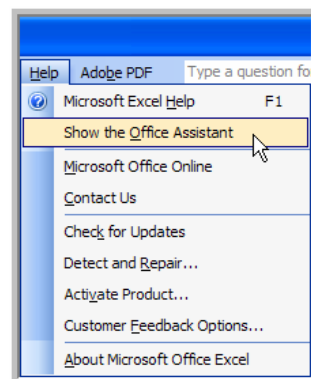


There are lots of little bits that you will become more familiar with as we go along. The active cell is surrounded by a border and is in Column A and Row 1. We will refer to cells as A1, B1 and so on.

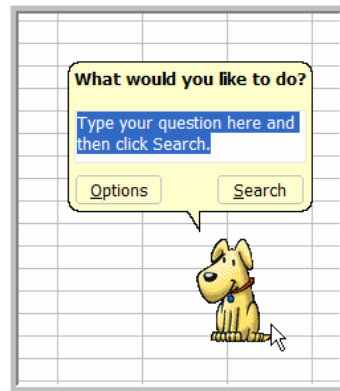
Across the top of the window is a **Menu bar**. Sliding the mouse over the items opens up a pull down menu, showing further options.



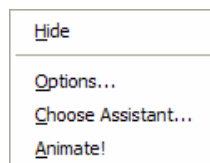
Perhaps the most important of all these is **Help**.



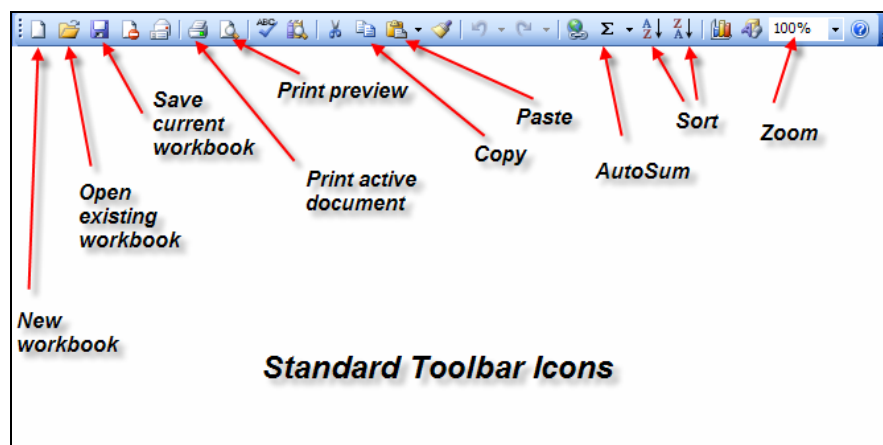
Microsoft Office has a cute (or annoying) feature. You can have a little assistant showing on the screen. If you click your assistant you can type in a question or search.



If you **right-click** on the critter you can choose to hide him or change his personality. We will hide him.

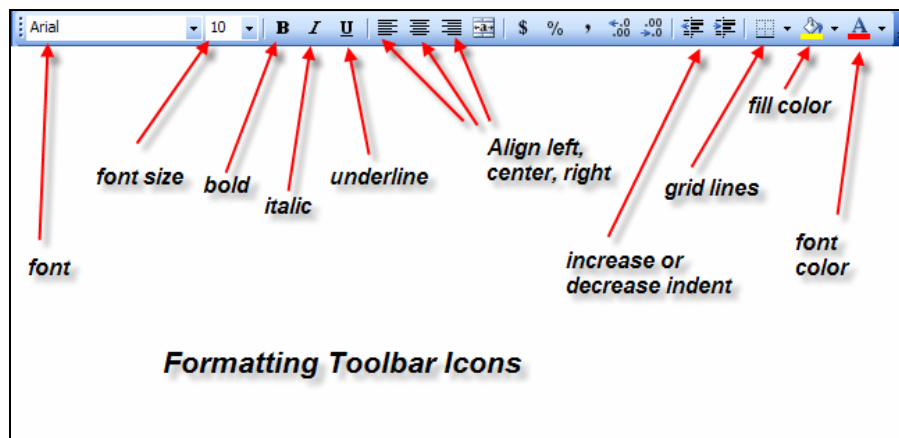


The **Standard Toolbar** has the usual Microsoft functions **New**, **Open**, **Save**, **Print**, **Print preview**, **Copy** and **Paste**. The **AutoSum** key is a feature of Excel, and the **Sort** buttons allow you to order data according to the magnitude of one of your columns.



The **Formatting Toolbar** has the usual functions. The use of **Grid lines** can clarify a worksheet, as can the use of colored fonts and filling in cells for emphasis.

## 4 Chapter 1



### 1.2 ENTERING DATA

We will use Excel to analyze data. To enter data into an Excel worksheet move the cursor to a cell and type. First enter *X* in cell A1 and *Y* in cell B1. Navigate by moving the cursor with the mouse, or use the **Tab** key to move right or left, or **Arrow** keys to move right, left, up or down.

|   | A | B | C |
|---|---|---|---|
| 1 | X | Y |   |
| 2 |   |   |   |
| 3 |   |   |   |
| 4 |   |   |   |

In A2, enter the number 1. Press **Enter** and it will take you to the next cell, fill in the rest as shown.

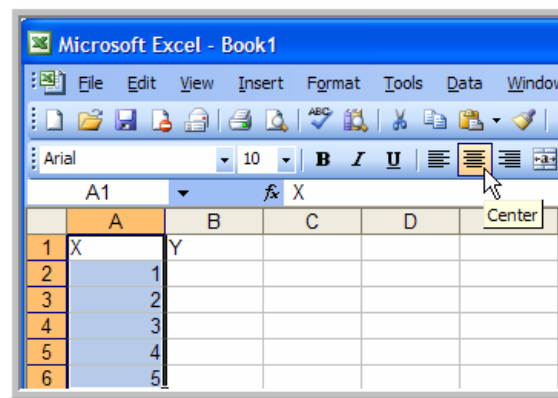
|   | A | B |
|---|---|---|
| 1 | X | Y |
| 2 |   | 1 |
| 3 |   | 2 |
| 4 |   | 3 |
| 5 |   | 4 |
| 6 |   | 5 |

To nicely center the data in the cells, highlight cells A1:A6. There are several ways to highlight the cells. For small areas the easiest way is to place cursor in A1, hold down the left mouse button and drag it across the area you wish to highlight. For larger areas, using a key-stroke combination is very convenient.

- To highlight a column—place cursor in A1. Hold down **Ctrl**-key and **Shift**-key simultaneously, which we will denote as **Ctrl-Shift**. Press the **down arrow** ↓ on the keyboard.
- To highlight a row—place cursor in A1. Press **Ctrl-Shift** and **right arrow** → on the keyboard.

- To highlight a region—place cursor in A1. Press **Ctrl-Shift** then the **down arrow** and then the **right arrow**.

After selecting A1:A6, click **Center**



The result is

|   | A | B |
|---|---|---|
| 1 | X | Y |
| 2 | 1 |   |
| 3 | 2 |   |
| 4 | 3 |   |
| 5 | 4 |   |
| 6 | 5 |   |

Repeat this process for B1:B6. This centering is just for appearance, and has no affect on any functionality.

|   | A | B | C |
|---|---|---|---|
| 1 | X | Y |   |
| 2 | 1 | 2 |   |
| 3 | 2 | 3 |   |
| 4 | 3 | 6 |   |
| 5 | 4 | 5 |   |
| 6 | 5 | 7 |   |
| 7 |   |   |   |

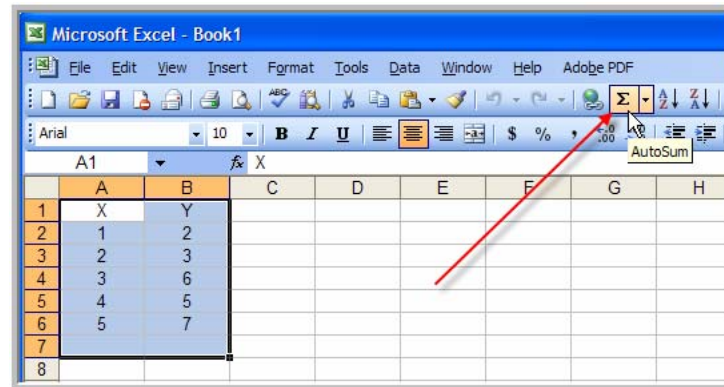
To highlight the entire worksheet, click the square in the upper left corner.

 A screenshot of an Excel worksheet showing columns A, B, and C, and rows 1 through 8. A red arrow points to the small square in the upper left corner (the intersection of the row and column headers). The text "To highlight worksheet" is overlaid on the image.
 

|   | A | B | C |
|---|---|---|---|
| 1 |   |   |   |
| 2 |   |   |   |
| 3 |   |   |   |
| 4 |   |   |   |
| 5 |   |   |   |
| 6 |   |   |   |
| 7 |   |   |   |
| 8 |   |   |   |

### 1.3 USING EXCEL FOR CALCULATIONS

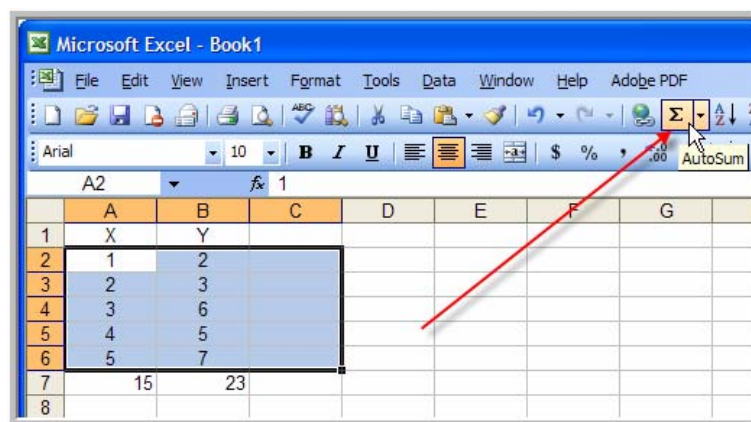
What is Excel good for? Its primary usefulness is to carry out repeated calculations. We can add, subtract, multiply and divide; and we can apply mathematical and statistical functions to the data in our worksheet. To illustrate, highlight Columns A & B, down to row 7, as shown below. Click **AutoSum**. This will sum the rows and place the sum in the final row.



The result its

|   | A  | B  | C |
|---|----|----|---|
| 1 | X  | Y  |   |
| 2 | 1  | 2  |   |
| 3 | 2  | 3  |   |
| 4 | 3  | 6  |   |
| 5 | 4  | 5  |   |
| 6 | 5  | 7  |   |
| 7 | 15 | 23 |   |
| 8 |    |    |   |

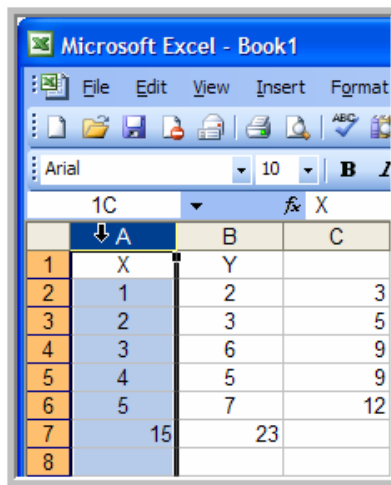
Carry out a similar sequence of steps to sum the columns. Highlight rows 2:6 and columns A:C. Click **AutoSum**.



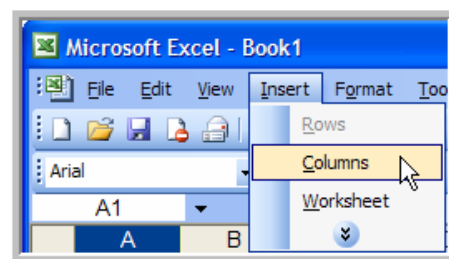
The column sum ( $X + Y$ ) is now in column C.

|   | A  | B  | C  |
|---|----|----|----|
| 1 | X  | Y  |    |
| 2 | 1  | 2  | 3  |
| 3 | 2  | 3  | 5  |
| 4 | 3  | 6  | 9  |
| 5 | 4  | 5  | 9  |
| 6 | 5  | 7  | 12 |
| 7 | 15 | 23 |    |

To **Insert Columns** hold the cursor over the A in column A. The cursor turns into a down arrow and the entire first column is highlighted.



On the Excel **Menu** select **Insert/Columns**. This will insert a new column to the left of the highlighted column.



Enter a column head, which will serve to identify what is in the first row, and enter “Sum” to identify the contents of row 7.

|   | A        | B  | C  | D  |
|---|----------|----|----|----|
| 1 | Variable | X  | Y  |    |
| 2 |          | 1  | 2  | 3  |
| 3 |          | 2  | 3  | 5  |
| 4 |          | 3  | 6  | 9  |
| 5 |          | 4  | 5  | 9  |
| 6 |          | 5  | 7  | 12 |
| 7 | Sum      | 15 | 23 |    |

## 8 Chapter 1

Add a header in column D.

|   | A        | B  | C  | D   |
|---|----------|----|----|-----|
| 1 | Variable | X  | Y  | Sum |
| 2 |          | 1  | 2  | 3   |
| 3 |          | 2  | 3  | 5   |
| 4 |          | 3  | 6  | 9   |
| 5 |          | 4  | 5  | 9   |
| 6 |          | 5  | 7  | 12  |
| 7 | Sum      | 15 | 23 |     |

### 1.3.1 Arithmetic operations

Standard arithmetical functions are defined as follows (from Excel **Help**, type **arithmetic operators**)

**Arithmetic operators:** To perform basic mathematical operations such as addition, subtraction, or multiplication; combine numbers; and produce numeric results, use the following arithmetic operators.

#### Arithmetic operator

+ (plus sign)

– (minus sign)

\* (asterisk)

/ (forward slash)

% (percent sign)

^ (caret)

#### Meaning (Example)

Addition (3+3)

Subtraction (3–1)Negation (–1)

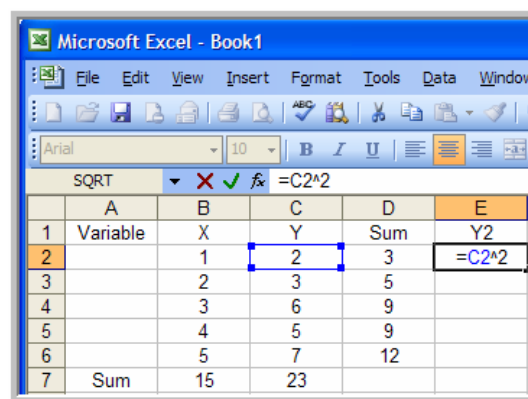
Multiplication (3\*3)

Division (3/3)

Percent (20%)

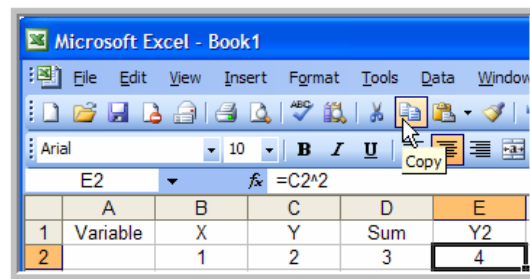
Exponentiation (3^2)

To create a new variable Y2, select cell E2 and enter the formula **=C2^2**. This command instructs Excel to square the value in cell C2.



The new value is Y2 = 4. Select this cell, and click **Copy** (or the shortcut **Ctrl+C**)





The border of the cell begins to rotate. Move the cursor to the lower right corner of the cell until the “plus” sign appears.

|   | A        | B  | C  | D   | E  |
|---|----------|----|----|-----|----|
| 1 | Variable | X  | Y  | Sum | Y2 |
| 2 |          | 1  | 2  | 3   | 4  |
| 3 |          | 2  | 3  | 5   |    |
| 4 |          | 3  | 6  | 9   |    |
| 5 |          | 4  | 5  | 9   |    |
| 6 |          | 5  | 7  | 12  |    |
| 7 | Sum      | 15 | 23 |     |    |

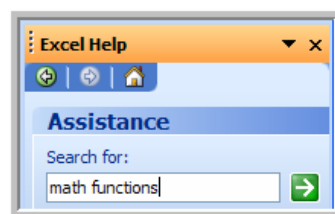
Hold the left mouse button down and **drag** the plus down to cell E6.

|   | A        | B | C | D   | E  |
|---|----------|---|---|-----|----|
| 1 | Variable | X | Y | Sum | Y2 |
| 2 |          | 1 | 2 | 3   | 4  |
| 3 |          | 2 | 3 | 5   | 9  |
| 4 |          | 3 | 6 | 9   | 36 |
| 5 |          | 4 | 5 | 9   | 25 |
| 6 |          | 5 | 7 | 12  | 49 |

Release the left button and you will find the new values of Y2. What you have done is **Copy** the formula in E2 to the new cells, and Excel calculates the square of each Y value.

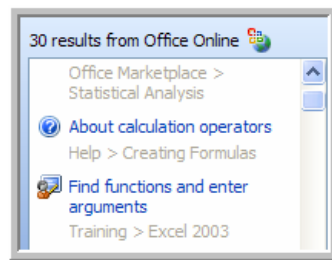
### 1.3.2 Mathematical functions

There are a large number of mathematical functions, most of which you will never use. Find the right function is fairly simple using **Help**. Search for math functions.

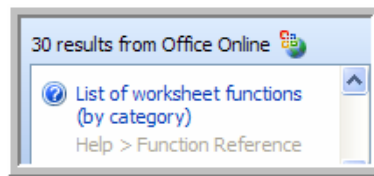


Among the results returned is “Find functions and enter arguments” which is an online training session video, which you might find useful.

## 10 Chapter 1



Another entry is “List of worksheet functions”. A partial list of the functions available is listed on the next page. These are taken from the Excel help result.



## Math and trigonometry functions

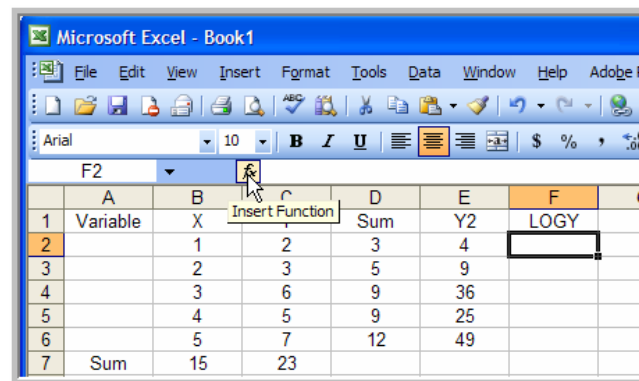
| Function    | Description                                                                        |
|-------------|------------------------------------------------------------------------------------|
| ABS         | Returns the absolute value of a number                                             |
| CEILING     | Rounds a number to the nearest integer or to the nearest multiple of significance  |
| COMBIN      | Returns the number of combinations for a given number of objects                   |
| EVEN        | Rounds a number up to the nearest even integer                                     |
| EXP         | Returns $e$ raised to the power of a given number                                  |
| FACT        | Returns the factorial of a number                                                  |
| FACTDOUBLE  | Returns the double factorial of a number                                           |
| FLOOR       | Rounds a number down, toward zero                                                  |
| GCD         | Returns the greatest common divisor                                                |
| INT         | Rounds a number down to the nearest integer                                        |
| LCM         | Returns the least common multiple                                                  |
| LN          | Returns the natural logarithm of a number                                          |
| LOG10       | Returns the base-10 logarithm of a number                                          |
| ODD         | Rounds a number up to the nearest odd integer                                      |
| PI          | Returns the value of pi                                                            |
| POWER       | Returns the result of a number raised to a power                                   |
| PRODUCT     | Multiplies its arguments                                                           |
| QUOTIENT    | Returns the integer portion of a division                                          |
| RAND        | Returns a random number between 0 and 1                                            |
| RANDBETWEEN | Returns a random number between the numbers you specify                            |
| ROUND       | Rounds a number to a specified number of digits                                    |
| ROUNDDOWN   | Rounds a number down, toward zero                                                  |
| ROUNDUP     | Rounds a number up, away from zero                                                 |
| SERIESSUM   | Returns the sum of a power series based on the formula                             |
| SIGN        | Returns the sign of a number                                                       |
| SQRT        | Returns a positive square root                                                     |
| SQRTPI      | Returns the square root of (number * pi)                                           |
| SUBTOTAL    | Returns a subtotal in a list or database                                           |
| SUM         | Adds its arguments                                                                 |
| SUMIF       | Adds the cells specified by a given criteria                                       |
| SUMPRODUCT  | Returns the sum of the products of corresponding array components                  |
| SUMSQ       | Returns the sum of the squares of the arguments                                    |
| SUMX2MY2    | Returns the sum of the difference of squares of corresponding values in two arrays |
| SUMX2PY2    | Returns the sum of the sum of squares of corresponding values in two arrays        |
| SUMXMY2     | Returns the sum of squares of differences of corresponding values in two arrays    |
| TRUNC       | Truncates a number to an integer                                                   |

## 12 Chapter 1

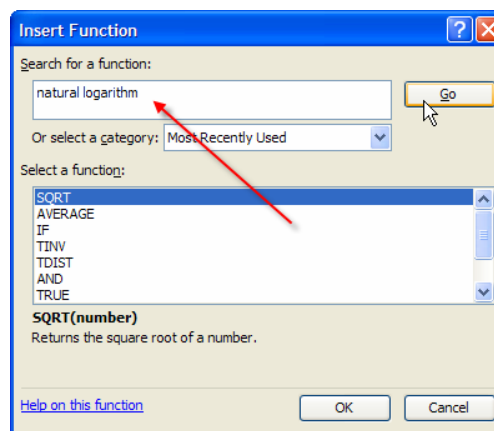
Of course there is no way to remember all these. Again however Excel makes the functions easy to use. There is an **Insert Function** (or **Paste Function**) button.



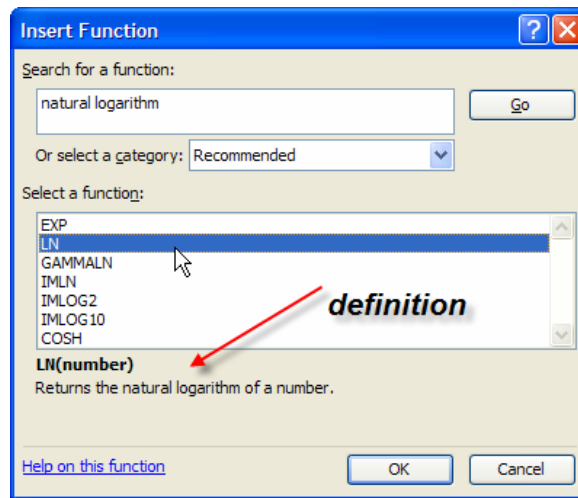
Create a new heading in F1, *LOGY*, which will contain the natural logarithm of Y. All logs used in *POE* are natural logs. Highlight F2 and click **Insert Function**.



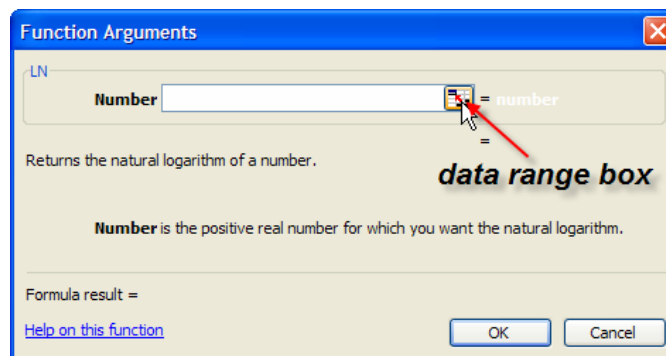
The **Insert Function** dialog box opens. Type in a description of the function you want, and press **Go**.



Excel will return some suggestions. Scroll down the list and note that the definitions of the functions appear at the bottom.



Click **OK**. A **Function Arguments** dialog box opens. Enter the number you wish to take the logarithm of, or to locate a cell click on the **Data range box**.

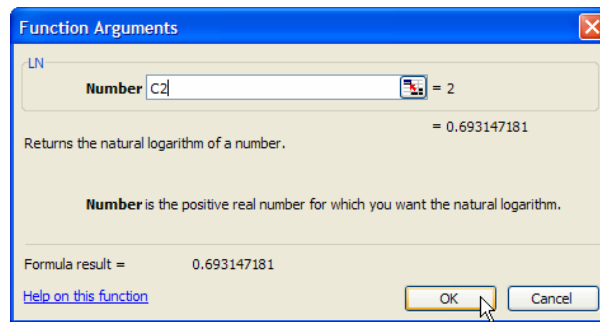


With this box open select the cell C2. This enters the command **=LN(C2)** into F2. Press **Enter**.

|   | A        | B  | C  | D   | E  | F       |
|---|----------|----|----|-----|----|---------|
| 1 | Variable | X  | Y  | Sum | Y2 | LOGY    |
| 2 |          | 1  | 2  | 3   | 4  | =LN(C2) |
| 3 |          | 2  | 3  | 5   | 9  |         |
| 4 |          | 3  | 6  | 9   |    |         |
| 5 |          | 4  | 5  | 9   |    |         |
| 6 |          | 5  | 7  | 12  |    |         |
| 7 | Sum      | 15 | 23 |     |    |         |

Back in the **Function Arguments** dialog box we find that the natural log of 2 is 0.693. Click **OK**.

## 14 Chapter 1



The value is returned to the worksheet in F2. Once again we can **Copy** the formula to compute the natural log of the remaining Y values.

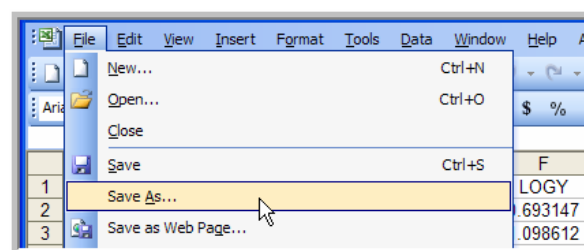
| E  | F        |
|----|----------|
| Y2 | LOGY     |
| 4  | 0.693147 |
| 9  |          |
| 36 |          |
| 25 |          |
| 49 |          |

|   | A        | B  | C  | D   | E  | F        |
|---|----------|----|----|-----|----|----------|
| 1 | Variable | X  | Y  | Sum | Y2 | LOGY     |
| 2 |          | 1  | 2  | 3   | 4  | 0.693147 |
| 3 |          | 2  | 3  | 5   | 9  | 1.098612 |
| 4 |          | 3  | 6  | 9   | 36 | 1.791759 |
| 5 |          | 4  | 5  | 9   | 25 | 1.609438 |
| 6 |          | 5  | 7  | 12  | 49 | 1.94591  |
| 7 | Sum      | 15 | 23 |     |    |          |

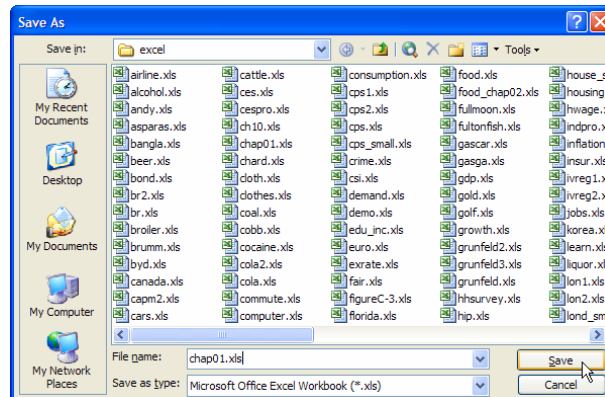
Once you know the function you wish, you can of course just enter the formula into a cell and press **Enter**.

|   | A        | B  | C  | D   | E  | F        | G |
|---|----------|----|----|-----|----|----------|---|
| 1 | Variable | X  | Y  | Sum | Y2 | LOGY     |   |
| 2 |          | 1  | 2  | 3   | 4  | 0.693147 |   |
| 3 |          | 2  | 3  | 5   | 9  | 1.098612 |   |
| 4 |          | 3  | 6  | 9   | 36 | 1.791759 |   |
| 5 |          | 4  | 5  | 9   | 25 | 1.609438 |   |
| 6 |          | 5  | 7  | 12  | 49 | 1.94591  |   |
| 7 | Sum      | 15 | 23 |     |    | =LN(C7)  |   |
| 8 |          |    |    |     |    |          |   |

Now that you have put lots of effort into this example, it is a good idea to save your work. On the Excel **Menu bar**, select **File/Save as**



In the resulting dialog box, find the folder in which you plan to save your work for *POE*. We will use the path **c:\data\excel**. The standard extension for Excel files is **\*.xls**. Name the file and click **Save**.

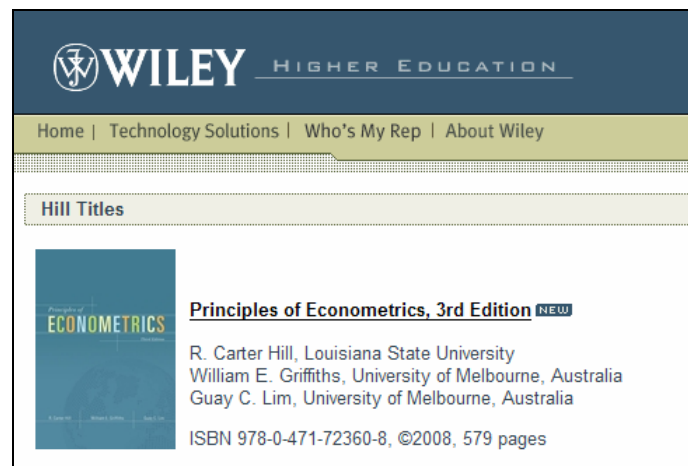


## 1.4 EXCEL FILES FOR PRINCIPLES OF ECONOMETRICS

The book *Principles of Econometrics*, 3e, uses many examples with data. These data files have been saved as workbooks and are available for you to download to your computer. There are about 150 such files. The data files and other supplementary materials can be downloaded from two web locations. You should download not only the **\*.xls** files, but also the definition files, which have the extension **\*.def**. Download these files from either the publisher John Wiley and Sons, or from the book website maintained by the authors.

### 1.4.1 John Wiley and Sons website

Using your web browser enter the address [www.wiley.com/college/hill](http://www.wiley.com/college/hill). Find, among the authors named “Hill” the book *Principles of Econometrics*.



## 16 Chapter 1

Click on the book title and follow the link to **Supplements**. Click on **Supplements**. There you will find links to many supplement materials, including a link that will allow you to download all the data files at once.

### 1.4.2 Principles of Econometrics website

Alternatively, you may wish to download individual files.

- Go to the site [www.bus.lsu.edu/hill/poe](http://www.bus.lsu.edu/hill/poe) for the data, errata and other supplements.
- For the Excel data files go to [www.bus.lsu.edu/hill/poe/excel.htm](http://www.bus.lsu.edu/hill/poe/excel.htm).

### 1.4.3 Definition files

There is a data definition file for each data file used in the book. These are simple “text” or “ASCII” files that can be opened with utilities like Notepad or Wordpad, or a word processor. Locate *food.def*. Its contents are:

```
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   |

The definition files contain variable names, variable definitions, and summary statistics.

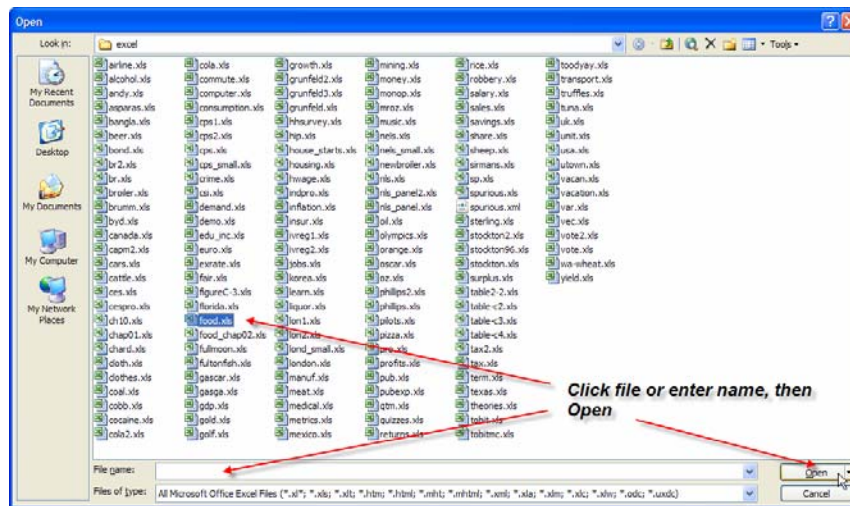
### 1.4.4 The food expenditure data

In the first few chapters you will use data on household food expenditure. Locate the file *food.xls* and open it. To illustrate, click on the **Open** icon on the **Menu**.



Navigate to the file you wish to open.

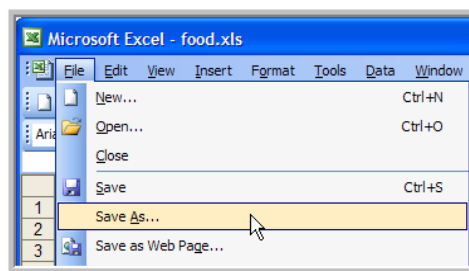




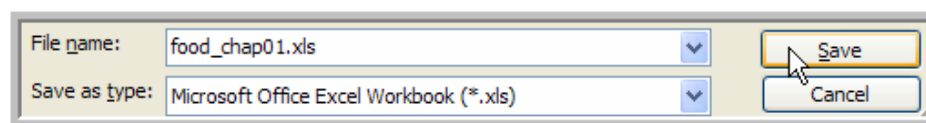
The worksheet should appear as

|    | A        | B      | C | D |
|----|----------|--------|---|---|
| 1  | food_exp | income |   |   |
| 2  | 115.22   | 3.69   |   |   |
| 3  | 135.98   | 4.39   |   |   |
| 4  | 119.34   | 4.75   |   |   |
| 5  | 114.96   | 6.03   |   |   |
| 6  | 187.05   | 12.47  |   |   |
| 7  | 243.92   | 12.98  |   |   |
| 8  | 267.43   | 14.2   |   |   |
| 9  | 238.71   | 14.76  |   |   |
| 10 | 295.94   | 15.32  |   |   |

So as to not alter the original file, you may want to save the file with a new name, such as *food\_chap01.xls*. Select **File>Save As** from the Excel Menu.

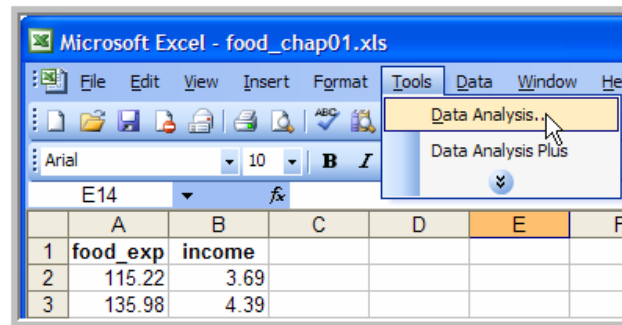


In the dialog box enter

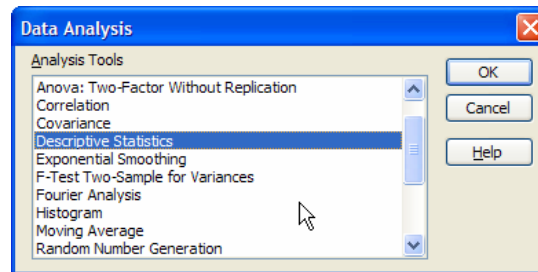


## 18 Chapter 1

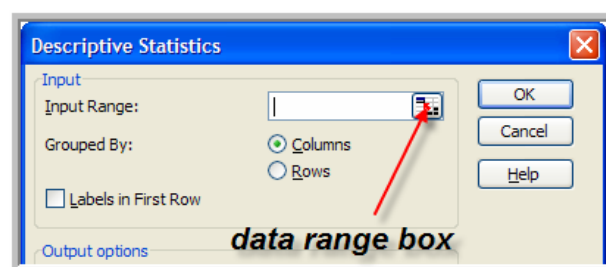
Compute the summary statistics, to make sure they match the ones in *food.def*. Select **Tools>Data Analysis**.



In the resulting dialog box choose **Descriptive Statistics**, then **OK**.



In the **Descriptive Statistics** dialog box we must enter the **Input Range** of the data. Click on the **Data Range box**.



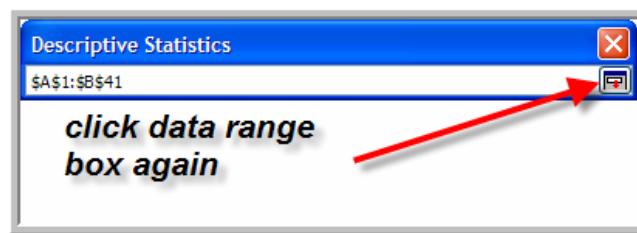
The following box will open.



While it is open, highlight columns A & B, rows 1 to 41. You can do this by

- Click A1, hold down left mouse button, and drag over desired area; or
- Click A1, hold down **Ctrl+Shift**. Press right arrow → and then down arrow ↓.

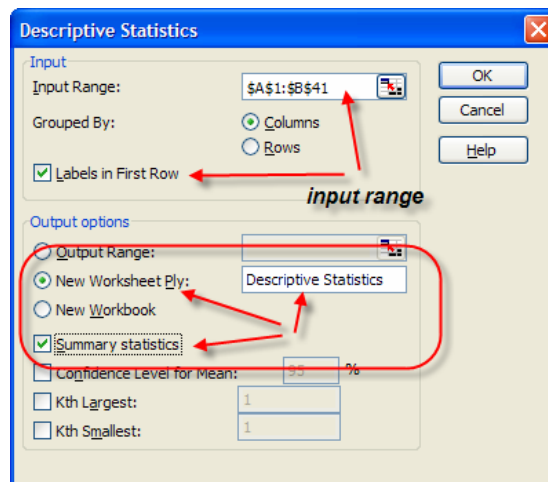
In the resulting window click the **Data Range box** again.



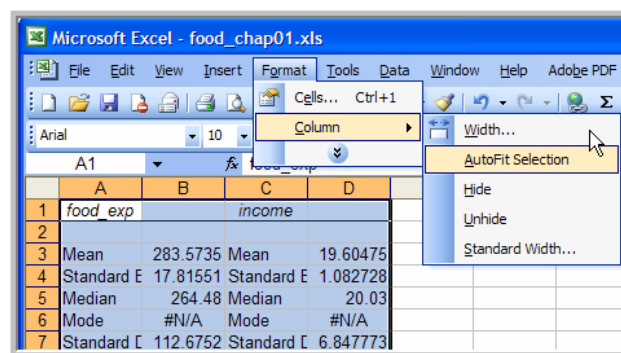
Note that now the input range is filled in.

- The Excel data range is \$A\$1:\$B\$41. This range locates the upper left corner of the highlighted area (A1) and lower right corner (B41). The \$ makes this an **Absolute Cell Reference**, that will not be changed if the data are moved.
- Tick the box **Labels in First Row** so that these cells will not be treated as data.
- Select the radio button **New Worksheet Ply** and enter a name for the new worksheet page.
- Tick the box **Summary Statistics** so that Excel will print the results.

Click **OK**



The resulting worksheet is not formatted well. Select **Format>Column>AutoFit Selection**



## 20 Chapter 1

One of the nice things about Microsoft Office is that information from one application can be transferred to another. With the **Descriptive Statistics** highlighted, enter **Ctrl+C** to copy. In an open document enter **Ctrl+V** to paste.

Now you have a nice table of statistics in your document that can be edited in the usual way.

| <i>food_exp</i>    |              | <i>income</i>      |              |
|--------------------|--------------|--------------------|--------------|
| Mean               | 283.5734993  | Mean               | 19.60475005  |
| Standard Error     | 17.81551026  | Standard Error     | 1.08272795   |
| Median             | 264.479996   | Median             | 20.0299995   |
| Mode               | #N/A         | Mode               | #N/A         |
| Standard Deviation | 112.6751802  | Standard Deviation | 6.847772819  |
| Sample Variance    | 12695.69623  | Sample Variance    | 46.89199259  |
| Kurtosis           | -0.002430221 | Kurtosis           | 0.48455582   |
| Skewness           | 0.511465877  | Skewness           | -0.651185498 |
| Range              | 477.949974   | Range              | 29.710002    |
| Minimum            | 109.709999   | Minimum            | 3.69         |
| Maximum            | 587.659973   | Maximum            | 33.400002    |
| Sum                | 11342.93997  | Sum                | 784.190002   |
| Count              | 40           | Count              | 40           |

You may wish to save this file by clicking



The food expenditure data will be used extensively in the next chapter.

# CHAPTER 2

## The Simple Linear Regression Model

### CHAPTER OUTLINE

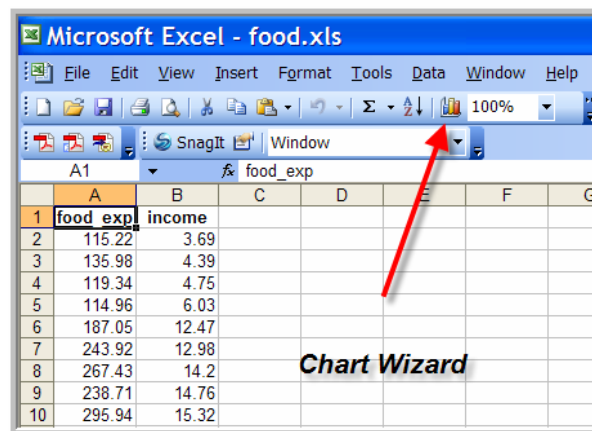
2.1 Plotting the Food Expenditure Data  
2.2 Estimating a Simple Regression  
2.3 Plotting a Simple Regression

2.4 Plotting the Least Squares Residuals  
2.5 Prediction Using Excel

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 Excel and show how to use the software to calculate the income elasticity of food expenditure, and to predict food expenditure from our regression results.

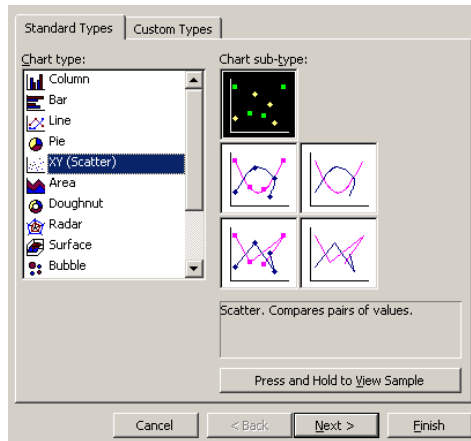
### 2.1 PLOTTING THE FOOD EXPENDITURE DATA

We will use **Chart Wizard** to scatter plot the data. Open *food.xls* file in Excel.

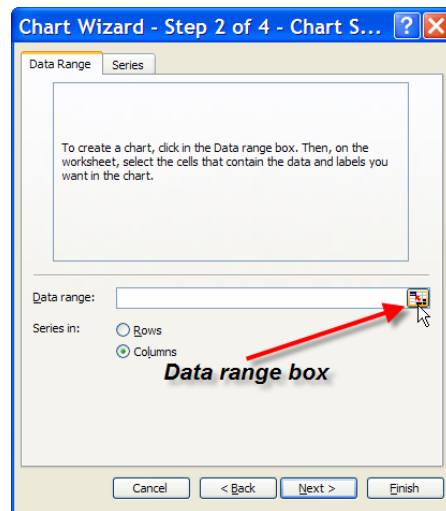


## 22 Chapter 2

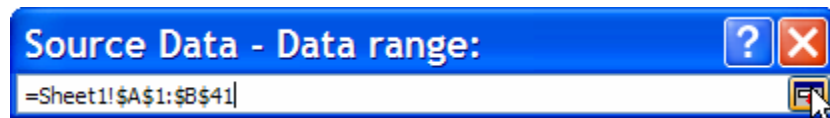
Place the cursor on the **Chart Wizard** icon and click. In the dialog box, choose the chart type as XY (Scatter) and click next.



To define the **Data Range**, highlight the data after clicking on the space provided.

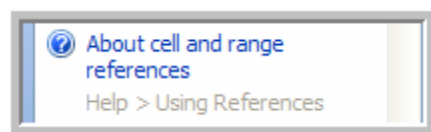


Select the data columns and click **Data range** again



### Aside: Referencing Cells

Select **Help**. Search the phrase “Cell Reference.” One of the resulting hits is



Click this link to find the following description of cell references:

A reference identifies a cell or a range of cells on a worksheet and tells Microsoft Excel where to look for the values or data you want to use in a formula. With references, you can use data contained in different parts of a worksheet in one formula or use the value from one cell in several formulas. You can also refer to cells on other sheets in the same workbook, and to other workbooks. References to cells in other workbooks are called links.

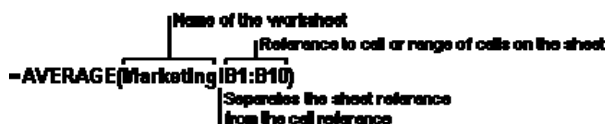
### The A1 reference style

By default, Excel uses the A1 reference style, which refers to columns with letters (A through IV, for a total of 256 columns) and refers to rows with numbers (1 through 65536). These letters and numbers are called row and column headings. To refer to a cell, enter the column letter followed by the row number. For example, B2 refers to the cell at the intersection of column B and row 2.

| To refer to                                                      | Use     |
|------------------------------------------------------------------|---------|
| The cell in column A and row 10                                  | A10     |
| The range of cells in column A and rows 10 through 20            | A10:A20 |
| The range of cells in row 15 and columns B through E             | B15:E15 |
| All cells in row 5                                               | 5:5     |
| All cells in rows 5 through 10                                   | 5:10    |
| All cells in column H                                            | H:H     |
| All cells in columns H through J                                 | H:J     |
| The range of cells in columns A through E and rows 10 through 20 | A10:E20 |

**Reference to another worksheet** In the following example, the AVERAGE worksheet function calculates the average value for the range B1:B10 on the worksheet named Marketing in the same workbook.

**-AVERAGE(Marketing!B1:B10)**



The diagram shows the formula **-AVERAGE(Marketing!B1:B10)** with three labels and arrows pointing to its parts:
 

- Name of the worksheet** points to **Marketing**.
- Reference to cell or range of cells on the sheet** points to **B1:B10**.
- Separates the sheet reference from the cell reference** points to the exclamation point **!**.

Note that the name of the worksheet and an exclamation point (!) precede the range reference.

### Aside: Relative vs. Absolute References

On the same **Help** page, you will find the following useful information:

### The difference between relative and absolute references

**Relative references** A relative cell reference in a formula, such as A1, is based on the relative position of the cell that contains the formula and the cell the reference refers to. If the position of

## 24 Chapter 2

the cell that contains the formula changes, the reference is changed. If you copy the formula across rows or down columns, the reference automatically adjusts. By default, new formulas use relative references. For example, if you copy a relative reference in cell B2 to cell B3, it automatically adjusts from =A1 to =A2.

|   | A | B   |
|---|---|-----|
| 1 |   |     |
| 2 |   | =A1 |
| 3 |   | =A2 |

Copied formula with relative reference

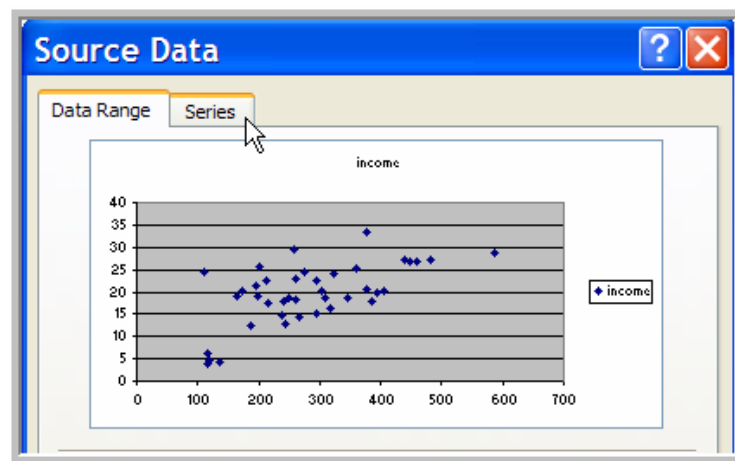
**Absolute references** An absolute cell reference in a formula, such as \$A\$1, always refer to a cell in a specific location. If the position of the cell that contains the formula changes, the absolute reference remains the same. If you copy the formula across rows or down columns, the absolute reference does not adjust. By default, new formulas use relative references, and you need to switch them to absolute references. For example, if you copy a absolute reference in cell B2 to cell B3, it stays the same in both cells =\$A\$1.

|   | A | B       |
|---|---|---------|
| 1 |   |         |
| 2 |   | =\$A\$1 |
| 3 |   | =\$A\$1 |

Copied formula with absolute reference

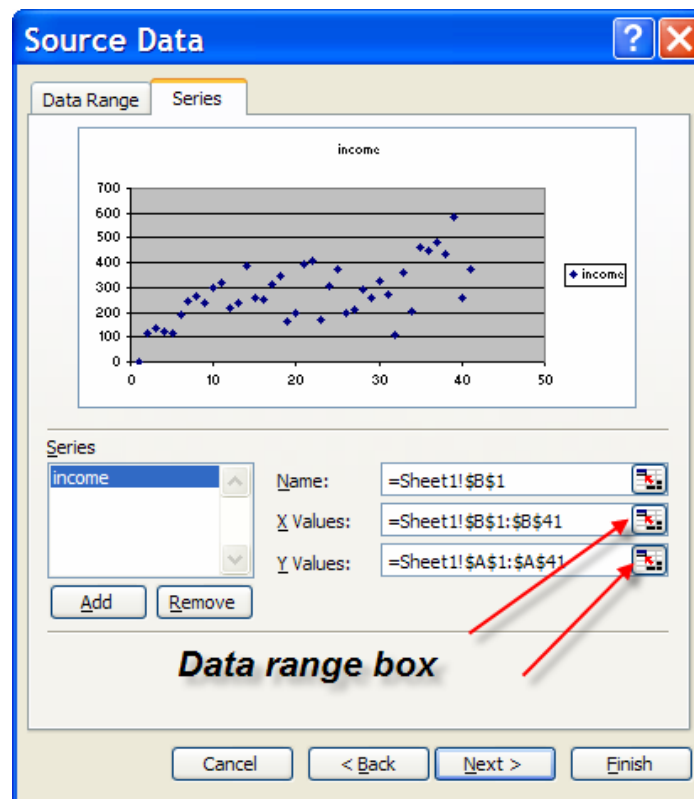
## 2.1 (Continued)

Excel assumes that the first column is the X-variable. Select the **Series** tab.

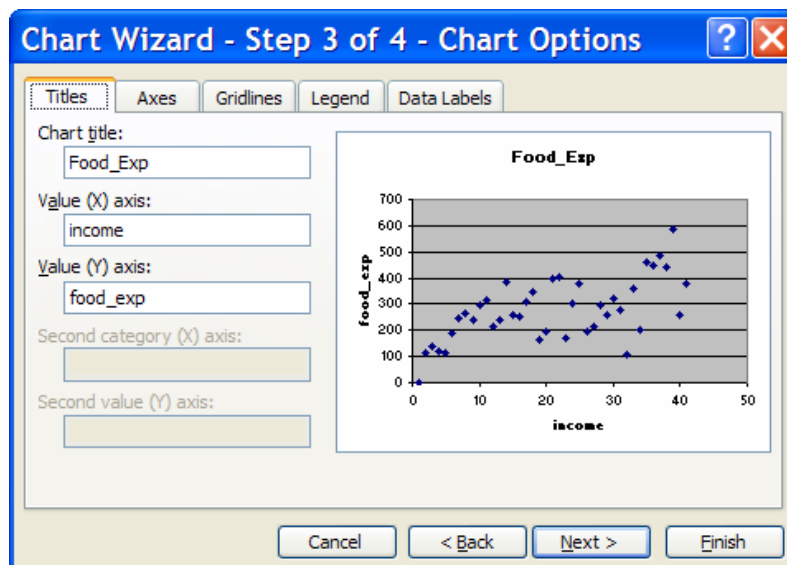


On the **Series tab** define the columns that are X and Y variables again using the **Data range box**.

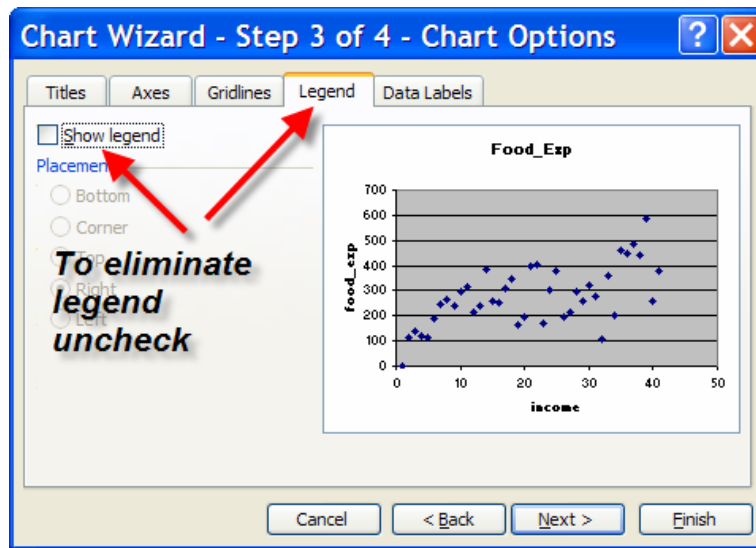




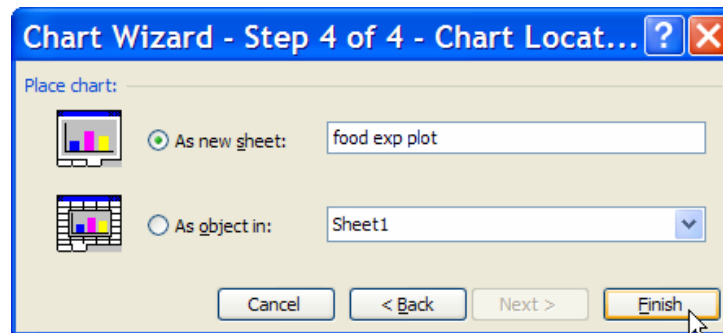
- Click Next.
- Add or modify labels and title and click Next.



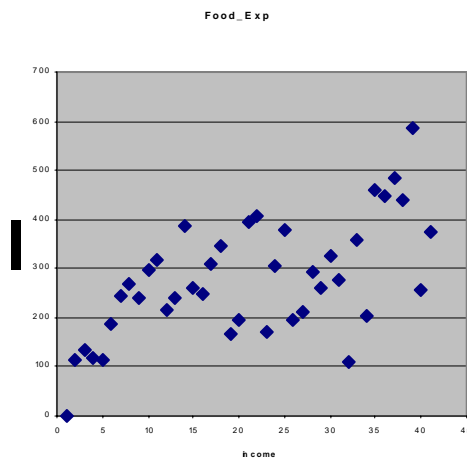
- The default legend is not informative. To delete it go the **Legend** tab and uncheck box.



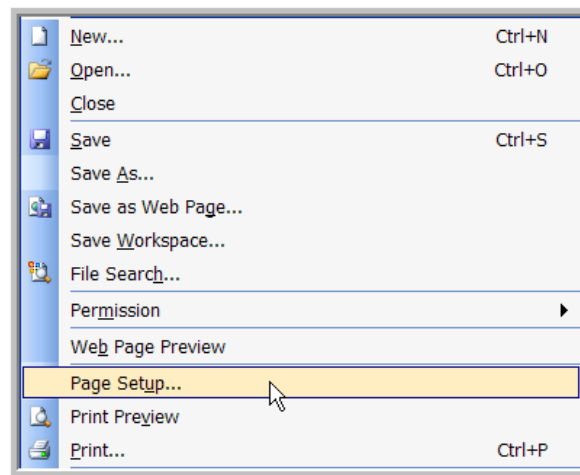
- The last step is to place your chart. You can have the chart on the current worksheet or in a new one. Make your choice and click **Finish**.



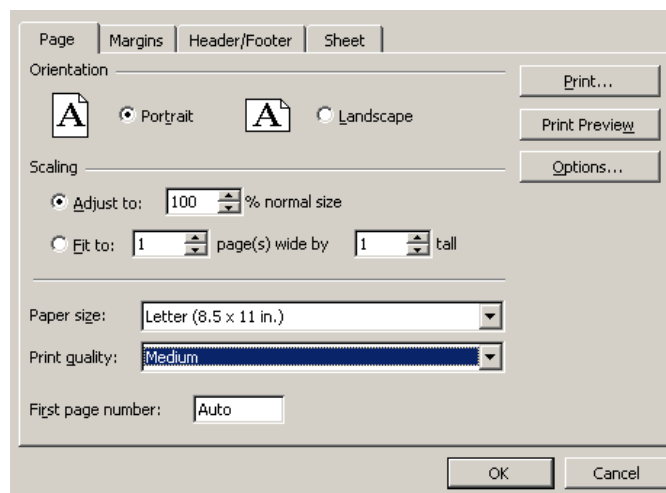
If you chose to print it in a new worksheet, your plot will appear in the new worksheet.



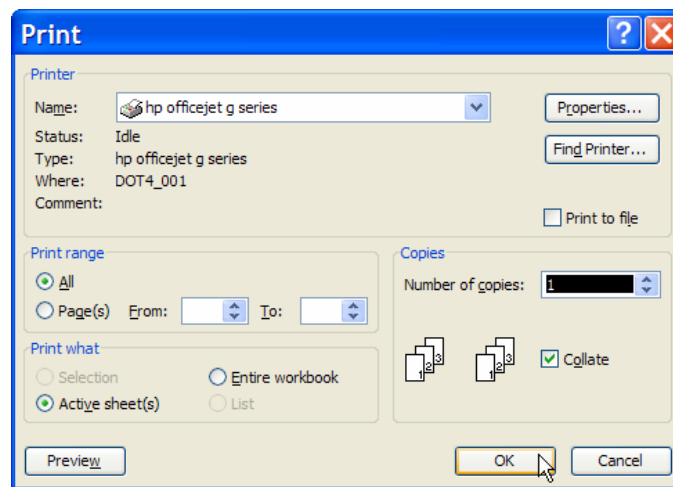
To print a worksheet, first click on **File/Page Setup**.



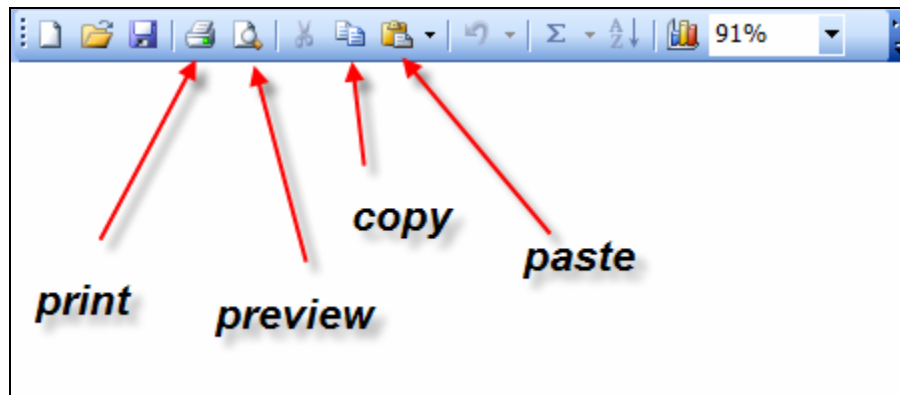
There you can choose the page layout, either **Portrait** or **Landscape**, and adjust the size. The **Print Preview** allows you to see what will be printed before actually doing so, which saves time and paper.



Click on **File/Print** to open the print dialog window. In the **Print** dialog box, make sure the printer is correctly specified. Here you can specify **Print range**, pages to be printed etc.



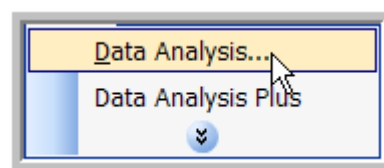
Alternatively, if the **Print Preview** is satisfactory, click on the printer icon on the main toolbar.



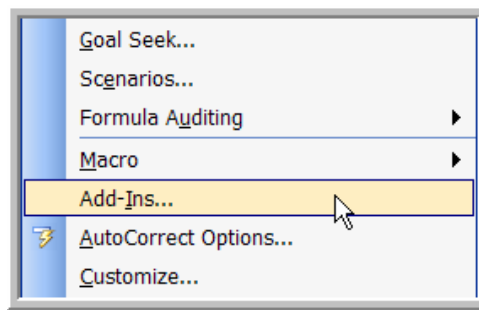
If you are creating a document, click inside the graph and enter **Ctrl+C** which copies the figure into the “clipboard.” Transfer to an open document, place the cursor in the document where you want the figure and enter **Ctrl+V** to paste the figure into the document, where it can be resized by dragging a corner inward. Give it a try.

## **2.2 ESTIMATING A SIMPLE REGRESSION**

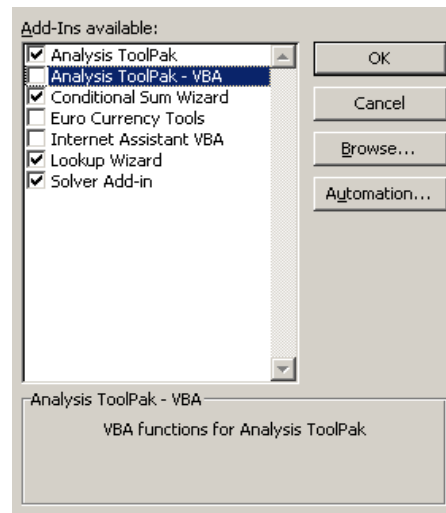
To estimate the parameters  $b_1$  and  $b_2$  of the food expenditure equation, place cursor in an empty cell and click on **Tools/Data Analysis**.



The **Data Analysis** tool may not automatically load with a default installation of the program, if **Data Analysis** tool doesn't appear on the menu, click on **Add-Ins** under **Tools**:

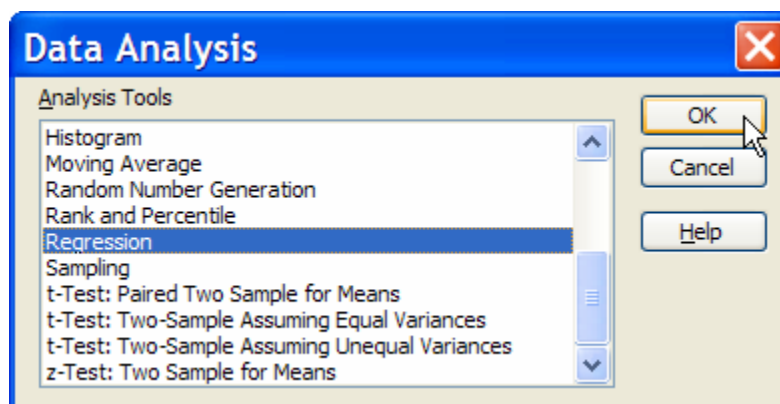


Check the box next to the **Analysis ToolPak** add-in and click OK.



Data Analysis should now appear in the **Tools menu**, and you should not have to run the add-ins again for this option.

When the **Data Analysis** dialog box appears, click on **Regression**, then **OK**.



The **Regression dialog box** will appear with lots of user-defined inputs and options.

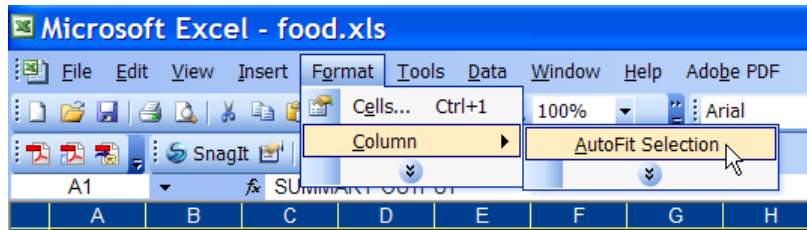
To define the input ranges, first click on the Input Y Range box. The box will minimize. Highlight the data in the y column, including the label. Do the same procedure to input the X Range.

Other options:

- Don't forget to check the **Label** box if you are including labels.
- Do not check the **Constant is Zero** box. This would suppress the intercept.
- The output can be sent to the current page or another page in the workbook. Name the new worksheet **Food Expenditure** and hit **Enter**.

Since you chose to place the output in a separate worksheet, a new worksheet will appear as a tab in the lower left corner of the work area. If you click on the **Food Expenditure** tab, you will

notice that the columns are not wide enough to show the cells completely. Highlight the data, or the entire sheet, and click on **Format/Column/AutoFit Selection**.



The output contains many items that you will learn about later. For now it is important to note that the *Coefficients* corresponding to *Intercept* and *income* are the least squares estimates  $b_1$  and  $b_2$ .

|    | A                     | B            | C              | D           | E           | F              | G           | H            | I           |
|----|-----------------------|--------------|----------------|-------------|-------------|----------------|-------------|--------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |             |             |                |             |              |             |
| 2  |                       |              |                |             |             |                |             |              |             |
| 3  | Regression Statistics |              |                |             |             |                |             |              |             |
| 4  | Multiple R            | 0.620485472  |                |             |             |                |             |              |             |
| 5  | R Square              | 0.385002221  |                |             |             |                |             |              |             |
| 6  | Adjusted R Square     | 0.368818069  |                |             |             |                |             |              |             |
| 7  | Standard Error        | 89.51700429  |                |             |             |                |             |              |             |
| 8  | Observations          | 40           |                |             |             |                |             |              |             |
| 9  |                       |              |                |             |             |                |             |              |             |
| 10 | ANOVA                 |              |                |             |             |                |             |              |             |
| 11 |                       | df           | SS             | MS          | F           | Significance F |             |              |             |
| 12 | Regression            | 1            | 190626.9788    | 190626.9788 | 23.78884107 | 1.94586E-05    |             |              |             |
| 13 | Residual              | 38           | 304505.1742    | 8013.294058 |             |                |             |              |             |
| 14 | Total                 | 39           | 495132.153     |             |             |                |             |              |             |
| 15 |                       |              |                |             |             |                |             |              |             |
| 16 |                       | Coefficients | Standard Error | t Stat      | P-value     | Lower 95%      | Upper 95%   | Lower 95.0%  | Upper 95.0% |
| 17 | Intercept             | 83.41600997  | 43.41016192    | 1.921577951 | 0.062182379 | -4.463267721   | 171.2952877 | -4.463267721 | 171.2952877 |
| 18 | income                | 10.2096425   | 2.093263461    | 4.877380554 | 1.94586E-05 | 5.972052202    | 14.4472328  | 5.972052202  | 14.4472328  |
| 19 |                       |              |                |             |             |                |             |              |             |
| 20 |                       |              |                |             |             |                |             |              |             |
| 21 |                       |              |                |             |             |                |             |              |             |
| 22 |                       |              |                |             |             |                |             |              |             |
| 23 |                       |              |                |             |             |                |             |              |             |
| 24 |                       |              |                |             |             |                |             |              |             |
| 25 |                       |              |                |             |             |                |             |              |             |
| 26 |                       |              |                |             |             |                |             |              |             |
| 27 |                       |              |                |             |             |                |             |              |             |
| 28 |                       |              |                |             |             |                |             |              |             |
| 29 |                       |              |                |             |             |                |             |              |             |
| 30 |                       |              |                |             |             |                |             |              |             |
| 31 |                       |              |                |             |             |                |             |              |             |
| 32 |                       |              |                |             |             |                |             |              |             |

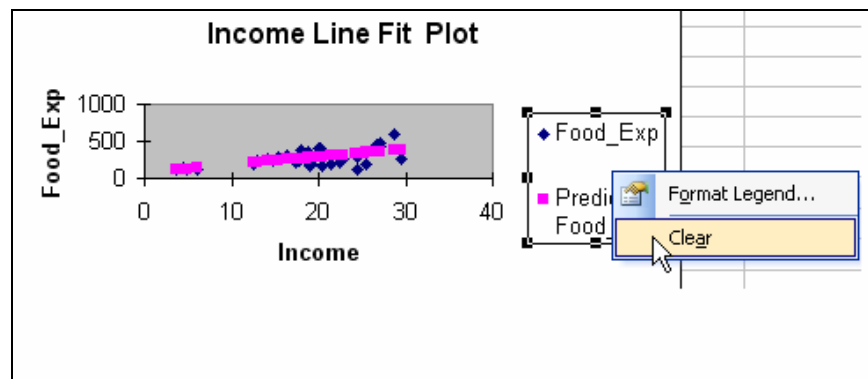
## 2.3 PLOTTING A SIMPLE REGRESSION

In order to plot the regression function we must re-estimate the food expenditure equation and choose the **Line Fit Plots** option in the regression dialog box.

Click **OK**. The graph will be produced, and placed on the same worksheet as the regression output. If you can't find it on the worksheet, click on the **File/Print Preview** or click on the Print Preview icon and look for it.

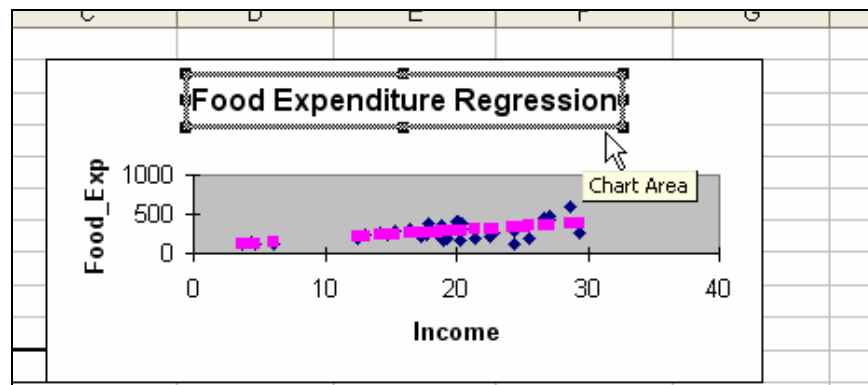
You will need to enhance your graph. Here are a few suggestions:

- Right click on the legend and delete, if desired.



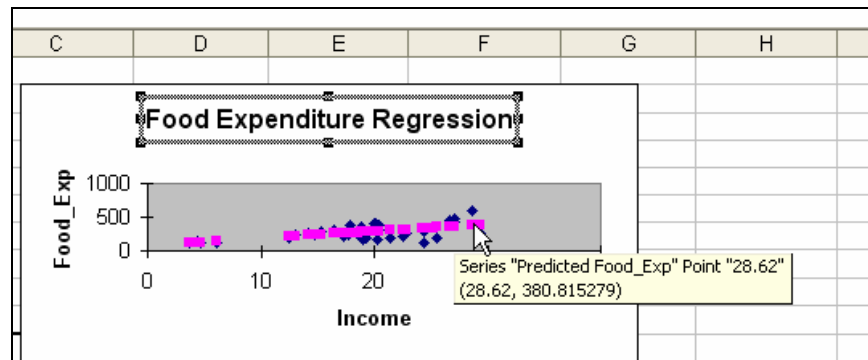
- Left click once on the title to highlight it. Left click once more to edit. Name appropriately. If you double click the box surrounding the title, a dialog box opens that allows you more formatting options.



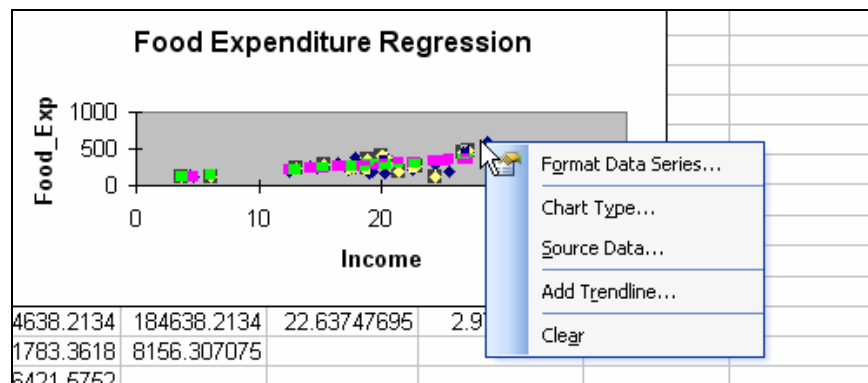


Repeat for the Y and X axes if you want to change the names of the variables.

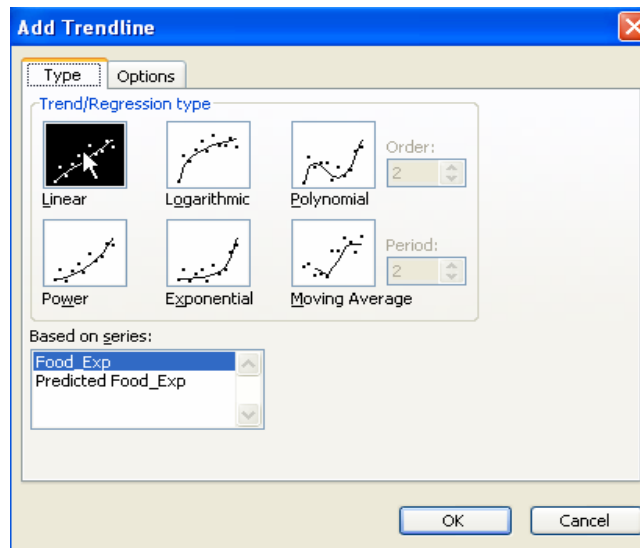
Notice that both the actual values of Y and the predicted values are plotted. To include the estimated regression function, place the cursor over one of the predicted Y points (the pink ones) until the caption "**Series Predicted Y**" appears.



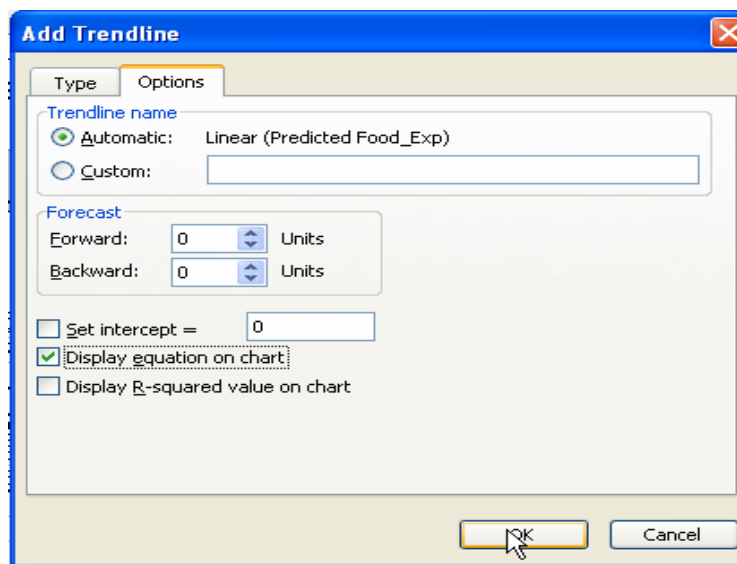
Right Click and choose Add **Trendline**.



Under the **Type** tab, choose **Linear**.

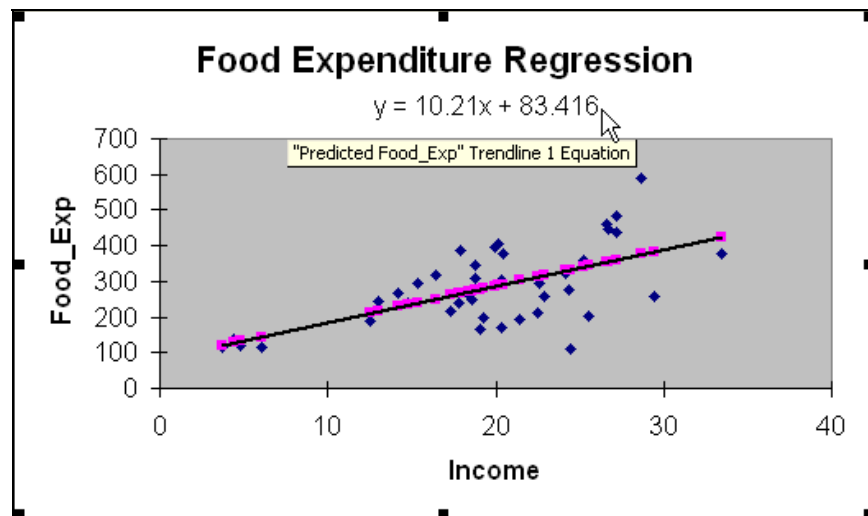


Under the **Options** tab, check the box next to **Display equation** on chart. Click **OK**.



If your figure is small, and begins to get cluttered, increase its size by clicking inside the border. Place the mouse arrow over the black square in the corner, until a double arrow appears. Then drag the mouse, with the left button held down, to increase (or decrease) the figure size.

Your figure should look something like the one below.



## 2.4 PLOTTING THE LEAST SQUARES RESIDUALS

The least squares residuals are defined as

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

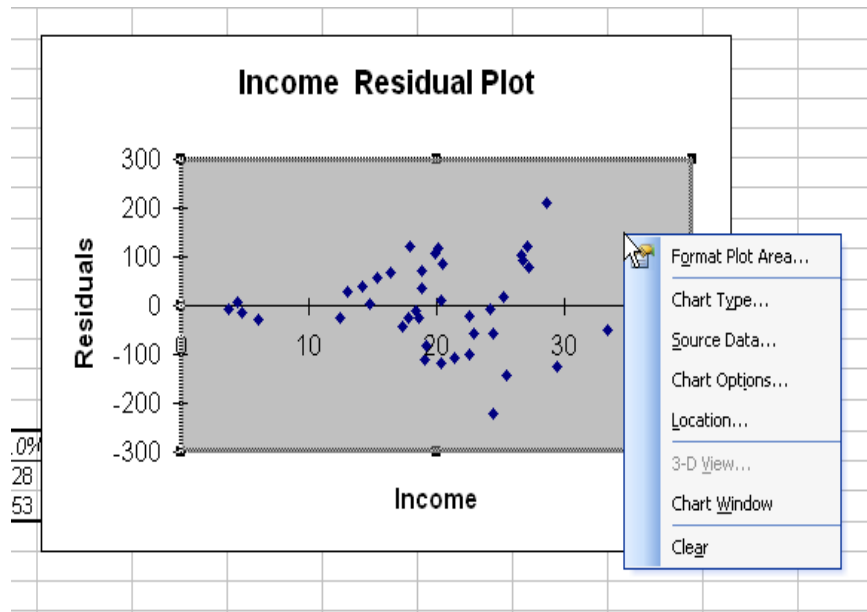
In order to plot the least squares residuals, we must re-estimate the food expenditure equation and choose the **Residual Plots** option in the regression dialog box.

The image shows a regression dialog box with the following sections:

- Input:**
  - Input Y Range: \$A\$1:\$A\$41
  - Input X Range: \$B\$1:\$B\$41
  - ☒ Labels
  - ☐ Constant is Zero
  - ☐ Confidence Level: 95 %
- Output options:**
  - ☐ Output Range:
  - ☒ New Worksheet Ply: Food Expenditure
  - ☐ New Workbook
- Residuals:**
  - ☐ Residuals
  - ☒ Residual Plots
  - ☐ Standardized Residuals
  - ☐ Norm Fit Plots
- Normal Probability:**
  - ☐ Normal Probability Plots

Buttons on the right: OK, Cancel, Help. A red arrow points to the "Residual Plots" checkbox.

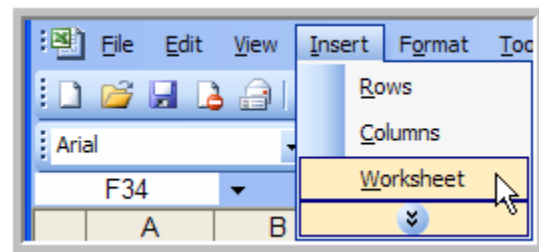
If you wish to enhance your graph, you can do so by right clicking on chart area or plot area.



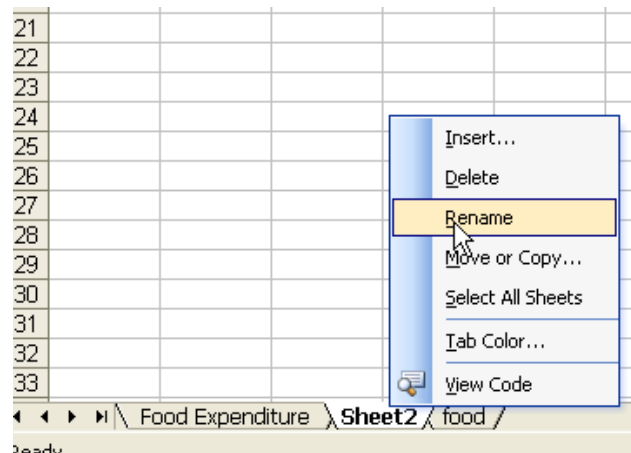
## 2.5 PREDICTION USING EXCEL

Obtaining predicted values from an estimated regression function can be done either by doing the computations or using the **TREND** function.

Insert a new worksheet to the workbook select **Insert/Worksheet** on the main menu.



Rename the worksheet **Predictions** by right clicking on the tab.



In the **Predictions** worksheet, create a template for prediction by copying the estimated coefficients from the regression and labeling them appropriately. Enter the income value for which you want to predict food expenditures. Create the formula for the predicted value of  $y$ ,  $\hat{y} = b_1 + b_2 \text{Income}$ , using the cell references.

|   | A                          | B           | C |
|---|----------------------------|-------------|---|
| 1 | Predicted Food Expenditure |             |   |
| 2 |                            |             |   |
| 3 | b1                         | 83.41600202 |   |
| 4 | b2                         | 10.20964297 |   |
| 5 | Income                     | 20          |   |
| 6 | Yhat                       | =B3+(B4*B5) |   |
| 7 |                            |             |   |
| 8 |                            |             |   |

The results in cell B6 will be 287.6088614.

|   | A                          | B           | C |
|---|----------------------------|-------------|---|
| 1 | Predicted Food Expenditure |             |   |
| 2 |                            |             |   |
| 3 | b1                         | 83.41600202 |   |
| 4 | b2                         | 10.20964297 |   |
| 5 | Income                     | 20          |   |
| 6 | Yhat                       | 287.6088614 |   |
| 7 |                            |             |   |
| 8 |                            |             |   |

Excel has a built in function that computes predicted values from simple regressions. The form of the **Trend function** is

**TREND(range of Y variable, range of X variable, value of  $x_0$ )**

## 38 Chapter 2

The value  $x_0$  is the value at which the prediction is desired. To use this function, return to the worksheet page containing the data. Type in the following command,

|    | A        | B      | C | D                        | E | F | G |
|----|----------|--------|---|--------------------------|---|---|---|
| 1  | Food_Exp | Income |   | Yhat                     |   |   |   |
| 2  | 115.22   | 3.69   |   | =TREND(A2:A41,B2:B41,20) |   |   |   |
| 3  | 135.98   | 4.39   |   |                          |   |   |   |
| 4  | 119.34   | 4.75   |   |                          |   |   |   |
| 5  | 114.96   | 6.03   |   |                          |   |   |   |
| 6  | 187.05   | 12.47  |   |                          |   |   |   |
| 7  | 243.92   | 12.98  |   |                          |   |   |   |
| 8  | 267.43   | 14.2   |   |                          |   |   |   |
| 9  | 238.71   | 14.76  |   |                          |   |   |   |
| 10 | 295.94   | 15.32  |   |                          |   |   |   |

The result will be in cell D2.

|   | A        | B      | C | D           | E | F |
|---|----------|--------|---|-------------|---|---|
| 1 | Food_Exp | Income |   | Yhat        |   |   |
| 2 | 115.22   | 3.69   |   | 287.6088614 |   |   |
| 3 | 135.98   | 4.39   |   |             |   |   |
| 4 | 119.34   | 4.75   |   |             |   |   |
| 5 | 114.96   | 6.03   |   |             |   |   |
| 6 | 187.05   | 12.47  |   |             |   |   |
| 7 | 243.92   | 12.98  |   |             |   |   |
| 8 | 267.43   | 14.2   |   |             |   |   |

As a final step, save your file. We recommend saving it under a new name, like *food\_chap02.xls*, so that the original data file will not be altered.

# CHAPTER 3

## Interval Estimation and Hypothesis Testing

### CHAPTER OUTLINE

#### 3.1 Interval Estimation

##### 3.1.1 Automatic interval estimates

##### 3.1.2 Constructing interval estimates

#### 3.2 Hypothesis Testing

##### 3.2.1 Right-Tail tests

##### 3.2.2 Left-Tail tests

##### 3.2.3 Two-Tail tests

In this chapter we continue to work with the simple linear regression model and our model of weekly food expenditure.

### 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 by

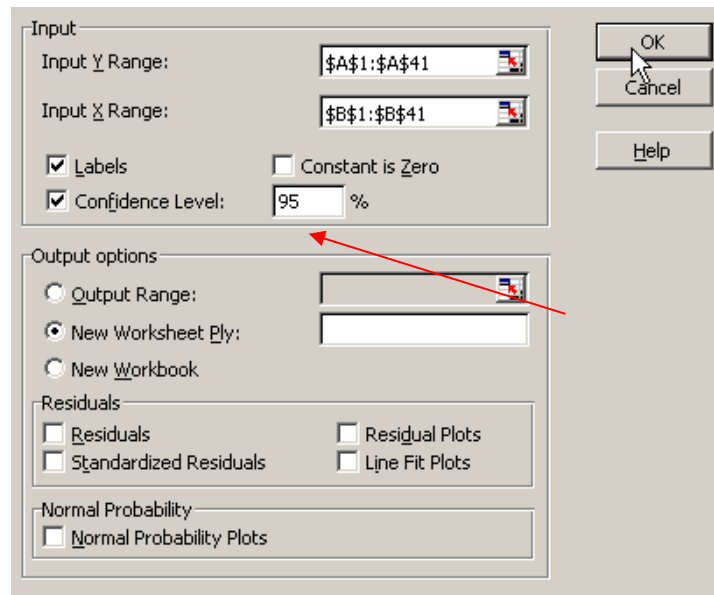
$$t = \frac{b_k - \beta_k}{se(b_k)} \sim t_{(N-2)} \quad \text{for } k = 1, 2$$

Using this result we can show that the interval  $b_k \pm t_c se(b_k)$  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  $P(t \geq t_c) = P(t \leq -t_c) = \alpha/2$

#### 3.1.1 Automatic interval estimates

To construct the confidence interval estimates, we will use the **Regression function** in Excel. To do that, open the workbook containing the food expenditure regression that we considered in

Chapter 2. Excel provides 95% confidence interval for the Least Squares estimates by checking the **Confidence Level** box in the **Regression Dialog** box.

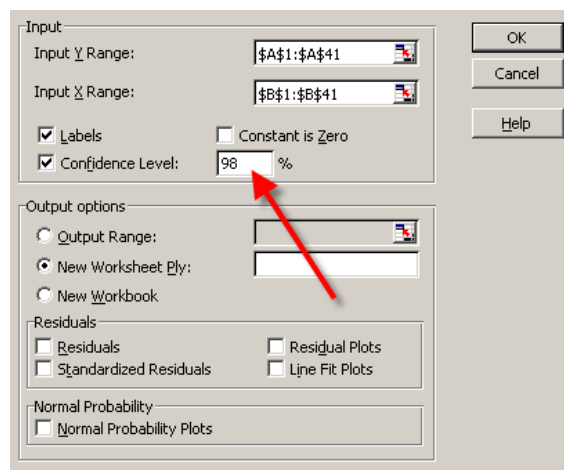


The image shows the Excel Regression Dialog Box. In the 'Input' section, 'Input Y Range' is '\$A\$1:\$A\$41' and 'Input X Range' is '\$B\$1:\$B\$41'. The 'Labels' checkbox is checked, and 'Constant is Zero' is unchecked. The 'Confidence Level' is set to '95 %'. In the 'Output options' section, 'New Worksheet Ply:' is selected. In the 'Residuals' section, 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' are all unchecked. In the 'Normal Probability' section, 'Normal Probability Plots' is unchecked. A red arrow points to the 'Confidence Level' box.

Excel will report the confidence interval in the **Summary Output** next to the coefficient estimates. The results show the lower and upper values for the 95% confidence interval for the coefficient estimates.

|    | A         | B                   | C                     | D             | E              | F                | G                | H                  | I                  |
|----|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|--------------------|--------------------|
| 15 |           |                     |                       |               |                |                  |                  |                    |                    |
| 16 |           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> | <i>Lower 95.0%</i> | <i>Upper 95.0%</i> |
| 17 | Intercept | 83.41600997         | 43.41016192           | 1.921577951   | 0.062182379    | -4.463267721     | 171.2952877      | -4.463267721       | 171.2952877        |
| 18 | Income    | 10.2096425          | 2.093263461           | 4.877380554   | 1.94586E-05    | 5.972052202      | 14.4472328       | 5.972052202        | 14.4472328         |
| 19 |           |                     |                       |               |                |                  |                  |                    |                    |
| 20 |           |                     |                       |               |                |                  |                  |                    |                    |
| 21 |           |                     |                       |               |                |                  |                  |                    |                    |

To have Excel calculate a different confidence interval, estimate a regression, and after checking the **Confidence Level** box, type in the desired level, for example, 98 for 98% confidence interval.



The image shows the Excel Regression Dialog Box with the 'Confidence Level' set to '98 %'. A red arrow points to the 'Confidence Level' box.

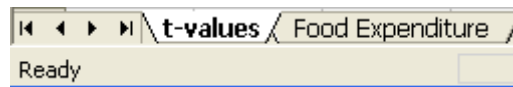


### 3.1.2 Constructing interval estimates

While using the automatic interval estimation feature in Excel is the quickest and easiest way to obtain interval estimates, a general template can be created for calculating interval estimates.

To construct the interval estimates we require the least square estimates  $b_k$ , their standard errors  $se(b_k)$  and the critical value for  $t$ -distribution,  $t_c$ . We already know we can find the least squares estimates  $b_k$ , their standard errors  $se(b_k)$  in summary output. We also need to find values  $t_c$ , such that  $\alpha/2$  of the probability is in either tail. As an example, the critical values that mark off  $\alpha/2 = .025$  of the probability in each tail of a  $t$ -distribution with 38 degrees of freedom. Checking Table 2 at the front of your book, we find that the value is  $t_c = 2.024$  on the positive side and, using the symmetry of the  $t$ -distribution,  $-2.024$  on the negative side. Excel makes it easy to compute critical values from the  $t$ -distribution using the **TINV** function.

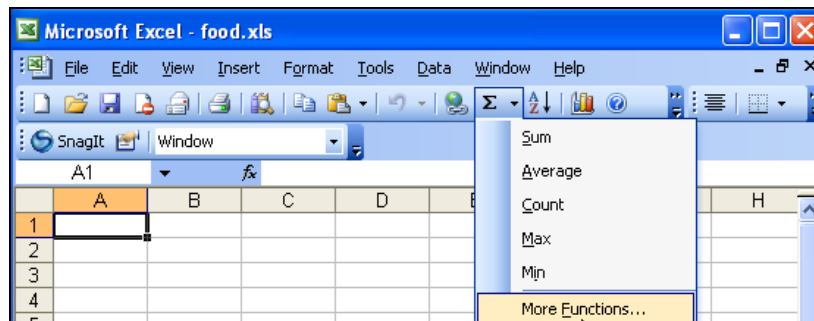
To generate 95% confidence interval, first open the workbook containing the food expenditure regression and insert a worksheet. Move the cursor over the tab with the default name, right click, and rename the sheet **t-values**.



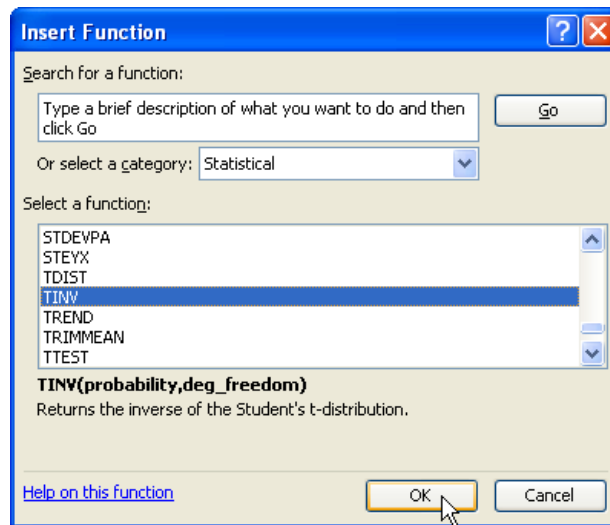
Go to the newly created worksheet and select a cell. Click on the drop down menu next to **Sum** ( $\Sigma$ ).



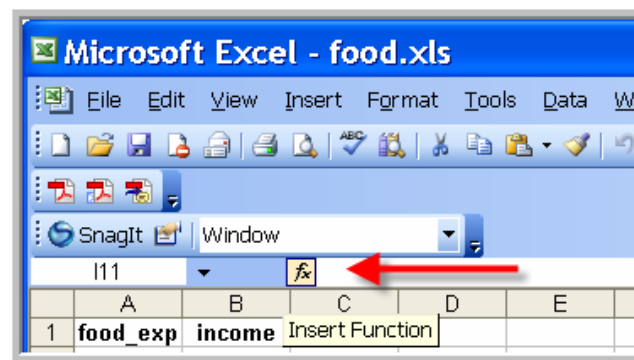
Find **TINV** from the **Statistical Function** category, and click **OK**.



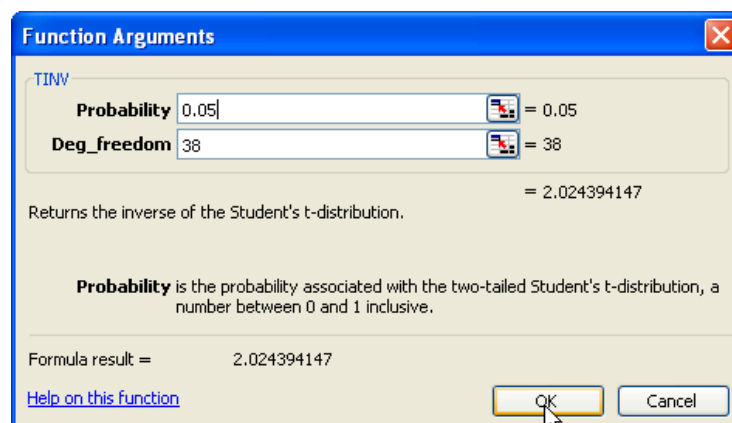
In case you don't see **Statistical Function** on your menu, choose **More Functions**. Then, find **TINV** from the **Statistical Function** category, and click **OK**.



Alternatively, click on the **Paste** or **Insert Function (f\*)** icon to open up the same search.



Either way, in the **TINV** dialog box, fill in the box as shown below.



The **Probability** value you need to fill in is the  $\alpha$  in *two-tails* of the  $t$ -distribution. Enter the degrees of freedom, 38, and click **OK**. The resulting value, 2.024394147 will appear in your worksheet when you click **OK**.

We can now create a **Confidence Interval Template**. Cells with bold border require user input, some obtained from the regression results.

|    | A                                               | B                   |
|----|-------------------------------------------------|---------------------|
| 1  | <i>Interval Estimation in Simple Regression</i> |                     |
| 2  |                                                 |                     |
| 3  | <b>Data Input</b>                               |                     |
| 4  | Sample Size                                     |                     |
| 5  | Confidence Level                                |                     |
| 6  | Least Square Estimate                           |                     |
| 7  | Standard Error                                  |                     |
| 8  | <b>Computed Values</b>                          |                     |
| 9  |                                                 | df =B4-2            |
| 10 |                                                 | t =TINV(1-B5,B4-2)  |
| 11 |                                                 | half-width =B10*B7  |
| 12 | <b>Confidence Interval</b>                      |                     |
| 13 |                                                 | Lower Limit =B6-B11 |
| 14 |                                                 | Upper Limit =B6+B11 |

Once you have the worksheet, the **Least Squares Estimates** and the **Standard Error** values can be typed in, or can be copied from the regression results on the food expenditure worksheet.

|    | A                                               | B                       |
|----|-------------------------------------------------|-------------------------|
| 1  | <i>Interval Estimation in Simple Regression</i> |                         |
| 2  |                                                 |                         |
| 3  | <b>Data Input</b>                               |                         |
| 4  | Sample Size                                     | 40                      |
| 5  | Confidence Level                                | 0.95                    |
| 6  | Least Square Estimate                           | 10.2096425              |
| 7  | Standard Error                                  | 2.093263461             |
| 8  | <b>Computed Values</b>                          |                         |
| 9  |                                                 | df 38                   |
| 10 |                                                 | t 2.024394147           |
| 11 |                                                 | half-width 4.237590298  |
| 12 | <b>Confidence Interval</b>                      |                         |
| 13 |                                                 | Lower Limit 5.972052202 |
| 14 |                                                 | Upper Limit 14.4472328  |

## 3.2 HYPOTHESIS TESTING

Inference in the linear regression model includes tests of hypotheses about parameters which also depend upon Student's  $t$ -distribution. One- or two-tail general tests can be calculated by methods similar to the confidence interval construction. The required ingredients are results from the least squares estimation and the ability to use the Excel functions **TINV** and **TDIST**.

### 3.2.1 Right-Tail tests

To test the hypothesis  $H_0: \beta_2 = 0$  against the alternative that it is positive ( $>0$ ), as described in Chapter 3.4.1a of *POE*, we use a one-tail significance test. For this purpose, we must compute the value of the critical values that define the rejection region, the test statistic and the  $p$ -value of the test.

- If we choose the  $\alpha = .05$  level of significance, the critical value is the 95<sup>th</sup> percentile of the  $t_{(38)}$  distribution which can be computed by the **TINV** function discussed earlier. **TINV**(0.10, 38) = 1.685954461. The value returned is the right tail critical value. Beware that the Excel function **TINV**(*probability, degrees of freedom*) returns the value such that the area in the *two-tails* of the  $t(df)$  distribution equals the given *probability* value. Thus if we want a critical value such that  $\alpha = .05$  is in the right tail, we must provide the TINV function with *probability* = .10.
- The test statistic is the ratio of the estimate  $b_2$  to its standard error,  $se(b_2)$ .
- For the  $p$ -value, we use the **TDIST** function where the amount of probability in the right-tail of a  $t$ -distribution with 38 degrees of freedom can be computed. To obtain this value, we find **TDIST** from the Statistical Function category under either **Sum** drop down menu, or **Paste Function**.

TDIST

X: 4.877380554 = 4.877380554

Deg\_freedom: 38 = 38

Tails: 1 = 1

Returns the Student's t-distribution. = 9.72931E-06

Tails specifies the number of distribution tails to return: one-tailed distribution = 1; two-tailed distribution = 2.

Formula result = 9.72931E-06

[Help on this function](#) OK Cancel

This work can be simplified by using a template that can be used when needed.

|    | A                                                            | B           |
|----|--------------------------------------------------------------|-------------|
| 1  | Right Tailed Test for Simple Regression Model                |             |
| 2  | <b>Data Input</b>                                            |             |
| 3  | Sample Size                                                  |             |
| 4  | Estimate                                                     |             |
| 5  | Standard Error                                               |             |
| 6  | Ho                                                           |             |
| 7  | Alpha                                                        |             |
| 8  | <b>Computed Values</b>                                       |             |
| 9  | df = B3-2                                                    |             |
| 10 | t = (B4-B6)/B5                                               |             |
| 11 | <b>Right-Tail Test</b>                                       |             |
| 12 | Right Critical Value = TINV(2*B7,B9)                         |             |
| 13 | Decision = IF(B10>=B16, "Reject Ho", "Do Not Reject Null")   |             |
| 14 | p-value = IF(B10>0, TDIST(B10,B9,1), 1-TDIST(ABS(B10),B9,1)) |             |
| 15 |                                                              |             |
| 16 |                                                              |             |
| 17 | Right Tailed Test for Simple Regression Model                |             |
| 18 | <b>Data Input</b>                                            |             |
| 19 | Sample Size                                                  | 40          |
| 20 | Estimate                                                     | 10.2096425  |
| 21 | Standard Error                                               | 2.093263461 |
| 22 | Ho                                                           | 0           |
| 23 | Alpha                                                        | 0.05        |
| 24 | <b>Computed Values</b>                                       |             |
| 25 | df                                                           | 38          |
| 26 | t                                                            | 4.877380554 |
| 27 | <b>Right-Tail Test</b>                                       |             |
| 28 | Right Critical Value                                         | 1.685954461 |
| 29 | Decision                                                     | Reject Ho   |
| 30 | p-value                                                      | 9.72931E-06 |
| 31 |                                                              |             |

The  $t$ -statistic value 4.8773 falls in the rejection region, and  $p$ -value is less than the level of significance  $\alpha$ , thus we reject this null hypothesis.

|                                               |  |             |
|-----------------------------------------------|--|-------------|
| Right Tailed Test for Simple Regression Model |  |             |
| <b>Data Input</b>                             |  |             |
| Sample Size                                   |  | 40          |
| Estimate                                      |  | 10.2096425  |
| Standard Error                                |  | 2.093263461 |
| Ho                                            |  | 5           |
| Alpha                                         |  | 0.05        |
| <b>Computed Values</b>                        |  |             |
| df                                            |  | 38          |
| t                                             |  | 2.488765794 |
| <b>Right-Tail Test</b>                        |  |             |
| Right Critical Value                          |  | 1.685954461 |
| Decision                                      |  | Reject Ho   |
| p-value                                       |  | 0.008658183 |

To test an economic hypothesis such as  $H_0 : \beta_2 \leq 5$  against  $H_1 : \beta_2 > 5$ , the same steps are followed except for the construction of the  $t$ -statistic. This can be accomplished by replacing the 0 with 5 in our template.

The  $t$ -statistic value 2.4887 falls in the rejection region, and  $p$ -value is smaller than the level of significance  $\alpha$ , thus we reject this null hypothesis.

### 3.2.2 Left-Tail tests

To test the significance, we test  $H_0 : \beta_2 \geq 0$  against  $H_1 : \beta_2 < 0$ . The value of the  $t$ -statistic for this null and alternative hypothesis is the same as for a right-tailed test.

- If we choose the  $\alpha = .05$  level of significance, the critical value is the 5<sup>th</sup> percentile of the  $t_{(38)}$  distribution which can be computed by the **TINV** function discussed earlier.  
 $-\text{TINV}(0.10, 38) = -1.685954461$ . The value returned is the right tail critical value.
- The test statistic is the ratio of the estimate  $b_2$  to its standard error,  $se(b_2)$ .
- The  $p$ -value, is the area to the left of the calculated  $t$ -statistic.

Let's plug in the left tail values into our template:

|    | A                                          | B                                                      |
|----|--------------------------------------------|--------------------------------------------------------|
| 1  | Left Tail Test for Simple Regression Model |                                                        |
| 2  | <b>Data Input</b>                          |                                                        |
| 3  | Sample Size                                |                                                        |
| 4  | Estimate                                   |                                                        |
| 5  | Standard Error                             |                                                        |
| 6  | Ho                                         |                                                        |
| 7  | Alpha                                      |                                                        |
| 8  | <b>Computed Values</b>                     |                                                        |
| 9  | df                                         | =B3-2                                                  |
| 10 | t                                          | =(B4-B6)/B5                                            |
| 11 | <b>Left-Tail Test</b>                      |                                                        |
| 12 | Left Critical Value                        | =TINV(2*B7,B9)                                         |
| 13 | Decision                                   | =IF(B10<=B12,"Reject Ho","Do Not Reject Null")         |
| 14 | p-value                                    | =IF(B10<0,TDIST(ABS(B10),B9,1),1-TDIST(ABS(B10),B9,1)) |
| 15 |                                            |                                                        |
| 16 | Left Tail Test for Simple Regression Model |                                                        |
| 17 | <b>Data Input</b>                          |                                                        |
| 18 | Sample Size                                | 40                                                     |
| 19 | Estimate                                   | 10.2096425                                             |
| 20 | Standard Error                             | 2.093263461                                            |
| 21 | Ho                                         | 0                                                      |
| 22 | Alpha                                      | 0.05                                                   |
| 23 | <b>Computed Values</b>                     |                                                        |
| 24 | df                                         | 38                                                     |
| 25 | t                                          | 4.877380554                                            |
| 26 | <b>Left-Tail Test</b>                      |                                                        |
| 27 | Left Critical Value                        | -1.685954461                                           |
| 28 | Decision                                   | Do Not Reject Null                                     |
| 29 | p-value                                    | 0.999990271                                            |

To test the null hypothesis that  $\beta_2 \geq 12$  against the alternative  $\beta_2 < 12$ , we use the template and plug in the input numbers. The  $t$ -statistic value  $-1.6859$  does not fall in the rejection region, and  $p$ -value is greater than the level of significance  $\alpha$ , thus we fail to reject this null hypothesis.

|                                            |                    |
|--------------------------------------------|--------------------|
| Left Tail Test for Simple Regression Model |                    |
| <b>Data Input</b>                          |                    |
| Sample Size                                | 40                 |
| Estimate                                   | 10.2096425         |
| Standard Error                             | 2.093263461        |
| Ho                                         | 12                 |
| Alpha                                      | 0.05               |
| <b>Computed Values</b>                     |                    |
| df                                         | 38                 |
| t                                          | -0.85529487        |
| <b>Left-Tail Test</b>                      |                    |
| Left Critical Value                        | -1.685954461       |
| Decision                                   | Do Not Reject Null |
| p-value                                    | 0.198874343        |

### 3.2.3 Two-Tail tests

For the two tail test of the null hypothesis that  $\beta_2 = 0$  against the alternative that  $\beta_2 \neq 0$  the same steps are taken. We can plug the necessary information into the template.

|    | A                                         | B                     |
|----|-------------------------------------------|-----------------------|
| 1  | Two Tail Test for Simple Regression Model |                       |
| 2  | <b>Data Input</b>                         |                       |
| 3  | Sample Size                               |                       |
| 4  | Estimate                                  |                       |
| 5  | Standard Error                            |                       |
| 6  | Ho                                        |                       |
| 7  | Alpha                                     |                       |
| 8  | <b>Computed Values</b>                    |                       |
| 9  | df                                        | =B3-2                 |
| 10 | t                                         | =(B4-B6)/B5           |
| 11 | <b>Two-Tail Test</b>                      |                       |
| 12 | Absolute Critical Value                   | =TINV(B7,B9)          |
| 13 | Decision                                  | Reject Ho             |
| 14 | p-value                                   | =TDIST(ABS(B10),B9,2) |
| 15 |                                           |                       |
| 16 | Two Tail Test for Simple Regression Model |                       |
| 17 | <b>Data Input</b>                         |                       |
| 18 | Sample Size                               | 40                    |
| 19 | Estimate                                  | 10.2096425            |
| 20 | Standard Error                            | 2.093263461           |
| 21 | Ho                                        | 0                     |
| 22 | Alpha                                     | 0.05                  |
| 23 | <b>Computed Values</b>                    |                       |
| 24 | df                                        | 38                    |
| 25 | t                                         | 4.877380554           |
| 26 | <b>Two-Tail Test</b>                      |                       |
| 27 | Absolute Critical Value                   | 2.024394147           |
| 28 | Decision                                  | Reject Ho             |
| 29 | p-value                                   | 1.94586E-05           |

Since the  $t$ -statistic value 4.8773 falls in the rejection region, and  $p$ -value is smaller than the level of significance  $\alpha$ , thus we reject this null hypothesis. This test is also carried out by Excel within the **Regression Summary Output** labeled  $t$ -Statistic and  $p$ -value.

|           | Coefficients | Standard Error | t Stat      | P-value     |
|-----------|--------------|----------------|-------------|-------------|
| Intercept | 83.41600997  | 43.41016192    | 1.921577951 | 0.062182379 |
| Income    | 10.2096425   | 2.093263461    | 4.877380554 | 1.94586E-05 |

To test the null hypothesis that  $\beta_2 = 12.5$  against the alternative  $\beta_2 \neq 12.5$ , we use the two-tailed test template and plug in the input numbers. The  $t$ -statistic value  $-1.6859$  does not fall in the rejection region, and  $p$ -value is greater than the level of significance  $\alpha$ , thus we fail to reject this null hypothesis.

|                                           |              |
|-------------------------------------------|--------------|
| Two Tail Test for Simple Regression Model |              |
| <b>Data Input</b>                         |              |
| Sample Size                               | 40           |
| Estimate                                  | 10.2096425   |
| Standard Error                            | 2.093263461  |
| Ho                                        | 12.5         |
| Alpha                                     | 0.05         |
| <b>Computed Values</b>                    |              |
| df                                        | 38           |
| t                                         | -1.094156346 |
| <b>Two-Tail Test</b>                      |              |
| Absolute Critical Value                   | 2.024394147  |
| Decision                                  | Reject Ho    |
| p-value                                   | 0.280773757  |



# **CHAPTER 4**

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

### **CHAPTER OUTLINE**

- 4.1 Prediction for the Food Expenditure Model
  - 4.1.1 Calculating the standard error of the forecast
  - 4.1.2 Prediction interval
- 4.2 Measuring Goodness-of-Fit
  - 4.2.1  $R^2$
  - 4.2.2 Covariance and correlation analysis
- 4.3 Residual Diagnostics
  - 4.3.1 The Jarque-Bera test
- 4.4 Modeling Issues
  - 4.4.1 Scaling the data
  - 4.4.2 The log-linear model
  - 4.4.3 The linear-log model
  - 4.4.4 The log-log model
- 4.5 More Examples
  - 4.5.1 Residual analysis with wheat data
  - 4.5.2 Log-linear model with wage data
  - 4.5.3 Generalized  $R^2$

### **4.1 PREDICTION IN THE FOOD EXPENDITURE MODEL**

We have already illustrated how to obtain the predicted values for the food expenditure for a household in Chapter 2. In this chapter, we will calculate the standard error of the forecasted value and construct a prediction interval.

#### **4.1.1 Calculating the standard error of the forecast**

Recall from Section 2.6 of this manual, the forecasted value of household food expenditure for a household with income of \$2000 per week is calculated as \$287.6088614. Now, we will compute the standard error of the forecasted value where forecast error is calculated as

$$f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)$$

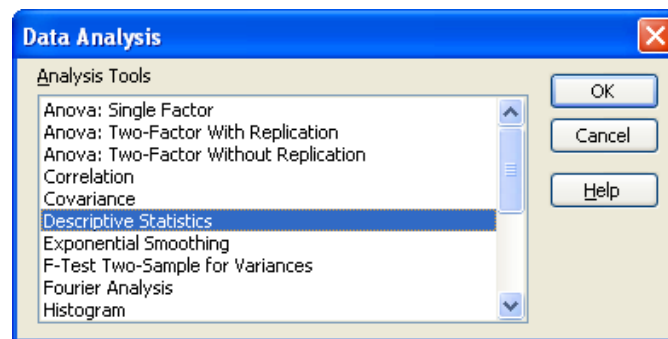
The estimated variance for the forecast error is

$$\text{var}(f) = \hat{\sigma}^2 \left[ 1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2 \text{var}(b_2)$$

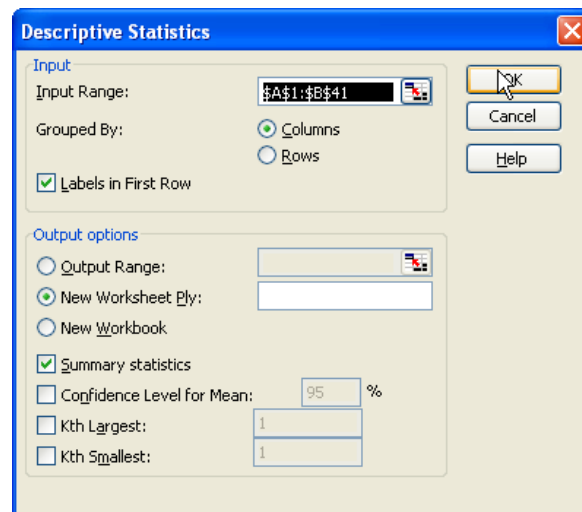
And the square root of the estimated variance is the standard error of the forecast.

$$se(f) = \sqrt{\text{var}(f)}$$

To calculate the forecast error for the food expenditure data, open the *food.xls* file. In addition to the summary regression output (see Section 2.3 of this manual), we will need the sample mean of income. To obtain the sample mean of income select **Tools/Data Analysis** and from the menu choose **Descriptive Statistics**.



In the dialog window specify the data range and ask for summary statistics.



Excel will provide the univariate summary statistics for Food Expenditure and Income variables. We will use the mean of income and sample size in the calculation of the standard error of the forecast.

The screenshot shows an Excel spreadsheet with two columns of data: 'Food\_Exp' (column B) and 'Income' (column C). The rows show various statistical measures for both variables. Two red arrows point to the 'Mean' values: 19.60475 for 'Income' and 784.19 for 'Food\_Exp'.

|    | A                  | B            | C                  | D        |
|----|--------------------|--------------|--------------------|----------|
| 1  | Food_Exp           |              | Income             |          |
| 2  |                    |              |                    |          |
| 3  | Mean               | 283.5734993  | Mean               | 19.60475 |
| 4  | Standard Error     | 17.81551026  | Standard Error     | 1.082728 |
| 5  | Median             | 264.479996   | Median             | 20.03    |
| 6  | Mode               | #N/A         | Mode               | #N/A     |
| 7  | Standard Deviation | 112.6751802  | Standard Deviation | 6.847773 |
| 8  | Sample Variance    | 12695.69623  | Sample Variance    | 46.89199 |
| 9  | Kurtosis           | -0.002430221 | Kurtosis           | 0.484556 |
| 10 | Skewness           | 0.511465877  | Skewness           | -0.65119 |
| 11 | Range              | 477.949974   | Range              | 29.71    |
| 12 | Minimum            | 109.709999   | Minimum            | 3.69     |
| 13 | Maximum            | 587.659973   | Maximum            | 33.4     |
| 14 | Sum                | 11342.93997  | Sum                | 784.19   |
| 15 | Count              | 40           | Count              | 40       |
| 16 |                    |              |                    |          |

Now, we can go back to **Regression Output**, and plug in the necessary numbers into our formula and calculate the standard error of the forecast. Recall from above that the formula is

$$\text{var}(f) = \hat{\sigma}^2 \left[ 1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2 \text{var}(b_2)$$

and

$$se(f) = \sqrt{\text{var}(f)}$$

The formula indicates that, the farther  $x_0$  is from the sample mean  $\bar{x}$ , the larger the variance of the prediction error, the smaller the sample size, the larger the forecast error and the less reliable the prediction is likely to be.

Microsoft Excel - food.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

SnagIt Window

B22  $\text{f} = \text{SQRT}(\$D\$13 + (\$D\$13/40) + (A22 - 19.60475005)^2 * \$C\$18^2)$

|    | A          | B              | C                  | D                  | E           |
|----|------------|----------------|--------------------|--------------------|-------------|
| 9  |            |                |                    |                    |             |
| 10 | ANOVA      |                |                    |                    |             |
| 11 |            | df             | SS                 | MS                 | F           |
| 12 | Regression | 1              | 190626.9788        | 190626.9788        | 23.78884107 |
| 13 | Residual   | 38             | 304505.1742        | <b>8013.294058</b> |             |
| 14 | Total      | 39             | 495132.153         |                    |             |
| 15 |            |                |                    |                    |             |
| 16 |            | Coefficients   | Standard Error     | t Stat             | P-value     |
| 17 | Intercept  | 83.41600997    | 43.41016192        | 1.921577951        | 0.062182379 |
| 18 | Income     | 10.2096425     | <b>2.093263461</b> | 4.877380554        | 1.94586E-05 |
| 19 |            |                |                    |                    |             |
| 20 |            |                |                    |                    |             |
| 21 | INCOME     | SE of forecast |                    |                    |             |
| 22 | 20.0000    | 90.6328        |                    |                    |             |
| 23 | 25.0000    | 104.6529       |                    |                    |             |
| 24 | 30.0000    | 110.2597       |                    |                    |             |

To calculate the standard error of the forecast for  $x_0$ , we will need the  $\sigma^2$  which is the **Mean Squared Residual (MSR)** from cell D13 in the **ANOVA** table, the  $\text{var}(b_2)$  which is the square of the standard error of income from cell C18 and  $x_0$  which is \$2000. Once we type in the formula, it is possible to make the same calculation for different values of income.

#### 4.1.2 Prediction interval

We construct a  $100(1 - \alpha)\%$  prediction interval as

$$\hat{y}_0 \pm t_c se(f)$$

Since forecasted value and the standard error of forecast have been already calculated, constructing the confidence interval is very straightforward. Recall that we can obtain the  $t_c$  values as shown in Section 3.1.2 and the forecasted values in Section 2.6 of this manual.

Microsoft Excel - food.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

SnagIt Window

C22 fx =+D22-2.02439416391\*B22

|    | A          | B              | C              | D                | E           |                |
|----|------------|----------------|----------------|------------------|-------------|----------------|
| 9  |            |                |                |                  |             |                |
| 10 | ANOVA      |                |                |                  |             |                |
| 11 |            | df             | SS             | MS               | F           | Significance F |
| 12 | Regression | 1              | 190626.9788    | 190626.9788      | 23.78884107 | 1.94586E-05    |
| 13 | Residual   | 38             | 304505.1742    | 8013.294058      |             |                |
| 14 | Total      | 39             | 495132.153     |                  |             |                |
| 15 |            |                |                |                  |             |                |
| 16 |            | Coefficients   | Standard Error | t Stat           | P-value     | Lower Bound    |
| 17 | Intercept  | 83.41600997    | 43.41016192    | 1.921577951      | 0.062182379 | -4.00000       |
| 18 | Income     | 10.2096425     | 2.093263461    | 4.877380554      | 1.94586E-05 | 5.00000        |
| 19 |            |                |                |                  |             |                |
| 20 |            |                |                |                  |             |                |
| 21 | INCOME     | SE of forecast | Lower Bound    | Forecasted Value | Upper Bound |                |
| 22 | 20.0000    | 90.6328        | 104.1322       | 287.6088         | 471.0854    |                |
| 23 | 25.0000    | 104.6529       | 126.7985       | 338.6571         | 550.5181    |                |
| 24 | 30.0000    | 110.2597       | 166.4962       | 389.7053         | 612.9144    |                |
| 25 |            |                |                |                  |             |                |

We can also create a worksheet, name it predictions and calculate standard error of forecast and the prediction intervals for specified values of income.

## 4.2 MEASURING GOODNESS-OF-FIT

The ANOVA table in the regression output provides the goodness-of-fit measures.

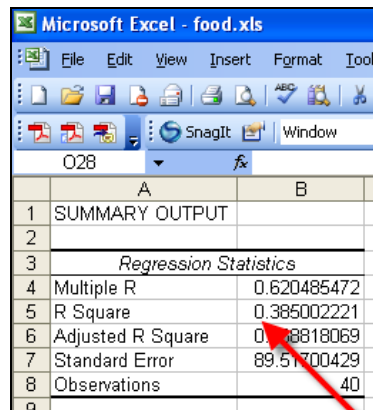
$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

|    |               |                 |             |                   |             |                             |
|----|---------------|-----------------|-------------|-------------------|-------------|-----------------------------|
| 9  |               |                 |             |                   |             |                             |
| 10 |               |                 |             |                   |             |                             |
| 11 |               | SSR (explained) |             | SSE (unexplained) |             |                             |
| 12 |               |                 |             |                   |             |                             |
| 13 | number of x's |                 |             |                   |             |                             |
| 14 | ANOVA         |                 |             |                   |             |                             |
| 15 |               | df              | SS          | MS                | F           | Significance F              |
| 16 | Regression    | 1               | 190626.9788 | 190626.9788       | 23.78884107 | 1.94586E-05                 |
| 17 | Residual      | 38              | 304505.1742 | 8013.294058       |             |                             |
| 18 | Total         | 39              | 495132.153  |                   |             |                             |
| 19 |               |                 |             |                   |             |                             |
| 20 | N-2           |                 |             |                   |             |                             |
| 21 | N-1           |                 |             |                   |             | SST (explained+unexplained) |

### 4.2.1 Calculating $R^2$

In the simple regression model, the  $R^2$  is the square of the sample correlation between the  $x$  and  $y$  variables. It is calculated as  $R^2 = SSR/SST = 1 - SSE/SST$  is reported in standard regression output.

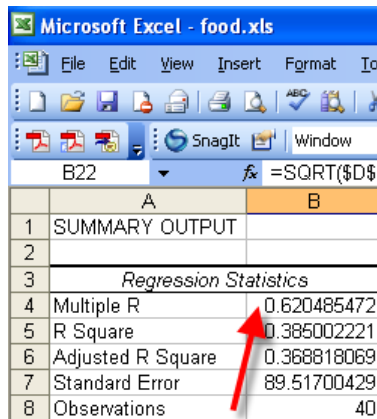


Microsoft Excel - food.xls

|   | A                     | B           |
|---|-----------------------|-------------|
| 1 | SUMMARY OUTPUT        |             |
| 2 |                       |             |
| 3 | Regression Statistics |             |
| 4 | Multiple R            | 0.620485472 |
| 5 | R Square              | 0.385002221 |
| 6 | Adjusted R Square     | 0.368818069 |
| 7 | Standard Error        | 89.51700429 |
| 8 | Observations          | 40          |

### 4.2.2 Covariance and correlation analysis

The covariance and correlation can tell us about the linear relationship between two variables, a primary concern of linear regression. Specifically, the covariance tells us the direction of the linear relationship, while the correlation is a measure of the strength (and direction) of the linear relationship. **Multiple R**, in the simple regression output, gives us the square root of  $R^2$  which is the correlation between  $x$  and  $y$ .

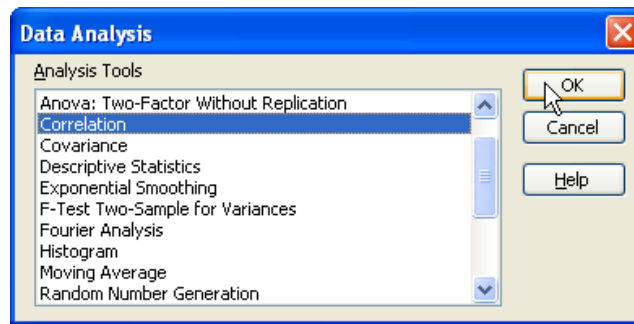


Microsoft Excel - food.xls

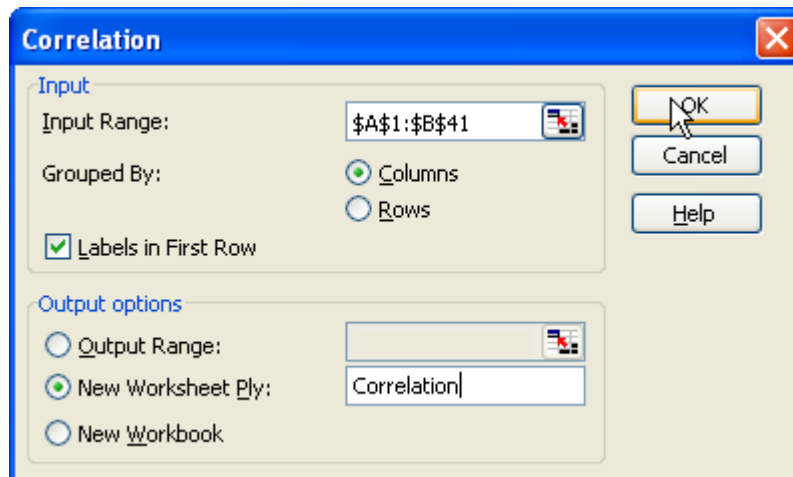
|   | A                     | B           |
|---|-----------------------|-------------|
| 1 | SUMMARY OUTPUT        |             |
| 2 |                       |             |
| 3 | Regression Statistics |             |
| 4 | Multiple R            | 0.620485472 |
| 5 | R Square              | 0.385002221 |
| 6 | Adjusted R Square     | 0.368818069 |
| 7 | Standard Error        | 89.51700429 |
| 8 | Observations          | 40          |

Formula bar: B22 =SQRT(\$D\$5)

A more general way to calculate covariance and correlation can be achieved by utilizing the **Data Analysis** under the **Tools** menu.



The sample correlation coefficient,  $r$ , measures the direction and strength of the linear relationship between two variables and is between  $-1$  and  $1$ . To obtain the sample correlation coefficient, choose correlation from the **Tool/Data Analysis** menu and click **OK**. The **Correlation dialog box** will appear. Fill in the appropriate input range and be sure to check the **Labels in First Row** box since labels are included. Place the output on a new worksheet named **Correlation** and click **OK**.

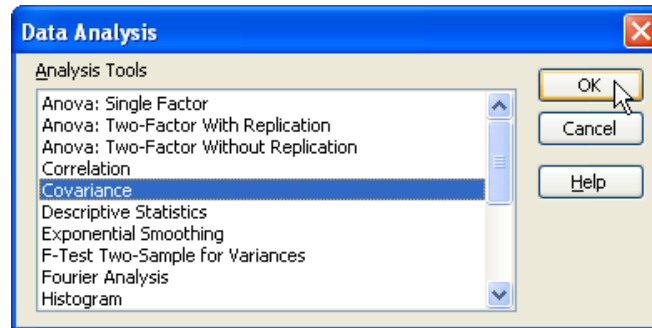


The correlations are:

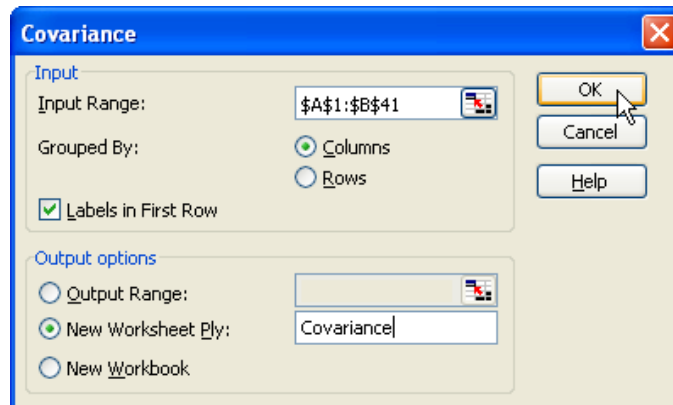
| Microsoft Excel - Book1      |          |          |        |
|------------------------------|----------|----------|--------|
| File Edit View Insert Format |          |          |        |
| M24                          |          |          |        |
|                              | A        | B        | C      |
| 1                            |          | Food_Exp | Income |
| 2                            | Food_Exp | 1        |        |
| 3                            | Income   | 0.620485 | 1      |
| 4                            |          |          |        |

The results will appear in the new worksheet, **Correlation**. You may need to format the worksheet by choosing **Format/Column/AutoFit Selection**. The estimated correlation between food expenditures and weekly income is  $0.620485$  which is the value given as **Multiple R** in the regression output summary. Values on the diagonal of the correlation matrix will always equal  $1$ .

To find the sample covariance matrix, click on **Tool/Data Analysis** menu again and highlight **Covariance** and click **OK**.



The **Covariance dialog box** will appear. Fill in the appropriate input range and be sure to check the **Labels in First Row** box since labels are included. Place the output on a new worksheet named **Covariance** and click **OK**.



The covariance matrix will appear on the new worksheet, but needs to be formatted. Choose **Format/Column/Auto Fit Selection**.

|   | A        | B        | C        |
|---|----------|----------|----------|
| 1 |          | Food_Exp | Income   |
| 2 | Food_Exp | 12378.3  |          |
| 3 | Income   | 466.7817 | 45.71969 |
| 4 |          |          |          |

The diagonal elements of the covariance matrix are the estimated sample variances. The covariance between food expenditures and weekly income is positive, suggesting a positive linear

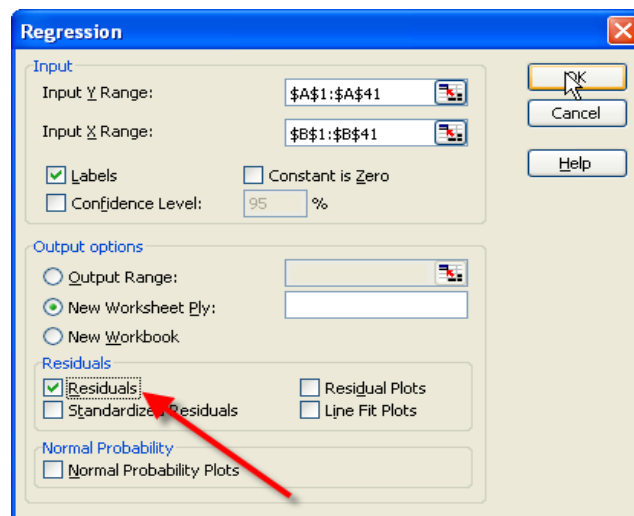


relationship. The value of the covariance, 466.7817, does not, however, tell you the strength of that linear relationship.

### 4.3 Residual Diagnostics

Every time a regression model is estimated, certain regression diagnostics need to be carried out. By analyzing the residuals of the fitted model, we may be able to detect model specification problems. A histogram of the residuals can suggest the distribution of the errors, and the Jarque-Bera test statistic can be used to formally test for normality. Both of these functions are important since hypothesis testing and interval estimations are based on distributional assumptions.

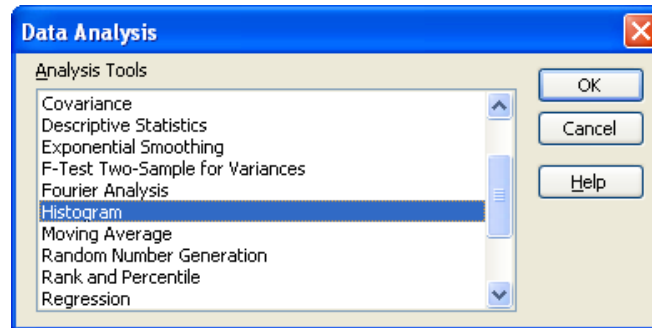
In order to create a histogram of the residuals, we need to rerun the food expenditures model and choose the **Residuals** output option.



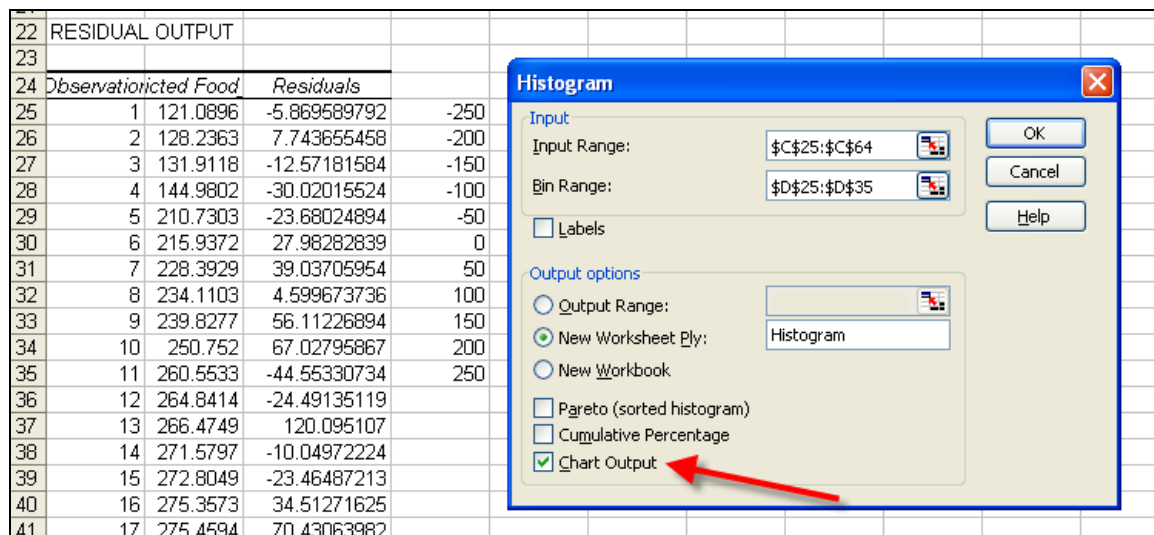
Excel will provide the residuals for each observation, in addition to the standard Regression output. Examine the values of the residuals, noting the lowest and highest values. Create a *BIN* column next to the residuals column and determine the category values for the histogram. In this column, enter the values -250, -200, -150, -100, -50, 0, 50, 100, 150, 200 and 250.

| RESIDUAL OUTPUT |                    |              |      |
|-----------------|--------------------|--------------|------|
| Observation     | Predicted Food Exp | Residuals    | BINS |
| 1               | 121.0895846        | -5.869584573 | -250 |
| 2               | 128.2363347        | 7.743665349  | -200 |
| 3               | 131.9118061        | -12.57180612 | -150 |
| 4               | 144.9801491        | -30.02014912 | -100 |
| 5               | 210.7302498        | -23.68024983 | -50  |
| 6               | 215.9371677        | 27.98283225  | 0    |
| 7               | 228.3929322        | 39.03706783  | 50   |
| 8               | 234.1103322        | 4.59966777   | 100  |
| 9               | 239.8277323        | 56.11226771  | 150  |
| 10              | 250.7520503        | 67.02794973  | 200  |
| 11              | 260.5533075        | -44.55330752 | 250  |

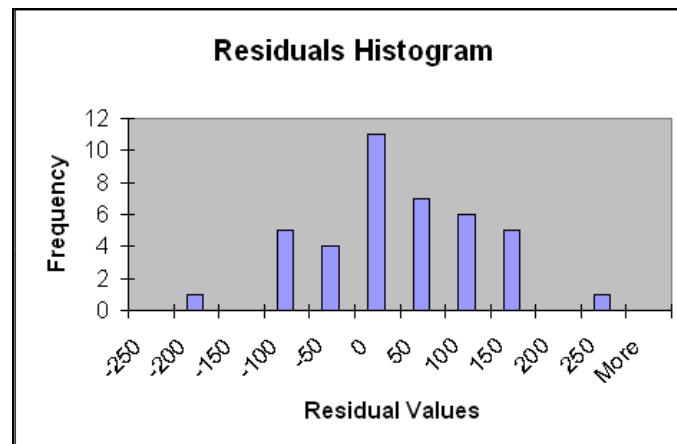
Once you have entered the *BIN* column, choose **Tools/Data Analysis/Histogram** from the menu bar, and click **OK**.



The **Histogram** dialog box will appear.



Fill in the data ranges, by highlighting the residuals for the **Input Range**, and highlight the values created in the *BIN* column for the **Bin Range**. Place the output on a new worksheet called **Histogram**. After checking the **Chart Output** box, click **OK**.



The **Bin** values and **Frequencies** appear, along with the histogram. Format the histogram graph as needed. (Remove legend, resize, rename title, etc). The residuals seem to be centered around zero, but the symmetry of the distribution seems questionable.

#### 4.3.1 Jarque-Bera test for normality

A formal test of normality, the **Jarque-Bera** test, uses skewness and kurtosis, which can be easily estimated with Excel.

As discussed earlier in Section 4.1.1, Excel has a tool for calculating the **Descriptive Statistics**, which can be found on the **Tools/Data Analysis** menu. Skewness and kurtosis can also be found among the statistics calculated.

**Skewness** is a measure of asymmetry of a distribution about its mean. For a sample  $x_1, x_2, \dots, x_T$  an empirical measure of skewness is

$$S = \frac{\Sigma (x_i - \bar{x})^3 / T}{\sigma_x^3}$$

where

$$\sigma_x = \sqrt{\frac{1}{T} \sum_{i=1}^T (x_i - \bar{x})^2}$$

**Kurtosis** measures the peakedness, or flatness, of a distribution. An empirical measure of Kurtosis is

$$K = \frac{\Sigma (x_i - \bar{x})^4 / T}{\sigma_x^4}$$

In the case of the least squares residuals,  $\hat{e}$ , the formulas simplify because  $\sum_{i=1}^T \hat{e}_i = 0$ , making  $\bar{\hat{e}} = 0$ . Thus the formulas for skewness and kurtosis of the least squares residuals are

$$S = \frac{\Sigma \hat{e}_i^3 / T}{\tilde{\sigma}^3}, \quad K = \frac{\Sigma \hat{e}_i^4 / T}{\tilde{\sigma}^4}$$

respectively where

$$\tilde{\sigma} = \sqrt{\frac{1}{T} \sum_{i=1}^T \hat{e}_i^2}$$

Skewness measures the symmetry of the data, a value of zero indicating perfect symmetry. Kurtosis refers to the "peakedness" of the distribution, with a value of 3 for a normal distribution. Using these measures, the test statistic for the Jarque-Bera test for normality is

$$JB = \frac{T}{6} \left( S^2 + \frac{(K-3)^2}{4} \right)$$

where  $S$  is skewness and  $K$  is kurtosis. This test statistic follows a chi-square distribution with 2 degrees of freedom. Below is preparation of the worksheet for this calculation for the food expenditures model.

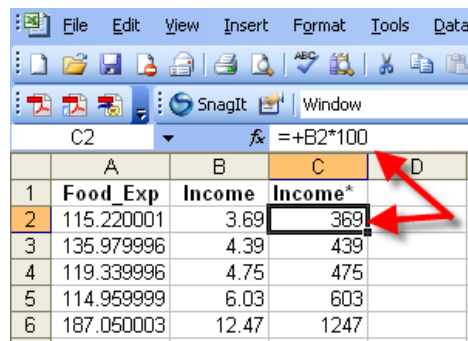
Return to the worksheet containing the regression results and the residuals. Copy the column containing the least squares residuals to a new worksheet and create three additional columns and label them 'ehatsquared', 'ehatcubed' and 'ehat to the fourth'. The formulas for these functions are ^2, ^3 and ^4 respectively and is illustrated in the below figure. Once you copy the formulas down the columns, you will be in a position to compute the  $S$ ,  $K$  and  $JB$  statistics. With formulas showing, this portion of the worksheet should look like

|   | A                | B                   | C                | D                         | E                     | F                            | G |
|---|------------------|---------------------|------------------|---------------------------|-----------------------|------------------------------|---|
| 1 | <i>Residuals</i> | <i>ehat-squared</i> | <i>ehatcubed</i> | <i>ehat to the fourth</i> | <b>NORMALITY TEST</b> |                              |   |
| 2 | -5.869589792     | =+A2^2              | =+A2^3           | =+A2^4                    | sigmatilda-square     | =+SUM(B2:B41)/40             |   |
| 3 | 7.743655458      | =+A3^2              | =+A3^3           | =+A3^4                    | sigmatilda            | =+SQRT(F2)                   |   |
| 4 | -12.57181584     | =+A4^2              | =+A4^3           | =+A4^4                    | Skewness              | =SUM(C2:C41)/(40*F3^3)       |   |
| 5 | -30.02015524     | =+A5^2              | =+A5^3           | =+A5^4                    | Kurtosis              | =SUM(D2:D41)/(40*F3^4)       |   |
| 6 | -23.68024894     | =+A6^2              | =+A6^3           | =+A6^4                    | JB                    | =+(40/6)*(F4^2+((F5-3)^2)/4) |   |
| 7 | 27.98282839      | =+A7^2              | =+A7^3           | =+A7^4                    | p-value               | =CHIDIST(F6,2)               |   |

## 4.4 MODELING ISSUES

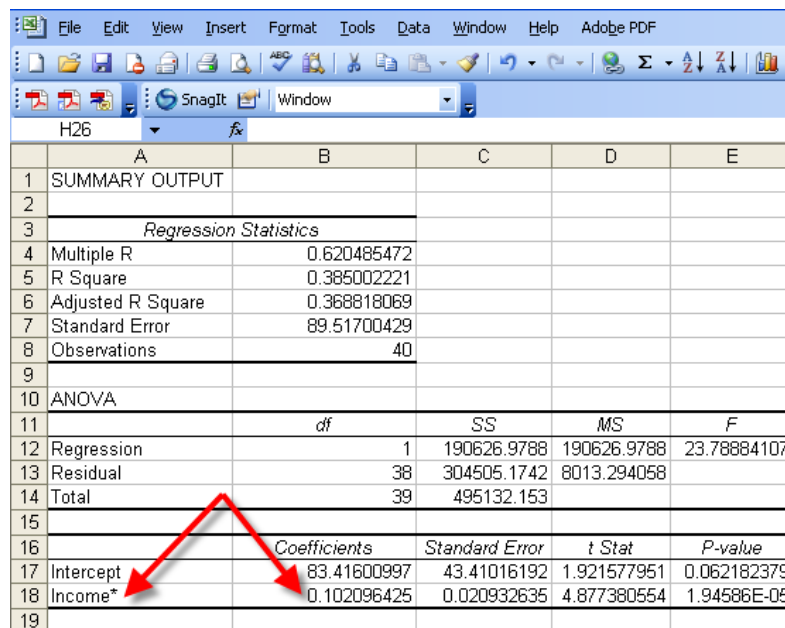
### 4.4.1 Scaling the data

Change the scale of the variables can easily be done on the worksheet containing the data. Label the empty column to the right of the independent variable *INCOME\**. In the first empty cell of this column, type **=B2\*100**. Copy the formula down the column.



|   | A          | B      | C       | D |
|---|------------|--------|---------|---|
| 1 | Food Exp   | Income | Income* |   |
| 2 | 115.220001 | 3.69   | 369     |   |
| 3 | 135.979996 | 4.39   | 439     |   |
| 4 | 119.339996 | 4.75   | 475     |   |
| 5 | 114.959999 | 6.03   | 603     |   |
| 6 | 187.050003 | 12.47  | 1247    |   |

Estimate a regression, using the new independent variable, *INCOME\** instead of the original independent variable, *INCOME*.



|    | A                            | B                   | C           | D           | E           |
|----|------------------------------|---------------------|-------------|-------------|-------------|
| 1  | SUMMARY OUTPUT               |                     |             |             |             |
| 2  |                              |                     |             |             |             |
| 3  | <i>Regression Statistics</i> |                     |             |             |             |
| 4  | Multiple R                   | 0.620485472         |             |             |             |
| 5  | R Square                     | 0.385002221         |             |             |             |
| 6  | Adjusted R Square            | 0.368818069         |             |             |             |
| 7  | Standard Error               | 89.51700429         |             |             |             |
| 8  | Observations                 | 40                  |             |             |             |
| 9  |                              |                     |             |             |             |
| 10 | ANOVA                        |                     |             |             |             |
| 11 |                              | df                  | SS          | MS          | F           |
| 12 | Regression                   | 1                   | 190626.9788 | 190626.9788 | 23.78884107 |
| 13 | Residual                     | 38                  | 304505.1742 | 8013.294058 |             |
| 14 | Total                        | 39                  | 495132.153  |             |             |
| 15 |                              |                     |             |             |             |
| 16 |                              | <i>Coefficients</i> |             |             |             |
| 17 | Intercept                    | 83.41600997         | 43.41016192 | 1.921577951 | 0.062182379 |
| 18 | Income*                      | 0.102096425         | 0.020932635 | 4.877380554 | 1.94586E-05 |
| 19 |                              |                     |             |             |             |

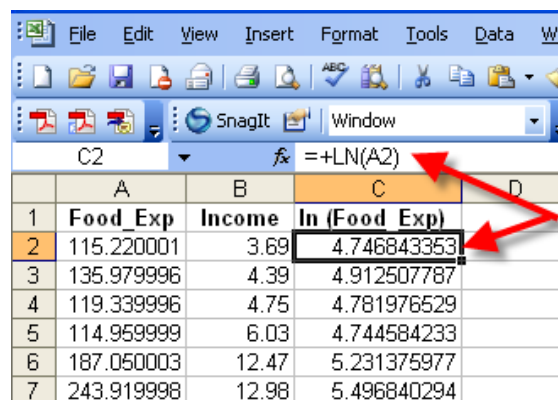
The coefficient on *INCOME* has changed and so has its standard error. Everything else in the regression output remained the same. When reporting results, be sure to note the appropriate units of measure for both food expenditures and weekly income.

#### 4.4.2 The log-linear model

The use of logarithmic transformations are very common with economic data. Transforming the dependent variable using the log function will result in changing the interpretation of the regression equation. To estimate the log-linear version of the food expenditure model, we need to transform the dependent variable.

$$\ln(y) = \beta_1 + \beta_2 x + e$$


First, create a new column and label it  $\ln(\text{Food\_Exp})$ . In this column, calculate the natural log of food expenditures by typing **=LN(A2)** in the first empty cell. Copy the formula down the column.



|   | A          | B      | C             | D |
|---|------------|--------|---------------|---|
| 1 | Food_Exp   | Income | ln (Food Exp) |   |
| 2 | 115.220001 | 3.69   | 4.746843353   |   |
| 3 | 135.979996 | 4.39   | 4.912507787   |   |
| 4 | 119.339996 | 4.75   | 4.781976529   |   |
| 5 | 114.959999 | 6.03   | 4.744584233   |   |
| 6 | 187.050003 | 12.47  | 5.231375977   |   |
| 7 | 243.919998 | 12.98  | 5.496840294   |   |

Estimate the regression model using the  $\ln(\text{Food\_Exp})$  as the dependent variable instead of  $\text{Food\_Exp}$ .

| $\ln(\text{Food\_Exp})$ |              |                |             |
|-------------------------|--------------|----------------|-------------|
|                         | Coefficients | Standard Error | t Stat      |
| Intercept               | 4.780239298  | 0.158959254    | 30.07210456 |
| Income                  | 0.040030081  | 0.007665108    | 5.222376421 |




The interpretation for will be, an increase in income of \$100 leads to a 4% increase in the food expenditure.

#### 4.4.3 The linear-log model

In linear-log model, the independent variable is transformed but not the dependent variable.

$$y = \beta_1 + \beta_2 \ln(x) + e$$

|    |                      |              |                |              |
|----|----------------------|--------------|----------------|--------------|
| 15 |                      |              |                |              |
| 16 |                      | Coefficients | Standard Error | t Stat       |
| 17 | Intercept            | -97.18641517 | 84.23744235    | -1.153719919 |
| 18 | $\ln(\text{Income})$ | 132.1658424  | 28.80461184    | 4.588357     |
| 19 |                      |              |                |              |
| 20 |                      |              |                |              |



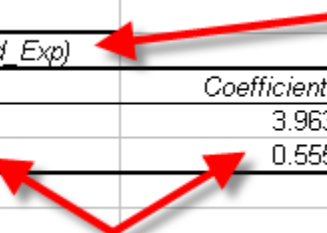
The interpretation for the linear-log model is, 1% increase in income, leads to about \$1.32 increase in the food expenditure.

#### 4.4.4 The log-log model

In log-log model, both the dependent and the independent variables are transformed. Estimated  $\beta_2$  represents the elasticity indicating % change in the  $y$  variable, when  $x$  variable increase by 1%.

$$\ln(y) = \beta_1 + \beta_2 \ln(x) + e$$

| $\ln(\text{Food\_Exp})$ |              |                |             |             |
|-------------------------|--------------|----------------|-------------|-------------|
|                         | Coefficients | Standard Error | t Stat      | P-value     |
| Intercept               | 3.963566918  | 0.294373       | 13.46443768 | 4.839E-16   |
| $\ln(\text{Income})$    | 0.555881175  | 0.100659514    | 5.522390809 | 2.57277E-06 |

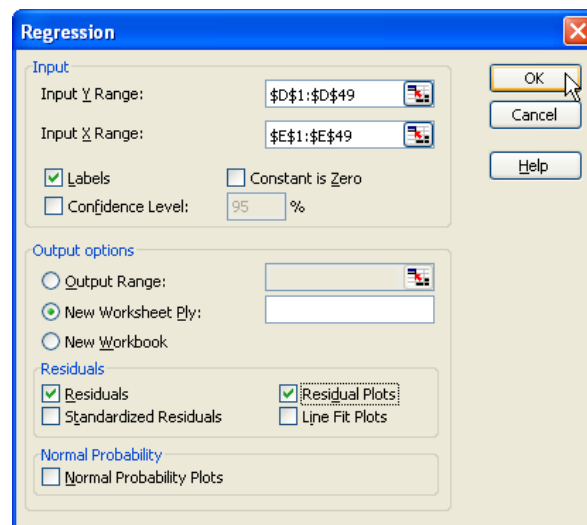


In the food expenditure model, a 1% increase in income, is expected to increase food expenditure by about 0.56%.

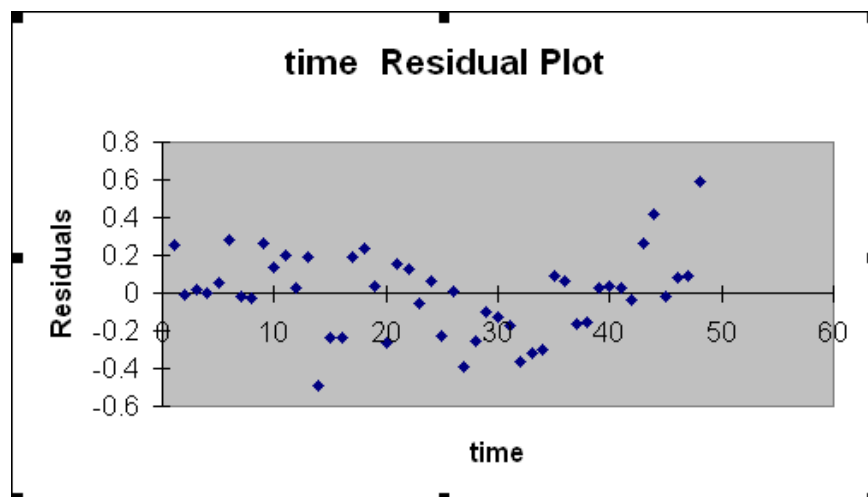
## 4.5 MORE EXAMPLES

### 4.5.1 Residual analysis with wheat data

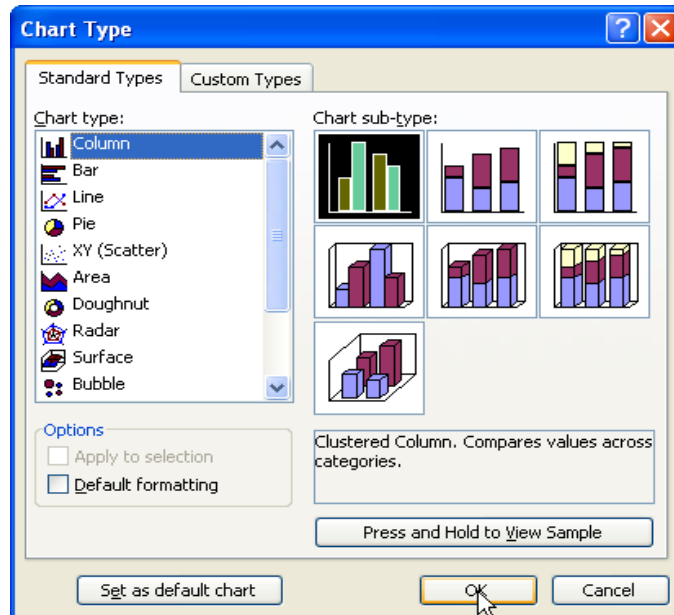
For this example, open the excel file *wa-wheat.xls*. The dataset contains 48 observations on average annual wheat yield in four different shires of Western Australia; Northampton, Chapman, Mullewa, Greenough and *TIME*. First, estimate the simple regression of Greenough on time, and ask for the residual plot.



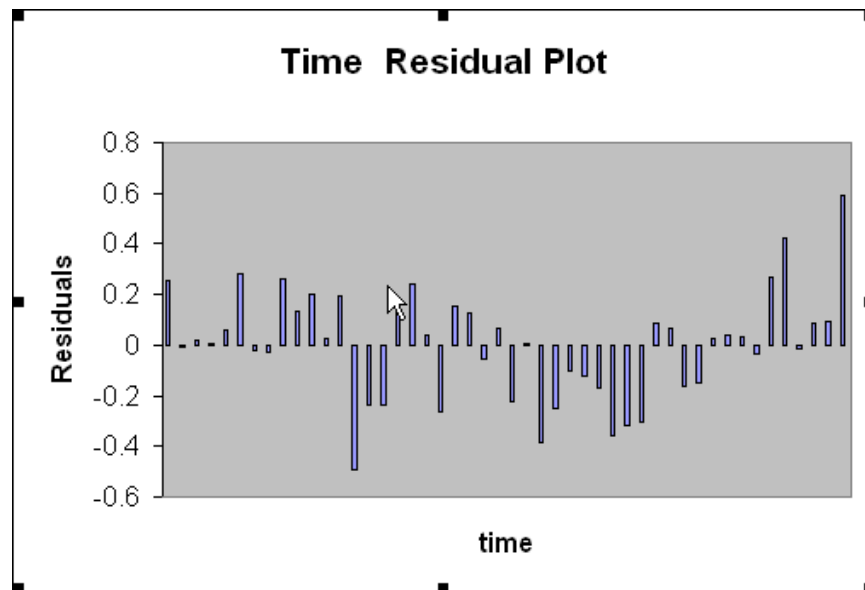
Excel will produce the scatter plot of the residuals through time in addition to the standard regression output.



You can format the plot or change the plot type by right-clicking on the picture and selecting **Chart Type**. You can then pick the desired chart type. Let's pick columns for this example.

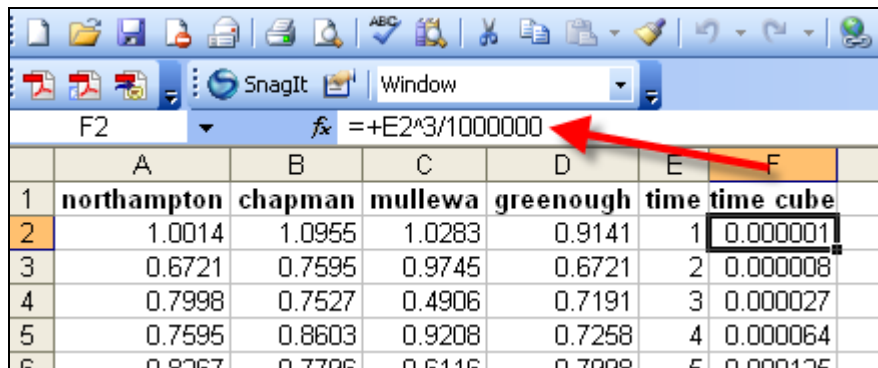


Excel will change the chart type of the residual plot, and provide a bar chart.



Based on the residual analysis, if you think that is the incorrect functional form, you can generate a new column in Excel and transform time variable and rerun the regression using the transformed variable as your independent variable. To generate the cubic equation results described in the text, generate a transformed column under column F and replace *TIME* with *TIME CUBE* as your independent variable.

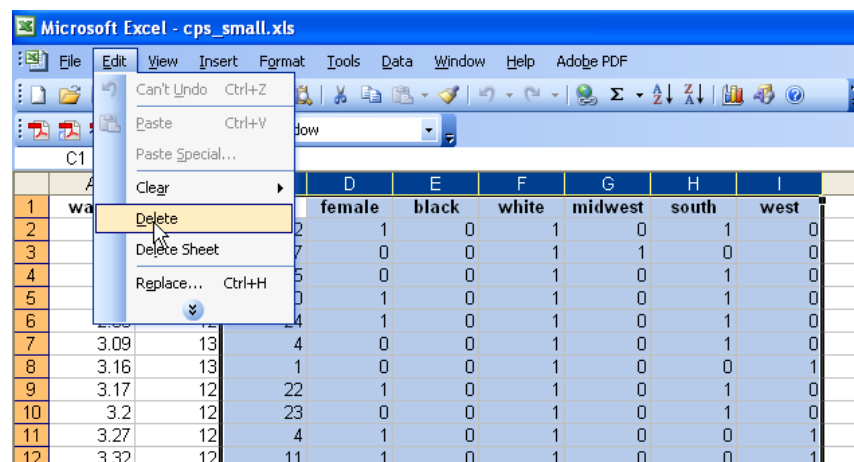




|   | A           | B       | C       | D         | E    | F         |
|---|-------------|---------|---------|-----------|------|-----------|
| 1 | northampton | chapman | mullewa | greenough | time | time cube |
| 2 | 1.0014      | 1.0955  | 1.0283  | 0.9141    | 1    | 0.000001  |
| 3 | 0.6721      | 0.7595  | 0.9745  | 0.6721    | 2    | 0.000008  |
| 4 | 0.7998      | 0.7527  | 0.4906  | 0.7191    | 3    | 0.000027  |
| 5 | 0.7595      | 0.8603  | 0.9208  | 0.7258    | 4    | 0.000064  |
| 6 | 0.8267      | 0.7706  | 0.6116  | 0.7008    | 5    | 0.000125  |

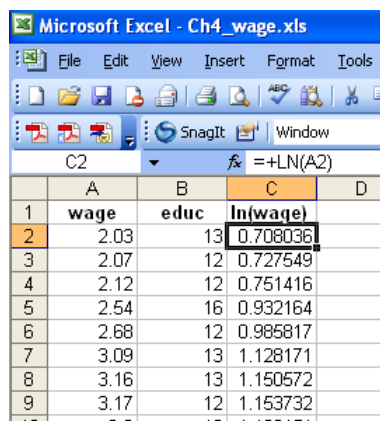
### 4.5.2 Log-linear model with wage data

We will use the *cps\_small* data set to illustrate a log-linear model. Open the *cps\_small.xls* data set. In this chapter, we will only use *EDUC* and *WAGES*, so delete the rest of the columns. You can delete the columns by highlighting the entire column and selecting **Delete** under **Edit**.



|    | A    | B  | C | D      | E     | F     | G       | H     | I    |
|----|------|----|---|--------|-------|-------|---------|-------|------|
| 1  | wa   |    |   | female | black | white | midwest | south | west |
| 2  |      |    |   | 1      | 0     | 1     | 0       | 1     | 0    |
| 3  |      |    |   | 0      | 0     | 1     | 1       | 0     | 0    |
| 4  |      |    |   | 0      | 0     | 1     | 0       | 1     | 0    |
| 5  |      |    |   | 1      | 0     | 1     | 0       | 1     | 0    |
| 6  |      |    |   | 1      | 0     | 1     | 0       | 1     | 0    |
| 7  | 3.09 | 13 |   | 4      | 0     | 1     | 0       | 1     | 0    |
| 8  | 3.16 | 13 |   | 1      | 0     | 1     | 0       | 0     | 1    |
| 9  | 3.17 | 12 |   | 22     | 1     | 1     | 0       | 1     | 0    |
| 10 | 3.2  | 12 |   | 23     | 0     | 1     | 0       | 1     | 0    |
| 11 | 3.27 | 12 |   | 4      | 1     | 1     | 0       | 0     | 1    |
| 12 | 3.32 | 12 |   | 11     | 1     | 1     | 0       | 0     | 1    |

After deleting the unwanted columns, save the Excel sheet as **Ch4\_Wage**. Then, transform the wage variable into log by typing the function in the first row below the label and copying down the entire column.



|    | A    | B    | C        | D |
|----|------|------|----------|---|
| 1  | wage | educ | ln(wage) |   |
| 2  | 2.03 | 13   | 0.708036 |   |
| 3  | 2.07 | 12   | 0.727549 |   |
| 4  | 2.12 | 12   | 0.751416 |   |
| 5  | 2.54 | 16   | 0.932164 |   |
| 6  | 2.68 | 12   | 0.985817 |   |
| 7  | 3.09 | 13   | 1.128171 |   |
| 8  | 3.16 | 13   | 1.150572 |   |
| 9  | 3.17 | 12   | 1.153732 |   |
| 10 | 3.2  | 12   | 1.157185 |   |
| 11 | 3.27 | 12   | 1.184148 |   |
| 12 | 3.32 | 12   | 1.200204 |   |

You can now estimate the log-linear function by using the new variable,  $\ln(WAGE)$  as the dependent variable.

|    | A                     | B            | C              | D           | E           | F              | G           | H           | I           |
|----|-----------------------|--------------|----------------|-------------|-------------|----------------|-------------|-------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |             |             |                |             |             |             |
| 2  |                       |              |                |             |             |                |             |             |             |
| 3  | Regression Statistics |              |                |             |             |                |             |             |             |
| 4  | Multiple R            | 0.4632717    |                |             |             |                |             |             |             |
| 5  | R Square              | 0.214620668  |                |             |             |                |             |             |             |
| 6  | Adjusted R Square     | 0.213833715  |                |             |             |                |             |             |             |
| 7  | Standard Error        | 0.490151106  |                |             |             |                |             |             |             |
| 8  | Observations          | 1000         |                |             |             |                |             |             |             |
| 9  |                       |              |                |             |             |                |             |             |             |
| 10 | ANOVA                 |              |                |             |             |                |             |             |             |
| 11 |                       | df           | SS             | MS          | F           | Significance F |             |             |             |
| 12 | Regression            | 1            | 65.52131245    | 65.52131245 | 272.7235332 | 2.39613E-54    |             |             |             |
| 13 | Residual              | 998          | 239.7676103    | 0.240248106 |             |                |             |             |             |
| 14 | Total                 | 999          | 305.2889227    |             |             |                |             |             |             |
| 15 |                       |              |                |             |             |                |             |             |             |
| 16 |                       | Coefficients | Standard Error | t Stat      | P-value     | Lower 95%      | Upper 95%   | Lower 95.0% | Upper 95.0% |
| 17 | Intercept             | 0.788374304  | 0.084897534    | 9.286186174 | 9.70516E-20 | 0.621776155    | 0.954972453 | 0.621776155 | 0.954972453 |
| 18 | educ                  | 0.103760809  | 0.006283072    | 16.51434326 | 2.39613E-54 | 0.091431261    | 0.116090356 | 0.091431261 | 0.116090356 |
| 19 |                       |              |                |             |             |                |             |             |             |

The forecasted value from the log-linear wage equation is

$$\ln y_{\text{hat}} = b_1 + b_2 x$$

In order to obtain a prediction for the dependent variables  $y$ , we need to use the exponential function to get the “natural predictor” back.

$$\hat{y}_n = \exp(\ln y_{\text{hat}}) = \exp(b_1 + b_2 x)$$

In large samples a more precise predictor is obtained by correcting that “natural predictor” by multiplying it by  $\exp(\hat{\sigma}^2 / 2)$ . Using the coefficient estimates, the raw forecasted values can be calculated as shown in the cell B23 of the below Excel output. Then, the natural predicted and the corrected predictors are calculated.

|    | A                     | B                              | C                   | D                                           |                      |
|----|-----------------------|--------------------------------|---------------------|---------------------------------------------|----------------------|
| 1  | SUMMARY OUTPUT        |                                |                     |                                             |                      |
| 2  |                       |                                |                     |                                             |                      |
| 3  | Regression Statistics |                                |                     |                                             |                      |
| 4  | Multiple R            | 0.4632717                      |                     |                                             |                      |
| 5  | R Square              | 0.214620668                    |                     |                                             |                      |
| 6  | Adjusted R Square     | 0.213833715                    |                     |                                             |                      |
| 7  | Standard Error        | 0.490151106                    |                     |                                             |                      |
| 8  | Observations          | 1000                           |                     |                                             |                      |
| 9  |                       |                                |                     |                                             |                      |
| 10 | ANOVA                 |                                |                     |                                             |                      |
| 11 |                       | df                             | SS                  | MS                                          |                      |
| 12 | Regression            | 1                              | 65.52131245         | 65.52131245                                 |                      |
| 13 | Residual              | 998                            | 239.7676103         | 0.240248106                                 |                      |
| 14 | Total                 | 999                            | 305.2889227         |                                             |                      |
| 15 |                       |                                |                     |                                             |                      |
| 16 |                       | Coefficients                   | Standard Error      | t Stat                                      |                      |
| 17 | Intercept             | 0.788374304                    | 0.084897534         | 9.286186174                                 |                      |
| 18 | educ                  | 0.103760809                    | 0.006283072         | 16.51434326                                 |                      |
| 19 |                       |                                |                     |                                             |                      |
| 20 | mean of x=            | 13.285                         |                     |                                             |                      |
| 21 |                       |                                |                     |                                             |                      |
| 22 | Education             | $\ln(\text{wage})$             | wage_n              | wage_c                                      | Corrected Prediction |
| 23 |                       | $1 = +\$B\$17 + \$B\$18 * A23$ | $= \text{EXP}(B23)$ | $= \text{EXP}(B23) * \text{EXP}(\$D\$13/2)$ |                      |
| 24 |                       |                                |                     |                                             |                      |

The **standard error of the forecast** can be found from the regression output using the formula provided in 4.1.1. After the predicted values and the standard error of forecast is found, it is very straight forward to construct the prediction interval. Remember that the corrected predictor will always be greater than the natural predictor since the correction factor is always greater than one.

#### 4.5.2 Generalized $R^2$

The generalized  $R^2$  is the appropriate measure of fit for this model is the square of the correlation between the “best” predictor and the wage variable. Remember that the corrected predictor and the natural predictor only differ by the constant so they have the same correlation to the wage variable. You can calculate the generalized  $R^2$  by using the **Tools>Data Analysis>Correlation** and choose the column of Wage and the column of predicted (either corrected or natural) wage or you can use the **CORREL** statistical function:

**CORREL(wage,wage\_c)**

# CHAPTER 5

## Multiple Linear Regression

### CHAPTER OUTLINE

5.1 Big Andy's Burger Barn

5.2 Prediction

5.3 Sampling Precision

5.4 Confidence Intervals

5.5 Hypothesis Testing

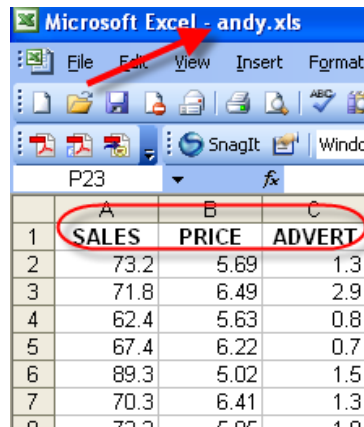
5.6 Goodness-of-Fit

### 5.1 BIG ANDY'S BURGER BARN

The multiple linear regression model is the extension of the simple model where there is more than one explanatory variable. We will use Big Andy's Burger Barn to transition to the multiple regression model. This is a multiple regression model where the dependent variable, *SALES* is a linear function of *PRICE* charged and the level of advertising, *ADVERT*.

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

Using Excel to perform multiple regression is just like estimating a simple regression model, except we will include all explanatory variables in the **X range**. Open *andy.xls*.



|   | A     | B     | C      |
|---|-------|-------|--------|
| 1 | SALES | PRICE | ADVERT |
| 2 | 73.2  | 5.69  | 1.3    |
| 3 | 71.8  | 6.49  | 2.9    |
| 4 | 62.4  | 5.63  | 0.8    |
| 5 | 67.4  | 6.22  | 0.7    |
| 6 | 89.3  | 5.02  | 1.5    |
| 7 | 70.3  | 6.41  | 1.3    |
| 8 | 72.7  | 5.05  | 1.0    |

The first column is the monthly sales in \$1,000 in a given city, the second column is the price of hamburgers (actually a price index for goods sold) measured in dollars and the third column is the advertising spending, also measured in \$1,000. To estimate this model, go to **Tools>Data Analysis>Regression** and fill in the regression dialog box so that the **Y Range** is the dependent variable, *SALES* and the **X Range** includes both the *PRICE* and the *ADVERT* columns. Make sure to check the **Labels box** and hit **OK**.

The results look very similar to what we've seen before, except now we have parameter estimates and other information on *PRICE* and the *ADVERT*. This portion of the output appears as

|    | A                            | B                   | C                     | D             | E              | F                     | G                |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|-----------------------|------------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |                       |                  |
| 2  |                              |                     |                       |               |                |                       |                  |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |                       |                  |
| 4  | Multiple R                   | 0.66952055          |                       |               |                |                       |                  |
| 5  | R Square                     | 0.448257766         |                       |               |                |                       |                  |
| 6  | Adjusted R Square            | 0.432931593         |                       |               |                |                       |                  |
| 7  | Standard Error               | 4.886124039         |                       |               |                |                       |                  |
| 8  | Observations                 | 75                  |                       |               |                |                       |                  |
| 9  |                              |                     |                       |               |                |                       |                  |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |                       |                  |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       | <i>Significance F</i> |                  |
| 12 | Regression                   | 2                   | 1396.538993           | 698.2694963   | 29.24785998    | 5.04086E-10           |                  |
| 13 | Residual                     | 72                  | 1718.942985           | 23.87420813   |                |                       |                  |
| 14 | Total                        | 74                  | 3115.481978           |               |                |                       |                  |
| 15 |                              |                     |                       |               |                |                       |                  |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i>      | <i>Upper 95%</i> |
| 17 | Intercept                    | 118.9136131         | 6.351637595           | 18.72172512   | 2.21429E-29    | 106.2518552           | 131.5753711      |
| 18 | PRICE                        | -7.907854804        | 1.095993037           | -7.215241826  | 4.42399E-10    | -10.09267696          | -5.723032645     |
| 19 | ADVERT                       | 1.862583787         | 0.683195483           | 2.726282349   | 0.008038199    | 0.500658501           | 3.224509073      |

## 5.2 PREDICTION

Using the estimated regression equation, we can now forecast values of sales for different values of price and advertising like we did in the simple regression case using the below formula.

$$\widehat{SALES} = b_1 + b_2 PRICE + b_3 ADVERT$$

$$= 118.9136131 - 7.907854804 * PRICE + 1.862583787 * ADVERT$$

You can transfer this formula to Excel in the following fashion:

|    | A                 | B                   | C                                 |
|----|-------------------|---------------------|-----------------------------------|
| 1  | SUMMARY OUTPUT    |                     |                                   |
| 16 |                   | <i>Coefficients</i> | <i>Standard Error</i>             |
| 17 | Intercept         | 118.9136131         | 6.351637595                       |
| 18 | PRICE             | -7.907854804        | 1.095993037                       |
| 19 | ADVERT            | 1.862583787         | 0.683195483                       |
| 20 |                   |                     |                                   |
| 21 |                   |                     |                                   |
| 22 | <b>PREDICTION</b> |                     |                                   |
| 23 | Price             | Advert Expenditure  | Sales Hat                         |
| 24 | 4.00              | 1.00                | =+\$B\$17+\$B\$18*A24+\$B\$19*B24 |
| 25 | 4.50              | 1.00                | =+\$B\$17+\$B\$18*A25+\$B\$19*B25 |
| 26 | 5.00              | 1.20                | =+\$B\$17+\$B\$18*A26+\$B\$19*B26 |
| 27 | 5.50              | 1.20                | =+\$B\$17+\$B\$18*A27+\$B\$19*B27 |

These formulas will yield the sales forecast for specified values of price and the advertising expenditure as shown below.

|    | A                 | B                   | C                     |
|----|-------------------|---------------------|-----------------------|
| 16 |                   | <i>Coefficients</i> | <i>Standard Error</i> |
| 17 | Intercept         | 118.9136131         | 6.351637595           |
| 18 | PRICE             | -7.907854804        | 1.095993037           |
| 19 | ADVERT            | 1.862583787         | 0.683195483           |
| 20 |                   |                     |                       |
| 21 |                   |                     |                       |
| 22 | <b>PREDICTION</b> |                     |                       |
| 23 | Price             | Advert Expenditure  | Sales Hat             |
| 24 | 4.00              | 1.00                | 89.14477771           |
| 25 | 4.50              | 1.00                | 85.19085031           |
| 26 | 5.00              | 1.20                | 81.60943966           |
| 27 | 5.50              | 1.20                | 77.65551226           |

### 5.3 SAMPLING PRECISION

To estimate the error variance, we will use the ANOVA (Analysis of Variance) table. Recall that the error variance of the regression equation is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N \hat{e}_i^2}{N - K} = \frac{SSE}{N - K}$$

The formula is the sum of squared errors divided by the degrees of freedom. This quantity is important and is reported automatically in the ANOVA table when a regression is estimated.

| ANOVA      | df | Sum of Squares Regression |             | Sum of Squares Error |             | Significance F |
|------------|----|---------------------------|-------------|----------------------|-------------|----------------|
|            |    | SSR                       | SSE         | MS                   | F           |                |
| Regression | 2  | 1396.538993               | 698.269496  | 29.24785998          | 5.04086E-10 |                |
| Residual   | 72 | 1718.942985               | 23.87420813 |                      |             |                |
| Total      | 74 | 3115.481978               |             |                      |             |                |

|           | Coefficients | Standard Error | t Stat       | P-value     | Lower 95%    | Upper 95%    |
|-----------|--------------|----------------|--------------|-------------|--------------|--------------|
| Intercept | 118.9136131  | 6.351637595    | 18.72172512  | 2.21429E-29 | 106.2518552  | 131.5753711  |
| PRICE     | -7.907854804 | 1.095993037    | -7.215241826 | 4.42399E-10 | -10.09267696 | -5.723032645 |
| ADVERT    | 1.862583787  | 0.683195483    | 2.726282349  | 0.008038199 | 0.500658501  | 3.224509073  |

The square root of the estimated regression variance is the **Standard Error** of the regression and is reported in the **Regression statistics**.

$$\text{Standard Error} = \sqrt{MSE} = \sqrt{23.87420813} = 4.886124039$$

|    | A                     | B           | C           | D           |
|----|-----------------------|-------------|-------------|-------------|
| 1  | SUMMARY OUTPUT        |             |             |             |
| 2  |                       |             |             |             |
| 3  | Regression Statistics |             |             |             |
| 4  | Multiple R            | 0.66952055  |             |             |
| 5  | R Square              | 0.448257766 |             |             |
| 6  | Adjusted R Square     | 0.432931593 |             |             |
| 7  | Standard Error        | 4.886124039 |             |             |
| 8  | Observations          | 75          |             |             |
| 9  |                       |             |             |             |
| 10 | ANOVA                 |             |             |             |
| 11 |                       | df          | SS          | MS          |
| 12 | Regression            | 2           | 1396.538993 | 698.2694963 |
| 13 | Residual              | 72          | 1718.942985 | 23.87420813 |
| 14 | Total                 | 74          | 3115.481978 |             |

The estimated least squares variance/covariance matrix can be represented as

$$\text{Cov}(b_1, b_2, b_3) = \begin{bmatrix} \text{var}(b_1) & \text{cov}(b_1, b_2) & \text{cov}(b_1, b_3) \\ \text{cov}(b_1, b_2) & \text{var}(b_2) & \text{cov}(b_2, b_3) \\ \text{cov}(b_1, b_3) & \text{cov}(b_2, b_3) & \text{var}(b_3) \end{bmatrix}$$

The estimated variance/covariance matrix of the least squares estimators are not directly reported by Excel. However, in the simple model they are easily obtained. The estimated variance of  $b_2$  is

$$\hat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum_{i=1}^T (x_i - \bar{x})^2}$$

The standard error of the estimated coefficient is

$$\text{se}(b_2) = \sqrt{\hat{\text{var}}(b_2)} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^T (x_i - \bar{x})^2}}$$

In the Excel output we are given the values of the standard errors for the least squares estimates. The **Standard Errors** are reported in the column next to the **Coefficient estimates**. The estimated variances can be obtained by squaring the standard errors.

|           | <i>Coefficients</i>       | <i>Standard Error</i> |
|-----------|---------------------------|-----------------------|
| Intercept | 118.9136131               | 6.351637595           |
| PRICE     | -7.907854804              | 1.095993037           |
| ADVERT    | 1.862583787               | 0.683195483           |
|           | 6.351637595^2=40.34330014 |                       |
| Var(b1)   | =+C17^2                   | 40.34330014           |
| Var(b2)   | =+C18^2                   | 1.201200738           |
| Var(b3)   | =+C19^2                   | 0.466756068           |

So, we can fill in the variances in the covariance matrix as follows:

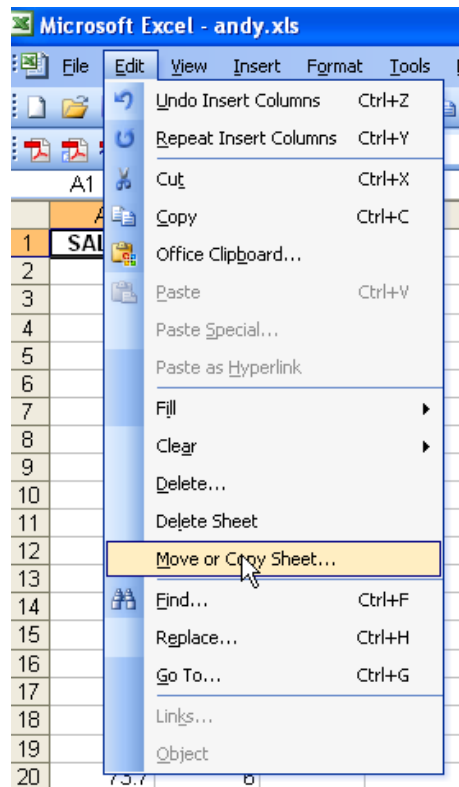
$$\text{Cov}(b_1, b_2, b_3) = \begin{bmatrix} 40.3433 & \text{cov}(b_1, b_2) & \text{cov}(b_1, b_3) \\ \text{cov}(b_1, b_2) & 1.2012 & \text{cov}(b_2, b_3) \\ \text{cov}(b_1, b_3) & \text{cov}(b_2, b_3) & .4668 \end{bmatrix}$$

The formula for  $\text{cov}(b_2, b_3)$  is:

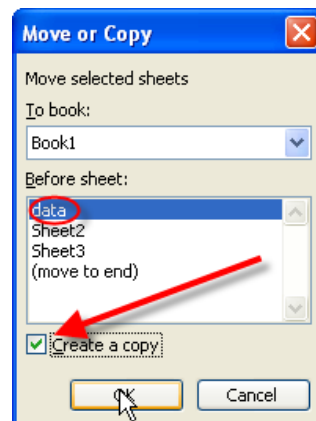


$$\text{cov}(b_2, b_3) = \frac{-r_{23} \hat{\sigma}^2}{(1 - r_{23}^2) \sqrt{\sum x_{t2}^{*2}} \sqrt{\sum x_{t3}^{*2}}} \text{ where } \sqrt{\sum x_{tk}^{*2}} = \sqrt{\sum_{t=1}^T (x_{tk} - \bar{x}_k)^2}.$$

Since the estimated covariances are not reported by Excel directly, we need to translate this formula into Excel. Let's start by creating a worksheet by copying our data worksheet. Choosing **Edit>Move or Copy Sheet**.



**Move/Copy** dialog box will open, choose the data sheet and make sure to click the **Create a copy** box.



You will see another tab added to your three existing worksheets. Rename the new sheet *covariance* by highlighting the name of the tab. First calculation will be the sum of square calculations in the denominator,

$$\sqrt{\sum x_{tk}^2} = \sqrt{\sum_{t=1}^T (x_{tk} - \bar{x}_k)^2}$$

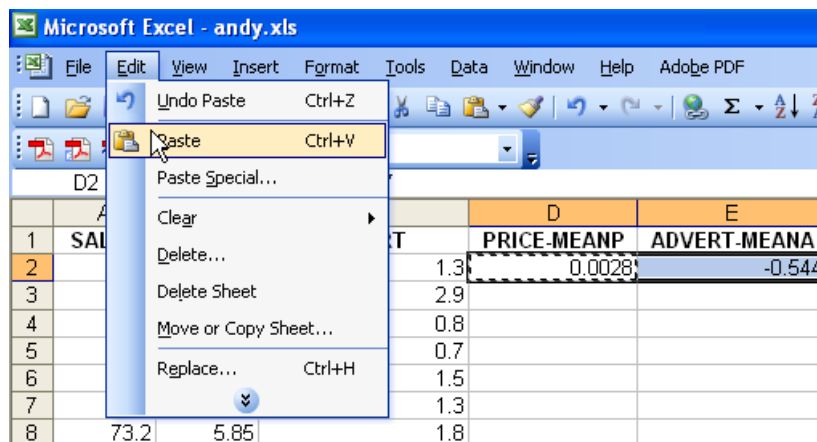
To accomplish this, go to the bottom of *PRICE* column and write the formula to get the average  $\bar{x}$ , and repeat it for *ADVERT* column.

|    | A    | B      | C     | D | E |
|----|------|--------|-------|---|---|
| 73 | 74.2 | 5.11   | 0.7   |   |   |
| 74 | 75.4 | 5.71   | 0.7   |   |   |
| 75 | 81.3 | 5.45   | 2     |   |   |
| 76 | 75   | 6.05   | 2.2   |   |   |
| 77 |      | 5.6872 | 1.844 |   |   |
| 78 |      |        |       |   |   |

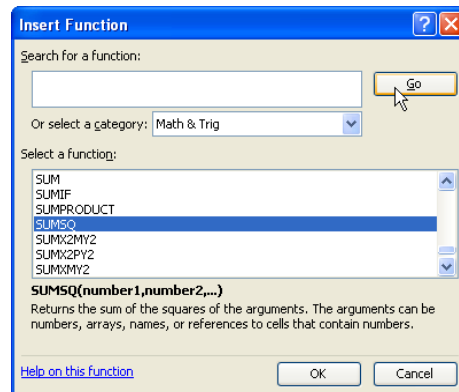
Next, create 2 new columns and name them *PRICE - MEANP* and *ADVERT- MEANA*. In these columns, we will store the difference from the mean.

|   | A     | B     | C      | D           | E            |
|---|-------|-------|--------|-------------|--------------|
| 1 | SALES | PRICE | ADVERT | PRICE-MEANP | ADVERT-MEANA |
| 2 | 73.2  | 5.69  | 1.3    | 0.0028      | -0.544       |
| 3 | 74.8  | 6.49  | 2.0    |             |              |

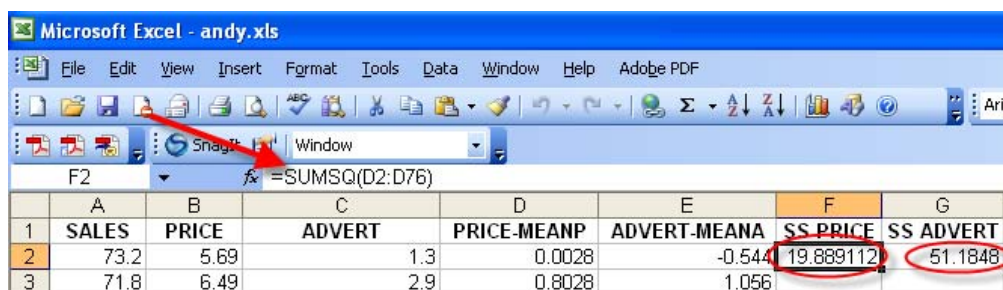
Recall that we store the mean of price in cell B77. By putting the \$ signs, we are making sure, EXCEL will use the same cell for the mean even when we copy our formula to another cell. Now, copy and paste the formula to the rest of the column and repeat the process for the *ADVERT-MEANA* column, too.



Once you have calculated the differences from means, we will now calculate the sum of squares. Recall that we can do **Sum of Squares** calculations using the **Insert** function



or simply by typing the formula = **SUMSQ(...)**.



Once you have the squares calculated, we can now create columns to calculate the deviation, all we have to do is to put the formula together. Recall that the formula is

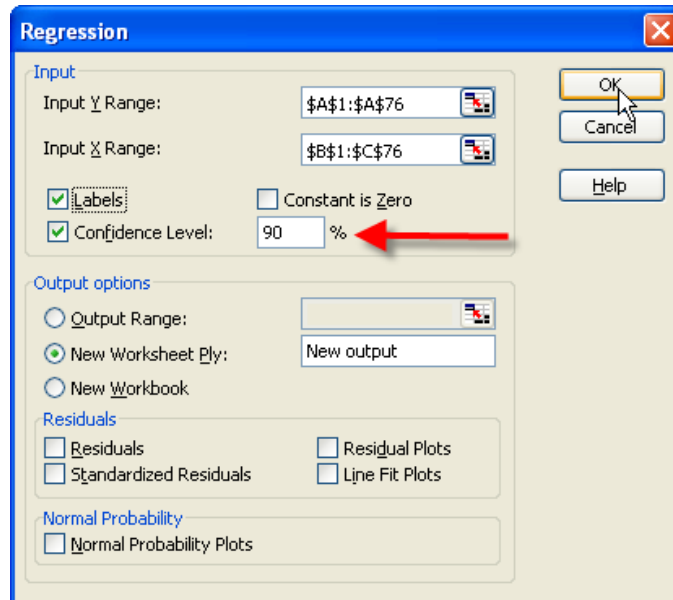
$$\text{cov}(b_2, b_3) = \frac{-r_{23}\hat{\sigma}^2}{(1-r_{23}^2)\sqrt{\sum x_{i2}^{*2}}\sqrt{\sum x_{i3}^{*2}}}$$

We already know that  $\hat{\sigma}^2 = MSE = 23.8742$  from our regression output. We have calculated the correlation coefficient between *PRICE* and *ADVERT* to be 0.0264. Plug all the numbers into the formula will give us the covariance of

$$\text{cov}(b_2, b_3) = -0.01974.$$

## 5.4 CONFIDENCE INTERVALS

The 95% confidence interval for each parameter is provided by default in the Excel regression output. If a different confidence interval is needed, Excel will also provide that. We can return to the worksheet containing the original *andy.xls* data. Run the regression using **Tools/Data Analysis/Regression**. We can then check the **Confidence Level** box and set the level to 90. Set all other desired options and click **OK**.



Both the 95% and 90% confidence intervals are reported for the  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ .

|    |           | Coefficients | Standard Error | t Stat       | P-value     | Lower 95%    | Upper 95%    | Lower 90.0%  | Upper 90.0%  |
|----|-----------|--------------|----------------|--------------|-------------|--------------|--------------|--------------|--------------|
| 16 |           |              |                |              |             |              |              |              |              |
| 17 | Intercept | 118.9136131  | 6.351637595    | 18.72172512  | 2.21429E-29 | 106.2518552  | 131.5753711  | 108.3299194  | 129.4973068  |
| 18 | PRICE     | -7.907854804 | 1.095993037    | -7.215241826 | 4.42399E-10 | -10.09267696 | -5.723032645 | -9.734101094 | -6.081608514 |
| 19 | ADVERT    | 1.862583787  | 0.683195483    | 2.726282349  | 0.008038199 | 0.500658501  | 3.224509073  | 0.72417946   | 3.000988115  |

The 90% confidence interval for  $\beta_3$  suggests that an additional \$1000 in advertising expenditures leads to an increase in total revenue that is between \$724 and \$3001. Note that a 95% interval is always wider than the 90% confidence interval.

## 5.5 HYPOTHESIS TESTING

The  $t$ -test to test the significance of  $\beta_2$ , that the coefficient is zero, against the two sided alternative that it is not, is

$$t = \frac{b_2 - 0}{se(b_2)} = \frac{-7.908}{1.096} = -7.215$$

Excel provides the  $t$ -stats and  $p$ -values in the regression output is the simplest way to test the significance of a single coefficient. The results are

|           | Coefficients | Standard Error | t Stat       | P-value     |
|-----------|--------------|----------------|--------------|-------------|
| Intercept | 118.9136131  | 6.351637595    | 18.72172512  | 2.21429E-29 |
| PRICE     | -7.907854804 | 1.095993037    | -7.215241826 | 4.42399E-10 |
| ADVERT    | 1.862583787  | 0.683195483    | 2.726282349  | 0.008038199 |

Recall that the  $t$ -stat is the coefficient divided by its standard error. Based on the reported  $p$ -values, both  $b_2$  and  $b_3$  are significant at the 5% as well as 1% level.

Sometimes we will need general tests about our parameters, such as the elasticity of demand where the null and alternative hypotheses are  $H_0: \beta_2 \geq 0$ : a decrease in price leads to a decrease in total revenue (demand is price inelastic) and  $H_1: \beta_2 < 0$ : a decrease in price leads to an increase in total revenue (demand is price elastic).

Or we may want to test an economic hypothesis such as

$$H_0: \beta_3 \leq 1$$

$$H_1: \beta_3 > 1$$

To test whether an increase in advertising expenditures is "worth it", that is, total revenues increase enough to cover the increased cost of the advertising, we use the  $t$ -statistic

$$t = \frac{b_3 - 1}{se(b_3)}$$

If the null hypothesis is true, it suggests that a dollar increase in advertising expenditures leads to less than a dollar increase in total revenue. In this case, it doesn't make sense to spend that extra dollar. On the other hand, if we reject the null hypothesis we conclude that there is sufficient statistical evidence to suggest that an increase in advertising expenditures is "worth it" in terms of the increase in total revenue. The value of the test statistic is

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

Since  $1.263 < 1.666$ , we do not reject the null hypothesis. The critical value for this one-tail test is obtained as

**Function Arguments**

**TINV**

**Probability** .10 = 0.1

**Deg\_freedom** 72 = 72

= 1.666293697

Returns the inverse of the Student's t-distribution.

**Probability** is the probability associated with the two-tailed Student's t-distribution, a number between 0 and 1 inclusive.

Formula result = 1.666293697

[Help on this function](#)

OK Cancel

## 5.6 GOODNESS-OF-FIT

The goodness-of-fit of the regression model is based on the ANOVA table. The coefficient of determination,  $R^2$  and the ANOVA table are reported for the multiple regression model as they are for the simple model. The  $R^2$  measures the percent variation in dependent variable explained by the regression model. We already know how to decompose the sums of squares as

$$SST = SSR + SSE$$

And the coefficient of determination,  $R^2$  is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The results from the regression are

|    | A                            | B           | C           | D           | E           |
|----|------------------------------|-------------|-------------|-------------|-------------|
| 1  | SUMMARY OUTPUT               |             |             |             |             |
| 2  |                              |             |             |             |             |
| 3  | <i>Regression Statistics</i> |             |             |             |             |
| 4  | Multiple R                   | 0.66952055  |             |             |             |
| 5  | R Square                     | 0.448257766 |             |             |             |
| 6  | Adjusted R Square            | 0.432931693 |             |             |             |
| 7  | Standard Error               | 4.886124039 |             |             |             |
| 8  | Observations                 | 75          |             |             |             |
| 9  |                              |             |             |             |             |
| 10 | <i>ANOVA</i>                 |             |             |             |             |
| 11 |                              | <i>df</i>   | <i>SS</i>   | <i>MS</i>   | <i>F</i>    |
| 12 | Regression                   | 2           | 1396.538993 | 698.2694963 | 29.24785998 |
| 13 | Residual                     | 72          | 1718.942985 | 23.87420813 |             |
| 14 | Total                        | 74          | 3115.481978 |             |             |

The value  $R^2 = .448$  means that 44.8% of the total variation in total revenue is explained by price and advertising expenditures. However, there can be a problem with this measure in the multiple regression model since addition of each additional explanatory variable will inflate the  $R^2$  even if there is no economic basis for the variables to appear in the model.

An alternative measure is the “adjusted  $R^2$ ”, denoted by Excel as **Adjusted R Square** is reported just below **R Square**. **Adjusted R Square** imposes a penalty for adding explanatory variables so it can never be larger than  $R^2$ . The Adjusted- $R^2$  can also be calculated from the ANOVA table.

$$\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$$

While this solves the problem associated with  $R^2$  (which has a particular interpretation!), the adjusted- $R^2$  has no interpretation. It is no longer the percent of the variation in total revenue that is explained by the model. It should NOT be used as a device for selecting appropriate explanatory variables; good economic theory should determine the model.

# CHAPTER 6

## Further Inference in the Multiple Regression Model

### CHAPTER OUTLINE

- |                                                   |                                                |
|---------------------------------------------------|------------------------------------------------|
| 6.1 The $F$ -test                                 | 6.5 Nonsample Information                      |
| 6.2 Testing the Overall Significance of the Model | 6.6 Model Specification                        |
| 6.3 An Extended Model                             | 6.6.1 Omitted variables                        |
| 6.4 Testing Economic Hypothesis                   | 6.6.2 Irrelevant variables                     |
| 6.4.1 The significance of advertising             | 6.6.3 Choosing the model                       |
| 6.4.2 Optimal level of advertising                | 6.7 Poor Data, Collinearity and Insignificance |

### 6.1 F-TEST

The  $t$ -test is used to test a specific null hypothesis, such as a single test of significance. With the multiple regression model, we might be interested in testing whether two or more explanatory variables are *jointly* important to the model. The  $F$ -test allows for testing joint hypotheses, and is based on a comparison of the sum of the squared errors from an unrestricted (full, or "original") model to the sum of squared errors from a model where the null hypothesis has been imposed. In the Big Andy's Burger Barn example, we estimated the model

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + e_i$$

We can use the  $t$ -test to test the hypothesis  $H_0 : \beta_2 = 0$  against  $H_1 : \beta_2 \neq 0$ . Another way to test this hypothesis is in terms of the models each hypothesis implies using an  $F$ -test. If the null hypothesis is true, then the restricted model is formed as

$$S_i = \beta_1 + \beta_3 A_i + e_i$$

The  $F$ -test compares the sums of squared errors from the restricted model and the unrestricted model. A large difference will signal that the restrictions are false. The  $F$ -statistic we will use is



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

where  $SSE_R$  and  $SSE_U$  are the sum squared errors from the ANOVA tables of the restricted and unrestricted models, respectively.  $J$  is the number of hypotheses or restrictions in the null hypothesis.  $N$  is the sample size of the unrestricted model, and  $K$  is the number of parameters in the unrestricted model. If the null hypothesis is true, this test statistic follows the  $F$ -distribution with  $J$  numerator and  $N-K$  denominator degrees of freedom.

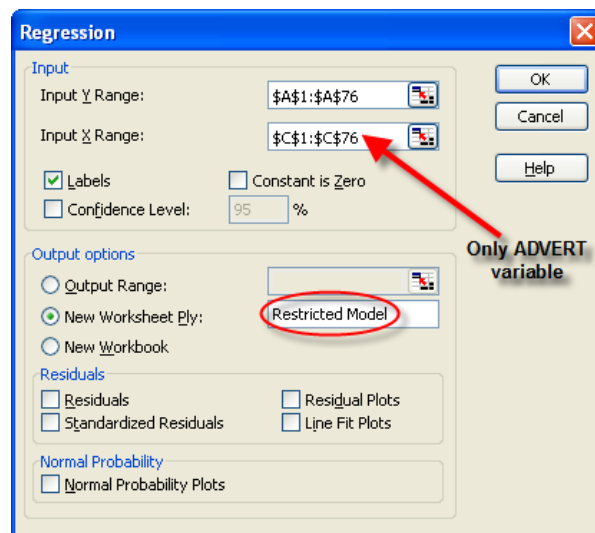
Unfortunately, Excel will not report these values automatically, so we will prepare a template for the  $F$ -test. Let's start with **Insert>Worksheet** in the *Andy.xls* file. Name the new worksheet **F Test**. Type a heading such as "**Hypothesis Testing - F-Test**" in cell A1. Type **Data Input** labels and **Computed Values** labels in column A. For appearances, left justify and set to bold font the labels **Data Input** and **Computed Values** and right justify the sub-labels as shown below.

|    | A                                  | B                                                 |
|----|------------------------------------|---------------------------------------------------|
| 1  | <b>Hypothesis Testing - F-Test</b> |                                                   |
| 2  |                                    |                                                   |
| 3  | <b>Data Input</b>                  |                                                   |
| 4  | J                                  |                                                   |
| 5  | N                                  |                                                   |
| 6  | K                                  |                                                   |
| 7  | SSE-RESTRICTED                     |                                                   |
| 8  | SSE-UNRESTRICTED                   |                                                   |
| 9  | ALPHA                              |                                                   |
| 10 |                                    |                                                   |
| 11 | <b>Computed Values</b>             |                                                   |
| 12 | df-numerator                       | =+B4                                              |
| 13 | df-denominator                     | =+B5-B6                                           |
| 14 | F                                  | =+((B7-B8)/B12)/(B8/B13)                          |
| 15 | Right Critical value               | =+FINV(B9,B12,B13)                                |
| 16 | Decision                           | =+IF(B14>B17,"Reject Null","Fail to Reject Null") |
| 17 | p-value                            | =+FDIST(B14,B12,B13)                              |

In column B, we will type the formulas necessary to calculate the  $F$ -statistic, the appropriate decision, and the  $p$ -value associated with the calculated  $F$ -statistic. The commands are similar to those used to create the  $t$ -test template in Chapter 5. To calculate the  $F$ -statistic for a particular test, see the formula in cell B14. The functions **FINV** and **FDIST** are used to find the  $F$ -critical value and the  $p$ -value associated with the calculated  $F$ -statistic, respectively. The syntax of these functions are **FINV( $\alpha$ ,df\_n,df\_d)** and **FDIST( $F$ -stat, df\_n,df\_d)**.

To obtain the information needed in the **Data Input** section of the template, we need two regressions; the unrestricted model and the restricted model. We will use the  $SSE$ 's from the ANOVA tables of each model. Now we can conduct the test for the restricted and unrestricted models mentioned above.

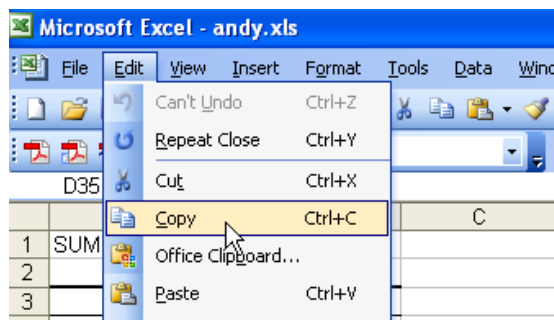
Since we already have the unrestricted regression model for *Andy.xls*, we only need to run the restricted model by going to **Tools>Data Analysis>Regression**. This time include only advertisement (*ADVERT*) as the explanatory variables and save the worksheet as "**Restricted Model**" and click **OK**.



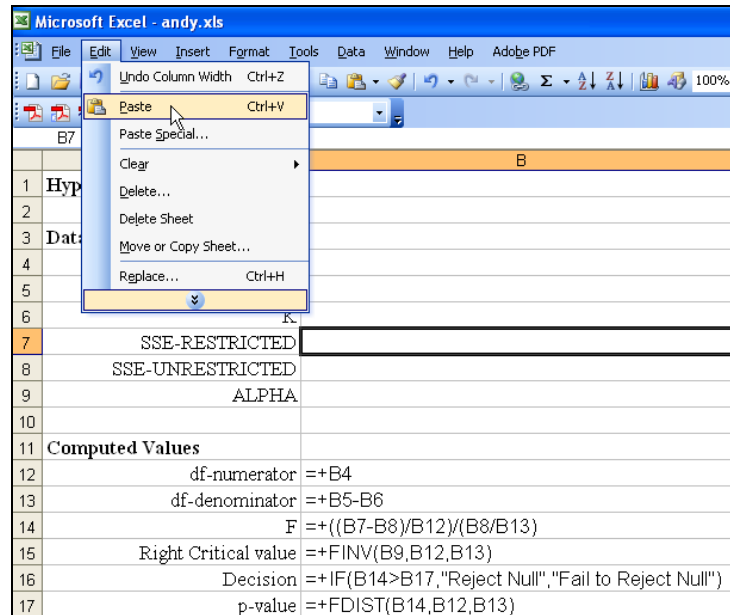
The regression output will be stored in the “**Restricted Model**” worksheet and will contain the ANOVA table.

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.222080315         |                       |               |                |
| 5  | R Square                     | 0.049319666         |                       |               |                |
| 6  | Adjusted R Square            | 0.036296648         |                       |               |                |
| 7  | Standard Error               | 6.3696922           |                       |               |                |
| 8  | Observations                 | 75                  |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 1                   | 153.6545312           | 153.6545312   | 3.787114875    |
| 13 | Residual                     | 73                  | 2961.827446           | 40.57297872   |                |
| 14 | Total                        | 74                  | 3115.481978           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 74.17972293         | 1.79898273            | 41.23426071   | 2.56286E-52    |
| 18 | ADVERT                       | 1.732615442         | 0.890323714           | 1.946051098   | 0.055498127    |

We can now transport the *SSEs* from the two models into our *F*-test template and test the hypothesis. First, highlight the cell that contains the *SSE* from the restricted model. With cell C13 highlighted, choose **Edit/Copy** from the main **Menu**. A scrolling border appears around the cell.



The value is now stored on the Excel's clipboard and be pasted as desired. Go to the “**Restricted Model**” worksheet and **Edit>Paste** the *SSE* number from the restricted model under the Data inputs at cell B7.



Now repeat this procedure for the unrestricted *SSE* from the original regression and paste it in cell B8. Fill in all the **Data Input** with appropriate information. Type **"0.05"** in cell B5 for testing at the 5% level. The computed values should now appear, and the appropriate decision reported.

|    | A                           | B           |
|----|-----------------------------|-------------|
| 1  | Hypothesis Testing - F-Test |             |
| 2  |                             |             |
| 3  | Data Input                  |             |
| 4  | J                           | 1           |
| 5  | N                           | 75          |
| 6  | K                           | 3           |
| 7  | SSE-RESTRICTED              | 2961.827    |
| 8  | SSE-UNRESTRICTED            | 1718.943    |
| 9  | ALPHA                       | 0.05        |
| 10 |                             |             |
| 11 | Computed Values             |             |
| 12 | df-numerator                | 1           |
| 13 | df-denominator              | 72          |
| 14 | F                           | 52.05969    |
| 15 | Right Critical value        | 3.973897    |
| 16 | Decision                    | Reject Null |
| 17 | p-value                     | 4.42E-10    |

Since the  $F$ -statistic  $>$   $F$ -critical value, and the  $p$ -value  $<$   $\alpha$ , we reject the null hypothesis at the 5% level and conclude the price does have an effect on total revenue.

## 6.2 TESTING THE OVERALL SIGNIFICANCE OF THE MODEL

In the application of the  $F$ -test, the model significance at the desired  $\alpha$ -level is determined. Consider a general linear model with  $K$  regressors with  $K-1$  explanatory variables and  $K$  unknown parameters.

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \dots + x_{ik}\beta_k + e_i$$

We test whether all of the coefficients on the  $K-1$  explanatory variables are jointly equal to zero, versus the alternative that at least one of coefficients is not zero. If the explanatory variables have no effect on the average value of  $y$ , then each of the slopes will be zero, leading to the following null and alternative hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_1 : \text{At least one of the } \beta_k \text{ is nonzero for } k=2,3,\dots,K$$

We can use the template created for any  $F$ -test of interest including this one. For jointly testing the significance of all of the explanatory variables in Andy's Burger Barn, we test that all the  $\beta$ 's are zero except  $\beta_1$ , the intercept. Note that when there are **NO** explanatory variables in the model other than the intercept. The  $SSE_R$  is equal to the  $SST$  from the unrestricted model. The results and appropriate decision are

|    | A                           | B           |
|----|-----------------------------|-------------|
| 1  | Hypothesis Testing - F-Test |             |
| 2  |                             |             |
| 3  | Data Input                  |             |
| 4  | J                           | 2           |
| 5  | N                           | 75          |
| 6  | K                           | 3           |
| 7  | SSE-RESTRICTED              | 3115.481978 |
| 8  | SSE-UNRESTRICTED            | 1718.942985 |
| 9  | ALPHA                       | 0.05        |
| 10 |                             |             |
| 11 | Computed Values             |             |
| 12 | df-numerator                | 2           |
| 13 | df-denominator              | 72          |
| 14 | F                           | 29.24785998 |
| 15 | Right Critical value        | 3.123907449 |
| 16 | Decision                    | Reject Null |
| 17 | p-value                     | 5.04086E-10 |

We reject the null hypothesis and conclude that our model is significant at the 5% level; price or advertising expenditures, or both have a significance effect on total revenue.

Alternatively, we can obtain the result to this  $F$ -test from Excel's ANOVA table. Notice that the  $F$  and  $p$ -values associated with the  $F$ -Test are identical.

|    |            |           |             |             |             |                       |
|----|------------|-----------|-------------|-------------|-------------|-----------------------|
| 9  |            |           |             |             |             |                       |
| 10 | ANOVA      |           |             |             |             |                       |
| 11 |            | <i>df</i> | <i>SS</i>   | <i>MS</i>   | <i>F</i>    | <i>Significance F</i> |
| 12 | Regression | 2         | 1396.538993 | 698.2694963 | 29.24785998 | 5.04086E-10           |
| 13 | Residual   | 72        | 1718.942985 | 23.87420813 |             |                       |
| 14 | Total      | 74        | 3115.481978 |             |             |                       |

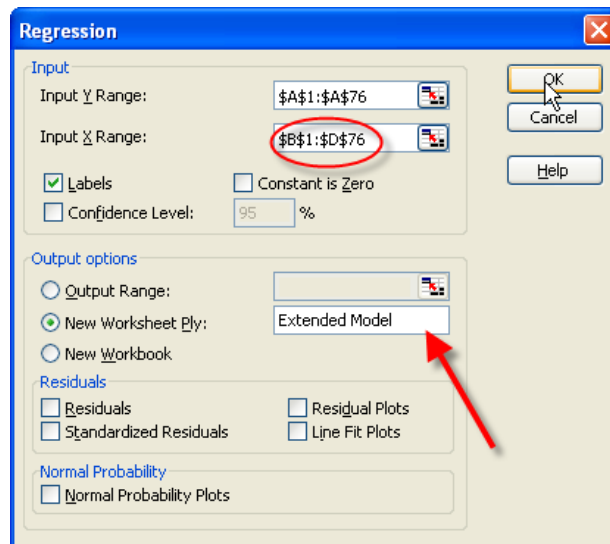
If we compare the  $F$ -statistic to a critical value, or more easily, compare the  $p$ -value to  $\alpha$ , we reject the null hypothesis and conclude the model is statistically "important".

### 6.3 AN EXTENDED MODEL

The concept of diminishing marginal returns is an important one in economics, and you should carefully consider this when modeling economic relationships. In our Sales for Andy's Burger Barn model, it seems reasonable that each and every dollar increase in advertising expenditures would not lead to the same increase in sales; that is, the possibility of diminishing marginal returns to advertising should be considered. To allow for this possibility, we include the explanatory variable squared in the model. For simplicity we will rename  $SALES = S$ ,  $PRICE = P$  and  $ADVERT = A$ .

$$S_i = \beta_1 + \beta_2 P_i + \beta_3 A_i + \beta_4 A_i^2 + e_i$$

To extend our model to include this additional regressor, open the worksheet *Andy.xls*. Label column D "A2" and in cell D2, type the formula  $=C2^2$ . Copy this formula down the column. Estimate the regression and save the output in the "Extended Model" worksheet.



We will get the following results:

|    | A                     | B            | C              | D            | E           | F              | G            |
|----|-----------------------|--------------|----------------|--------------|-------------|----------------|--------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |                |              |
| 2  |                       |              |                |              |             |                |              |
| 3  | Regression Statistics |              |                |              |             |                |              |
| 4  | Multiple R            | 0.712906125  |                |              |             |                |              |
| 5  | R Square              | 0.508235142  |                |              |             |                |              |
| 6  | Adjusted R Square     | 0.487456346  |                |              |             |                |              |
| 7  | Standard Error        | 4.645283161  |                |              |             |                |              |
| 8  | Observations          | 75           |                |              |             |                |              |
| 9  |                       |              |                |              |             |                |              |
| 10 | ANOVA                 |              |                |              |             |                |              |
| 11 |                       | df           | SS             | MS           | F           | Significance F |              |
| 12 | Regression            | 3            | 1583.397427    | 527.7991422  | 24.4593153  | 5.59997E-11    |              |
| 13 | Residual              | 71           | 1532.084551    | 21.57865565  |             |                |              |
| 14 | Total                 | 74           | 3115.481978    |              |             |                |              |
| 15 |                       |              |                |              |             |                |              |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     | Lower 95%      | Upper 95%    |
| 17 | Intercept             | 109.7190398  | 6.79904566     | 16.1374177   | 1.87037E-25 | 96.16212798    | 123.2759516  |
| 18 | PRICE                 | -7.640000543 | 1.045938915    | -7.304442384 | 3.23648E-10 | -9.725543479   | -5.554457608 |
| 19 | ADVERT                | 12.15123398  | 3.556164048    | 3.416949784  | 0.0010516   | 5.060444353    | 19.2420236   |
| 20 | A <sup>2</sup>        | -2.767962762 | 0.940624059    | -2.942687607 | 0.004392671 | -4.643513842   | -0.892411683 |

Price  $P$  is now insignificant at the 5% level when squared advertising expenditures ( $A^2$ ) are included in the model. The estimated coefficient on  $A^2$  is negative as expected, and comparing the  $p$ -value to (any level of)  $\alpha$  we conclude that it is significantly different from zero. This suggests that there are diminishing returns to advertising.

## 6.4 TESTING SOME ECONOMIC HYPOTHESES

Using this expanded model for total revenue, we will now examine several hypotheses of interest, using both  $t$ -tests and  $F$ -tests.

### 6.4.1 The significance of advertising

To test for the significance of all levels of advertising requires a joint hypothesis test; we must now consider the significance of both  $\beta_3$  and  $\beta_4$  using  $H_0: \beta_3 = 0, \beta_4 = 0$  against the alternative that at least one of the coefficients is not zero. We already have the unrestricted model in the previous section, we will need to estimate the restricted model, where  $PRICE$  will be the only explanatory variable. The ANOVA results are

|    | A                     | B           | C           | D           | E        | F              |
|----|-----------------------|-------------|-------------|-------------|----------|----------------|
| 1  | SUMMARY OUTPUT        |             |             |             |          |                |
| 2  |                       |             |             |             |          |                |
| 3  | Regression Statistics |             |             |             |          |                |
| 4  | Multiple R            | 0.625540559 |             |             |          |                |
| 5  | R Square              | 0.391300991 |             |             |          |                |
| 6  | Adjusted R Square     | 0.382962648 |             |             |          |                |
| 7  | Standard Error        | 5.09685747  |             |             |          |                |
| 8  | Observations          | 75          |             |             |          |                |
| 9  |                       |             |             |             |          |                |
| 10 | ANOVA                 |             |             |             |          |                |
| 11 |                       | df          | SS          | MS          | F        | Significance F |
| 12 | Regression            | 1           | 1219.091184 | 1219.091184 | 46.92791 | 1.97077E-09    |
| 13 | Residual              | 73          | 1896.390793 | 25.97795607 |          |                |
| 14 | Total                 | 74          | 3115.481978 |             |          |                |

We can now transfer the *SSEs* from this restricted model and the model we labeled “**Extended Model**” and use our template to test the significance of level of advertising at 5% significance level.

|    | A                                  | B                                  |
|----|------------------------------------|------------------------------------|
| 1  | <b>Hypothesis Testing - F-Test</b> |                                    |
| 2  |                                    |                                    |
| 3  | <b>Data Input</b>                  |                                    |
| 4  | J                                  | 2                                  |
| 5  | N                                  | 75                                 |
| 6  | K                                  | 4                                  |
| 7  | SSE-RESTRICTED                     | 1896.390793                        |
| 8  | SSE-UNRESTRICTED                   | 1532.084551                        |
| 9  | ALPHA                              | 0.05                               |
| 10 |                                    |                                    |
| 11 | <b>Computed Values</b>             | <b>SSE from the Extended Model</b> |
| 12 | df-numerator                       | 2                                  |
| 13 | df-denominator                     | 71                                 |
| 14 | F                                  | 8.441356314                        |
| 15 | Right Critical value               | 3.125764237                        |
| 16 | Decision                           | Reject Null                        |
| 17 | p-value                            | 0.000514161                        |

We reject the null hypothesis and conclude that advertising expenditures do significantly affect total revenue.

#### 6.4.2 Optimal level of advertising

We have already illustrated that the returns to advertising diminish. Then the optimal level of advertising will occur at the point where the marginal cost is equal to the marginal benefit of advertising. In other words, the optimum level is when the next dollar spent on advertising equals only one more dollar of sales. Taking the derivative of expected *SALES* with respect to Advertising expenditures *A* will give us the marginal benefit, which is

$$\frac{\Delta E(S)}{\Delta A} = \beta_3 + 2\beta_4 A = 1$$

Solving for *A* gives  $A^* = (1 - b_3)/2b_4$ , where  $b_3$  and  $b_4$  are the least square estimates. We can then substitute in our estimates for  $\beta_3$  and  $\beta_4$  and solve for the optimal level of advertising

$$A^* = \frac{1 - 12.15123398}{2(-2.767962762)} = 2.014$$

which is \$2014.

Suppose Andy wants to test if the optimal level of advertising is \$1,900. If we substitute 1.9 (since advertising data is in \$1000), leads to the following hypothesis:

$$H_0: \beta_3 + 2\beta_4(1.9) = 1 \text{ against the alternative } H_0: \beta_3 + 2\beta_4(1.9) \neq 1$$

or equivalently

$$H_0: \beta_3 + 3.8\beta_4 = 1 \text{ against the alternative } H_0: \beta_3 + 3.8\beta_4 \neq 1$$

A  $t$ -test could be used to test this single hypothesis, using the test statistic,

$$t = \frac{(b_3 + 3.8b_4) - 1}{se(b_3 + 3.8b_4)}$$

However, this test would require a calculation using the covariance between  $b_3$  and  $b_4$ . Since Excel does not report the estimated covariance matrix for the LS estimators, we will instead use the  $F$ -test. We can construct the restricted model by plug in the restriction ( $\beta_3 + 3.8\beta_4 = 1 \Rightarrow \beta_3 = 1 - 3.8\beta_4$ ) into our equation ( $S = \beta_1 + \beta_2 P + \beta_3 A + \beta_4 A^2$ ) such that the restricted model will become:

$$S - A = \beta_1 + \beta_2 P + \beta_4 (A^2 - 3.8A) + e$$

To run the restricted model, open *Andy.xls*.

- Highlight column B. Insert a column by choosing **Insert>Columns** from the menu bar. Label this new column *S-A* indicating *Sales - Advertising*.
- In the first empty cell of this column, type **=A2-D2** and copy this formula down the column. This column will represent our new dependent variable for the restricted model.
- Next, highlight column D and insert a new column. Label this *A2-3.8A*. In the first empty cell, type **=F2-(3.8\*E2)** and copy the formula down the column. This column will be our new explanatory variable.
- Observe that we inserted the *A2-3.8A* column next to the *PRICE* column since in Excel, the columns used for the **X-Range** (the explanatory variables) must be in adjacent columns. [So, if you ever find that you want to run a regression and the explanatory variables are not all in adjacent columns, simply highlight and move things around as needed.](#) Now we are ready to run the restricted model. Choose **Tools>Data Analysis>Regression**. Use *S-A* as the **Y-Range**. Use *PRICE* and *A2-3.8A* as the **X-Range**. Perform this regression as usual.

The regression summary output is



|    | A                     | B            | C              | D            | E           | F              | G            |
|----|-----------------------|--------------|----------------|--------------|-------------|----------------|--------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |                |              |
| 2  |                       |              |                |              |             |                |              |
| 3  | Regression Statistics |              |                |              |             |                |              |
| 4  | Multiple R            | 0.693339057  |                |              |             |                |              |
| 5  | R Square              | 0.480719048  |                |              |             |                |              |
| 6  | Adjusted R Square     | 0.466294577  |                |              |             |                |              |
| 7  | Standard Error        | 4.64322439   |                |              |             |                |              |
| 8  | Observations          | 75           |                |              |             |                |              |
| 9  |                       |              |                |              |             |                |              |
| 10 | ANOVA                 |              |                |              |             |                |              |
| 11 |                       | df           | SS             | MS           | F           | Significance F |              |
| 12 | Regression            | 2            | 1437.013271    | 718.5066356  | 33.32663303 | 5.6818E-11     |              |
| 13 | Residual              | 72           | 1552.286357    | 21.55953273  |             |                |              |
| 14 | Total                 | 74           | 2989.299628    |              |             |                |              |
| 15 |                       |              |                |              |             |                |              |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     | Lower 95%      | Upper 95%    |
| 17 | Intercept             | 110.3589599  | 6.763803393    | 16.31610996  | 6.84193E-26 | 96.87556446    | 123.8423554  |
| 18 | PRICE                 | -7.603104222 | 1.044780309    | -7.277227713 | 3.39617E-10 | -9.685835675   | -5.520372768 |
| 19 | A2-3.8A               | -2.876514915 | 0.93349559     | -3.081444569 | 0.002917717 | -4.737404337   | -1.015625493 |

This sum squared residual can now be used in the  $F$ -test template. Recall that we are testing  $H_0: \beta_3 + 2\beta_4(1.9) = 1$  or that \$1,900 is the optimal level of advertising or not. After copying  $SSE$ s for the restricted and unrestricted model, set  $J=1$  since we have only a single null hypothesis. The results from this test are

|    | A                           | B                   |
|----|-----------------------------|---------------------|
| 1  | Hypothesis Testing - F-Test |                     |
| 2  |                             |                     |
| 3  | Data Input                  |                     |
| 4  | J                           | 1                   |
| 5  | N                           | 75                  |
| 6  | K                           | 4                   |
| 7  | SSE-RESTRICTED              | 1552.286357         |
| 8  | SSE-UNRESTRICTED            | 1532.084551         |
| 9  | ALPHA                       | 0.05                |
| 10 |                             |                     |
| 11 | Computed Values             |                     |
| 12 | df-numerator                | 1                   |
| 13 | df-denominator              | 71                  |
| 14 | F                           | 0.9361939           |
| 15 | Right Critical value        | 3.975810047         |
| 16 | Decision                    | Fail to Reject Null |
| 17 | p-value                     | 0.336543031         |

We cannot reject the hypothesis that \$1,900 is the optimal level of weekly advertising expenditures at the 5% level and conclude that Andy's advertising strategy is compatible with the data.

We can also conduct a joint test of two of Big Andy's suppositions. Let's say in addition to proposing that the optimal level of monthly advertising expenditure is \$1,900, Andy is assuming that  $P = 6$  will yield sales revenue of \$80. The joint hypothesis will be:

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

$H_1$ : at least one hypothesis is not true

Since  $J = 2$ , we must perform an  $F$ -test. The restricted model is found by substituting both the hypotheses in the null on the model and rearranging terms to form a model suitable for estimation. It can be shown that the equation used to estimate the restricted model is

$$S - A - 80 - 1.9 = \beta_2(P - 6) + \beta_4(A^2 - 3.8A + 3.61) + e$$

To test this joint hypothesis in Excel, return to the worksheet containing the original data in *andy.xls*.

- Create three new columns labeled *YSTAR*,  $P-6$ , and *ASTAR*.
- In the first empty cell of *YSTAR*, type the formula **=B2-78.1**, where cell B2 contains  $S-A$ .
- In the first empty cell of  $P-6$ , type **=C2-6**, where cell C2 contains *PRICE*.
- In the first empty cell of *ASTAR*, type **=F2-(3.8\*E2)+3.61**, where cell F2 contains A2 and E2 contains A.
- Highlight the three cells containing these new formulas. Place the cursor on the lower right-hand corner of this selection until it turns into a cross-hatch. Left click, hold, and drag down to row 79. Release and the values appear in the cells.
- Estimate a regression using *YSTAR* as the **Y-Range**. Use  $P-6$ , and *ASTAR* as the **X-Range**.
- Use the *SSE* from this restricted model and *SSE* from the unrestricted model to conduct the  $F$ -test, where  $K = 4$ ,  $J = 2$ .

## 6.5 NONSAMPLE INFORMATION

Often times we have information about a particular model that does not come directly from the data. The information may come from past experience or from economic tenets. If correct the nonsample information improves the precision with which you can estimate the remaining parameters. To illustrate how we might go about incorporating the nonsample information, consider a model designed to explain the demand for beer, we will use a model of demand for beer ( $Q$ ) based on its price ( $PB$ ), the price of other liquor ( $PL$ ), the price of all other remaining goods and services ( $PR$ ), and income ( $I$ ). The nonsample information is that consumers do not suffer from "money illusion"; that is, when all prices and income go up by the same proportion, there is no change in quantity demanded.

We will use a log-log functional form for the model, and then impose restrictions that incorporate our nonsample information. The unrestricted model is

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

and will impose the restriction  $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$ .

Rearranging this restriction, we have  $\beta_4 = -\beta_2 - \beta_3 - \beta_5$ , which can be substituted into the unrestricted model. After some manipulation, the equation we will estimate is

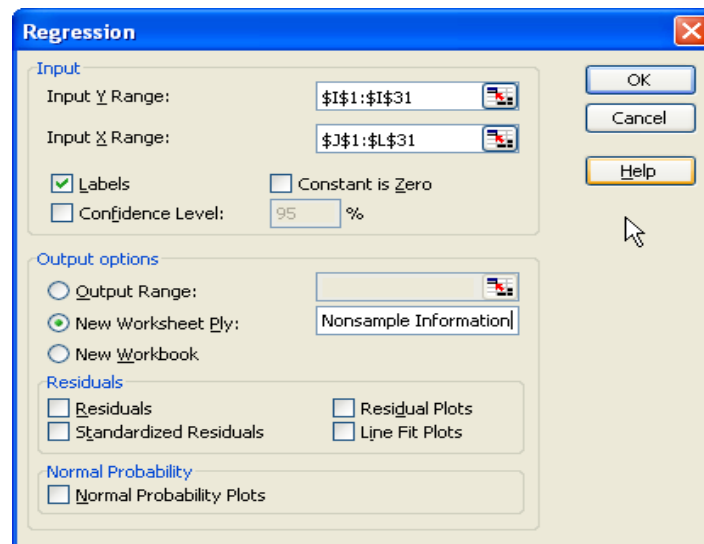
$$\ln(Q_t) = \beta_1 + \beta_2 \ln\left(\frac{PB_t}{PR_t}\right) + \beta_3 \ln\left(\frac{PL_t}{PR_t}\right) + \beta_5 \ln\left(\frac{I_t}{PR_t}\right) + e$$

This equation incorporates the properties of logarithms, as well as the nonsample information. We will get the **Restricted Least Square Estimates** by incorporating this restriction.

Open the data file *beer.xls*. Note that the original data have associated labels already.

- Label the next 3 columns *PB/PR*, *PL/PR*, and *I/PR*.
- For *PB/PR*, in cell F2, type **=B2/D2**. Highlight cell F2, left-click on the lower right hand corner, hold, and drag down the next two cells to right.
- Repeat the process for *PL/PR*, and *I/PR*.
- Label the next four columns for the logs of the data, labeling them *lnQ*, *lnPB/PR*, *lnPL/PR*, and *lnI/PR*. Calculate the natural log of *Q* in cell I2 by typing **=LN(A2)**, where cell A2 contains the first observation on *Q*.
- Calculate the natural log of *PB/PR* in cell J2 by typing **=LN(F2)**. Highlight cell J2, left-click on the lower right hand corner, hold, and drag across the next two cells to right. This copies the formula for the other variables.
- Highlight the section of cells containing the log formulas (I2 to L2). Left-click on the lower right hand corner of the selection, hold, and drag down the column to copy the formulas down to row 31.

We can now run the regression by choosing **Tools>Data Analysis>Regression** to estimate the regression model. Use *lnQ* as the **Y-Range** and *lnPB/PR*, *lnPL/PR*, and *lnI/PR* as the **X-Range**. Include labels by checking the **Labels** box. Store the regression output on a new worksheet named “**Nonsample Information**”.



and click **OK** for the output.

|    | A                     | B            | C              | D            | E           | F              | G            |
|----|-----------------------|--------------|----------------|--------------|-------------|----------------|--------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |                |              |
| 2  |                       |              |                |              |             |                |              |
| 3  | Regression Statistics |              |                |              |             |                |              |
| 4  | Multiple R            | 0.898859761  |                |              |             |                |              |
| 5  | R Square              | 0.80794887   |                |              |             |                |              |
| 6  | Adjusted R Square     | 0.785789124  |                |              |             |                |              |
| 7  | Standard Error        | 0.061675593  |                |              |             |                |              |
| 8  | Observations          | 30           |                |              |             |                |              |
| 9  |                       |              |                |              |             |                |              |
| 10 | ANOVA                 |              |                |              |             |                |              |
| 11 |                       | df           | SS             | MS           | F           | Significance F |              |
| 12 | Regression            | 3            | 0.416070592    | 0.138690197  | 36.46020488 | 1.83399E-09    |              |
| 13 | Residual              | 26           | 0.098900847    | 0.003803879  |             |                |              |
| 14 | Total                 | 29           | 0.514971439    |              |             |                |              |
| 15 |                       |              |                |              |             |                |              |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     | Lower 95%      | Upper 95%    |
| 17 | Intercept             | -4.797797376 | 3.71390504     | -1.291847079 | 0.207775913 | -12.43183844   | 2.836243691  |
| 18 | lnPB/PR               | -1.299386484 | 0.165737623    | -7.840021241 | 2.57799E-08 | -1.640065044   | -0.958707925 |
| 19 | lnPL/PR               | 0.186815879  | 0.284383258    | 0.656915882  | 0.517008126 | -0.397742275   | 0.771374032  |
| 20 | lnI/PR                | 0.945828579  | 0.427046831    | 2.214812313  | 0.035742225 | 0.068021255    | 1.823635904  |

Recall that our restriction was  $\beta_4 = -\beta_2 - \beta_3 - \beta_5$ . To get back the estimate of  $\beta_4$ , we need one more step. Unfortunately, while some statistical packages have options to automatically estimate restricted least squares estimates, Excel does not.

In cell A21, type  $b_4^*$  as a label. Next, in cell B21 type  $= -B18-B19-B20$  to calculate  $b_4^* = -b_2^* - b_3^* - b_5^*$ , where cells B18, B19, and B20 contain the respective estimates. The restricted least squares estimates are

|    | A                     | B                  | C              |
|----|-----------------------|--------------------|----------------|
| 1  | SUMMARY OUTPUT        |                    |                |
| 2  |                       |                    |                |
| 3  | Regression Statistics |                    |                |
| 4  | Multiple R            | 0.898859761        |                |
| 5  | R Square              | 0.80794887         |                |
| 6  | Adjusted R Square     | 0.785789124        |                |
| 7  | Standard Error        | 0.061675593        |                |
| 8  | Observations          | 30                 |                |
| 9  |                       |                    |                |
| 10 | ANOVA                 |                    |                |
| 11 |                       | df                 | SS             |
| 12 | Regression            | 3                  | 0.416070592    |
| 13 | Residual              | 26                 | 0.098900847    |
| 14 | Total                 | 29                 | 0.514971439    |
| 15 |                       |                    |                |
| 16 |                       | Coefficients       | Standard Error |
| 17 | Intercept             | -4.797797376       | 3.71390504     |
| 18 | lnPB/PR               | -1.299386484       | 0.165737623    |
| 19 | lnPL/PR               | 0.186815879        | 0.284383258    |
| 20 | lnI/PR                | 0.945828579        | 0.427046831    |
| 21 |                       |                    |                |
| 22 | <b>b4*</b>            | <b>0.166742026</b> |                |

Recall that the log-log model specification provides estimates of elasticities, not marginal effects. Substituting these results back into our specification, we have

$$\ln(Q_i) = -4.7978 - 1.2994 \ln(PB) + 0.1868 \ln(PL) + 0.1667 \ln(PR) + 0.9458 \ln(I)$$

From the results, we find that demand for beer is price elastic ( $b_2 < -1$ ), does not seem to be affected by the price of other liquor ( $\beta_3$  is not statistically significant), and might be an inferior good ( $\beta_5 < 1$ ), although this would have to be formally tested.

## 6.6 MODEL SPECIFICATION

Three essential features of model choice are (1) choice of functional form, (2) choice of explanatory variables (regressors) to be included in the model, and (3) whether the multiple regression model assumptions MR1–MR6, listed in Chapter 5, hold. In this section, we will explore the first two.

### 6.6.1 Omitted variables

If you omit relevant variables from your model, then least squares estimator will be biased. To introduce the omitted variable problem, we will consider a sample of married couples where both husbands and wives work. Open *edu\_inc.xls* and first regress family income (*FAMINC*) on both husband's (*HE*) and wife's education (*WE*). The results are

|    | A                     | B            | C              | D            | E           |
|----|-----------------------|--------------|----------------|--------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |
| 2  |                       |              |                |              |             |
| 3  | Regression Statistics |              |                |              |             |
| 4  | Multiple R            | 0.401622273  |                |              |             |
| 5  | R Square              | 0.16130045   |                |              |             |
| 6  | Adjusted R Square     | 0.157353629  |                |              |             |
| 7  | Standard Error        | 40497.85958  |                |              |             |
| 8  | Observations          | 428          |                |              |             |
| 9  |                       |              |                |              |             |
| 10 | ANOVA                 |              |                |              |             |
| 11 |                       | df           | SS             | MS           | F           |
| 12 | Regression            | 2            | 1.34055E+11    | 67027380194  | 40.86844417 |
| 13 | Residual              | 425          | 6.97033E+11    | 1640076630   |             |
| 14 | Total                 | 427          | 8.31087E+11    |              |             |
| 15 |                       |              |                |              |             |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     |
| 17 | Intercept             | -5533.630698 | 11229.53295    | -0.492774786 | 0.622426214 |
| 18 | HE                    | 3131.509312  | 802.9079857    | 3.900209449  | 0.00011168  |
| 19 | WE                    | 4522.641199  | 1066.326646    | 4.241328128  | 2.72885E-05 |

Omitting wife's education and regressing family income (*FAMINC*) on only husband's (*HE*) yields:

|    | A                     | B            | C              | D           | E           |
|----|-----------------------|--------------|----------------|-------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |             |             |
| 2  |                       |              |                |             |             |
| 3  | Regression Statistics |              |                |             |             |
| 4  | Multiple R            | 0.354684413  |                |             |             |
| 5  | R Square              | 0.125801033  |                |             |             |
| 6  | Adjusted R Square     | 0.123748923  |                |             |             |
| 7  | Standard Error        | 41297.49176  |                |             |             |
| 8  | Observations          | 428          |                |             |             |
| 9  |                       |              |                |             |             |
| 10 | ANOVA                 |              |                |             |             |
| 11 |                       | df           | SS             | MS          | F           |
| 12 | Regression            | 1            | 1.04552E+11    | 1.04552E+11 | 61.3032526  |
| 13 | Residual              | 426          | 7.26536E+11    | 1705482826  |             |
| 14 | Total                 | 427          | 8.31087E+11    |             |             |
| 15 |                       |              |                |             |             |
| 16 |                       | Coefficients | Standard Error | t Stat      | P-value     |
| 17 | Intercept             | 26191.26868  | 8541.108357    | 3.066495305 | 0.002303803 |
| 18 | HE                    | 5155.483577  | 658.4573486    | 7.829639366 | 3.92051E-14 |

And including *WE* and number of preschool age children (*KL6*) yields:

|    | A                     | B                   | C                     | D             | E              |
|----|-----------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT        |                     |                       |               |                |
| 2  |                       |                     |                       |               |                |
| 3  | Regression Statistics |                     |                       |               |                |
| 4  | Multiple R            | 0.420919613         |                       |               |                |
| 5  | R Square              | 0.177173321         |                       |               |                |
| 6  | Adjusted R Square     | 0.171351434         |                       |               |                |
| 7  | Standard Error        | 40160.0814          |                       |               |                |
| 8  | Observations          | 428                 |                       |               |                |
| 9  |                       |                     |                       |               |                |
| 10 | ANOVA                 |                     |                       |               |                |
| 11 |                       | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression            | 3                   | 1.47247E+11           | 49082167249   | 30.43228498    |
| 13 | Residual              | 424                 | 6.83841E+11           | 1612832138    |                |
| 14 | Total                 | 427                 | 8.31087E+11           |               |                |
| 15 |                       |                     |                       |               |                |
| 16 |                       | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept             | -7755.331329        | 11162.93447           | -0.694739484  | 0.487599098    |
| 18 | HE                    | 3211.525676         | 796.7026365           | 4.031021775   | 6.58407E-05    |
| 19 | WE                    | 4776.907489         | 1061.16372            | 4.501574447   | 8.72703E-06    |
| 20 | KL6                   | -14310.92032        | 5003.928369           | -2.859937086  | 0.004446558    |

### 6.6.2 Irrelevant variables

Including irrelevant variables in your model impairs the precision of least squares estimator. Least squares will be unbiased, but standard errors of the coefficients will be larger than necessary. In this example, two extraneous variables (*XTRA\_X5* and *XTRA\_X6*) are added to the model. The results are:

|    | A                     | B                   | C                     | D             | E              |
|----|-----------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT        |                     |                       |               |                |
| 2  |                       |                     |                       |               |                |
| 3  | Regression Statistics |                     |                       |               |                |
| 4  | Multiple R            | 0.421659186         |                       |               |                |
| 5  | R Square              | 0.177796469         |                       |               |                |
| 6  | Adjusted R Square     | 0.168054721         |                       |               |                |
| 7  | Standard Error        | 40239.88895         |                       |               |                |
| 8  | Observations          | 428                 |                       |               |                |
| 9  |                       |                     |                       |               |                |
| 10 | ANOVA                 |                     |                       |               |                |
| 11 |                       | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression            | 5                   | 1.47764E+11           | 29552878521   | 18.2509822     |
| 13 | Residual              | 422                 | 6.83323E+11           | 1619248663    |                |
| 14 | Total                 | 427                 | 8.31087E+11           |               |                |
| 15 |                       |                     |                       |               |                |
| 16 |                       | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept             | -7558.615278        | 11195.41057           | -0.675153022  | 0.4999483      |
| 18 | HE                    | 3339.791791         | 1250.039195           | 2.671749657   | 0.007837757    |
| 19 | WE                    | 5868.678132         | 2278.066794           | 2.576165961   | 0.010329332    |
| 20 | KL6                   | -14200.18311        | 5043.719575           | -2.815418839  | 0.005099603    |
| 21 | XTRA_X5               | 888.8441205         | 2242.490245           | 0.396364766   | 0.692036298    |
| 22 | XTRA_X6               | -1067.186625        | 1981.684891           | -0.538524883  | 0.590498671    |

Notice how much larger the estimated standard errors became compared to the last regression in the previous section.

### 6.6.3 Choosing the model

Choosing an appropriate functional form for the model is very important. Although theoretical considerations should be the primary guide to functional form selection, you can also use the **RESET** (Regression Specification Error Test) test as a check to determine if you are making an obvious error in specifying your function or not. RESET is basically an *F*-test where the

restricted model is the "original" model and the unrestricted model is a polynomial approximation including the predicted  $y_i$ 's squared, cubed, etc., as explanatory variables. The general foundation of the test is that if the model is improved by artificially including powers of the predicted values, then the original model must not have been adequate.

RESET is a simple test with the null hypothesis is that your functional form is correct, the alternative is that it is not. We will talk about two variants of the RESET test; RESET(1) and RESET(2). The first adds only  $\hat{y}^2$  to the model and tests its significance using the t-test. The second adds both  $\hat{y}^2$  and  $\hat{y}^3$  and tests their joint significance.

Estimate the regression assuming the functional form is correct and obtain the coefficient estimates, calculate the predicted values. We will illustrate these tests using the family income regression where the family income is the dependent variables and the education of the husband and wife are the explanatory variables.

$$FAMINC = \beta_1 + \beta_2 HE + \beta_3 WE + e$$

The unrestricted model includes the squares of the predicted  $y$  for RESET(1) and both squares and cubes of the predicted  $y$  for RESET(2).

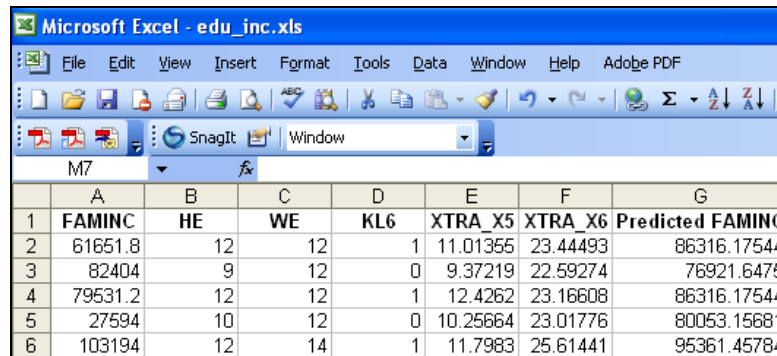
$$\text{RESET (1)} \quad y_i = \beta_1 + \beta_2 HE_i + \beta_3 WE_i + \gamma_1 \hat{y}_i^2 + e$$

$$\text{RESET(2)} \quad y_i = \beta_1 + \beta_2 HE_i + \beta_3 WE_i + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + e$$

Open *edu\_inc.xls*. First estimate the “**Restricted Model**” where *FAMINC* is the **Y-Range**, and *HE* and *WE* are the **X-Range**. Place the output on a new worksheet called “**Restricted**”. Check the **Residuals** box. This is needed to obtain the predicted  $y$ -values, then click **OK**.

From the regression output, we will use the *SSE* to be used for the *F*-test. Below the regular regression output, you will observe the residual output. The residual option in the regression

dialog box provides the predicted values and the errors associated with each observation. Since we will use the squares and cubes of the predicted column, we need to copy and paste this column next to our data. Highlight cells **B25 to B453**, which contain the predicted *FAMINC*. Choose **Edit>Copy** from the menu bar. Move to the worksheet containing the original data. Go to cell **G1** and **Edit>Paste**. Your data now looks like:



|   | A       | B  | C  | D   | E        | F        | G                |
|---|---------|----|----|-----|----------|----------|------------------|
| 1 | FAMINC  | HE | WE | KL6 | XTRA_X5  | XTRA_X6  | Predicted FAMINC |
| 2 | 61651.8 | 12 | 12 | 1   | 11.01355 | 23.44493 | 86316.17544      |
| 3 | 82404   | 9  | 12 | 0   | 9.37219  | 22.59274 | 76921.6475       |
| 4 | 79531.2 | 12 | 12 | 1   | 12.4262  | 23.16608 | 86316.17544      |
| 5 | 27594   | 10 | 12 | 0   | 10.25664 | 23.01776 | 80053.15681      |
| 6 | 103194  | 12 | 14 | 1   | 11.7983  | 25.61441 | 95361.45784      |

Now create two new columns, *YHAT2* and *YHAT3* in columns H and I. However, before we create the square and the cube of the predicted values, we will need to make an adjustment for the RESET test to work. We will first create a column for the adjustment; in cell H1 type *adjusted yhat*. In cell H2, type **=G2/10,000** and copy it down the entire column.

| FAMINC   | HE | WE | KL6 | XTRA_X5  | XTRA_X6  | Predicted FAMINC | adjusted yhat |
|----------|----|----|-----|----------|----------|------------------|---------------|
| 61651.8  | 12 | 12 | 1   | 11.01355 | 23.44493 | 86316.17544      | 8.631617544   |
| 82404    | 9  | 12 | 0   | 9.37219  | 22.59274 | 76921.6475       | 7.69216475    |
| 79531.2  | 12 | 12 | 1   | 12.4262  | 23.16608 | 86316.17544      | 8.631617544   |
| 27594    | 10 | 12 | 0   | 10.25664 | 23.01776 | 80053.15681      | 8.005315681   |
| 103194   | 12 | 14 | 1   | 11.7983  | 25.61441 | 95361.45784      | 9.536145784   |
| 73691.1  | 11 | 12 | 0   | 11.4462  | 24.16109 | 83184.66613      | 8.318466613   |
| 79954.56 | 12 | 16 | 0   | 11.69595 | 26.28513 | 104406.7402      | 10.44067402   |
| 71442    | 8  | 12 | 0   | 5.067864 | 16.52149 | 73790.13819      | 7.379013819   |
| 77130.9  | 4  | 12 | 0   | 4.254293 | 17.90169 | 61264.10094      | 6.126410094   |
| 77206.5  | 12 | 12 | 0   | 10.87064 | 21.94787 | 86316.17544      | 8.631617544   |
| 122094   | 12 | 12 | 0   | 13.67145 | 25.39433 | 86316.17544      | 8.631617544   |
| 108486   | 14 | 11 | 0   | 10.16985 | 22.11444 | 88056.55286      | 8.805655286   |

Now, using the adjusted yhat, we will create two new cells. **Insert>Column** two new columns next to variable *WE* and label them *yhat2* and *yhat3*, respectively. Notice that the adjusted yhat cell has moved to column J. In cell D2, type **=J2^2**. In cell E2, type **=J2^3**.

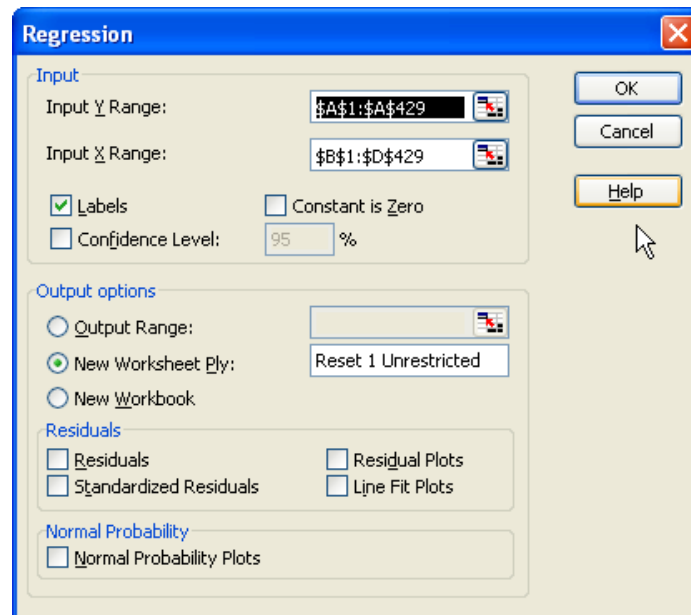
| A        | B  | C  | D          | E           | F   | G        | H        | I                | J             |
|----------|----|----|------------|-------------|-----|----------|----------|------------------|---------------|
| FAMINC   | HE | WE | yhat2      | yhat3       | KL6 | XTRA_X5  | XTRA_X6  | Predicted FAMINC | adjusted yhat |
| 61651.8  | 12 | 12 | =J2^2      | =J2^3       | 1   | 11.01355 | 23.44493 | 86316.17544      | 8.631617544   |
| 82404    | 9  | 12 | 87.8379454 | 823.2339135 | 0   | 9.37219  | 22.59274 | 76921.6475       | 7.69216475    |
| 79531.2  | 12 | 12 | 154.410422 | 1918.734626 | 1   | 12.4262  | 23.16608 | 86316.17544      | 8.631617544   |
| 27594    | 10 | 12 | 105.198644 | 1078.98451  | 0   | 10.25664 | 23.01776 | 80053.15681      | 8.005315681   |
| 103194   | 12 | 14 | 139.199765 | 1642.31989  | 1   | 11.7983  | 25.61441 | 95361.45784      | 9.536145784   |
| 73691.1  | 11 | 12 | 131.015449 | 1499.628766 | 0   | 11.4462  | 24.16109 | 83184.66613      | 8.318466613   |
| 79954.56 | 12 | 16 | 136.795246 | 1599.950362 | 0   | 11.69595 | 26.28513 | 104406.7402      | 10.44067402   |
| 71442    | 8  | 12 | 25.6832455 | 130.1591954 | 0   | 5.067864 | 16.52149 | 73790.13819      | 7.379013819   |
| 77130.9  | 4  | 12 | 18.0990089 | 76.998487   | 0   | 4.254293 | 17.90169 | 61264.10094      | 6.126410094   |
| 77206.5  | 12 | 12 | 118.170749 | 1284.591314 | 0   | 10.87064 | 21.94787 | 86316.17544      | 8.631617544   |

Highlight both cells D2 and E2. **Copy** the formulas down to row 429.



Now we have all the explanatory variables ready. Recall that when you perform a regression in Excel, all of the explanatory variables must be in adjacent cells, that is why we inserted the new columns next to the other explanatory variables.

For RESET(1), we will only use *YHAT2*.



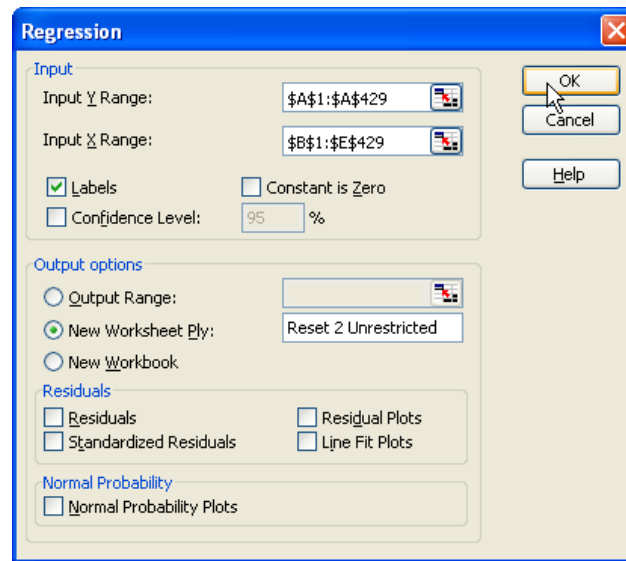
The image shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' set to '\$A\$1:\$A\$429' and 'Input X Range' set to '\$B\$1:\$D\$429'. The 'Labels' checkbox is checked, and 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected with the name 'Reset 1 Unrestricted'. The 'Residuals' section has 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' all unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. Buttons for 'OK', 'Cancel', and 'Help' are on the right.

We can now conduct the RESET(1). Since we are testing the significance of *yhat2*, we can either do a *t*-test or and *F*-test using the *SSE* from the unrestricted model (above) and the restricted model.

|           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
|-----------|---------------------|-----------------------|---------------|----------------|
| Intercept | 80147.89156         | 43868.10724           | 1.827019593   | 0.068399627    |
| HE        | -2013.988986        | 2669.983808           | -0.754307555  | 0.451083279    |
| WE        | -3490.936858        | 4106.98392            | -0.85000028   | 0.4095804723   |
| yhat2     | 934.2866496         | 462.5243462           | 2.01997291    | 0.044014621    |

The *t*-statistic from the regression output of RESET(1) Unrestricted is 2.02 and the *p*-value is 0.044. Since  $t\text{-stat} > t\text{-critical}$  and  $p\text{-value} < \alpha$ , we reject the null hypothesis; we have evidence to support model misspecification. Also, since in a single hypotheses test,  $F = t^2$ ,  $F = 4.08$ .

For RESET(2), go to **Tools>Data Analysis>Regression**.



This time use *FAMINC* in the **Y-Range**, and *HE*, *WE*, *YHAT2* and *YHAT3* as the **X-Range**. Place the output on a new worksheet called “**Reset 2 Unrestricted**”.


| ANOVA      |                     |                       |               |                |
|------------|---------------------|-----------------------|---------------|----------------|
|            | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| Regression | 4                   | 1.44217E+11           | 36054151749   | 22.20345943    |
| Residual   | 423                 | 6.86871E+11           | 1623807852    |                |
| Total      | 427                 | 8.31087E+11           |               |                |
|            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept  | 279165.3441         | 142130.1137           | 1.964153386   | 0.050166679    |
| HE         | -22010.89846        | 13844.86128           | -1.589824413  | 0.112621469    |
| WE         | -31547.04547        | 19497.27245           | -1.618023524  | 0.106402912    |
| yhat2      | 8521.538202         | 5175.350858           | 1.646562414   | 0.100390716    |
| yhat3      | -291.9788247        | 198.3673197           | -1.471909915  | 0.14178879     |

This time we are testing the joint significance of *YHAT2* and *YHAT3*. In other words, we test the hypotheses  $H_0: \gamma_1 = \gamma_2 = 0$  versus the alternative that  $\gamma_1 \neq 0$ ,  $\gamma_2 \neq 0$ , or both do not equal zero. For the joint test, open the template labeled “*F-Test*”.

- Fill in the **Data Input** values for the *F*-test.  $J = 2$ ,  $T = 428$ , and  $K = 4$ .
- Next, we will copy the *SSEs* from the restricted and unrestricted models and paste it into the template. The *SSE* from the unrestricted model is cell C13 of the output above. So right-click on cell C13 from the regression output.
- Choose **Copy**, place the cursor in cell B8 of the *F*-test template, right click and choose **Paste**.
- Next, return to the “Restricted Model” regression output and copy the *SSE* from the output and paste it in cell C7.

Below are the results of the *F*-test; we reject the null hypotheses and conclude that the specification is inadequate.

| Hypothesis Testing - F-Test |  |             |
|-----------------------------|--|-------------|
| Data Input                  |  |             |
| J                           |  | 2           |
| N                           |  | 428         |
| K                           |  | 4           |
| SSE-RESTRICTED              |  | 6.97033E+11 |
| SSE-UNRESTRICTED            |  | 6.86871E+11 |
| ALPHA                       |  | 0.05        |
| Computed Values             |  |             |
| df-numerator                |  | 2           |
| df-denominator              |  | 424         |
| F                           |  | 3.136414778 |
| Right Critical value        |  | 3.01699839  |
| Decision                    |  | Reject Null |
| p-value                     |  | 0.044447774 |



Beware that the RESET results are different than those in your book. Although these results are valid, EXCEL will not be able to provide you with the same RESET values despite the adjustment we made above.

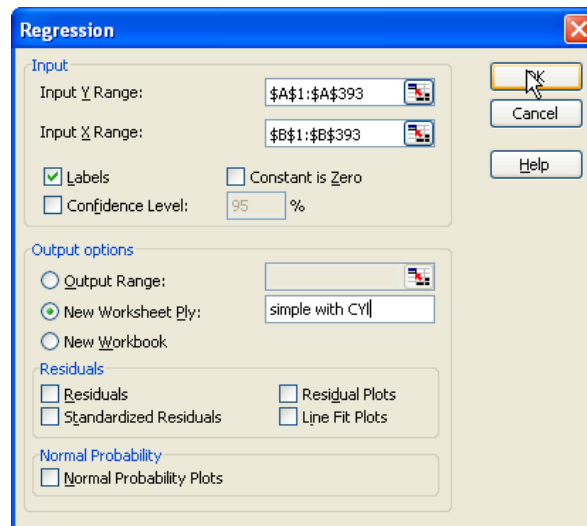
## 6.7 POOR DATA, COLLINEARITY AND INSIGNIFICANCE

When two or more explanatory variables are correlated, or collinear, the multiple regression model is unable to isolate individual affects on the dependent variable. Collinearity can cause high standard errors for the least squares estimators, resulting in  $t$ -tests that suggest the parameters are not significantly different from zero. Some strange results can occur, and we should be careful in interpreting our results when collinearity is present.

When there are one or more exact linear relationships between any of the explanatory variables, the least squares estimation process does not work. Many statistical packages will not even provide results, and will issue some type of error message. Excel does produce results, without any warnings or error messages but issues zeros for the standard errors.

More commonly, we face situations where the collinearity is not perfect, but can be "harmful". When linear relationships between our explanatory variables are strong enough, high standard errors, low  $t$ -statistics, and unstable estimates result. We should, therefore, look to see if our results are being affected by collinearity. There are several things to look at when trying to determine the existence of this type of problem, correlation and something we call an auxiliary regression.

To explore the ways of identifying collinearity, we will use *cars.xls*. Open the data set and first estimate the model of miles per gallon (*MPG*) as a function of the number of cylinders (*CYL*) in the engine.



**Regression**

**Input**

Input Y Range: \$A\$1:\$A\$393

Input X Range: \$B\$1:\$B\$393

☒ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

**Output options**

☐ Output Range:

☒ New Worksheet Ply: simple with CYL

☐ New Workbook

**Residuals**

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

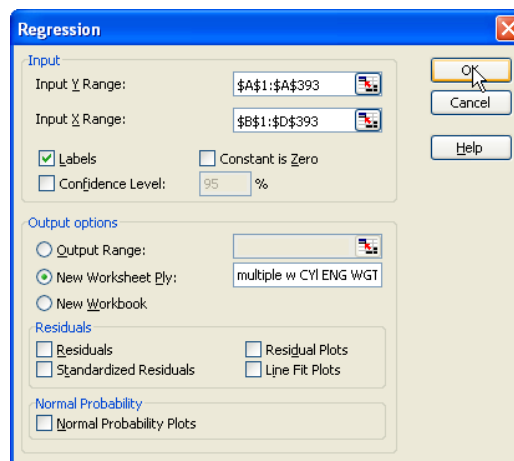
☐ Normal Probability Plots

OK Cancel Help

The output from this simple regression indicates a strong negative linear relationship between the *MGP* and *CYL*.

|    | A                     | B            | C              | D            |
|----|-----------------------|--------------|----------------|--------------|
| 1  | SUMMARY OUTPUT        |              |                |              |
| 2  |                       |              |                |              |
| 3  | Regression Statistics |              |                |              |
| 4  | Multiple R            | 0.777617509  |                |              |
| 5  | R Square              | 0.60468899   |                |              |
| 6  | Adjusted R Square     | 0.603675372  |                |              |
| 7  | Standard Error        | 4.913589267  |                |              |
| 8  | Observations          | 392          |                |              |
| 9  |                       |              |                |              |
| 10 | ANOVA                 |              |                |              |
| 11 |                       | df           | SS             | MS           |
| 12 | Regression            | 1            | 14403.08286    | 14403.08286  |
| 13 | Residual              | 390          | 9415.910199    | 24.14335948  |
| 14 | Total                 | 391          | 23818.99306    |              |
| 15 |                       |              |                |              |
| 16 |                       | Coefficients | Standard Error | t Stat       |
| 17 | Intercept             | 42.9155052   | 0.834866841    | 51.4040121   |
| 18 | CYL                   | -3.558078341 | 0.145675537    | -24.42467981 |

Now add the car's engine displacement in cubic inches (*ENG*) and weight (*WGT*) to the model.



**Regression**

**Input**

Input Y Range: \$A\$1:\$A\$393

Input X Range: \$B\$1:\$D\$393

☒ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

**Output options**

☐ Output Range:

☒ New Worksheet Ply: multiple w CYL ENG WGT

☐ New Workbook

**Residuals**

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

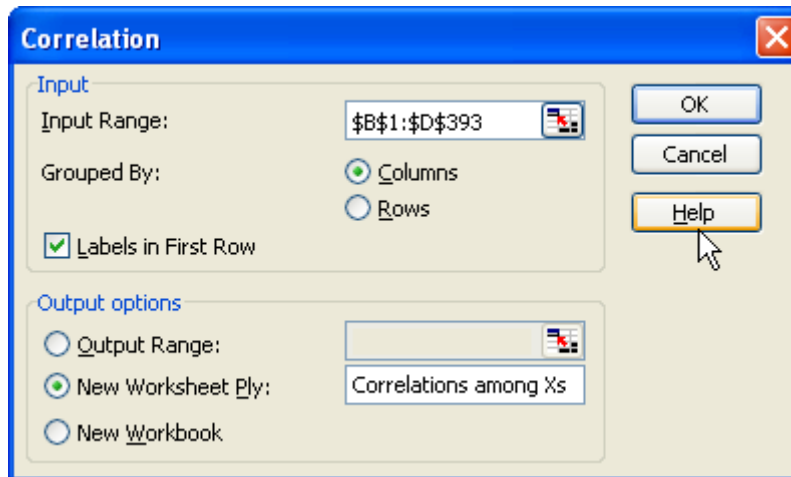
☐ Normal Probability Plots

OK Cancel Help

The output now shows a very different relationship between the *MPG* and *CYL*.

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.836237128         |                       |               |                |
| 5  | R Square                     | 0.699292534         |                       |               |                |
| 6  | Adjusted R Square            | 0.696967476         |                       |               |                |
| 7  | Standard Error               | 4.296530924         |                       |               |                |
| 8  | Observations                 | 392                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 3                   | 16656.444             | 5552.148001   | 300.7635141    |
| 13 | Residual                     | 388                 | 7162.549057           | 18.46017798   |                |
| 14 | Total                        | 391                 | 23818.99306           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 44.37096115         | 1.480685053           | 29.96650844   | 5.3199E-103    |
| 18 | CYL                          | -0.267796747        | 0.413067325           | -0.648312588  | 0.517166276    |
| 19 | ENG                          | -0.01267396         | 0.008250068           | -1.536224886  | 0.125298269    |
| 20 | WGT                          | -0.005707884        | 0.000713919           | -7.995142549  | 1.50112E-14    |

The value of the coefficient of *CYL* is much smaller, the standard error much larger and the variable is not significant anymore. We can also test other hypothesis such as the significance of *ENG* and joint significance of *CYL* and *ENG*. We will observe that *CYL* and *ENG* are individually not significant at 5% level but jointly they are. Joint test can be conducted by estimating the restricted model where both of these parameters are zero and using the *F*-test template as described in the previous section. This can happen when we can not isolate their individual impact. The independent variables *CYL*, *ENG* and *WGT* are highly correlated with each other and therefore highly collinear. A simple way to check the strength of possible linearity is to look at the correlation matrix. Correlation measures the direction and strength of a linear relationship between two variables. Choose **Tools>Data Analysis>Correlation** and indicate the data range in the dialog box.



The correlation matrix is

|   | A   | B        | C        | D   |
|---|-----|----------|----------|-----|
| 1 |     | ENG      | CYL      | WGT |
| 2 | ENG | 1        |          |     |
| 3 | CYL | 0.950823 | 1        |     |
| 4 | WGT | 0.932994 | 0.897527 | 1   |

These results show the high degree of linearity between all of the variables. However, correlation only measures the pair-wise linearity between variables. A more complex linear relationship between several variables at a time is not detected by correlation.

To detect more complex linear relationships, we will use the coefficient of determination,  $R^2$ , introduced in chapter six. Recall that  $R^2$  is interpreted as the percent of the total variation in the dependent variable that is explained by the model, or the explanatory variables. This interpretation is very helpful now.

An auxiliary regression is a multiple regression, but one of the original explanatory variables is used as the dependent variable. We are not concerned with any of the regression output except the  $R^2$ , because it measures how much of the variation in that one explanatory variable is explained, or being determined by, the other explanatory variables. This is, then, a measure of collinearity.

We can estimate the auxiliary regressions where each explanatory variable is regressed on all the others.

Regressing *ENG* on *CYL* and *WGT*

|    | A                     | B            | C              | D            | E           |
|----|-----------------------|--------------|----------------|--------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |
| 2  |                       |              |                |              |             |
| 3  | Regression Statistics |              |                |              |             |
| 4  | Multiple R            | 0.967809154  |                |              |             |
| 5  | R Square              | 0.936654558  |                |              |             |
| 6  | Adjusted R Square     | 0.936328874  |                |              |             |
| 7  | Standard Error        | 26.40496601  |                |              |             |
| 8  | Observations          | 392          |                |              |             |
| 9  |                       |              |                |              |             |
| 10 | ANOVA                 |              |                |              |             |
| 11 |                       | df           | SS             | MS           | F           |
| 12 | Regression            | 2            | 4010374.266    | 2005187.133  | 2875.965578 |
| 13 | Residual              | 389          | 271219.4474    | 697.2222299  |             |
| 14 | Total                 | 391          | 4281593.714    |              |             |
| 15 |                       |              |                |              |             |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     |
| 17 | Intercept             | -151.5981681 | 4.870943525    | -31.12295746 | 1.2538E-107 |
| 18 | CYL                   | 35.78843715  | 1.77531476     | 20.15892502  | 2.1609E-62  |
| 19 | WGT                   | 0.050436196  | 0.003565214    | 14.1467502   | 5.99868E-37 |

Regressing *CYL* on *WGT* and *ENG*

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.951263558         |                       |               |                |
| 5  | R Square                     | 0.904902357         |                       |               |                |
| 6  | Adjusted R Square            | 0.904413423         |                       |               |                |
| 7  | Standard Error               | 0.527378352         |                       |               |                |
| 8  | Observations                 | 392                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 2                   | 1029.499563           | 514.7497815   | 1850.766256    |
| 13 | Residual                     | 389                 | 108.1917635           | 0.278127927   |                |
| 14 | Total                        | 391                 | 1137.691327           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 2.215643749         | 0.14287129            | 15.50797046   | 1.4281E-42     |
| 18 | ENG                          | 0.014276314         | 0.000708188           | 20.15892502   | 2.1609E-62     |
| 19 | WGT                          | 0.000161476         | 8.72468E-05           | 1.850795374   | 0.064956881    |

Regressing *WGT* on *ENG* and *CYL*

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.933600095         |                       |               |                |
| 5  | R Square                     | 0.871609138         |                       |               |                |
| 6  | Adjusted R Square            | 0.870949031         |                       |               |                |
| 7  | Standard Error               | 305.1365283         |                       |               |                |
| 8  | Observations                 | 392                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 2                   | 245881392.2           | 122940696.1   | 1320.405322    |
| 13 | Residual                     | 389                 | 36219129.05           | 93108.30091   |                |
| 14 | Total                        | 391                 | 282100521.2           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 1372.356942         | 78.84466956           | 17.40583034   | 1.36052E-50    |
| 18 | CYL                          | 54.05697362         | 29.20742853           | 1.850795374   | 0.064956881    |
| 19 | ENG                          | 6.735339693         | 0.476105084           | 14.1467502    | 5.99868E-37    |

Check the  $R^2$  from each of these regressions. Any above 80% indicates strong collinearity which may adversely affect the precision with which you can estimate the parameters of the model containing all of these correlated variables. In our example, the  $R^2$ s are approximately 94%, 90% and 87% respectively which are all well above the 80% threshold. Therefore, it is not surprising that it is difficult to isolate the individual contributions of displacement and number of cylinders to a car's gas mileage.

# CHAPTER 7

## Nonlinear Relationships

### CHAPTER OUTLINE

- |                                                       |                                                  |
|-------------------------------------------------------|--------------------------------------------------|
| 7.1 Nonlinear Relationships                           | 7.3 Applying Dummy Variables                     |
| 7.1.1 Summarize data and estimate regression          | 7.3.1 Interactions between qualitative factors   |
| 7.1.2 Calculating a marginal effect                   | 7.3.2 Adding regional dummy variables            |
| 7.2 Dummy Variables                                   | 7.3.3 Testing the equivalence of two regressions |
| 7.2.1 Creating dummy variables                        | 7.4 Interactions Between Continuous Variables    |
| 7.2.2 Estimating a dummy variable regression          | 7.5 Dummy Variables in Log-linear Models         |
| 7.2.3 Testing the significance of the dummy variables |                                                  |
| 7.2.4 Further calculations                            |                                                  |

### 7.1 NONLINEAR RELATIONSHIPS

The least squares estimation procedure we have been using is based on the assumption that the model is linear in the parameters, though not necessarily linear in the variables. We saw an example of nonlinearity in variable in Chapter 6 in the sales model with diminishing marginal returns to advertising expenditures. To allow for this effect, we included the square of advertising expenditures as another explanatory variable. By transforming the advertisement variable, we captured the diminishing marginal returns without violating the assumptions of the linear regression model.

Models that have parameters that are nonlinear require *nonlinear least squares estimation*. Although Excel is a powerful spreadsheet, it is not designed to be a complete econometric software package, and consequently it does not have the capabilities to estimate models that are nonlinear in the parameters. If you encounter such a problem, use econometric software such as Stata, EVIEWS, Shazam, or SAS.

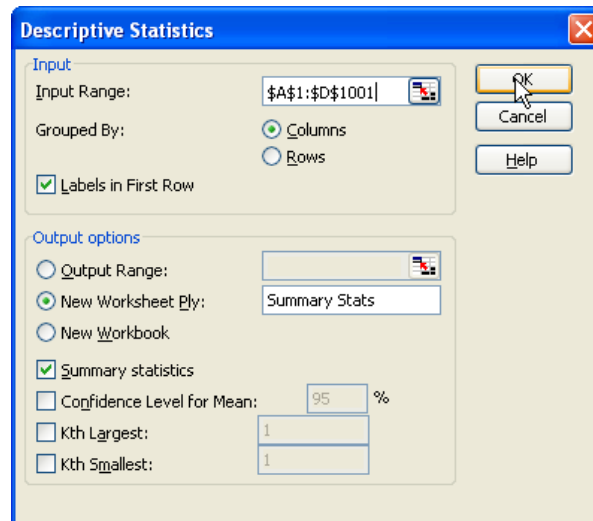
#### **7.1.1 Summarize data and estimate regression**

The following example will illustrate the flexibility that the polynomial terms can add to the linear regression model. We will use the wage equation:



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

Get the summary statistics, **Tools>Data Analysis>Descriptive Statistics**.



This will provide the summary statistics for *WAGE*, *EDUC*, *EXPER* and *EXPER2* variables.

|    | A                  | B           | C                  | D            | E                  | F            | G                  | H           |
|----|--------------------|-------------|--------------------|--------------|--------------------|--------------|--------------------|-------------|
| 1  | wage               |             | educ               |              | exper              |              | exper2             |             |
| 2  |                    |             |                    |              |                    |              |                    |             |
| 3  | Mean               | 10.21302002 | Mean               | 13.285       | Mean               | 18.78        | Mean               | 480.676     |
| 4  | Standard Error     | 0.197536118 | Standard Error     | 0.078050412  | Standard Error     | 0.357979557  | Standard Error     | 15.45983006 |
| 5  | Median             | 8.79        | Median             | 13           | Median             | 18           | Median             | 324         |
| 6  | Mode               | 4.42        | Mode               | 12           | Mode               | 16           | Mode               | 256         |
| 7  | Standard Deviation | 6.246640531 | Standard Deviation | 2.468170753  | Standard Deviation | 11.3188213   | Standard Deviation | 488.8827522 |
| 8  | Sample Variance    | 39.02051792 | Sample Variance    | 6.091866867  | Sample Variance    | 128.1157157  | Sample Variance    | 239006.3454 |
| 9  | Kurtosis           | 7.051484348 | Kurtosis           | 1.538729824  | Kurtosis           | -0.651950158 | Kurtosis           | 1.379091453 |
| 10 | Skewness           | 1.956193012 | Skewness           | -0.211964003 | Skewness           | 0.333047432  | Skewness           | 1.331582256 |
| 11 | Range              | 58.159999   | Range              | 17           | Range              | 52           | Range              | 2704        |
| 12 | Minimum            | 2.03        | Minimum            | 1            | Minimum            | 0            | Minimum            | 0           |
| 13 | Maximum            | 60.189999   | Maximum            | 18           | Maximum            | 52           | Maximum            | 2704        |
| 14 | Sum                | 10213.02002 | Sum                | 13285        | Sum                | 18780        | Sum                | 480676      |
| 15 | Count              | 1000        | Count              | 1000         | Count              | 1000         | Count              | 1000        |

To estimate the wage equation, open *cps\_small.xls*. Highlight column D and insert a column using **Insert>Column**. Name the new column “*EXPER2*” and enter the formula **=C2^2** and copy it down.

|   | A    | B    | C     | D      | E      | F     | G     | H       | I     | J    |
|---|------|------|-------|--------|--------|-------|-------|---------|-------|------|
| 1 | wage | educ | exper | exper2 | female | black | white | midwest | south | west |
| 2 | 2.03 | 13   | 2     | 4      | 1      | 0     | 1     | 0       | 1     | 0    |
| 3 | 2.07 | 12   | 7     | 49     | 0      | 0     | 1     | 1       | 0     | 0    |
| 4 | 2.12 | 12   | 35    | 1225   | 0      | 0     | 1     | 0       | 1     | 0    |
| 5 | 2.54 | 16   | 20    | 400    | 1      | 0     | 1     | 0       | 1     | 0    |
| 6 | 2.68 | 12   | 24    | 576    | 1      | 0     | 1     | 0       | 1     | 0    |

Estimate the model using **Tools>Data Analysis>Regression**.

Estimate a regression and use *WAGE* as the **Y-Range** and *EDUC*, *EXPER*, *EXPER2* as the **X-Range**. Name the output worksheet “**Polynomial**”.

|    | A                            | B                   | C                     | D             | E              | F                     | G                |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|-----------------------|------------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |                       |                  |
| 2  |                              |                     |                       |               |                |                       |                  |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |                       |                  |
| 4  | Multiple R                   | 0.520513263         |                       |               |                |                       |                  |
| 5  | R Square                     | 0.270934057         |                       |               |                |                       |                  |
| 6  | Adjusted R Square            | 0.268738075         |                       |               |                |                       |                  |
| 7  | Standard Error               | 5.341743073         |                       |               |                |                       |                  |
| 8  | Observations                 | 1000                |                       |               |                |                       |                  |
| 9  |                              |                     |                       |               |                |                       |                  |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |                       |                  |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       | <i>Significance F</i> |                  |
| 12 | Regression                   | 3                   | 10561.41522           | 3520.47174    | 123.3771891    | 5.96228E-68           |                  |
| 13 | Residual                     | 996                 | 28420.08218           | 28.53421906   |                |                       |                  |
| 14 | Total                        | 999                 | 38981.4974            |               |                |                       |                  |
| 15 |                              |                     |                       |               |                |                       |                  |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i>      | <i>Upper 95%</i> |
| 17 | Intercept                    | -9.817697044        | 1.054963555           | -9.30619546   | 8.19064E-20    | -11.88790328          | -7.747490804     |
| 18 | educ                         | 1.210071834         | 0.0702378             | 17.22821381   | 2.03903E-58    | 1.072240787           | 1.347902881      |
| 19 | exper                        | 0.340949174         | 0.051431361           | 6.629207719   | 5.52183E-11    | 0.240022916           | 0.441875432      |
| 20 | exper2                       | -0.005093062        | 0.001197941           | -4.251512714  | 2.32293E-05    | -0.00744384           | -0.002742284     |

### 7.1.2 Calculating a marginal effect

Since the wage equation is nonlinear in variable “*EXPER*”, the marginal effect (slope) must be calculated as follows:

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

The marginal effect needs to be evaluated at a specific point, such as the median. You can get the median from the summary statistics. The median for the *EXPER* variable is 18 from the output above. The marginal effect at the median is:

$$\frac{\partial E(WAGE)}{\partial EXPER} = 0.340949174 - 0.005093062 * 2 * 18 = 0.157598937$$

## 7.2 DUMMY VARIABLES

Dummy variables are binary (or indicator) variables that indicate the presence or absence of a characteristic. In this section, we will use dummy variables in a real estate example. Open *utown.xls*.

### 7.2.1 Creating dummy variables

In many examples in *POE* dummy variables have already been created and are ready to use. An important issue in the real estate industry is how to accurately predict the price of a house, based on several of its characteristics, including the ever-important "location, location, location". Economists commonly use a "hedonic" model of pricing based on several characteristics such as size, location, number of bedrooms, age, etc. Using a dummy variable,  $D_i$ , which is equal to 1 if the house is in a desirable neighborhood and is equal to 0 if the house is not in a desirable neighborhood captures the qualitative factor of location. Including this variable in the regression model will allow the intercept to be different for houses in desirable areas compared to the intercept for houses not in desirable areas. This variable has been stored as variable *UTOWN*.

We can also allow for different slopes for houses in different areas by including an interaction variable, the product of the dummy variable and one of the continuous explanatory variables.

### 7.2.2 Estimating a dummy variable regression

Estimating a dummy variables model is no different than estimating any other regression model. We will use the interaction slope dummy variable between the size of the house *sqft* and the dummy variable for university town, *utown*. This will allow for the extra square footage of living space in a good neighborhood affecting the price differently than a house not in a good neighborhood. The full model we will estimate is

$$\begin{aligned} PRICE = & \beta_1 + \delta_1 UTOWN + \beta_2 SQFT + \gamma(SQFT \times UTOWN) \\ & + \beta_3 AGE + \delta_2 POOL + \delta_3 FPLACE + e \end{aligned}$$

where  $PRICE$  = the price of the house, in dollars  
 $UTOWN$  = 1 for houses near the university (desirable), 0 otherwise  
 $SQFT$  = square feet of living area  
 $AGE$  = age of house in years  
 $POOL$  = 1 if house has a pool, 0 otherwise  
 $FPLACE$  = 1 if house has a fireplace, 0 otherwise.

Note that this model contains two continuous explanatory variables ( $SQFT$  and  $AGE$ ), and three dummy variables, capturing the qualitative characteristics of location; presence of a pool, and a

fireplace in addition to the utown variable. Let's now estimate this model. Open *utown.xls*. Label column G *sqftXutown*. In cell G2 create the interaction variable by typing **=B2\*D2**.

|   | A       | B     | C   | D     | E    | F      | G          |
|---|---------|-------|-----|-------|------|--------|------------|
|   | price   | sqft  | age | utown | pool | fplace | sqftXutown |
| 1 |         |       |     |       |      |        |            |
| 2 | 205.452 | 23.46 | 6   | 0     | 0    | 1      | =B2*D2     |
| 3 | 185.328 | 20.03 | 5   | 0     | 0    | 1      |            |
| 4 | 248.422 | 27.77 | 6   | 0     | 0    | 0      |            |
| 5 | 154.69  | 20.17 | 1   | 0     | 0    | 0      |            |

Copy this formula down the column to row 1001.

|   | D     | E    | F      | G          |
|---|-------|------|--------|------------|
|   | Utown | Pool | Fplace | sqftXUtown |
| 6 | 0     | 0    | 1      | 0          |
| 5 | 0     | 0    | 1      |            |
| 6 | 0     | 0    | 0      |            |
| 1 | 0     | 0    | 0      |            |
| 0 | 0     | 0    | 1      |            |

Zeros appear in column G down to row 482. Then sqft values appear after that. This is because the variable *UTOWN* is equal to zero through row 482, then one after that.

Estimate a regression using **Tools>Data Analysis>Regression**, using column A as the **Y-Range** and columns B through G as the **X-Range**. Don't forget the include labels by checking the **Labels** box. Save to a new worksheet called **Dummy and Interaction**.

**Regression**

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

The parameter estimates and corresponding *t*-stats and *p*-values are

|    |            |                     |                       |               |                |
|----|------------|---------------------|-----------------------|---------------|----------------|
| 16 |            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept  | 24.49998329         | 6.191721216           | 3.956893801   | 8.13332E-05    |
| 18 | sqft       | 7.612176612         | 0.245176458           | 31.04774691   | 1.8674E-148    |
| 19 | age        | -0.190086388        | 0.051204606           | -3.712290812  | 0.000216812    |
| 20 | utown      | 27.45295601         | 8.42258204            | 3.259446555   | 0.001154208    |
| 21 | pool       | 4.377164078         | 1.196691609           | 3.65772104    | 0.000267836    |
| 22 | fplace     | 1.64917557          | 0.971956791           | 1.696758113   | 0.090055792    |
| 23 | sqftXutown | 1.29940476          | 0.332047741           | 3.913307036   | 9.72454E-05    |

### 7.2.3 Testing the significance of the dummy variables

To test the significance of the University Town location, we test the individual null hypothesis using the  $t$ -test. From the output above, we conclude that all of the parameters are significant, using a one-tailed 5% level test. (Remember that Excel reports  $p$ -values for a two-tailed test). To conclude that the  $\delta_1$  is statistically different from zero means that there is a shift in the intercept for houses near the university. Similarly, concluding that  $\gamma$  is different from zero means that the marginal effect of the size of the house is different for houses near the university. But Excel doesn't know that we have allowed the intercept and the coefficient on  $SQFT$  to differ across observations. It is our responsibility to correctly determine the intercept and slope estimates. Looking at the original model that we estimated

Alternatively, we can test the significance of the location by testing the joint null hypothesis  $H_0: \delta_1 = 0, \gamma_1 = 0$  against the alternative that at least one coefficient is not zero. To construct the  $F$ -test, we will run the following restricted model:

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \delta_2 POOL + \delta_3 FPLACE + e$$

And compare the  $SSE$ s from the restricted model and the unrestricted model (labeled **Dummy and Interaction**) using the  $F$ -test template. Since we are testing two hypothesis,  $J = 2$  with the other input data of  $N = 1,000$  and  $K = 7$ .

### 7.2.4 Further calculations

When the  $UTOWN$  is equal to one, the intercept is  $\beta_1 + \delta_1$  and when  $UTOWN$  is zero the constant is simply  $\beta_1$ . Similarly, the coefficient on  $SQFT$  when  $UTOWN$  is one is equal to  $\beta_2 + \gamma$ , and is equal to  $\beta_2$  when  $UTOWN$  is zero. We can calculate the estimates for these parameters on the Excel regression output worksheet.

In cell A26, type the label *NewIntercept* and in cell A27, type the label *New Beta2*. In cell B26, type **=B17+B20** to calculate the intercept when  $UTOWN$  is equal to one and type **=B18+B23** to calculate the coefficient estimate for  $SQFT$  when  $UTOWN$  is equal to one in cell B27.

|    |               |                     |
|----|---------------|---------------------|
| 16 |               | <i>Coefficients</i> |
| 17 | Intercept     | 24.49998329         |
| 18 | sqft          | 7.612176612         |
| 19 | age           | -0.190086388        |
| 20 | utown         | 27.45295601         |
| 21 | pool          | 4.377164078         |
| 22 | fplace        | 1.64917557          |
| 23 | sqftXutown    | 1.29940476          |
| 24 |               |                     |
| 25 |               |                     |
| 26 | New Intercept | =+B17+B20           |
| 27 | New Beta2     | =B18+B23            |

|    |               |                     |
|----|---------------|---------------------|
| 16 |               | <i>Coefficients</i> |
| 17 | Intercept     | 24.49998329         |
| 18 | sqft          | 7.612176612         |
| 19 | age           | -0.190086388        |
| 20 | utown         | 27.45295601         |
| 21 | pool          | 4.377164078         |
| 22 | fplace        | 1.64917557          |
| 23 | sqftXutown    | 1.29940476          |
| 24 |               |                     |
| 25 |               |                     |
| 26 | New Intercept | 51.9529393          |
| 27 | New Beta2     | 8.911581373         |

The estimated regression functions for the houses near the university is

$$\hat{PRICE} = 24.5 + 27.453_1(1) + 7.6122SQFT(1) + 1.2994SQFT + \\ -0.1901AGE + 4.3772POOL + 1.6492FPLACE$$

$$\hat{PRICE} = 51.953 + 8.9116SQFT + -0.1901AGE + 4.3772POOL + 1.6492FPLACE$$

### 7.3 APPLYING DUMMY VARIABLES

In this section we will illustrate a variety of applications of dummy variables using *cps\_small.xls*.

#### 7.3.1 Interactions between qualitative factors

First we will consider the interaction between two dummy variables, black and female in the following model:

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

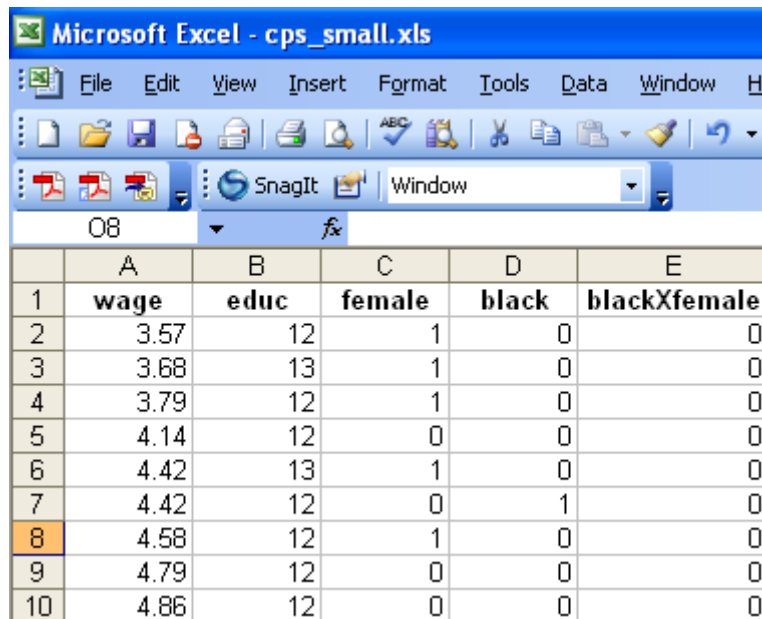
Open *cps\_small.xls*. Label column K *blackXfemale*. In cell K2 create the interaction variable by typing **=F2\*E2**.

|   | A    | B    | C     | D      | E      | F     | G     | H       | I     | J    | K            |
|---|------|------|-------|--------|--------|-------|-------|---------|-------|------|--------------|
| 1 | wage | educ | exper | exper2 | female | black | white | midwest | south | west | blackXfemale |
| 2 | 3.57 | 12   | 0     | 0      | 1      | 0     | 1     | 1       | 0     | 0    | =+F2*E2      |
| 3 | 3.68 | 13   | 0     | 0      | 1      | 0     | 1     | 0       | 0     | 0    |              |
| 4 | 3.79 | 12   | 0     | 0      | 1      | 0     | 1     | 0       | 0     | 1    |              |

Copy this formula down the column to row 1001.

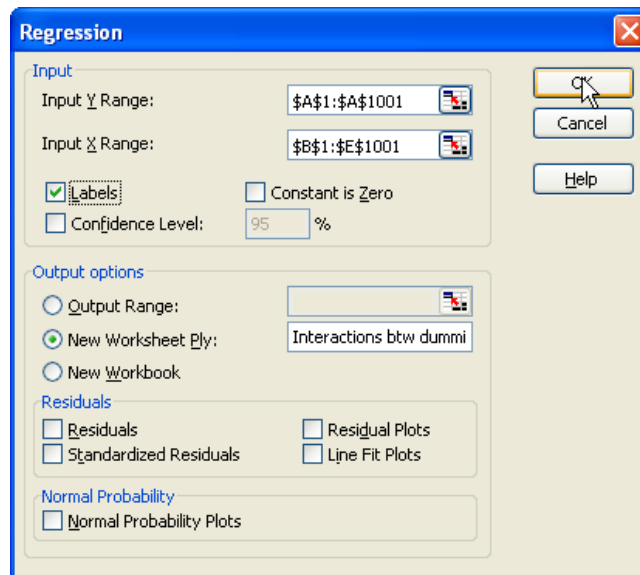
|    | A     | B    | C     | D      | E      | F     | G     | H       | I     | J    | K            |
|----|-------|------|-------|--------|--------|-------|-------|---------|-------|------|--------------|
| 1  | wage  | educ | exper | exper2 | female | black | white | midwest | south | west | blackXfemale |
| 2  | 3.57  | 12   | 0     | 0      | 1      | 0     | 1     | 1       | 0     | 0    | 0            |
| 3  | 3.68  | 13   | 0     | 0      | 1      | 0     | 1     | 0       | 0     | 0    | 0            |
| 4  | 3.79  | 12   | 0     | 0      | 1      | 0     | 1     | 0       | 0     | 1    | 0            |
| 5  | 4.14  | 12   | 0     | 0      | 0      | 0     | 1     | 0       | 0     | 1    | 0            |
| 6  | 4.42  | 13   | 0     | 0      | 1      | 0     | 1     | 1       | 0     | 0    | 0            |
| 7  | 4.42  | 12   | 0     | 0      | 0      | 1     | 0     | 0       | 0     | 0    | 0            |
| 8  | 4.58  | 12   | 0     | 0      | 1      | 0     | 1     | 1       | 0     | 0    | 0            |
| 9  | 4.79  | 12   | 0     | 0      | 0      | 0     | 1     | 0       | 1     | 0    | 0            |
| 10 | 4.86  | 12   | 0     | 0      | 0      | 0     | 1     | 1       | 0     | 0    | 0            |
| 11 | 5.16  | 13   | 0     | 0      | 1      | 0     | 1     | 0       | 0     | 1    | 0            |
| 12 | 5.16  | 12   | 0     | 0      | 0      | 0     | 1     | 0       | 0     | 1    | 0            |
| 13 | 5.16  | 16   | 0     | 0      | 1      | 0     | 1     | 0       | 0     | 0    | 0            |
| 14 | 5.53  | 12   | 0     | 0      | 0      | 0     | 1     | 0       | 1     | 0    | 0            |
| 15 | 8.49  | 16   | 0     | 0      | 1      | 0     | 1     | 0       | 0     | 0    | 0            |
| 16 | 12.89 | 16   | 0     | 0      | 0      | 0     | 1     | 0       | 1     | 0    | 0            |
| 17 | 3.16  | 13   | 1     | 1      | 0      | 0     | 1     | 0       | 0     | 1    | 0            |
| 18 | 3.68  | 12   | 1     | 1      | 1      | 1     | 0     | 0       | 0     | 0    | 1            |
| 19 | 4.04  | 11   | 1     | 1      | 1      | 1     | 0     | 0       | 0     | 0    | 1            |

Recall that you need the **X-Range** variables next to each other. Insert columns and organize the columns by copying and pasting.



|    | A    | B    | C      | D     | E            |
|----|------|------|--------|-------|--------------|
| 1  | wage | educ | female | black | blackXfemale |
| 2  | 3.57 | 12   | 1      | 0     | 0            |
| 3  | 3.68 | 13   | 1      | 0     | 0            |
| 4  | 3.79 | 12   | 1      | 0     | 0            |
| 5  | 4.14 | 12   | 0      | 0     | 0            |
| 6  | 4.42 | 13   | 1      | 0     | 0            |
| 7  | 4.42 | 12   | 0      | 1     | 0            |
| 8  | 4.58 | 12   | 1      | 0     | 0            |
| 9  | 4.79 | 12   | 0      | 0     | 0            |
| 10 | 4.86 | 12   | 0      | 0     | 0            |

Estimate a regression using **Tools>Data Analysis>Regression**, using column A as the **Y-Range** and columns B through E as the **X-Range**. Don't forget the include labels by checking the **Labels** box. Save to a new worksheet called "Interactions btw dummies".



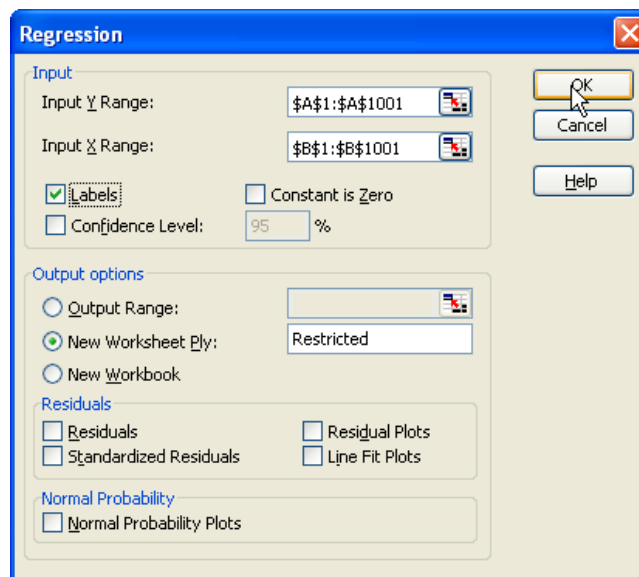
The image shows the 'Regression' dialog box in Microsoft Excel. The 'Input' section has 'Input Y Range' set to '\$A\$1:\$A\$1001' and 'Input X Range' set to '\$B\$1:\$E\$1001'. The 'Labels' checkbox is checked. The 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected, with the name 'Interactions btw dummi'. The 'Residuals' section has 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' all unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. Buttons for 'OK', 'Cancel', and 'Help' are on the right.

The output will be:

|    | A                            | B                   | C                     | D             | E              | F                     | G                |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|-----------------------|------------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |                       |                  |
| 2  |                              |                     |                       |               |                |                       |                  |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |                       |                  |
| 4  | Multiple R                   | 0.498160217         |                       |               |                |                       |                  |
| 5  | R Square                     | 0.248163601         |                       |               |                |                       |                  |
| 6  | Adjusted R Square            | 0.245141144         |                       |               |                |                       |                  |
| 7  | Standard Error               | 5.427244562         |                       |               |                |                       |                  |
| 8  | Observations                 | 1000                |                       |               |                |                       |                  |
| 9  |                              |                     |                       |               |                |                       |                  |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |                       |                  |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       | <i>Significance F</i> |                  |
| 12 | Regression                   | 4                   | 9673.788784           | 2418.447196   | 82.106554      | 2.92685E-60           |                  |
| 13 | Residual                     | 995                 | 29307.70862           | 29.45498353   |                |                       |                  |
| 14 | Total                        | 999                 | 38981.4974            |               |                |                       |                  |
| 15 |                              |                     |                       |               |                |                       |                  |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i>      | <i>Upper 95%</i> |
| 17 | Intercept                    | -3.230327242        | 0.967499411           | -3.33884156   | 0.000872491    | -5.128900665          | -1.38175382      |
| 18 | educ                         | 1.116823459         | 0.069714406           | 16.01998091   | 1.46665E-51    | 0.980019325           | 1.253627593      |
| 19 | female                       | -2.552070371        | 0.359685641           | -7.095280104  | 2.44952E-12    | -3.257899842          | -1.846240899     |
| 20 | black                        | -1.831239486        | 0.895726469           | -2.044418189  | 0.041175021    | -3.588969205          | -0.073509766     |
| 21 | blackXfemale                 | 0.587905222         | 1.216953881           | 0.483095729   | 0.629134045    | -1.800185425          | 2.975995869      |

To test the joint hypothesis  $H_0 : \delta_1 = 0, \delta_2 = 0, \gamma_1 = 0$ , we will need to estimate the restricted model and carry out an  $F$ -Test. Estimate the **Restricted** model, assuming the null hypothesis is correct.





The image shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' set to '\$A\$1:\$A\$1001' and 'Input X Range' set to '\$B\$1:\$B\$1001'. The 'Labels' checkbox is checked, and 'Constant is Zero' is unchecked. The 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected with the name 'Restricted'. The 'Residuals' section has 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' all unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. The 'OK', 'Cancel', and 'Help' buttons are on the right.

Estimate a regression using **Tools>Data Analysis>Regression**, using column A as the **Y-Range** and columns B only as the **X-Range**. Save to a new worksheet called “**Restricted**.”

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.449850568         |                       |               |                |
| 5  | R Square                     | 0.202365533         |                       |               |                |
| 6  | Adjusted R Square            | 0.2015663           |                       |               |                |
| 7  | Standard Error               | 5.581692977         |                       |               |                |
| 8  | Observations                 | 1000                |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 1                   | 7888.511508           | 7888.511508   | 253.1996931    |
| 13 | Residual                     | 998                 | 31092.98589           | 31.15529649   |                |
| 14 | Total                        | 999                 | 38981.4974            |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | -4.912180625        | 0.966787511           | -5.080930988  | 4.48182E-07    |
| 18 | educ                         | 1.138517173         | 0.07154973            | 15.91224978   | 5.59313E-51    |

SSE(Restricted)

We can now open the *F*-test template and fill in the **Input** data.

|    | A                           | B           |
|----|-----------------------------|-------------|
| 1  | Hypothesis Testing - F-Test |             |
| 2  |                             |             |
| 3  | Data Input                  |             |
| 4  | J                           | 3           |
| 5  | N                           | 1000        |
| 6  | K                           | 5           |
| 7  | SSE-RESTRICTED              | 31092.98589 |
| 8  | SSE-UNRESTRICTED            | 29307.70862 |
| 9  | ALPHA                       | 0.05        |
| 10 |                             |             |
| 11 | Computed Values             |             |
| 12 | df-numerator                | 3           |
| 13 | df-denominator              | 995         |
| 14 | F                           | 20.2034547  |
| 15 | Right Critical value        | 2.613848392 |
| 16 | Decision                    | Reject Null |
| 17 | p-value                     | 1.02707E-12 |

### 7.3.2 Adding regional dummy variables

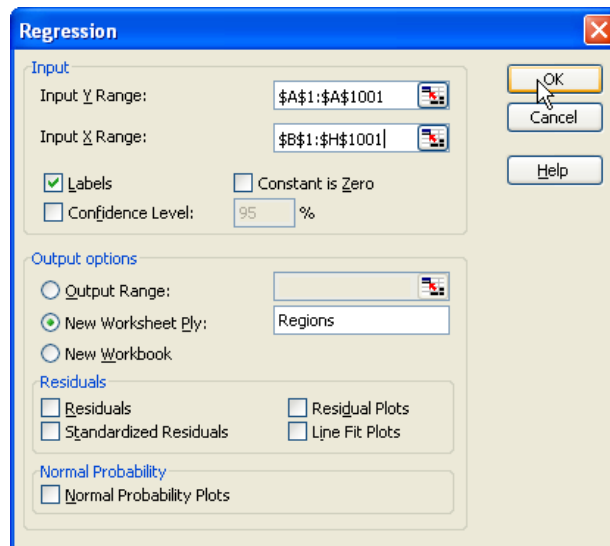
Next, we can add dummy variables with several categories such as the regional dummies. Our model becomes:

$$\begin{aligned}
 WAGE = & \beta_1 + \beta_2 EDUC + \delta_1 SOUTH + \delta_2 MIDWEST + \delta_3 WEST \\
 & + \delta_4 BLACK + \delta_5 FEMALE + \gamma_1 (BLACK \times FEMALE) + e
 \end{aligned}$$

Since the dummies are already present in the file, we simply estimate the model using **Tools>Data Analysis>Regression**, after arranging the columns for the **X-Range** variables.

|    | A    | B    | C       | D     | E    | F      | G     | H            | I     | J      | K     |
|----|------|------|---------|-------|------|--------|-------|--------------|-------|--------|-------|
| 1  | wage | educ | midwest | south | west | female | black | blackXfemale | exper | exper2 | white |
| 2  | 3.57 | 12   | 1       | 0     | 0    | 1      | 0     | 0            | 0     | 0      | 1     |
| 3  | 3.68 | 13   | 0       | 0     | 0    | 1      | 0     | 0            | 0     | 0      | 1     |
| 4  | 3.79 | 12   | 0       | 0     | 1    | 1      | 0     | 0            | 0     | 0      | 1     |
| 5  | 4.14 | 12   | 0       | 0     | 1    | 0      | 0     | 0            | 0     | 0      | 1     |
| 6  | 4.42 | 13   | 1       | 0     | 0    | 1      | 0     | 0            | 0     | 0      | 1     |
| 7  | 4.42 | 12   | 0       | 0     | 0    | 0      | 1     | 0            | 0     | 0      | 0     |
| 8  | 4.58 | 12   | 1       | 0     | 0    | 1      | 0     | 0            | 0     | 0      | 1     |
| 9  | 4.79 | 12   | 0       | 1     | 0    | 0      | 0     | 0            | 0     | 0      | 1     |
| 10 | 4.86 | 12   | 1       | 0     | 0    | 0      | 0     | 0            | 0     | 0      | 1     |

Use column A as the **Y-Range** and columns B through H as the **X-Range**. Save to a new worksheet called “**Regions**.”



The output will provide us with the unrestricted model to test the joint hypothesis of  $H_0 : \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ .

| Microsoft Excel - cps_small.xls |                       |              |                |              |
|---------------------------------|-----------------------|--------------|----------------|--------------|
| Q19                             |                       |              |                |              |
|                                 | A                     | B            | C              | D            |
| 1                               | SUMMARY OUTPUT        |              |                |              |
| 2                               |                       |              |                |              |
| 3                               | Regression Statistics |              |                |              |
| 4                               | Multiple R            | 0.503446397  |                |              |
| 5                               | R Square              | 0.253458275  |                |              |
| 6                               | Adjusted R Square     | 0.248190339  |                |              |
| 7                               | Standard Error        | 5.416271996  |                |              |
| 8                               | Observations          | 1000         |                |              |
| 9                               |                       |              |                |              |
| 10                              | ANOVA                 |              |                |              |
| 11                              |                       | df           | SS             | MS           |
| 12                              | Regression            | 7            | 9880.18308     | 1411.454726  |
| 13                              | Residual              | 992          | 29101.31432    | 29.33600234  |
| 14                              | Total                 | 999          | 38981.4974     |              |
| 15                              |                       |              |                |              |
| 16                              |                       | Coefficients | Standard Error | t Stat       |
| 17                              | Intercept             | -2.455685517 | 1.050990434    | -2.336544117 |
| 18                              | educ                  | 1.102461555  | 0.069986191    | 15.75255832  |
| 19                              | midwest               | -0.499562279 | 0.505628233    | -0.988003134 |
| 20                              | south                 | -1.24428066  | 0.479427352    | -2.595347668 |
| 21                              | west                  | -0.546183401 | 0.515397714    | -1.059731906 |
| 22                              | female                | -2.500920325 | 0.35997464     | -6.947490311 |
| 23                              | black                 | -1.607663765 | 0.90343217     | -1.779506884 |
| 24                              | blackXfemale          | 0.6464628    | 1.215207538    | 0.53197728   |

Under the null hypothesis, the restricted model is the full (unrestricted) model from the previous section. If we fill in the **Data Inputs** in the *F*-Test template for this test, we fail to reject the null hypothesis at 5% level.

|    | A                           | B                   |
|----|-----------------------------|---------------------|
| 1  | Hypothesis Testing - F-Test |                     |
| 2  |                             |                     |
| 3  | Data Input                  |                     |
| 4  | J                           | 3                   |
| 5  | N                           | 1000                |
| 6  | K                           | 8                   |
| 7  | SSE-RESTRICTED              | 29307.70862         |
| 8  | SSE-UNRESTRICTED            | 29101.31432         |
| 9  | ALPHA                       | 0.05                |
| 10 |                             |                     |
| 11 | Computed Values             |                     |
| 12 | df-numerator                | 3                   |
| 13 | df-denominator              | 992                 |
| 14 | F                           | 2.345176344         |
| 15 | Right Critical value        | 2.613875482         |
| 16 | Decision                    | Fail to Reject Null |
| 17 | p-value                     | 0.071445679         |

### 7.3.3 Testing the equivalence of two regressions

To test the equivalence of the wage equations for the south region versus the reminder of the country, we create an interaction variable for each variable in the regression model with the dummy variable **south**. In other words, our equation

$$WAGE = \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma_1 (BLACK \times FEMALE) + e$$

becomes

$$\begin{aligned}
 WAGE = & \beta_1 + \beta_2 EDUC + \delta_1 BLACK + \delta_2 FEMALE + \gamma_1 (BLACK \times FEMALE) + \theta_1 SOUTH \\
 & + \theta_2 (EDUC \times SOUTH) + \theta_3 (BLACK \times SOUTH) + \theta_4 (FEMALE \times SOUTH) \\
 & + \theta_5 (BLACK \times FEMALE \times SOUTH) + e
 \end{aligned}$$

First we have to create the interaction variables as explained in the previous section.

|   | A    | B    | C     | D      | E            | F     | G          | H           | I            | J                  |
|---|------|------|-------|--------|--------------|-------|------------|-------------|--------------|--------------------|
|   | wage | educ | black | female | blackXfemale | south | educXsouth | blackXsouth | femaleXsouth | blackXfemaleXsouth |
| 1 |      |      |       |        |              |       |            |             |              |                    |
| 2 | 3.57 | 12   | 0     | 1      | 0            | 0     | =B2*F2     | =C2*F2      | =+D2*F2      | =E2*F2             |
| 3 | 3.68 | 13   | 0     | 1      | 0            | 0     |            |             |              |                    |
| 4 | 3.79 | 12   | 0     | 1      | 0            | 0     |            |             |              |                    |
| 5 | 4.14 | 12   | 0     | 0      | 0            | 0     |            |             |              |                    |
| 6 | 4.42 | 13   | 0     | 1      | 0            | 0     |            |             |              |                    |
| 7 | 4.42 | 12   | 1     | 0      | 0            | 0     |            |             |              |                    |

After labeling and writing the formulas, copy the formulas for all 1,000 observations.

|   | A    | B    | C     | D      | E            | F     | G          | H           | I            | J                  |
|---|------|------|-------|--------|--------------|-------|------------|-------------|--------------|--------------------|
|   | wage | educ | black | female | blackXfemale | south | educXsouth | blackXsouth | femaleXsouth | blackXfemaleXsouth |
| 1 |      |      |       |        |              |       |            |             |              |                    |
| 2 | 3.57 | 12   | 0     | 1      | 0            | 0     | 0          | 0           | 0            | 0                  |
| 3 | 3.68 | 13   | 0     | 1      | 0            | 0     |            |             |              |                    |
| 4 | 3.79 | 12   | 0     | 1      | 0            | 0     |            |             |              |                    |
| 5 | 4.14 | 12   | 0     | 0      | 0            | 0     |            |             |              |                    |
| 6 | 4.42 | 13   | 0     | 1      | 0            | 0     |            |             |              |                    |
| 7 | 4.42 | 12   | 1     | 0      | 0            | 0     |            |             |              |                    |
| 8 | 4.58 | 12   | 0     | 1      | 0            | 0     |            |             |              |                    |

Once the interaction variables are created, we estimate the model by **Tools>Data Analysis>Regression**, using column A as the **Y-Range** and columns B through J as the **X-Range**. Save to a new worksheet called “FULL.”

**Regression**

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

The results will give us the unrestricted model results.

|    | A                     | B            | C              | D            | E           |
|----|-----------------------|--------------|----------------|--------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |
| 2  |                       |              |                |              |             |
| 3  | Regression Statistics |              |                |              |             |
| 4  | Multiple R            | 0.505698656  |                |              |             |
| 5  | R Square              | 0.255731131  |                |              |             |
| 6  | Adjusted R Square     | 0.24896505   |                |              |             |
| 7  | Standard Error        | 5.413480648  |                |              |             |
| 8  | Observations          | 1000         |                |              |             |
| 9  |                       |              |                |              |             |
| 10 | ANOVA                 |              |                |              |             |
| 11 |                       | df           | SS             | MS           | F           |
| 12 | Regression            | 9            | 9968.78134     | 1107.642489  | 37.7960513  |
| 13 | Residual              | 990          | 29012.715      | 29.30577272  |             |
| 14 | Total                 | 999          | 38981.4974     |              |             |
| 15 |                       |              |                |              |             |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     |
| 17 | Intercept             | -3.577535917 | 1.1513322      | -3.107301191 | 0.001941864 |
| 18 | educ                  | 1.165847216  | 0.082408377    | 14.1471931   | 1.61003E-41 |
| 19 | black                 | -0.431164992 | 1.348248998    | -0.319796264 | 0.749190255 |
| 20 | female                | -2.754044353 | 0.425705319    | -6.46936796  | 1.54458E-10 |
| 21 | blackXfemale          | 0.067319743  | 1.906317869    | 0.035314018  | 0.971836465 |
| 22 | south                 | 1.302260038  | 2.114734976    | 0.615802951  | 0.538166125 |
| 23 | educXsouth            | -0.191725337 | 0.154239615    | -1.243035629 | 0.214149011 |
| 24 | blackXsouth           | -1.744431965 | 1.826694979    | -0.954966201 | 0.339827908 |
| 25 | femaleXsouth          | 0.91193858   | 0.795976102    | 1.145685878  | 0.252202013 |
| 26 | blackXfemaleXsouth    | 0.542832938  | 2.511153729    | 0.21616874   | 0.828900751 |

To test the hypothesis that there is no difference between the model for the south and the rest of the nation, we have to test the joint hypothesis  $H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0$  against the alternative that at least one of the five hypotheses is not true. Under the null hypothesis, the restricted model will be

|    | A                     | B            | C              | D            | E           |
|----|-----------------------|--------------|----------------|--------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |
| 2  |                       |              |                |              |             |
| 3  | Regression Statistics |              |                |              |             |
| 4  | Multiple R            | 0.498160217  |                |              |             |
| 5  | R Square              | 0.248163601  |                |              |             |
| 6  | Adjusted R Square     | 0.245141144  |                |              |             |
| 7  | Standard Error        | 5.427244562  |                |              |             |
| 8  | Observations          | 1000         |                |              |             |
| 9  |                       |              |                |              |             |
| 10 | ANOVA                 |              |                |              |             |
| 11 |                       | df           | SS             | MS           | F           |
| 12 | Regression            | 4            | 9673.78878     | 2418.447196  | 82.106554   |
| 13 | Residual              | 995          | 29307.70862    | 29.45498353  |             |
| 14 | Total                 | 999          | 38981.4974     |              |             |
| 15 |                       |              |                |              |             |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     |
| 17 | Intercept             | -3.230327242 | 0.967499411    | -3.33884156  | 0.000872491 |
| 18 | educ                  | 1.116823459  | 0.069714406    | 16.01998091  | 1.46665E-51 |
| 19 | black                 | -1.831239486 | 0.895726469    | -2.044418189 | 0.041175021 |
| 20 | female                | -2.552070371 | 0.359685641    | -7.095280104 | 2.44952E-12 |
| 21 | blackXfemale          | 0.587905222  | 1.216953881    | 0.483095729  | 0.629134045 |

If we input the **Data** into the *F*-Test template, we will get:

|    | A                                  | B                   |
|----|------------------------------------|---------------------|
| 1  | <b>Hypothesis Testing - F-Test</b> |                     |
| 2  |                                    |                     |
| 3  | <b>Data Input</b>                  |                     |
| 4  | J                                  | 5                   |
| 5  | N                                  | 1000                |
| 6  | K                                  | 10                  |
| 7  | SSE-RESTRICTED                     | 29307.70862         |
| 8  | SSE-UNRESTRICTED                   | 29012.715           |
| 9  | ALPHA                              | 0.05                |
| 10 |                                    |                     |
| 11 | <b>Computed Values</b>             |                     |
| 12 | df-numerator                       | 5                   |
| 13 | df-denominator                     | 990                 |
| 14 | F                                  | 2.013211682         |
| 15 | Right Critical value               | 2.223143069         |
| 16 | Decision                           | Fail to Reject Null |
| 17 | p-value                            | 0.074378974         |

## 7.4 INTERACTIONS BETWEEN CONTINUOUS VARIABLES

When we include the product of two continuous explanatory variables in a model, we alter the relationship between each of them and the dependent variable. Reporting and interpreting the results require care.

The model we use here is based on a "life-cycle" model of the effects of age and income on a person's expenditures on pizza. We believe that as a person ages, the marginal effect of income will probably change (the marginal propensity to spend on pizza probably falls). Since we assume that the effect of income depends on age, we include an interaction variable that is the product of these two variables. The model we will estimate is

$$PIZZA = \beta_1 + \beta_2 AGE + \beta_3 INCOME + \beta_4 (INCOME \times AGE) + e$$

where *PIZZA* = individual's annual expenditure on pizza, in dollars

*AGE* = the age of the individual in years

*Y* = the annual income of the individual, in dollars

Open *pizza.xls* to estimate the above model and create the interaction variable after arranging the order of the explanatory variables.

Microsoft Excel - pizza.xls

File Edit View Insert Format Tools Data

SnagIt Window

NORMINV  $\times$   $\checkmark$   $\wedge$  =B2\*C2

|   | A     | B   | C      | D          |
|---|-------|-----|--------|------------|
| 1 | pizza | age | income | incomeXage |
| 2 | 109   | 25  | 15000  | =B2*C2     |
| 3 | 0     | 45  | 30000  |            |
| 4 | 0     | 20  | 12000  |            |
| 5 | 108   | 28  | 20000  |            |
| 6 | 220   | 25  | 15000  |            |
| 7 | 189   | 35  | 30000  |            |

Copy the formula to all 40 cells.

Microsoft Excel - pizza.xls

File Edit View Insert Format Tools Data

SnagIt Window

G2  $\times$   $\checkmark$  0

|    | A     | B   | C      | D          |
|----|-------|-----|--------|------------|
| 1  | pizza | age | income | incomeXage |
| 2  | 109   | 25  | 15000  | 375000     |
| 3  | 0     | 45  | 30000  | 1350000    |
| 4  | 0     | 20  | 12000  | 240000     |
| 5  | 108   | 28  | 20000  | 560000     |
| 6  | 220   | 25  | 15000  | 375000     |
| 7  | 189   | 35  | 30000  | 1050000    |
| 8  | 64    | 40  | 12000  | 480000     |
| 9  | 262   | 22  | 12000  | 264000     |
| 10 | 64    | 30  | 28000  | 840000     |
| 11 | 35    | 21  | 22000  | 462000     |
| 12 | 94    | 40  | 44000  | 1760000    |

Estimate the model.

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | <i>Regression Statistics</i> |                     |                       |               |                |
| 2  | Multiple R                   | 0.622349295         |                       |               |                |
| 3  | R Square                     | 0.387318645         |                       |               |                |
| 4  | Adjusted R Square            | 0.336261866         |                       |               |                |
| 5  | Standard Error               | 126.996134          |                       |               |                |
| 6  | Observations                 | 40                  |                       |               |                |
| 7  |                              |                     |                       |               |                |
| 8  | <i>ANOVA</i>                 |                     |                       |               |                |
| 9  |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 10 | Regression                   | 3                   | 367043.25             | 122347.75     | 7.586037514    |
| 11 | Residual                     | 36                  | 580608.65             | 16128.01806   |                |
| 12 | Total                        | 39                  | 947651.9              |               |                |
| 13 |                              |                     |                       |               |                |
| 14 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 15 | Intercept                    | 161.465432          | 120.6634096           | 1.338147434   | 0.189239689    |
| 16 | age                          | -2.977423365        | 3.352100814           | -0.88822608   | 0.380315589    |
| 17 | income                       | 0.009073877         | 0.003669598           | 2.472716864   | 0.01826628     |
| 18 | incomeXage                   | -0.000160211        | 8.67343E-05           | -1.847147792  | 0.072957528    |



The effect of age and income are no longer given by their estimated coefficients. Instead we must calculate the marginal effects.

To find the marginal effects of age and income, we take the first derivative of the pizza function, with respect to the variable of interest. We find that the marginal effect of age on pizza expenditures is  $b_2 + b_4Y$  and the effect of income is  $b_3 + b_4AGE$ . These will be estimated and calculated using Excel. In cell A21, type the label *AGE* and in cell A22, type the label *INCOME*. In cell A24, type the label *age effect* and in cell A25, type the label *income effect*. In cell B24, type **=B\$16+(B18\*B22)** and in cell B25, type **=B\$17+(B\$18\*B21)**. This template can now be used to calculate the marginal effects, given different levels of age and/or income. Below is an illustration of two examples.

|    |               |                  |                  |
|----|---------------|------------------|------------------|
| 19 |               |                  |                  |
| 20 | <b>INPUT</b>  | ⊕ EX1            | EX2              |
| 21 | AGE           | 30               | 50               |
| 22 | INCOME        | 90,000           | 25,000           |
| 23 | <b>OUTPUT</b> |                  |                  |
| 24 | age effect    | =B\$16+B\$18*B22 | =B\$16+B\$18*C22 |
| 25 | income effect | =B\$17+B\$18*B21 | =B\$17+B\$18*C21 |

The formulas provide us with the marginal effect of *AGE* for incomes 90,00 and 25,000 and marginal effect of *INCOME* at age 30 and 50.

|    |               |              |              |
|----|---------------|--------------|--------------|
| 19 |               |              |              |
| 20 | <b>INPUT</b>  | EX1          | EX2          |
| 21 | AGE           | 30           | 50           |
| 22 | INCOME        | 90,000       | 25,000       |
| 23 | <b>OUTPUT</b> |              |              |
| 24 | age effect    | -17.39642745 | -6.982702279 |
| 25 | income effect | 0.004267542  | 0.001063319  |
| 26 |               |              |              |

## 7.5 DUMMY VARIABLES IN LOG-LINEAR MODELS

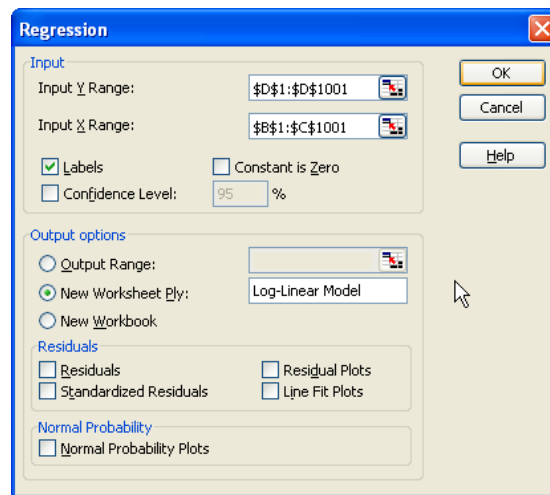
Consider the model:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \gamma_1 FEMALE + e$$

The calculation of the exact effect of a dummy variable in log-linear model is slightly more complicated. We will use *cps\_small.xls* data to illustrate the calculations. Label cell D1, *ln(wage)* and type the formula in D2 **=ln(A2)**. Copy the formula down all the cells.

|   | A    | B    | C      | D        |
|---|------|------|--------|----------|
| 1 | wage | educ | female | ln(wage) |
| 2 | 3.57 | 12   | 1      | 1.272566 |
| 3 | 3.68 | 13   | 1      | 1.302913 |
| 4 | 3.79 | 12   | 1      | 1.332366 |
| 5 | 4.14 | 12   | 0      | 1.420696 |
| 6 | 4.42 | 13   | 1      | 1.48614  |
| 7 | 4.42 | 12   | 0      | 1.48614  |
| 8 | 4.58 | 12   | 1      | 1.521699 |
| 9 | 4.79 | 12   | 0      | 1.56653  |

Estimate the log-linear wage equation by using  $\ln(\text{wage})$  as the **Y-Range** and  $\text{EDUC}$  and  $\text{FEMALE}$  as the **X-Range** variables.



The image shows the 'Regression' dialog box in Microsoft Excel. The 'Input' section has 'Input Y Range' set to '\$D\$1:\$D\$1001' and 'Input X Range' set to '\$B\$1:\$C\$1001'. The 'Labels' checkbox is checked, and 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected, with the output range set to 'Log-Linear Model'. The 'Residuals' section has 'Residuals' and 'Standardized Residuals' unchecked, while 'Residual Plots' and 'Line Fit Plots' are checked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked.

The results are:

| Microsoft Excel - cps_small.xls                               |                       |              |                |              |
|---------------------------------------------------------------|-----------------------|--------------|----------------|--------------|
| File Edit View Insert Format Tools Data Window Help Adobe PDF |                       |              |                |              |
| NORMINV $=100*(EXP(B19)-1)$                                   |                       |              |                |              |
|                                                               | A                     | B            | C              | D            |
| 1                                                             | SUMMARY OUTPUT        |              |                |              |
| 2                                                             |                       |              |                |              |
| 3                                                             | Regression Statistics |              |                |              |
| 4                                                             | Multiple R            | 0.516562833  |                |              |
| 5                                                             | R Square              | 0.266837161  |                |              |
| 6                                                             | Adjusted R Square     | 0.265366423  |                |              |
| 7                                                             | Standard Error        | 0.473814303  |                |              |
| 8                                                             | Observations          | 1000         |                |              |
| 9                                                             |                       |              |                |              |
| 10                                                            | ANOVA                 |              |                |              |
| 11                                                            |                       | df           | SS             | MS           |
| 12                                                            | Regression            | 2            | 81.46242936    | 40.73121468  |
| 13                                                            | Residual              | 997          | 223.8264934    | 0.224499993  |
| 14                                                            | Total                 | 999          | 305.2889227    |              |
| 15                                                            |                       |              |                |              |
| 16                                                            |                       | Coefficients | Standard Error | t Stat       |
| 17                                                            | Intercept             | 0.929035785  | 0.08374832     | 11.09318719  |
| 18                                                            | educ                  | 0.102565817  | 0.006075312    | 16.88239575  |
| 19                                                            | female                | -0.252603282 | 0.029976973    | -8.426577262 |

We can calculate the exact effect of the female dummy given the output as:

|    |                                        |              |
|----|----------------------------------------|--------------|
| 17 | Intercept                              | 0.929035785  |
| 18 | educ                                   | 0.102565817  |
| 19 | female                                 | -0.252603282 |
| 20 |                                        |              |
| 21 |                                        |              |
| 22 |                                        |              |
| 23 | <b>marginal effect of female dummy</b> |              |
| 24 | $=100*(EXP(B19)-1)$                    |              |

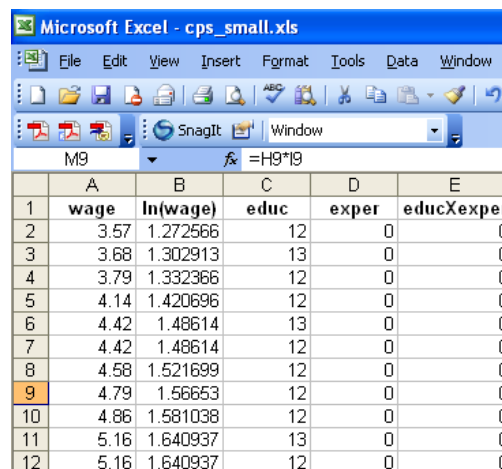
The marginal effect will be a nonlinear function of the parameters:

|    |                                        |
|----|----------------------------------------|
| 22 |                                        |
| 23 | <b>marginal effect of female dummy</b> |
| 24 | -22.32240183                           |
| 25 |                                        |

Similarly, we can calculate other nonlinear marginal effects when we include interaction terms. Estimate the following model:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \gamma_1 (EDUC \times EXPER) + e$$

First, we need to create the interaction variable and estimate this log-linear model using the  $\ln(wage)$  as **Y-Range** and  $EDUC$ ,  $EXPER$  and  $educXexper$  as the **X-Range** variables.



|    | A    | B        | C    | D     | E          |
|----|------|----------|------|-------|------------|
| 1  | wage | ln(wage) | educ | exper | educXexper |
| 2  | 3.57 | 1.272566 | 12   | 0     | 0          |
| 3  | 3.68 | 1.302913 | 13   | 0     | 0          |
| 4  | 3.79 | 1.332366 | 12   | 0     | 0          |
| 5  | 4.14 | 1.420696 | 12   | 0     | 0          |
| 6  | 4.42 | 1.48614  | 13   | 0     | 0          |
| 7  | 4.42 | 1.48614  | 12   | 0     | 0          |
| 8  | 4.58 | 1.521699 | 12   | 0     | 0          |
| 9  | 4.79 | 1.56653  | 12   | 0     | 0          |
| 10 | 4.86 | 1.581038 | 12   | 0     | 0          |
| 11 | 5.16 | 1.640937 | 13   | 0     | 0          |
| 12 | 5.16 | 1.640937 | 12   | 0     | 0          |

Given the coefficient estimates from the regression, we can calculate the approximate marginal effect of experience using

|    |                               |                     |
|----|-------------------------------|---------------------|
| 16 |                               | <i>Coefficients</i> |
| 17 | Intercept                     | 0.152782545         |
| 18 | educ                          | 0.134086013         |
| 19 | exper                         | 0.024916391         |
| 20 | educXexper                    | -0.000962375        |
| 21 |                               |                     |
| 22 |                               |                     |
| 23 | marginal effect of experience |                     |
| 24 | =100*(B19+B20*16)             |                     |
| 25 |                               |                     |

The marginal effect will be again a nonlinear function of the parameters for education of 16 years.

|    |                               |
|----|-------------------------------|
| 22 |                               |
| 23 | marginal effect of experience |
| 24 | 0.951838471                   |
| 25 |                               |

# **CHAPTER 8**

## Heteroskedasticity

### **Chapter Outline**

- |                                             |                                     |
|---------------------------------------------|-------------------------------------|
| 8.1 The Nature of Heteroskedasticity        | 8.4 Detecting Heteroskedasticity    |
| 8.2 Using the Least Squares Estimator       | 8.4.1 Residual plots                |
| 8.3 The Generalized Least Squares Estimator | 8.4.2 The Goldfeld-Quandt test      |
| 8.3.1 Transforming the model                | 8.4.3 Testing the variance function |
| 8.3.2 Estimating the variance function      |                                     |
| 8.3.3 A heteroskedastic partition           |                                     |

### **8.1 THE NATURE OF HETEROSKEDASTICITY**

In simple and multiple linear regression models of earlier chapters, we had assumed all the assumptions of the Classical Linear Regression (CLRM) model have been met.

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

where  $y_i$  is the dependent variable,  $x_i$  is the  $i^{\text{th}}$  observation on the independent variable,  $\beta_1$  and  $\beta_2$  are the unknown parameters and  $e_i$  is the random error. The error assumptions of CLRM are:

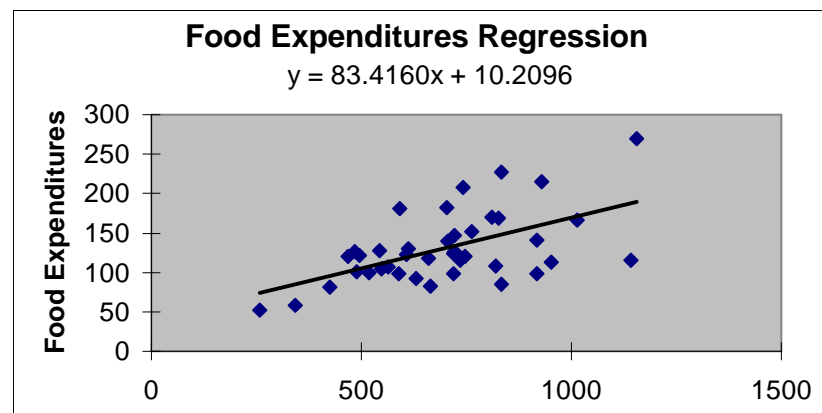
$$E(e_i) = 0 \quad \text{Var}(e_i) = \sigma^2 \quad \text{Cov}(e_i) = 0$$

One of the above mentioned assumptions of the linear regression model is that the variance of the error term and of the dependent variable is constant across all of the observations. If this assumption is not met, OLS estimator is still unbiased and consistent but the least square standard errors and the tests based on these standard errors are neither consistent nor asymptotically valid.

## 8.2 USING THE LEAST SQUARES ESTIMATOR

In Chapter 3, we introduced a model of food expenditures as a function of income. At each income level, a household chooses its level of food expenditures. It seems reasonable that households with higher incomes have more choices, anywhere from cheap and simple to extravagant food. This suggests that the variance for high-income households is greater than that of low-income households. The consequences of estimating the parameters using the least square estimator, if heteroskedasticity is present, are that the standard errors that Excel reports are wrong. Therefore, the  $t$ -statistics,  $p$ -values, and confidence intervals are also wrong. Let's return to that model and our Excel results to reconsider our assumptions about the variance of the model.

Open *food.xls*. Estimate the regression, using food expenditures as the **Y-Range** and income as the **X-Range**. Choose **Line Fit Plots** under the Residuals option to produce a chart of the estimated regression line. After some formatting, the graph should look like



As income increases, the data points are scattered further and further from the estimated regression line. The residuals, the vertical distances between the regression line and the individual observations, are getting larger as income increases. The graph above, therefore, suggests that  $\sigma^2$  is not constant, but is greater for larger income households indicating heteroskedasticity.

|    |           |              |                |             |
|----|-----------|--------------|----------------|-------------|
| 16 |           | Coefficients | Standard Error | t Stat      |
| 17 | Intercept | 83.41600202  | 43.41016314    | 1.921577714 |
| 18 | Income    | 10.20964297  | 2.093263531    | 4.877380614 |

Since, with heteroskedasticity, the reported standard errors are incorrect, we need a method for determining the correct standard errors for our least squares estimators. We can then recalculate our  $t$ -stats and confidence intervals with these heteroskedasticity adjusted standard errors.

One such adjustment is **White's Heteroskedasticity Consistent Standard Errors**, which for the simple regression model is.

$$\hat{\text{var}}(b_2) = \frac{\sum [(x_i - \bar{x})^2 \hat{e}_i^2]}{\left[ \sum (x_i - \bar{x})^2 \right]^2}$$

The square root of this is the estimated standard error for  $b_2$ . Some statistical packages will calculate White's standard errors automatically. Excel does not, but we can use the standard Excel functions to calculate it "by hand".

Return to the data contained in *food.xls*. Estimate another regression, choosing the **Residuals** option.

- From the regression output, **C**opy the residuals from the residual output to the worksheet containing the original data and **P**aste these residuals in column C. Include the label.
- Label column D " $\hat{e}^2$ " and square the values in column C here by typing **=C^2** in cell D2 and copy the formula down the column.
- Label column E " $\bar{x}$ ". In cell E2 type **=AVERAGE(B2:B41)** where column B contains the income.
- Label column F " $(x-\bar{x})^2$ ". In cell F2 type **=(B2-\$E\$2)^2**. Recall that the dollar sign anchors the cell containing " $\bar{x}$ ".
- Label column G "numerator". In cell G2 type **=F2\*D2**. Highlight cells F2 and G2, and copy these formulas down the columns.

|   | A               | B             | C                | D             | E                 | F                 | G                |
|---|-----------------|---------------|------------------|---------------|-------------------|-------------------|------------------|
| 1 | <b>Food_Exp</b> | <b>Income</b> | <b>Residuals</b> | <b>ehat^2</b> | <b>xbar</b>       | <b>(x-xbar)^2</b> | <b>numerator</b> |
| 2 | 115.22          | 3.69          | -5.869584573     | =C2^2         | =+AVERAGE(B2:B41) | =(B2-\$E\$2)^2    | =F2*D2           |
| 3 | 135.98          | 4.39          | 7.743665349      | 59.96435      |                   |                   |                  |

- In cell F42, type **=SUM(F2:F41)** to sum the column;
- in cell G42, type **=SUM(G2:G41)**.
- Label B44, **White's var( $b_2$ )**, and B45 **White's se( $b_2$ )**.
- In cell C44, type **=G42/(F42^2)**
- and in cell C45, type **=SQRT(C44)**.

The results are

|    |                        |       |                    |          |  |                    |                 |
|----|------------------------|-------|--------------------|----------|--|--------------------|-----------------|
| 37 | 482.55                 | 27.14 | 122.0442878        | 14894.81 |  | 56.77999256        | 845727.1        |
| 38 | 438.29                 | 27.16 | 77.58009497        | 6018.671 |  | 57.08180256        | 343556.6        |
| 39 | 587.66                 | 28.62 | 212.0440162        | 44962.66 |  | 81.27473256        | 3654328.6       |
| 40 | 257.95                 | 29.4  | -125.6295053       | 15782.77 |  | 95.94692256        | 1514308.5       |
| 41 | 375.73                 | 33.4  | -48.68807716       | 2370.529 |  | 190.3089226        | 451132.79       |
| 42 |                        |       |                    |          |  | <b>1828.787598</b> | <b>10398342</b> |
| 43 |                        |       |                    |          |  |                    |                 |
| 44 | <b>White's var(b2)</b> |       | <b>3.109120669</b> |          |  |                    |                 |
| 45 | <b>White's se(b2)</b>  |       | <b>1.76326988</b>  |          |  |                    |                 |

Excel's regression output reports the standard error for  $b_2$  as 2.093263531 which is incorrect due to the existence of heteroskedasticity. White's standard error is actually different so are the corrected  $t$ -stat and confidence intervals. While it is usually the case that the corrected standard errors, it is not always true as you can see from this example. You should recalculate and report the corrected  $t$ -stat and confidence interval, do not report those produced in the regression output.

### **8.3 THE GENERALIZED LEAST SQUARES ESTIMATOR**

Since the least squares is inefficient in heteroskedastic models, we may wish to use **Generalized Least Square (GLS)** which is the Best Linear Unbiased Estimator. GLS estimator works by transforming the model into a homoskedastic one and applying OLS to the transformed model.

#### **8.3.1 Transforming the model**

Since we have  $\text{var}(e_i) = \sigma_i^2$ , we can get constant error variance by dividing  $e_i$  by  $\sigma_i$ . To transform the model, we will weigh the observations using  $\sigma_i$ . For the food expenditure model,

$$\frac{y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{x_i}{\sigma_i} + \frac{e_i}{\sigma_i}$$

We saw in the graph of the regression that the estimated error terms seem to get larger as income increases. So we can assume the variance of the model is proportional to income and can be modeled as  $\text{var}(e_i) = \sigma_i^2 x_i$ . It can be shown that, if we transform our original data by dividing all observations by the square root of  $x_i$ , the new, transformed model is homoskedastic and we can estimate the new model using least squares where the  $t$ -stats and confidence intervals are correct.

To estimate the food expenditure model, where we want to weight the data by  $1/\sqrt{x_i}$ . Go back to the worksheet containing the original data on food expenditures and income.

- Label column C  $SQRT(x)$  and columns D, E and F,  $INT^*$ ,  $X^*$  and  $Y^*$  respectively.
- In cell C2, type **=SQRT(B2)**, where cell B2 contains the first observation on income.
- In cell D2, type **=1/C2**. This creates a new intercept term, not equal to one anymore.
- In cell E2, type **=B2/C2**.
- In cell F2, type **=A2/C2**.
- Highlight cells C2 through F2. Copy the formulas down the column.



|   | A        | B      | C           | D           | E        | F        |
|---|----------|--------|-------------|-------------|----------|----------|
| 1 | Food_Exp | Income | SQRT(x)     | INT*        | X*       | Y*       |
| 2 | 115.22   | 3.69   | =SQRT(B2)   | =1/C2       | =+B2/C2  | =+A2/C2  |
| 3 | 135.98   | 4.39   | 2.095232684 | 0.47727396  | 2.095233 | 64.89971 |
| 4 | 119.34   | 4.75   | 2.179449472 | 0.458831468 | 2.179449 | 54.75695 |
| 5 | 114.96   | 6.03   | 2.455605832 | 0.407231481 | 2.455606 | 46.81533 |
| 6 | 187.05   | 12.47  | 3.531288717 | 0.283182736 | 3.531289 | 52.96933 |
| 7 | 243.92   | 12.98  | 3.602776707 | 0.277563691 | 3.602777 | 67.70334 |

Estimate a regression, using  $Y^*$  as the **Y-Range** and  $INT^*$  and  $X^*$  as the **X-Range**.

- Include labels and check the **L**abels box.
- Check the Constant is **Z**ero box since we now have our new, transformed intercept term.
- Place the output on the worksheet named GLS and click **OK**.

**Regression**

Input

Input Y Range: \$F\$1:\$F\$41

Input X Range: \$D\$1:\$E\$41

☒ Labels

☒ Constant is Zero

☐ Confidence Level: 95 %

Output options

☐ Output Range:

☒ New Worksheet Ply: GLS

☐ New Workbook

Residuals

☒ Residuals

☐ Standardized Residuals

☐ Residual Plots

☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Make sure this box is checked

The regression results are:

|    |           |              |                |             |             |             |             |
|----|-----------|--------------|----------------|-------------|-------------|-------------|-------------|
| 16 |           | Coefficients | Standard Error | t Stat      | P-value     | Lower 95%   | Upper 95%   |
| 17 | Intercept | 0            | #N/A           | #N/A        | #N/A        | #N/A        | #N/A        |
| 18 | INT*      | 78.68408018  | 23.78872165    | 3.307621206 | 0.00206413  | 30.52633132 | 126.841829  |
| 19 | X*        | 10.45100906  | 1.385891228    | 7.541002387 | 4.61376E-09 | 7.645418968 | 13.25659915 |

The estimates,  $b_1$  and  $b_2$ , differ from the original regression results. However, the interpretations are the same. Transforming the data in the manner we did changed a heteroskedastic model to a homoskedastic model; not the meanings of our estimates. The GLS standard errors are lower than those calculated using White's approximation. This is to be expected because the GLS procedure is more efficient and provides smaller standard errors, higher  $t$ -stats and narrower confidence intervals.

### 8.3.2 Estimating the variance function

In the above example, the observation's standard error (or what it is proportional to) is known. In most cases this information will not be known and we will have to estimate it. This turns the GLS estimator into **feasible GLS (FGLS)**.

The first step is to choose a model for variance that is some function of the independent variables. A common model of the variance uses the exponential function:

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_{i2} + \dots + \alpha_s z_{is})$$

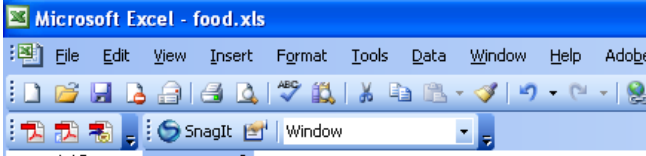
where the  $z_{is}$  are independent variables and  $\alpha$ 's are the unknown parameters. Taking the natural logarithm, substituting the least squares residuals for the unobservable  $\sigma_i^2$ , and adding an error term gives you a regression model that can be estimated for  $\alpha_i$ .

$$\ln(\hat{e}_i^2) = \ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 z_{i2} + v_i$$

where the  $\hat{e}_i^2$  are from least squares estimation of the original heteroskedastic regression model. Let  $z_i = \log(\text{income})$ .

To estimate this model using the food expenditure data, go back to the worksheet containing the original data on food expenditures and income.

- Label column C  $\log(\text{income})$  and type **=LN(B2)** in cell C2.
- Copy and paste the residuals from the initial regression to column D.
- Label columns E and F,  $\text{resid2}$  and  $\log(\text{resid2})$ , respectively.
- In cell E2, type **=D2^2**, where cell D2 contains the first observation on residuals.
- In cell F2, type **=LN(E2)**.
- Highlight cells E2 and F2 and copy the formulas down the column.



|    | A        | B      | C           | D         | E        | F           |
|----|----------|--------|-------------|-----------|----------|-------------|
| 1  | Food Exp | Income | log(income) | Residuals | resid2   | log(resid2) |
| 2  | 115.22   | 3.69   | =LN(B2)     | -5.86958  | =D2^2    | =LN(E2)     |
| 3  | 135.98   | 4.39   | 1.479329227 | 7.743665  | 59.96435 | 4.0937503   |
| 4  | 119.34   | 4.75   | 1.558144618 | -12.5718  | 158.0503 | 5.0629134   |
| 5  | 114.96   | 6.03   | 1.796747011 | -30.0201  | 901.2094 | 6.8037376   |
| 6  | 187.05   | 12.47  | 2.52332576  | -23.6802  | 560.7542 | 6.3292827   |
| 7  | 243.92   | 12.98  | 2.563409711 | 27.98283  | 783.0389 | 6.6631824   |
| 8  | 267.43   | 14.2   | 2.653241965 | 39.03707  | 1523.893 | 7.3290233   |
| 9  | 238.71   | 14.76  | 2.691920819 | 4.599668  | 21.15694 | 3.0519682   |
| 10 | 295.94   | 15.32  | 2.729159164 | 56.11227  | 3148.587 | 8.0547089   |
| 11 | 317.78   | 16.39  | 2.796671393 | 67.02795  | 4492.746 | 8.4102194   |

Estimate a regression, using  $\log(\text{resid2})$  as the **Y-Range** and  $\log(\text{income})$  as the **X-Range**.

Include labels and check the **Labels** box. Place the output on the worksheet named **FGLS** and click OK. The regression results are:

| A1 | fx                           | SUMMARY OUTPUT      |                       |               |                |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
|    | A                            | B                   | C                     | D             | E              |
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.572361183         |                       |               |                |
| 5  | R Square                     | 0.327597324         |                       |               |                |
| 6  | Adjusted R Square            | 0.309902517         |                       |               |                |
| 7  | Standard Error               | 1.720854788         |                       |               |                |
| 8  | Observations                 | 40                  |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 1                   | 54.82554506           | 54.82554506   | 18.51375486    |
| 13 | Residual                     | 38                  | 112.5309656           | 2.961341201   |                |
| 14 | Total                        | 39                  | 167.3565107           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.937795963         | 1.583105624           | 0.592377381   | 0.557106636    |
| 18 | log(income)                  | 2.329238722         | 0.541335797           | 4.302761307   | 0.000113872    |

### 8.3.3 A heteroskedastic partition

To illustrate a model with a heteroskedastic partition, we use a model of wages as a function of education and experience. In addition a dummy variable is included that is equal to one if the individual lives in a metropolitan area. This is an intercept dummy variable indicating people living in metropolitan areas make higher wages relative to those living in rural with similar level of experience and education.

First, open *cps2.xls* and highlight columns D through I and **Edit>Delete** the columns.

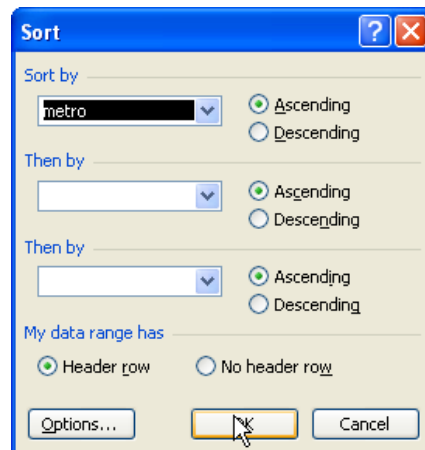
$$wage_i = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 metro_i + e_i$$

Estimate the model using **Tools>Data Analysis>Regression** with *WAGE* as the **Y-Range** and *EDUC*, *EXPER* and *METRO* as the **X-Range**.

|    |           |                     |                       |               |                |
|----|-----------|---------------------|-----------------------|---------------|----------------|
| 15 |           |                     |                       |               |                |
| 16 |           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept | -9.913984218        | 1.075662517           | -9.216630734  | 1.77326E-19    |
| 18 | educ      | 1.233963998         | 0.069961261           | 17.63781812   | 8.42574E-61    |
| 19 | exper     | 0.133243681         | 0.015231619           | 8.747834543   | 9.13789E-18    |
| 20 | metro     | 1.524104206         | 0.431090949           | 3.535458611   | 0.000425795    |

Next, estimate each subset (metro and rural) separately using least squares and save the standard errors of the regression.

- Highlight all the columns of the data, including labels, choose **Data>Sort** from the menu bar.
- Click on the **down-arrow** in the **Sort By** box. Choose *METRO*.
- Check either **Ascending** or **Descending** option.
- Check the **Header Row** option, since we included the labels and click **OK**.



Looking at the *METRO* column, you will observe that the column will contain zeros till row 194 and after that row, *METRO* will be one.

|     |       |    |    |   |
|-----|-------|----|----|---|
| 182 | 16.21 | 12 | 29 | 0 |
| 183 | 16.28 | 18 | 26 | 0 |
| 184 | 16.37 | 12 | 22 | 0 |
| 185 | 16.94 | 13 | 29 | 0 |
| 186 | 18.12 | 14 | 12 | 0 |
| 187 | 18.17 | 16 | 9  | 0 |
| 188 | 18.33 | 12 | 40 | 0 |
| 189 | 21.25 | 16 | 9  | 0 |
| 190 | 22.1  | 16 | 21 | 0 |
| 191 | 25.42 | 18 | 34 | 0 |
| 192 | 26.98 | 12 | 45 | 0 |
| 193 | 27.26 | 16 | 3  | 0 |
| 194 | 2.07  | 12 | 7  | 1 |
| 195 | 2.12  | 12 | 35 | 1 |
| 196 | 2.54  | 16 | 20 | 1 |
| 197 | 2.68  | 12 | 24 | 1 |
| 198 | 3.09  | 13 | 4  | 1 |
| 199 | 3.17  | 12 | 22 | 1 |
| 200 | 3.2   | 12 | 23 | 1 |
| 201 | 3.27  | 12 | 4  | 1 |
| 202 | 3.32  | 12 | 11 | 1 |
| 203 | 3.32  | 13 | 3  | 1 |
| 204 | 3.34  | 18 | 15 | 1 |
| 205 | 3.39  | 13 | 7  | 1 |

Now we must estimate two regressions, using the first half of the data in the first one (*METRO* = 0), and the second half of the data in the second regression (*METRO* = 1). The only output we're interested in from these regressions is the estimated variance of the model.

Estimate a regression on the data, using cells A2 through A193 for the **Y-Range** and cells B2 through C193 for the **X-Range**. DO NOT include labels and don't include the *METRO* variable in column D. Place the output on a worksheet named "**regression 1**."

The first Excel Regression dialog box is shown. In the 'Input' section, the 'Input Y Range' is '\$A\$2:\$A\$193' and the 'Input X Range' is '\$B\$2:\$C\$193'. The 'Labels' checkbox is unchecked, and the 'Constant is Zero' checkbox is also unchecked. The 'Confidence Level' is set to '95 %'. In the 'Output options' section, the 'New Worksheet Ply:' radio button is selected, and the worksheet name is 'Rural'. In the 'Residuals' section, all checkboxes ('Residuals', 'Standardized Residuals', 'Residual Plots', 'Line Fit Plots') are unchecked. In the 'Normal Probability' section, the 'Normal Probability Plots' checkbox is unchecked. The 'OK' button is highlighted.

Repeat the procedures above, but now include cells A194 through A1001 as the **Y-Range** and cells B194 through C1001 as the **X-Range**. Save output to a worksheet named “**regression 2.**”

The second Excel Regression dialog box is shown. In the 'Input' section, the 'Input Y Range' is '\$A\$194:\$A\$1001' and the 'Input X Range' is '\$B\$194:\$C\$1001'. The 'Labels' checkbox is unchecked, and the 'Constant is Zero' checkbox is also unchecked. The 'Confidence Level' is set to '95 %'. In the 'Output options' section, the 'New Worksheet Ply:' radio button is selected, and the worksheet name is 'Metro'. In the 'Residuals' section, all checkboxes ('Residuals', 'Standardized Residuals', 'Residual Plots', 'Line Fit Plots') are unchecked. In the 'Normal Probability' section, the 'Normal Probability Plots' checkbox is unchecked. The 'OK' button is highlighted.

In order to obtain a homoskedastic model, we will transform the data by using the estimated variances from each partition. The estimated variances of the models are what we are interested in; no other output from the regressions is important at this point. We divide the data on the first 193 observations by the square root of the estimated variance from a regression using just those observations, and divide the last 808 observations by the square root of the estimated variance obtained from the regression using just these data. Then we “pool” the transformed data and perform generalized least squares estimation to obtain the correct estimates.

The simplest thing to do at this point is to simply write down the values of the **MS Residuals** from the ANOVA tables of the regressions; metro and rural. From **metro**, the value is 31.8237318; from **rural**, it is 15.24298659. Transform the first 193 observations using the square root of the estimated variance from **rural** and the rest of the observations using the square root of

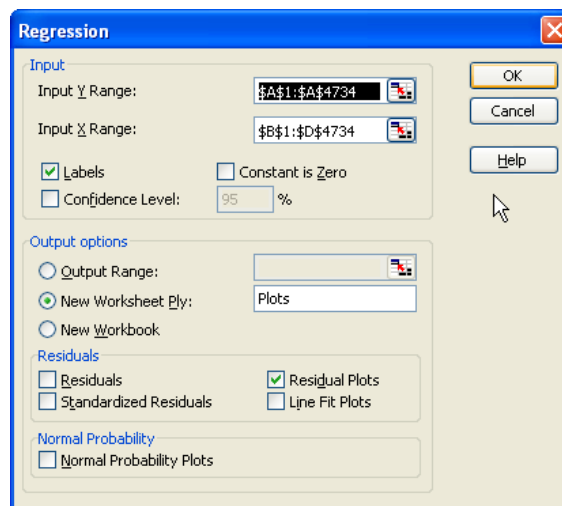
the estimated variance from **metro** as shown in Section 8.3.1 of this manual. Estimate a regression, using *wage\** as the **Y-Range** and *int\**, *edu\**, *exper\** and *metro\** as the **X-Range** where the stars represent the transformed values for the intercept (*int*), years of education (*edu*), years of experience (*exper*) and dummy variable for metropolitan area (*metro*), respectively. Include all 1000 observations. Include labels and suppress the intercept by checking the **Constant is Zero** box.

## 8.4 DETECTING HETEROSKEDASTICITY

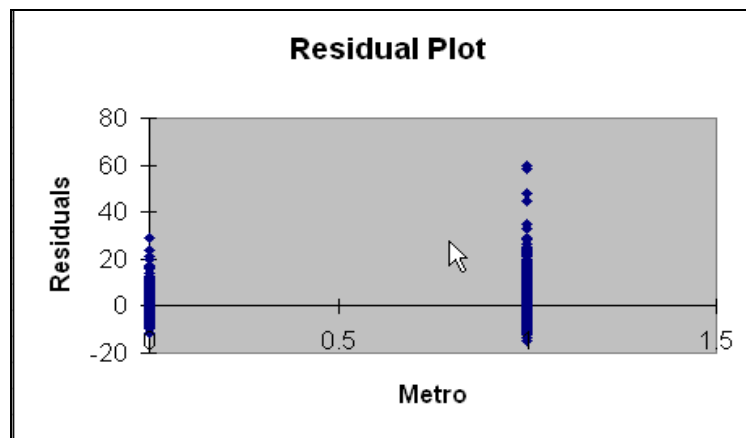
### 8.4.1 Residual plots

If the regression errors are homoskedastic, when we plot the residuals, there should be no systematic pattern evident. If the errors are heteroskedastic, we may be able to detect a particular pattern in the residuals and perhaps even discover the form of the heteroskedasticity.

Let's use the wage model to plot the least square residuals against the metro dummy. Open *cps.xls*. Delete all the columns except *wage*, *educ*, *exper*, *metro* and save the file as *cps\_modified.xls*. Estimate a regression, *wage* as the **Y-Range** and *educ*, *exper* and *metro* as the **X-Range**. Under the **Residuals options**, choose **Residual Plots**. This will produce a graph of the residuals, plotted against each of the explanatory variables.



Let's look at the plot against the *metro* dummy variable. After formatting, it should look similar to this.



In this plot, the least squares residuals are grouped according to rural ( $metro=0$ ) or metropolitan ( $metro=1$ ). Wider variation of residuals at  $metro=1$  indicates higher variance for these observations and evidence of **groupwise heteroskedasticity**.

### 8.4.2 The Goldfeld-Quandt test

Although plots are helpful in diagnosing heteroskedasticity, the Goldfeld-Quandt formally tests for equal variances. It is a type of  $F$ -test and the steps are to order the data based on variance, split the data into parts, compute the estimated variances, calculate the test statistic  $GQ = \hat{\sigma}_1^2 / \hat{\sigma}_2^2$ , and compare to an  $F$ -critical value, based on  $T_1 - K$  numerator and  $T_2 - K$  denominator degrees of freedom. If heteroskedasticity is present, the **GQ** test statistic should be large, and we would reject a null hypothesis of equal variances.

Returning to the wage model, where **metro**  $F$  is 31.824 and **rural**  $F$  is 15.243. Open a new worksheet and name it “**GQ test**”. Create the following template to use for any Goldfeld-Quandt Test.

| Microsoft Excel - food.xls                                    |                                                       |
|---------------------------------------------------------------|-------------------------------------------------------|
| File Edit View Insert Format Tools Data Window Help Adobe PDF |                                                       |
| H11                                                           |                                                       |
| A                                                             | B                                                     |
| 1                                                             | Goldfeld-Quandt Test for Equal Variance               |
| 2                                                             | Input Data                                            |
| 3                                                             | T1                                                    |
| 4                                                             | T2                                                    |
| 5                                                             | K                                                     |
| 6                                                             | SigmaHatSquared1                                      |
| 7                                                             | SigmaHatSquared2                                      |
| 8                                                             | Alpha 0.05                                            |
| 9                                                             | Computed Values                                       |
| 10                                                            | df-numerator =B3-B5                                   |
| 11                                                            | df-denominator =B4-B5                                 |
| 12                                                            | GQ =B6/B7                                             |
| 13                                                            | 2-tailed Test: Right Critical Value =FINV(B8,B10,B11) |
| 14                                                            | Decision =IF(B12>B13,"Reject Ho","Fail to Reject Ho") |
| 15                                                            | p-value =FDIST(B12,B10,B11)                           |

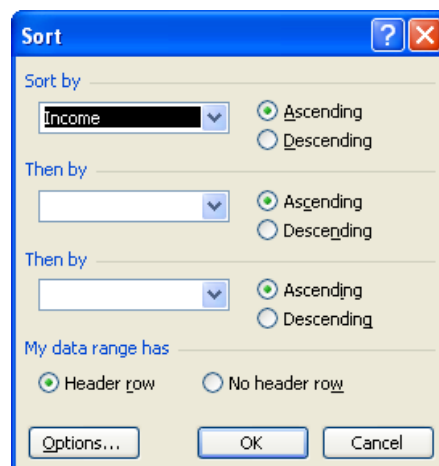
Fill in the **Input Data** as **T1**=808 and **T2**=193, **K**=2, and **Alpha** should be .05, for testing at the 5% level. From the **metro** worksheet, highlight cell D13 (the MS residual from the ANOVA table), right-click and choose **C**opy. Go back to GQ test worksheet and **P**aste the value into cell B6. Remember to always place the larger of the estimated variances in the numerator of the formula for the **GQ** test statistic. Copy cell D13 from the **rural** worksheet to cell B7 of **GQ** test. The resulting template is

|    | A                                              | B           |
|----|------------------------------------------------|-------------|
| 1  | <b>Goldfeld-Quandt Test for Equal Variance</b> |             |
| 2  | <b>Input Data</b>                              |             |
| 3  | T1                                             | 808         |
| 4  | T2                                             | 193         |
| 5  | K                                              | 2           |
| 6  | SigmaHatSquared1                               | 31.8237318  |
| 7  | SigmaHatSquared2                               | 15.24298659 |
| 8  | Alpha                                          | 0.05        |
| 9  | <b>Computed Values</b>                         |             |
| 10 | df-numerator                                   | 806         |
| 11 | df-denominator                                 | 191         |
| 12 | GQ                                             | 2.087762238 |
| 13 | e-tailed Test:Right Critical Value             | 1.213925811 |
| 14 | Decision                                       | Reject Ho   |
| 15 | p-value                                        | 1.32044E-09 |

Let's test for heteroskedasticity in the food expenditures model. Formally, we test the null hypothesis  $H_0: \sigma_i^2 = \sigma^2$  against the alternative  $H_1: \sigma_i^2 = \sigma^2 x_i$ .

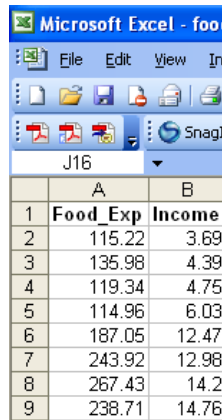
Since the variance is an increasing function of income we have to sort the data in an ascending order by *income*. Open the original data on *food expenditures* and *income*, **food.xls**.

- Highlight both columns of the data, including labels, choose **D**ata>**S**ort from the menu bar.
- Click on the down-arrow in the **Sort By** box. Choose *income*.
- Check the **A**scending option;
- Check the **H**ead**R**ow option, since we included the labels and click OK.





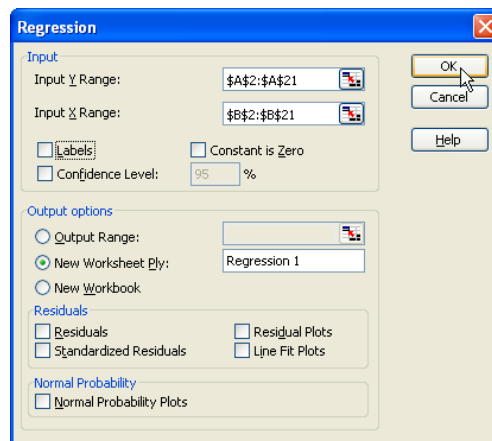
Looking at the column containing *income*, the numbers should be increasing as you look down the column.



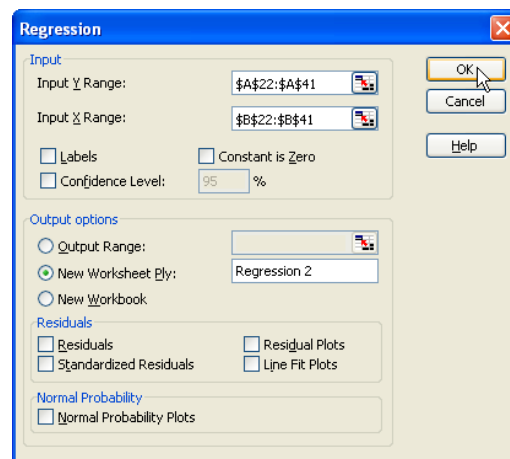
|   | A               | B             |
|---|-----------------|---------------|
| 1 | <b>Food Exp</b> | <b>Income</b> |
| 2 | 115.22          | 3.69          |
| 3 | 135.98          | 4.39          |
| 4 | 119.34          | 4.75          |
| 5 | 114.96          | 6.03          |
| 6 | 187.05          | 12.47         |
| 7 | 243.92          | 12.98         |
| 8 | 267.43          | 14.2          |
| 9 | 238.71          | 14.76         |

Now we must estimate two regressions, using the first part of the data in the first one, and the second part of the data in the second regression. Because there is no natural dividing point, we will break the ordered sample into two equal halves. The only output we're interested in from these regressions is the estimated variance of the model.

Estimate a regression on the data, using cells A2 through A21 for the **Y-Range** and cells B2 through B21 for the **X-Range**. Do NOT include labels. Place the output on a worksheet named “**regression 1.**”



Repeat the procedures above, but now include cells A22 through A41 as the **Y-Range** and cells B22 through B1 as the **X-Range**. Save output to a worksheet named “**regression 2.**”



Fill in the Input data. **T1** and **T2** both equal 20, **K**=2 and **Alpha** should be .05, for testing at the 5% level.

- From the **regression 1** worksheet, highlight cell D13 (the MS residual from the ANOVA table), right-click and choose Copy.
- Go back to GQ test worksheet and Paste the value into cell B7. Remember to always place the larger of the estimated variances in the numerator of the formula for the GQ test statistic.
- Copy cell D13 from the **regression 2** worksheet to cell B6 of GQ test.

The resulting template is

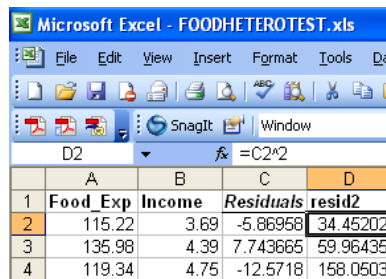
|    | A                                              | B           |
|----|------------------------------------------------|-------------|
| 1  | <b>Goldfeld-Quandt Test for Equal Variance</b> |             |
| 2  | <b>Input Data</b>                              |             |
| 3  | T1                                             | 20          |
| 4  | T2                                             | 20          |
| 5  | K                                              | 2           |
| 6  | SigmaHatSquared1                               | 12921.92662 |
| 7  | SigmaHatSquared2                               | 3574.77175  |
| 8  | Alpha                                          | 0.05        |
| 9  | <b>Computed Values</b>                         |             |
| 10 | df-numerator                                   | 18          |
| 11 | df-denominator                                 | 18          |
| 12 | GQ                                             | 3.614755716 |
| 13 | e-tailed Test:Right Critical Value             | 2.217197134 |
| 14 | Decision                                       | Reject Ho   |
| 15 | p-value                                        | 0.00459643  |

We reject the null hypothesis and conclude that heteroskedasticity IS present. If we assume proportional heteroskedasticity, we would proceed as in section 11.3. If we couldn't assume any particular form of the heteroskedasticity, then we should at least calculate White's standard errors and report the corrected *t*-statistics and confidence intervals.

### 8.4.3 Testing the variance function

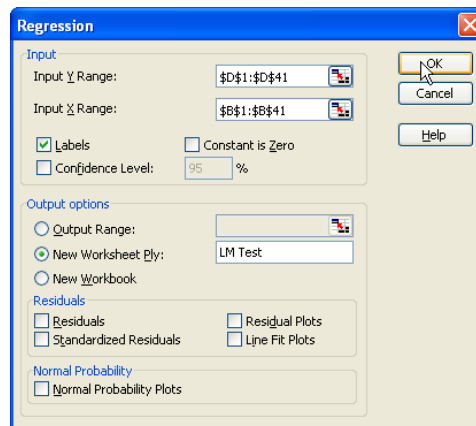
There are many other tests to test for the existence of heteroskedasticity. In this section, two more of these tests will be discussed; **Breusch-Pagan** and **White's** tests. In both of the tests the null is homoskedastic errors and alternative is heteroskedastic errors. Both of the tests also require auxiliary regressions.

First, we will illustrate the Breusch-Pagan test using the food expenditure data. Open the original series, *food.xls* and estimate the regression, click the residual option to save the residuals. Then, copy and paste the residuals to column C in the data sheet. Label D1 “*resid2*” and type  $=C2^2$  in D2.



|   | A        | B      | C         | D        |
|---|----------|--------|-----------|----------|
| 1 | Food_Exp | Income | Residuals | resid2   |
| 2 | 115.22   | 3.69   | -5.86958  | 34.45202 |
| 3 | 135.98   | 4.39   | 7.743665  | 59.96435 |
| 4 | 119.34   | 4.75   | -12.5718  | 158.0503 |

Next, regress residuals on *income*.



The test statistic ( $LM$ ) is calculated by multiplying the  $N \cdot R^2$  and has a  $\chi^2_{s-1}$  distribution.

|    | A                            | B                   |
|----|------------------------------|---------------------|
| 1  | SUMMARY OUTPUT               |                     |
| 2  |                              |                     |
| 3  | <i>Regression Statistics</i> |                     |
| 4  | Multiple R                   | 0.429663369         |
| 5  | R Square                     | 0.184610611         |
| 6  | Adjusted R Square            | 0.163152996         |
| 7  | Standard Error               | 9946.64208          |
| 8  | Observations                 | 40                  |
| 9  |                              |                     |
| 10 | ANOVA                        |                     |
| 11 |                              | <i>df</i>           |
| 12 | Regression                   | 1                   |
| 13 | Residual                     | 38                  |
| 14 | Total                        | 39                  |
| 15 |                              |                     |
| 16 |                              | <i>Coefficients</i> |
| 17 | Intercept                    | -5762.369835        |
| 18 | Income                       | 682.232583          |
| 19 |                              |                     |
| 20 |                              |                     |
| 21 |                              |                     |
| 22 | LM=                          | 7.384424443         |
| 23 | CHI-CRITICAL                 | 3.841459149         |
| 24 | P-VALUE                      | 0.006579112         |

White's test is just a minor modification to Breush-Pagan test. Add one more column to the data and label it *income2*. This column will contain the squared income variable. After putting in the formula and copying it down, regress *resid2* on *income* and *income2*. Recall that you need to get the *income* and *income2* in adjacent columns to be able to run the regression. The test statistic is again the product of  $N$  and  $R^2$  which has a  $\chi^2_{s-1}$  distribution.

|    | A                            | B                   |
|----|------------------------------|---------------------|
| 1  | SUMMARY OUTPUT               |                     |
| 2  |                              |                     |
| 3  | <i>Regression Statistics</i> |                     |
| 4  | Multiple R                   | 0.434599775         |
| 5  | R Square                     | 0.188876964         |
| 6  | Adjusted R Square            | 0.145032476         |
| 7  | Standard Error               | 10053.75429         |
| 8  | Observations                 | 40                  |
| 9  |                              |                     |
| 10 | ANOVA                        |                     |
| 11 |                              | <i>df</i>           |
| 12 | Regression                   | 2                   |
| 13 | Residual                     | 37                  |
| 14 | Total                        | 39                  |
| 15 |                              |                     |
| 16 |                              | <i>Coefficients</i> |
| 17 | Intercept                    | -2908.78281         |
| 18 | income2                      | 11.16528894         |
| 19 | Income                       | 291.7457265         |
| 20 |                              |                     |
| 21 | <b>N* R Square</b>           |                     |
| 22 |                              |                     |
| 23 |                              |                     |
| 24 |                              |                     |
| 25 | LM=                          | 7.555078561         |
| 26 | CHI-CRITICAL                 | 3.841459149         |
| 27 | p-value                      | 0.2287893           |

The results for both of the tests indicate a rejection of the null hypothesis of no heteroskedasticity.

# **CHAPTER 9**

## Dynamic Models, Autocorrelation, and Forecasting

### **CHAPTER OUTLINE**

- 9.1 Lags in the Error Term
- 9.2 Area Response for Sugar
- 9.3 Estimating an AR(1) Model
  - 9.3.1 Least squares

- 9.4 Detecting Autocorrelation
  - 9.4.1 The Durbin-Watson test
  - 9.4.2 An LM test
- 9.5 Autoregressive Models
- 9.6 Finite Distributed Lags
- 9.7 ARDL Model

### **9.1 LAGS IN THE ERROR TERM**

The multiple linear regression model of Chapters 5 and 6 assumed that the observations are not correlated with one another. This assumption is not realistic if the observations are drawn sequentially in time. With times-series data, where the observations follow a natural ordering through time, there is a possibility that successive errors are correlated with each other. Shocks to a model may take time to work out and effects may carry over to successive time periods. The result is that the error term in period  $t$  can affect the error term in period  $t+1$ , or  $t+2$ , and so on. Somehow, we must take these lasting effects into account.

In the first example the supply response for an agricultural crop is modeled as a log-log linear model where area planted (acres) depends on price. The first dynamic model we will consider is one with a lag in the error term.

$$\ln(A_t) = \beta_1 + \beta_2 \ln(P_t) + e_t$$

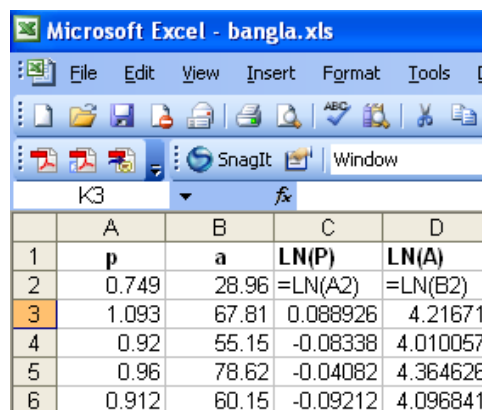
$$e_t = \rho e_{t-1} + v_t$$

Where  $\rho$  (rho) is the parameter that describes the relation between  $e_t$  and  $e_{t-1}$ ,  $v_t$  is the new random error term. For stability (and stationarity) of the model,  $-1 < \rho < 1$ .

## 9.2 AREA RESPONSE FOR SUGAR

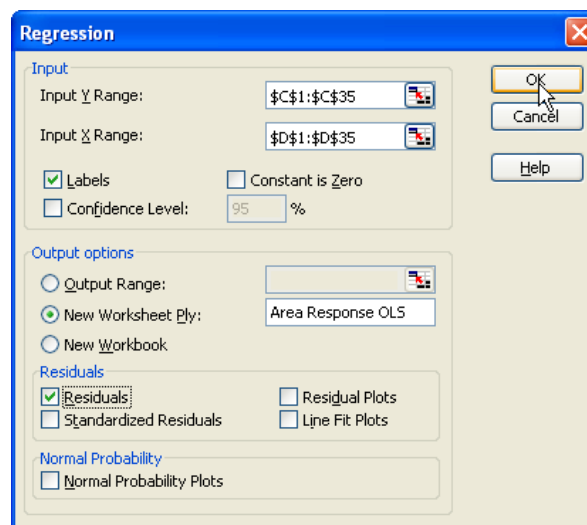
First, we will estimate the above mentioned model using least squares. Open *bangla.xls*.

- **Label** columns C and D as  $LN(P)$  and  $LN(A)$  respectively.
- In cell C2, type **=LN(A2)**, copy this formula to cell D2.
- Highlight cells C2 and D2, and copy the formulas down the columns.



|   | A     | B     | C        | D        |
|---|-------|-------|----------|----------|
| 1 | p     | a     | LN(P)    | LN(A)    |
| 2 | 0.749 | 28.96 | =LN(A2)  | =LN(B2)  |
| 3 | 1.093 | 67.81 | 0.088926 | 4.21671  |
| 4 | 0.92  | 55.15 | -0.08338 | 4.010057 |
| 5 | 0.96  | 78.62 | -0.04082 | 4.364626 |
| 6 | 0.912 | 60.15 | -0.09212 | 4.096841 |

Estimate a regression, using  $LN(A)$  as the **Y-Range** and  $LN(P)$  as the **X-Range**. Include labels. Check the **Residuals** option so that the estimated errors are produced and click **OK**.



**Regression**

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

The least squares estimates are:

|    |           |                     |                     |               |                |
|----|-----------|---------------------|---------------------|---------------|----------------|
| 16 |           | <i>Coefficients</i> | <i>Standard Err</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept | 3.893255748         | 0.061345            | 63.46486      | 3.12E-35       |
| 18 | LN(P)     | 0.776118781         | 0.277467            | 2.797154      | 0.008653       |

Since we are using times-series data, we should explore the possibility of autocorrelation. Visually, we can plot the residuals against time and see if we can detect any patterns.

Return to the worksheet containing the original data.

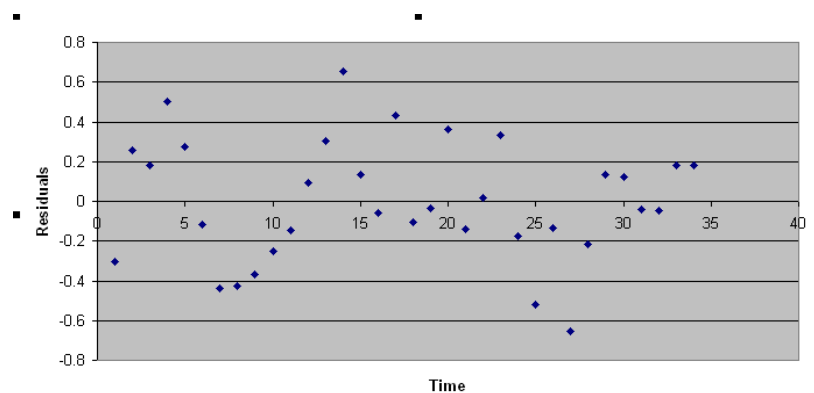
- Label cell E1  $t$ , for time.
- Type “1” in cell E2.
- Type  $=E1+1$  in cell E3.
- Highlight cells E2 and E3, place cursor on the lower right hand corner of the highlighted area until it turns into a cross-hatch.
- Left-click and drag down the column to fill in the values in ascending order.

|   | A     | B     | C        | D        | E |
|---|-------|-------|----------|----------|---|
| 1 | p     | a     | LN(P)    | LN(A)    | t |
| 2 | 0.749 | 28.96 | -0.28902 | 3.365916 | 1 |
| 3 | 1.093 | 67.81 | 0.088926 | 4.21671  | 2 |
| 4 | 0.92  | 55.15 | -0.08338 | 4.010057 | 3 |
| 5 | 0.96  | 78.62 | -0.04082 | 4.364626 | 4 |
| 6 | 0.912 | 60.15 | -0.09212 | 4.096841 | 5 |
| 7 | 1.054 | 45.54 | 0.052592 | 3.818591 | 6 |
| 8 | 1.079 | 33.62 | 0.076035 | 3.515121 | 7 |
| 9 | 1.525 | 44.58 | 0.421994 | 3.797285 | 8 |

Return to the worksheet containing the regression output.

- **C**opy residuals to the worksheet containing the original data and **P**aste in column F.
- Create an XY Scatter graph with  $t$  on the horizontal axis and *Residuals* on the vertical axis.

The results will look like





Looking at the values of the residuals (ehat) and at the graph above, there seems to be a tendency for negative values to follow negative values, and positive values to follow positive values. This is consistent with positive correlation between successive terms. While such a conclusion about autocorrelation is subjective, we will later look at a more formal test. For now, however, it does appear that there is a problem.

We can also check the correlation between the ehat and the lagged values. Rename the residual column ehat, create a new column and call it ehat\_1. Lag the ehat column by copying F2-F35 and paste it to G3. Delete F2 and G36 to make the columns even. Go to **Tools>Data Analysis>Correlation**. Put ehat and ehat\_1 in the range and hit **OK**. The output indicates about .40 correlation between the errors one period apart.

|   | A      | B        | C      |
|---|--------|----------|--------|
| 1 |        | ehat     | ehat_1 |
| 2 | ehat   | 1        |        |
| 3 | ehat_1 | 0.403997 | 1      |

### **9.3 ESTIMATING AN AR(1) MODEL**

When the errors follow an AR(1) model,  $e_t = \rho e_{t-1} + v_t$ , the least squares assumption 4 is violated. Least squares is unbiased and consistent, but no longer efficient. The reported standard errors are no longer correct, leading to statistically invalid hypothesis tests and confidence intervals.

#### **9.3.1 Least Squares**

In the previous chapter, we transformed our data so that we could move from a heteroskedastic model to a homoskedastic. The same type of procedure can be used to correct for first order autoregressive errors AR(1). Our objective is to transform the model

$$y_t = \beta_1 + \beta_2 x_t + e_t \text{ where } e_t = \rho e_{t-1} + v_t$$

such that the autocorrelated term  $e_t$  is replaced by the uncorrelated error term  $v_t$ . After some substitution and rearranging, the transformed model we obtain is

$$y_t - \rho y_{t-1} = \beta_1(1 - \rho) + \beta_2(x_t - \rho x_{t-1}) + v_t$$

All we need to do is to transform the dependent variable, intercept, and explanatory variable as above and proceed with the generalized least squares estimation. The first problem is  $\rho$  is unknown and must be estimated and observe that we now have  $T-1$  observations since we lose the "first" one.

Estimate the original model and store the residuals. Since

$$\hat{\rho} = \frac{\sum_{t=2}^T \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^T \hat{e}_{t-1}^2}$$

and we already have our residuals from the least squares estimation, we're ready to go!

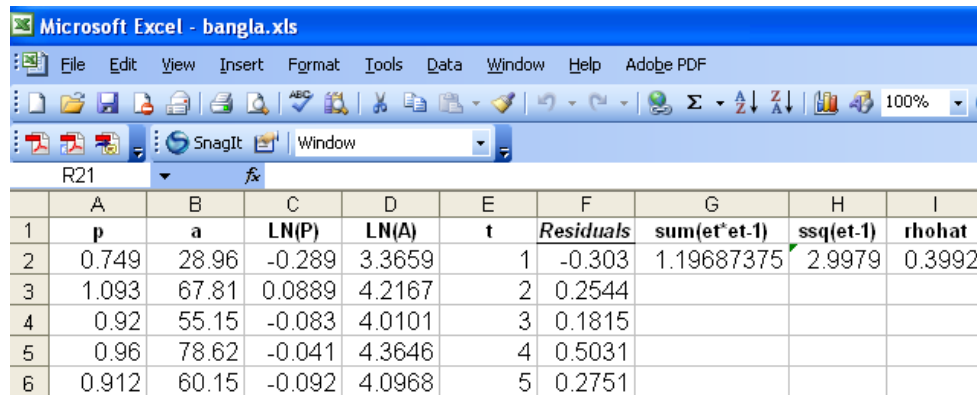
Return to the worksheet containing the original data and the residuals. Label cells G1, H1, and I1 as *sum(et\*et-1)*, *ssq(et-1)*, and *rhohat* respectively. In cell G2, type

**=SUMPRODUCT(F3:F35,F2:F34)**

This corresponds to the numerator in the formula above. In cell H2, type

**=SUMSQ(F2:F34)**

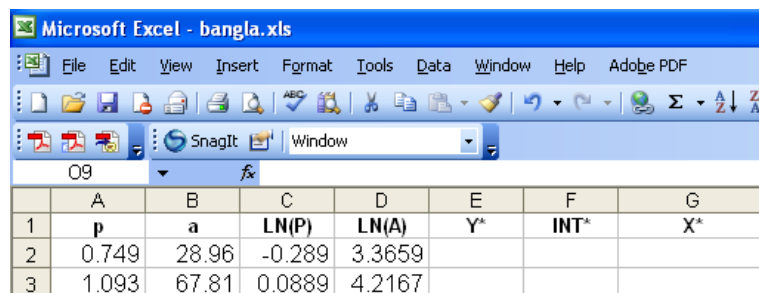
This calculates the denominator. Finally, in cell I2, divide G2 by H2 by typing **=G2/H2**. The result is 0.3992.



|   | A     | B     | C      | D      | E | F         | G            | H         | I      |
|---|-------|-------|--------|--------|---|-----------|--------------|-----------|--------|
| 1 | p     | a     | LN(P)  | LN(A)  | t | Residuals | sum(et*et-1) | ssq(et-1) | rhohat |
| 2 | 0.749 | 28.96 | -0.289 | 3.3659 | 1 | -0.303    | 1.19687375   | 2.9979    | 0.3992 |
| 3 | 1.093 | 67.81 | 0.0889 | 4.2167 | 2 | 0.2544    |              |           |        |
| 4 | 0.92  | 55.15 | -0.083 | 4.0101 | 3 | 0.1815    |              |           |        |
| 5 | 0.96  | 78.62 | -0.041 | 4.3646 | 4 | 0.5031    |              |           |        |
| 6 | 0.912 | 60.15 | -0.092 | 4.0968 | 5 | 0.2751    |              |           |        |

Now, to deal with the issue of transforming the first observation for the transformed model, we have  $y_1 = \beta_1 + x_1\beta_2 + e_1$  with an error variance of  $\text{var}(e_1) = \sigma_e^2 = \sigma_v^2 / (1 - \rho^2)$ . The transformation that gets to where we want (a variance of  $\sigma_v^2$ ) is multiplication of the terms in the model, for the first observation, by  $\sqrt{1 - \rho^2}$ .

Use the worksheet with columns *P*, *A*, *LN(P)*, *LN(A)*. Label columns E, F, and G as *y\**, *int\** and *x\** respectively.



|   | A     | B     | C      | D      | E  | F    | G  |
|---|-------|-------|--------|--------|----|------|----|
| 1 | p     | a     | LN(P)  | LN(A)  | Y* | INT* | X* |
| 2 | 0.749 | 28.96 | -0.289 | 3.3659 |    |      |    |
| 3 | 1.093 | 67.81 | 0.0889 | 4.2167 |    |      |    |

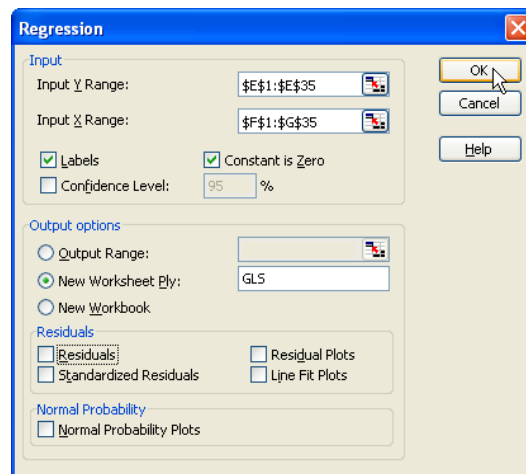
- In cell E2, type **=SQRT(1-(0.3992^2))\*D2**.
- In cell F2, type **=SQRT(1-(0.3992^2))**.
- In cell G2, type **=SQRT(1-(0.3992^2))\*C2**.

The first observation for the transformed model is now complete. For the remaining observations:

- In cell E3, type **=D3-(0.3992\*D2)**.
- In cell F3, type **=1-0.3992**.
- In cell G3, type **=C3-(0.3992\*C2)**.
- Highlight cells E3 through G3. Copy the formulas down the columns.

|   | A     | B     | C      | D      | E                      | F                   | G                      |
|---|-------|-------|--------|--------|------------------------|---------------------|------------------------|
| 1 | p     | a     | LN(P)  | LN(A)  | Y*                     | INT*                | X*                     |
| 2 | 0.749 | 28.96 | -0.289 | 3.3659 | =SQRT(1-(0.3992^2))*D2 | =SQRT(1-(0.3992^2)) | =SQRT(1-(0.3992^2))*C2 |
| 3 | 1.093 | 67.81 | 0.0889 | 4.2167 | =D3-(0.3992*D2)        | =1-0.3992           | =C3-(0.3992*C2)        |
| 4 | 0.92  | 55.15 | -0.083 | 4.0101 |                        |                     |                        |

Run a regression, using  $y^*$  as the **Y-Range** and  $int^*$  and  $x^*$  as the **X-Range**. Include labels as usual AND suppress the intercept by checking the **Constant is Zero** box. Place output on a new worksheet names **GLS** and click **OK**.



The generalized least squares results are:

| 16 |           | Coefficients | Standard Error | t Stat      | P-value     |
|----|-----------|--------------|----------------|-------------|-------------|
| 17 | Intercept | 0            | #N/A           | #N/A        | #N/A        |
| 18 | INT*      | 3.873889527  | 0.081947116    | 47.27304313 | 3.50629E-31 |
| 19 | X*        | 0.945993569  | 0.240753509    | 3.929303356 | 0.000427015 |

Once again, interpretations of the estimates are as usual. The price elasticity of sugar cane area response seems to be one.

## 9.4 DETECTING AUTOCORRELATION

We will consider two formal tests to test for the existence and extent of autocorrelation; Durbin Watson and *LM* (Lagrange Multiplier) serial correlation tests. Both tests test the hypothesis  $H_0: \rho = 0$  versus the alternative  $H_1: \rho > 0$ . If  $\rho$  is zero, then no transformation is necessary and ordinary least squares estimation is BLUE.

### 9.4.1 The Durbin-Watson test

The Durbin-Watson test statistic uses the residuals from the least squares procedure and is closely related to  $\hat{\rho}$ . The statistic is

$$d = \frac{\sum_{t=2}^T (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^T \hat{e}_t^2}$$

It is approximately equal to  $2(1 - \hat{\rho})$ .

If the estimated value of  $\rho$  is zero, the Durbin-Watson test statistic equals 2. If the estimated value of  $\rho$  is one, the Durbin-Watson test statistic equals 0. Therefore, a low value of the DW test statistic suggests the null hypothesis should be rejected. The distribution of the DW test statistic is difficult and Excel cannot compute the  $p$ -value associated with  $d$ , but tables are available for performing the hypothesis test [see [www.bus.lsu.edu/hill/poe](http://www.bus.lsu.edu/hill/poe)], now called the Durbin-Watson bounds test.

$$\sum_{t=2}^T (\hat{e}_t - \hat{e}_{t-1})^2 = \text{sum of squared differences} = \text{SUMXMY2(F3:F35,F2:F34)} = 3.54386803$$

$$\sum_{t=1}^T \hat{e}_t^2 = \text{sum of squared residuals} = \text{SUMSQ(F2:F35)} = 3.0316$$

Using the formula above, we calculate the Durbin-Watson statistic to be  $d \cong 1.169$  as follows: The least squares residuals for *bangle.xls* are stored in cells F2:F35 in the spreadsheet. The numerator and the denominator of the DW statistic can be obtained using the Excel functions **SUMXMY2**, and **SUMSQ**, respectively.

The decision rule for the Durbin-Watson bounds test is

- if  $d > \text{upper bound}$ , fail to reject the null hypothesis of no serial correlation,
- if  $d < \text{lower bound}$ , reject the null hypothesis and conclude that positive autocorrelation is present,
- if  $\text{lower bound} < d < \text{upper bound}$ , the test is inconclusive.

With  $T=34$  and  $K=2$ , the lower bound is 1.393 and the upper bound is 1.514. Since  $d <$  the lower bound, we reject the null hypothesis that  $\rho$  is zero, and find evidence of positive autocorrelation.

### 9.4.2 An LM test

Another way to test for autocorrelation is to use test whether residuals are correlated with one another using the *LM* test. This test is based on the auxiliary regression where you regress least square residuals on the original regressors and lagged residual(s). If the auxiliary regression explains sufficient variation in  $\hat{e}_t$ , then we conclude there is autocorrelation.

Now, let's see how it works in Excel. Return to the worksheet containing the original data for *bangla.xls*, include the logs of the data, and the residuals.

- Insert a new column and label it *lagged residuals*.
- Write 0 into cell E2.
- In cell E3, type **=G2**, where G2 contains the first residual from the regression output and copy the formula down the column.

|   | A     | B     | C      | D           | E      | F                |
|---|-------|-------|--------|-------------|--------|------------------|
| 1 | p     | a     | LN(A)  | Residuals   | LN(P)  | Lagged residuals |
| 2 | 0.749 | 28.96 | 3.3659 | -0.30302924 | -0.289 | 0                |
| 3 | 1.093 | 67.81 | 4.2167 | 0.254436598 | 0.0889 | -0.30302924      |
| 4 | 0.92  | 55.15 | 4.0101 | 0.181515067 | -0.083 | 0.254436598      |
| 5 | 0.96  | 78.62 | 4.3646 | 0.503053127 | -0.041 | 0.181515067      |
| 6 | 0.912 | 60.15 | 4.0968 | 0.275078133 | -0.092 | 0.503053127      |

Once we created the lagged values of the residuals, we can now run the regression, using *residuals* as the **Y-Range** and *LN(P)* and *lagged residuals* as the **X-Range**. Since the explanatory variables need to be next to each other, you will have to move some of the columns around. Do **NOT** suppress the intercept and name the worksheet *LM test*. Click **OK**.

The image shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' set to '\$D\$1:\$D\$35' and 'Input X Range' set to '\$E\$1:\$F\$35'. The 'Labels' checkbox is checked, and 'Constant is Zero' is unchecked. The 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected, with the name 'LM test' entered. The 'Residuals' section has 'Residuals' and 'Standardized Residuals' checked, while 'Residual Plots' and 'Line Fit Plots' are unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. The 'OK' button is highlighted with a mouse cursor.

|    | A                     | B            | C              | D            |
|----|-----------------------|--------------|----------------|--------------|
| 1  | SUMMARY OUTPUT        |              |                |              |
| 2  |                       |              |                |              |
| 3  | Regression Statistics |              |                |              |
| 4  | Multiple R            | 0.401259455  |                |              |
| 5  | R Square              | 0.16100915   |                |              |
| 6  | Adjusted R Square     | 0.106880708  |                |              |
| 7  | Standard Error        | 0.286438676  |                |              |
| 8  | Observations          | 34           |                |              |
| 9  |                       |              |                |              |
| 10 | ANOVA                 |              |                |              |
| 11 |                       | df           | SS             | MS           |
| 12 | Regression            | 2            | 0.488110715    | 0.244055357  |
| 13 | Residual              | 31           | 2.543460566    | 0.082047115  |
| 14 | Total                 | 33           | 3.031571281    |              |
| 15 |                       |              |                |              |
| 16 |                       | Coefficients | Standard Error | t Stat       |
| 17 | Intercept             | -0.00811623  | 0.057185889    | -0.141927146 |
| 18 | LN(P)                 | 0.091601353  | 0.260933616    | 0.351052326  |
| 19 | Lagged residuals      | 0.407821387  | 0.167202391    | 2.439088245  |

Recall that Excel reports a  $p$ -value based on a two-tailed test. We would conclude that  $\rho > 0$ , testing at the 5% level, but we cannot reject the null hypothesis  $H_0: \rho = 0$  at the 1% level. The  $LM$  scalar is calculated by  $N \cdot R^2$  where  $N$  and  $R^2$  are the number of observations and the coefficient of determination in the auxiliary regression and has a  $\chi^2_{(1)}$  distribution if the null hypothesis is true. The test statistic value in this case is  $34 \cdot 0.16 = 5.44$ . We reject the null hypothesis and conclude that there is significant autocorrelation.

## 9.5 AUTOREGRESSIVE MODELS

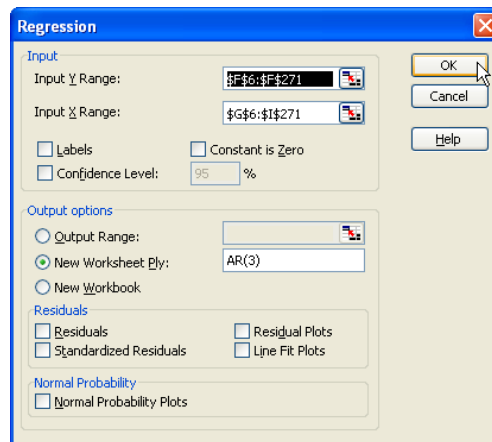
In addition to incorporating lagged values of error term into a regression equation, we can add lagged values of dependent and/or independent variables. Autoregressive models include lags of the dependent variable as regressors. The  $AR(p)$  model has  $p$  lags of  $y_t$  as regressors.

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \dots + \theta_p y_{t-p} + v_t$$

We will use the *inflation.xls* to estimate an  $AR(3)$  model of inflation rate. Open the data file and add three columns and label them *INFLAG1*, *INFLAG2* and *INFLAG3*. Copy the *INFLN* column and paste it to *INFLAG1* starting from cell G4, to *INFLAG2* starting from cell H5 and finally to *INFLAG3* starting from cell I6 as shown below.

|   | A    | B     | C    | D        | E     | F        | G        | H        | I        |
|---|------|-------|------|----------|-------|----------|----------|----------|----------|
| 1 | YEAR | MONTH | WAGE | PCWAGE   | CPI   | INFLN    | INFLAG1  | INFLAG2  | INFLAG3  |
| 2 | 1983 | 12    | 8.32 |          | 101.4 |          |          |          |          |
| 3 | 1984 | 1     | 8.37 | 0.599163 | 102.1 | 0.687963 |          |          |          |
| 4 | 1984 | 2     | 8.36 | -0.11955 | 102.6 | 0.488521 | 0.687963 |          |          |
| 5 | 1984 | 3     | 8.4  | 0.477328 | 102.9 | 0.291971 | 0.488521 | 0.687963 |          |
| 6 | 1984 | 4     | 8.44 | 0.47506  | 103.3 | 0.387973 | 0.291971 | 0.488521 | 0.687963 |
| 7 | 1984 | 5     | 8.43 | -0.11855 | 103.5 | 0.193424 | 0.387973 | 0.291971 | 0.488521 |
| 8 | 1984 | 6     | 8.47 | 0.473374 | 103.7 | 0.19305  | 0.193424 | 0.387973 | 0.291971 |

Then, estimate the model using *INFLN* as the **Y-Range** and *INFLAG1* *INFLAG2* and *INFLAG3* as the **X-Range** variables. Since there are missing observations for each column, make sure to start from row 6, the first row for which there are no missing observations.



The result is:

|    | A                            | B                   | C                     | D             |
|----|------------------------------|---------------------|-----------------------|---------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |
| 2  |                              |                     |                       |               |
| 3  | <i>Regression Statistics</i> |                     |                       |               |
| 4  | Multiple R                   | 0.35957599          |                       |               |
| 5  | R Square                     | 0.129294893         |                       |               |
| 6  | Adjusted R Square            | 0.119324987         |                       |               |
| 7  | Standard Error               | 0.197246727         |                       |               |
| 8  | Observations                 | 266                 |                       |               |
| 9  |                              |                     |                       |               |
| 10 | ANOVA                        |                     |                       |               |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression                   | 3                   | 1.513669915           | 0.504556638   |
| 13 | Residual                     | 262                 | 10.19344305           | 0.038906271   |
| 14 | Total                        | 265                 | 11.70711296           |               |
| 15 |                              |                     |                       |               |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept                    | 0.188335077         | 0.025290482           | 7.44687586    |
| 18 | INF_LAG1                     | 0.373292594         | 0.06148081            | 6.071692822   |
| 19 | INF_LAG2                     | -0.21791892         | 0.064472475           | -3.380030331  |
| 20 | INF_LAG3                     | 0.101254114         | 0.061267998           | 1.65264277    |

## 9.6 FINITE DISTRIBUTED LAG MODELS

Finite distributed lag models contain independent variables and their lags as regressors.

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \dots + \beta_q x_{t-q} + v_t$$

Using the *inflation.xls*, let's model the inflation rate as a function of the percentage change in wages and three lags of wages. Open the original file and add three columns. Label them *pcwage\_L1*, *pcwage\_L2* and *pcwage\_L3*. Copy the *PCWAGE* column and paste it to *pcwage\_L1* starting from cell F4, to *pcwage\_L2* starting from cell G5 and finally to *pcwage\_L3* starting from cell H6 as shown below.

|   | A    | B     | C     | D        | E        | F         | G         | H         |
|---|------|-------|-------|----------|----------|-----------|-----------|-----------|
| 1 | YEAR | MONTH | CPI   | INFLN    | PCWAGE   | pcwage_L1 | pcwage_L2 | pcwage_L3 |
| 2 | 1983 | 12    | 101.4 |          |          |           |           |           |
| 3 | 1984 | 1     | 102.1 | 0.687963 | 0.599163 |           |           |           |
| 4 | 1984 | 2     | 102.6 | 0.488521 | -0.11955 | 0.599163  |           |           |
| 5 | 1984 | 3     | 102.9 | 0.291971 | 0.477328 | -0.119546 | 0.599163  |           |
| 6 | 1984 | 4     | 103.3 | 0.387973 | 0.47506  | 0.477328  | -0.119546 | 0.599163  |
| 7 | 1984 | 5     | 103.5 | 0.193424 | -0.11855 | 0.47506   | 0.477328  | -0.119546 |
| 8 | 1984 | 6     | 103.7 | 0.19305  | 0.473374 | -0.118554 | 0.47506   | 0.477328  |
| 9 | 1984 | 7     | 104.1 | 0.384986 | 0.471143 | 0.473374  | -0.118554 | 0.47506   |

Observe that we now have  $T - n$  complete observations since we lose observations when we create the lags. Now run the regression with  $T - n$  observations.

The least squares results are:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.22007857          |                       |               |                |
| 5  | R Square                     | 0.048434577         |                       |               |                |
| 6  | Adjusted R Square            | 0.033851199         |                       |               |                |
| 7  | Standard Error               | 0.206596984         |                       |               |                |
| 8  | Observations                 | 266                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 4                   | 0.567029064           | 0.141757266   | 3.321217936    |
| 13 | Residual                     | 261                 | 11.1400839            | 0.042682314   |                |
| 14 | Total                        | 265                 | 11.70711296           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.121873261         | 0.048654677           | 2.504862205   | 0.012860034    |
| 18 | pcwage                       | 0.156088849         | 0.088501629           | 1.763683336   | 0.078955262    |
| 19 | L1                           | 0.10749781          | 0.085055032           | 1.263861846   | 0.207407246    |
| 20 | L2                           | 0.049485287         | 0.08525777            | 0.580419672   | 0.562132418    |
| 21 | L3                           | 0.199014534         | 0.087885431           | 2.264476969   | 0.024365444    |



## 9.6 AUTOREGRESSIVE DISTRIBUTED LAG MODELS (ARDL)

Finally, we consider a model that contains both finite distributed lags and is autoregressive.

$$y_t = \delta + \delta_0 x_t + \delta_1 x_{t-1} + \dots + \delta_q x_{t-q} + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \dots + \theta_p y_{t-p} + v_t$$

ARDL ( $p, q$ ) model has  $p$  lags of the dependent variable,  $y_t$  and  $q$  lags of the independent variable  $x_t$ . Let's illustrate ARDL (2,3) model for the inflation data. Modify the *inflation.xls* to incorporate the two lags of *INFLN* and three lags of *PCWAGE*

|    | A    | B     | C     | D        | E        | F         | G         | H         | I        | J        |
|----|------|-------|-------|----------|----------|-----------|-----------|-----------|----------|----------|
| 1  | YEAR | MONTH | CPI   | INFLN    | PCWAGE   | pcwage_L1 | pcwage_L2 | pcwage_L3 | inf_L1   | inf_L2   |
| 2  | 1983 | 12    | 101.4 |          |          |           |           |           |          |          |
| 3  | 1984 | 1     | 102.1 | 0.687963 | 0.599163 |           |           |           |          |          |
| 4  | 1984 | 2     | 102.6 | 0.488521 | -0.11955 | 0.599163  |           |           | 0.687963 |          |
| 5  | 1984 | 3     | 102.9 | 0.291971 | 0.477328 | -0.119546 | 0.599163  |           | 0.488521 | 0.687963 |
| 6  | 1984 | 4     | 103.3 | 0.387973 | 0.47506  | 0.477328  | -0.119546 | 0.599163  | 0.291971 | 0.488521 |
| 7  | 1984 | 5     | 103.5 | 0.193424 | -0.11855 | 0.47506   | 0.477328  | -0.119546 | 0.387973 | 0.291971 |
| 8  | 1984 | 6     | 103.7 | 0.19305  | 0.473374 | -0.118554 | 0.47506   | 0.477328  | 0.193424 | 0.387973 |
| 9  | 1984 | 7     | 104.1 | 0.384986 | 0.471143 | 0.473374  | -0.118554 | 0.47506   | 0.19305  | 0.193424 |
| 10 | 1984 | 8     | 104.4 | 0.28777  | 0        | 0.471143  | 0.473374  | -0.118554 | 0.384986 | 0.19305  |
| 11 | 1984 | 9     | 104.7 | 0.286944 | 0.468934 | 0         | 0.471143  | 0.473374  | 0.28777  | 0.384986 |

Observe the missing observations due to lags. Run the regression starting from the 6<sup>th</sup> observation using the *INFLN* as **Y-Range** and *PCWAGE*, its three lags, and two lags of *INFLN* as the **X-Range**

**Regression**

**Input**

Input Y Range:

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level:  %

**Output options**

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

**Residuals**

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

The least squares results are:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.40904996          |                       |               |                |
| 5  | R Square                     | 0.16732187          |                       |               |                |
| 6  | Adjusted R Square            | 0.148032029         |                       |               |                |
| 7  | Standard Error               | 0.194005303         |                       |               |                |
| 8  | Observations                 | 266                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 6                   | 1.958856029           | 0.326476005   | 8.674092797    |
| 13 | Residual                     | 259                 | 9.748256934           | 0.037638058   |                |
| 14 | Total                        | 265                 | 11.70711296           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.098876586         | 0.046806866           | 2.112437633   | 0.0356062      |
| 18 | PCWAGE                       | 0.114903242         | 0.083390039           | 1.37790129    | 0.169423076    |
| 19 | WL1                          | 0.037733717         | 0.081245468           | 0.464440889   | 0.642722064    |
| 20 | WL2                          | 0.059274631         | 0.081173529           | 0.730221197   | 0.465914956    |
| 21 | WL3                          | 0.236130508         | 0.082944072           | 2.846864185   | 0.004768882    |
| 22 | INFL1                        | 0.353640184         | 0.06041118            | 5.853886373   | 1.44819E-08    |
| 23 | INFL2                        | -0.197561316        | 0.06042121            | -3.269734524  | 0.001222346    |

# CHAPTER 10

## Random Regressors and Moment-Based Estimation

### CHAPTER OUTLINE

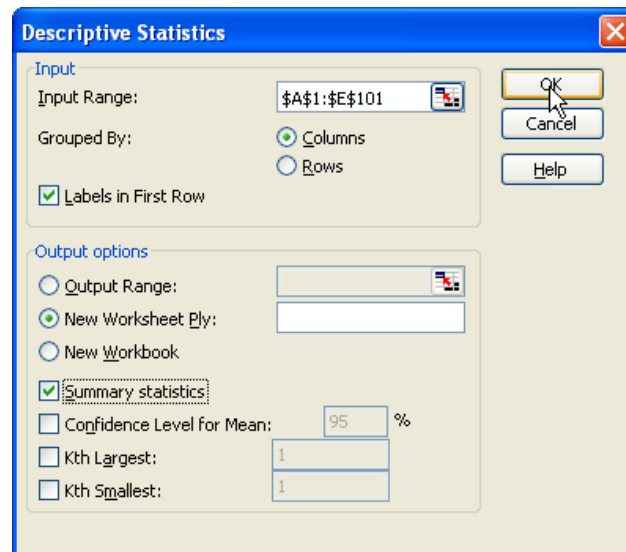
- 10.1 Least Squares with Simulated Data
- 10.2 Instrumental Variables Estimation with Simulated Data
  - 10.2.1 Correction of IV standard errors
  - 10.2.2 Corrected standard errors for simulated data
- 10.3 The Hausman Test: Simulated Data
- 10.4 Testing for Weak Instruments: Simulated Data
- 10.5 Testing for Validity of Surplus Instruments: Simulated Data
- 10.6 Estimation using Mroz Data
  - 10.6.1 Least squares regression
  - 10.6.2 Two-stage least squares
- 10.7 Testing for Endogeneity of Education
- 10.8 Testing for Weak Instruments
- 10.9 Testing for Validity of Surplus Instruments

### 10.1 LEAST SQUARES WITH SIMULATED DATA

When the explanatory variables are random, the relationship between  $x$  and the error term,  $e$ , is crucial in deciding whether ordinary least squares estimation is appropriate. If  $x$  and  $e$  are uncorrelated, then least squares can, and should, be used. However, if the  $\text{cov}(x, e) \neq 0$ , then the least squares estimator is biased and inconsistent. In this case, instrumental variables (IV) / two-stage least squares estimation process gives us a consistent but inefficient estimator. IV is not directly available in Excel as a built-in function. However, we will show that it is easy to perform IV estimation using Excel functions you are already familiar with.

We will use *ch10.xls* to explore the properties of the least squares estimator when  $\text{cov}(x, e) \neq 0$ . The data set contains simulated data of sample size 100,  $\rho_{xe} = 0.6$ , where  $E(y) = \beta_1 + \beta_2 x = 1 + 1x$  and  $x_i, e_i \sim N(0,1)$ .

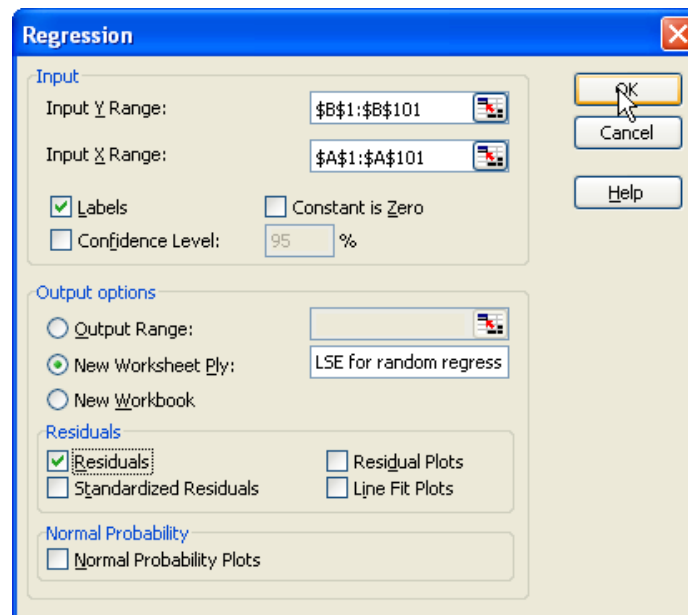
First, get the summary statistics for the dataset where  $z_1$ ,  $z_2$ , and  $z_3$  are the instrumental variables we will consider later. Go to **Tools>Data Analysis>Descriptive Statistics**



Below is a sample of the descriptive statistics provided by Excel.

|   | A         | B           | C         | D           | E         | F           | G         | H           | I         | J           |
|---|-----------|-------------|-----------|-------------|-----------|-------------|-----------|-------------|-----------|-------------|
| 1 | x         |             | y         |             | z1        |             | z2        |             | z3        |             |
| 2 |           |             |           |             |           |             |           |             |           |             |
| 3 | Mean      | 0.2391607   | Mean      | 1.3862871   | Mean      | 0.0342066   | Mean      | 0.1205662   | Mean      | 0.062191    |
| 4 | Std. Dev. | 0.956655159 | Std. Dev. | 1.838818655 | Std. Dev. | 0.893107752 | Std. Dev. | 1.027655692 | Std. Dev. | 1.100069225 |
| 5 | Minimum   | -1.6484     | Minimum   | -2.96671    | Minimum   | -2.61576    | Minimum   | -2.29936    | Minimum   | -2.98265    |
| 6 | Maximum   | 2.76628     | Maximum   | 6.72735     | Maximum   | 2.09323     | Maximum   | 2.10895     | Maximum   | 3.1457      |
| 7 | Count     | 100         | Count     | 100         | Count     | 100         | Count     | 100         | Count     | 100         |

Next estimate the simple regression of  $y$  on  $x$  using  $y$  as the **Y-Range** and  $x$  as the **X-Range**. Choose the **Residuals** option for future reference and place the output in a worksheet named **LSE for random regress**. Click **OK**.



The results are:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.88621922          |                       |               |                |
| 5  | R Square                     | 0.785384505         |                       |               |                |
| 6  | Adjusted R Square            | 0.783194551         |                       |               |                |
| 7  | Standard Error               | 0.856197583         |                       |               |                |
| 8  | Observations                 | 100                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 1                   | 262.9028692           | 262.9028692   | 358.6305905    |
| 13 | Residual                     | 98                  | 71.84128144           | 0.7230743     |                |
| 14 | Total                        | 99                  | 334.7441507           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.978893255         | 0.088280975           | 11.08838296   | 5.35253E-19    |
| 18 | x                            | 1.703431396         | 0.089949962           | 18.93754447   | 1.61396E-34    |

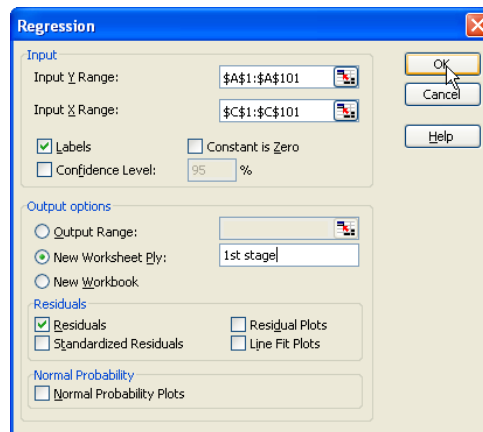
Notice that the estimated slope is 1.7, when in fact the true slope of the artificial data is 1. These results are due to the correlation between the error term and explanatory variable that has been built into the data.

## 10.2 INSTRUMENTAL VARIABLES ESTIMATION WITH SIMULATED DATA

*Ch10.xls* also contains two instrumental variables both of which are correlated to  $x$  yet uncorrelated with the error term. Since our problem is that the explanatory variable we are using is correlated with the error term, using these instrumental variables should help solve our problem.

We will estimate a **reduced form** or **1<sup>st</sup> stage** regression  $x_i = \alpha_1 + \alpha_2 z_{1i} + e_i$  to obtain the predicted values of  $x_i$ ,  $\hat{x}_i$ . Then, we use  $\hat{x}_i$  as an instrumental variable in the equation  $y_i = \beta_1 + \beta_2 \hat{x}_i + e_i$ . Estimation of this equation is the "**2<sup>nd</sup> stage**". Thus the name, **2-stage least squares**.

Return to the worksheet containing the simulated data, *ch10.xls*. Run a regression, using  $x$  as the **Y-Range** and  $z$  as the **X-Range**. Choose the **Residuals** option to obtain  $\hat{x}_i$ . Place output on a worksheet names **1<sup>st</sup> stage**.



The "*Predicted x*" from the RESIDUAL OUTPUT is our instrumental variable.

|    | A                            | B                   | C                     | D             |
|----|------------------------------|---------------------|-----------------------|---------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |
| 2  |                              |                     |                       |               |
| 3  | <i>Regression Statistics</i> |                     |                       |               |
| 4  | Multiple R                   | 0.533152714         |                       |               |
| 5  | R Square                     | 0.284251816         |                       |               |
| 6  | Adjusted R Square            | 0.276948263         |                       |               |
| 7  | Standard Error               | 0.81346731          |                       |               |
| 8  | Observations                 | 100                 |                       |               |
| 9  |                              |                     |                       |               |
| 10 | ANOVA                        |                     |                       |               |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression                   | 1                   | 25.75427204           | 25.75427204   |
| 13 | Residual                     | 98                  | 64.84944826           | 0.661729064   |
| 14 | Total                        | 99                  | 90.60372029           |               |
| 15 |                              |                     |                       |               |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept                    | 0.219625715         | 0.081406977           | 2.69787337    |
| 18 | z1                           | 0.57108895          | 0.091541632           | 6.23856256    |
| 19 |                              |                     |                       |               |
| 20 |                              |                     |                       |               |
| 21 |                              |                     |                       |               |
| 22 | RESIDUAL OUTPUT              |                     |                       |               |
| 23 |                              |                     |                       |               |
| 24 | <i>Observation</i>           | <i>Predicted x</i>  | <i>Residuals</i>      |               |
| 25 | 1                            | 0.597543328         | -1.279973328          |               |
| 26 | 2                            | 0.677518519         | 1.804681481           |               |

**Copy** cells B24 through B124 containing *Predicted x* over to the worksheet containing the original data. Now run another regression, using *y* as the **Y-Range** and *Predicted x* as the **X-Range**. Place output on a worksheet named **2<sup>nd</sup> stage**.

**Regression**

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☒ Residuals ☐ Residual Plots

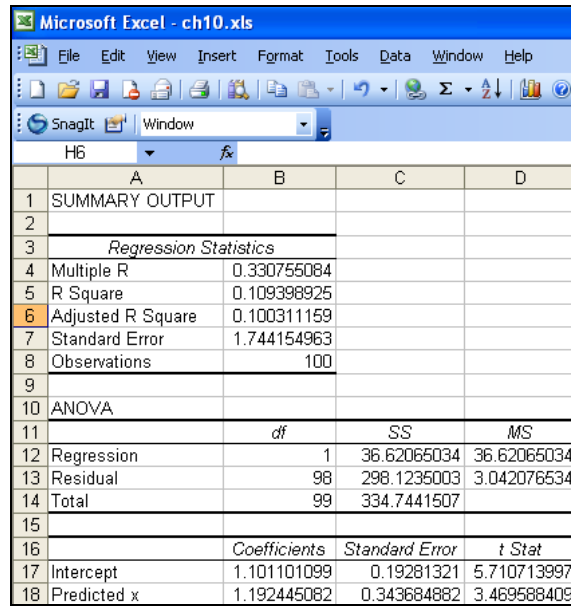
☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

The results of this Two-Stage Least Squares Estimation are:



The screenshot shows a Microsoft Excel window titled 'ch10.xls'. The active cell is H6. The table content is as follows:

|    | A                            | B                   | C                     | D             |
|----|------------------------------|---------------------|-----------------------|---------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |
| 2  |                              |                     |                       |               |
| 3  | <i>Regression Statistics</i> |                     |                       |               |
| 4  | Multiple R                   | 0.330755084         |                       |               |
| 5  | R Square                     | 0.109398925         |                       |               |
| 6  | Adjusted R Square            | 0.100311159         |                       |               |
| 7  | Standard Error               | 1.744154963         |                       |               |
| 8  | Observations                 | 100                 |                       |               |
| 9  |                              |                     |                       |               |
| 10 | <i>ANOVA</i>                 |                     |                       |               |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression                   | 1                   | 36.62065034           | 36.62065034   |
| 13 | Residual                     | 98                  | 298.1235003           | 3.042076534   |
| 14 | Total                        | 99                  | 334.7441507           |               |
| 15 |                              |                     |                       |               |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept                    | 1.101101099         | 0.19281321            | 5.710713997   |
| 18 | Predicted x                  | 1.192445082         | 0.343684882           | 3.469588409   |

The IV estimate of the slope 1.19 is closer to the true value of 1 but the standard errors are incorrect.

### 10.2.1 Correction of IV standard errors

In the simple linear regression model  $y_i = \beta_1 + \beta_2 x_i + e_i$  the 2SLS estimator is the least squares estimator applied to  $y_i = \beta_1 + \beta_2 \hat{x}_i + e_i$  where  $\hat{x}_i$  is the predicted value from a reduced form equation. So, the 2SLS estimators are

$$\hat{\beta}_2 = \frac{\sum (\hat{x}_i - \bar{x})(y_i - \bar{y})}{\sum (\hat{x}_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

In large samples the 2SLS estimators have approximate normal distributions. In the simple regression model

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum (\hat{x}_i - \bar{x})^2}\right)$$

The error variance  $\sigma^2$  should be estimated using the estimator

$$\hat{\sigma}_{2SLS}^2 = \frac{\sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2}{N - 2}$$

with the quantity in the numerator being the sum of squared 2SLS residuals, or  $SSE_{2SLS}$ . The problem with doing 2SLS with two least squares regressions is that in the second estimation the estimated variance is

$$\hat{\sigma}_{wrong}^2 = \frac{\sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 \hat{x}_i)^2}{N - 2}$$

The numerator is the  $SSE$  from the regression of  $y_i$  on  $\hat{x}_i$ , which is  $SSE_{wrong}$ .

Thus correct 2SLS standard error is

$$se(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}_{2SLS}^2}{\sum (\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}_{2SLS}^2}}{\sqrt{\sum (\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{2SLS}}{\sqrt{\sum (\hat{x}_i - \bar{x})^2}}$$

and the “wrong” standard error, calculated in the 2<sup>nd</sup> least squares estimation, is

$$se_{wrong}(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}_{wrong}^2}{\sum (\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}_{wrong}^2}}{\sqrt{\sum (\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{wrong}}{\sqrt{\sum (\hat{x}_i - \bar{x})^2}}$$

Given that we have the “wrong” standard error in the 2<sup>nd</sup> regression, we can adjust it using a correction factor

$$se(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}_{2SLS}^2}{\hat{\sigma}_{wrong}^2}} se_{wrong}(\hat{\beta}_2) = \frac{\hat{\sigma}_{2SLS}}{\hat{\sigma}_{wrong}} se_{wrong}(\hat{\beta}_2)$$

### 10.2.2 Corrected standard errors for simulated data

To correct the inflated standard errors and modify the  $t$ -statistic and  $p$ -value with the correct standard error, return to the worksheet containing the original data.

- **Insert** a new column and label it *residuals*= $y-b_1-b_2x$ .
- Calculate *residuals* using the coefficients from the 2SLS estimation. In other words, in cell G2, type **=y-1.101101099-1.192445082\*A2**, where A2 contains the first observation on  $x$ . **Copy** the formula down the column.
- Label cell H1 *sighat\_2sls*, cell H2 *sighat\_wrong*, and cell H3 *correction factor*.
- In cell I1, type **=SQRT(SUMSQ(G2:G101)/98)** where 98 is T-K. This cell calculates the square root of correct mean of squared errors which is the *sighat\_2sls*.
- In cell I2, type **=SQRT(3.042)**, where 3.042 is the wrong MSE and its square root it the *sighat\_wrong*.
- In cell I3, type **=I1/I2** which calculates the correction factor.



|   | A        | B        | C        | D        | E        | F            | G                      | H            | I                        | J      | K |
|---|----------|----------|----------|----------|----------|--------------|------------------------|--------------|--------------------------|--------|---|
| 1 | x        | y        | z1       | z2       | z3       | Predicted x  | residuals=y - b1 - b2x | sighat_2sls  | =SQRT(SUMSQ(G2:G101)/98) |        |   |
| 2 | -0.68243 | -0.55321 | 0.66175  | -1.03449 | -1.69849 | 0.597543328  | =B2-1.1011-1.1924*A2   | sighat_wrong | =SQRT(3.042)             |        |   |
| 3 | 2.4822   | 4.7472   | 0.80179  | 1.31019  | 2.44266  | 0.677518519  |                        | 0.68632472   |                          |        |   |
| 4 | 0.14934  | 1.53093  | -1.95637 | -1.1287  | 0.90907  | -0.897634098 |                        | 0.251756984  | correction factor        | =I1/I2 |   |
| 5 | 0.4729   | 1.83448  | -0.87638 | 0.44545  | 0.38953  | -0.280864558 |                        | 0.16949404   |                          |        |   |
| 6 | -1.08409 | -0.47212 | -0.3553  | -0.5262  | 0.34793  | 0.016718079  |                        | -0.280551084 |                          |        |   |

This is our correct model standard error.

|   | A        | B        | C        | D        | E        | F            | G                      | H            | I        |
|---|----------|----------|----------|----------|----------|--------------|------------------------|--------------|----------|
| 1 | x        | y        | z1       | z2       | z3       | Predicted x  | residuals=y - b1 - b2x | sighat_2sls  | 0.987177 |
| 2 | -0.68243 | -0.55321 | 0.66175  | -1.03449 | -1.69849 | 0.597543328  | -0.840580468           | sighat_wrong | 1.744133 |
| 3 | 2.4822   | 4.7472   | 0.80179  | 1.31019  | 2.44266  | 0.677518519  |                        | 0.68632472   | 0.565998 |
| 4 | 0.14934  | 1.53093  | -1.95637 | -1.1287  | 0.90907  | -0.897634098 |                        | 0.251756984  |          |
| 5 | 0.4729   | 1.83448  | -0.87638 | 0.44545  | 0.38953  | -0.280864558 |                        | 0.16949404   |          |

Copy the correction factor, return to the worksheet **2<sup>nd</sup> stage** and paste after the regression output. Then, create the corrected standard errors by multiplying the current standard errors with the correction factor. The correct *t*-stat will be obtained by dividing the coefficient estimates by the corrected standard errors. And you can calculate the correct *p*-value by typing **=TDIST(D20,98,2)** for a two-tailed test where **D20** is the corrected *t*-stat.

|                   | Coefficients | Standard Error           | t Stat           |
|-------------------|--------------|--------------------------|------------------|
| Intercept         | 1.101101099  | 0.19281321               | 5.710713997      |
| Predicted x       | 1.192445082  | 0.343684882              | 3.469588409      |
|                   |              |                          |                  |
|                   |              |                          |                  |
| correction factor | 0.56599844   |                          |                  |
|                   |              |                          |                  |
|                   | Coefficients | Corrected Standard Error | Corrected t Stat |
| Intercept         | 1.101101099  | =+B21*C17                | =B24/C24         |
| Predicted x       | 1.192445082  | =+C18*B21                | =B25/C25         |

The results will be:

|                   | Coefficients | Standard Error           | t Stat           |
|-------------------|--------------|--------------------------|------------------|
| Intercept         | 1.101101099  | 0.19281321               | 5.710713997      |
| Predicted x       | 1.192445082  | 0.343684882              | 3.469588409      |
|                   |              |                          |                  |
|                   |              |                          |                  |
| correction factor | 0.56599844   |                          |                  |
|                   |              |                          |                  |
|                   | Coefficients | Corrected Standard Error | Corrected t Stat |
| Intercept         | 1.101101099  | 0.109131976              | 10.08962851      |
| Predicted x       | 1.192445082  | 0.194525107              | 6.130031753      |

The correct *t*-stat and *p*-value are quite different from those reported originally by Excel.

### 10.3 THE HAUSMAN TEST: SIMULATED DATA

Since we don't always know if the explanatory variables are endogenous or not, we can empirically check it using the Hausman test. We want to test  $H_0: \text{cov}(x, e) = 0$  against the alternative  $H_1: \text{cov}(x, e) \neq 0$ . The Hausman Test is a formal test of these hypotheses and can be based on using the residuals from the 1<sup>st</sup> stage estimation of

$$x_i = \gamma_1 + \theta_1 z_{i1} + \theta_2 z_{i2} + v_i$$

Denote the least squares residuals from the reduced form as  $\hat{v}_i$ . Include them in an artificial regression

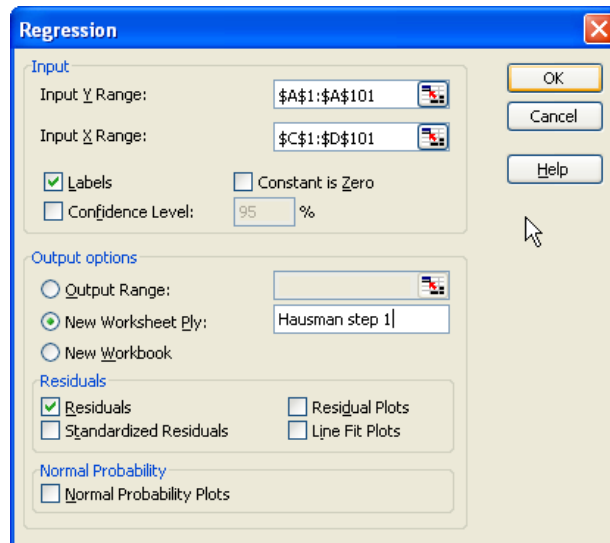
$$y_i = \beta_1 + \beta_2 x_i + \delta \hat{v}_i + e_i$$

Estimate this model by least squares and test the significance of  $\hat{v}_i$  using a standard  $t$ -test.

$$\begin{aligned} H_0 : \delta &= 0 && \text{no correlation between } x \text{ and } e \\ H_1 : \delta &\neq 0 && \text{correlation between } x \text{ and } e \end{aligned}$$

If the null hypothesis is true, the ordinary least squares estimators are more efficient and should be used. If the null hypothesis is not true, the instrumental variables estimator is consistent and should be used.

Return to the worksheet containing the data **ch10.xls**. Estimate a regression, using  $x$  as the **Y-Range** and  $z_1$  and  $z_2$  as the **X-Range**. Make sure to click the residual box to get the residuals. Place the results on a worksheet named **Hausman step 1**.



**Copy** cells C25 through C125 which contain the *Residuals* to the worksheet containing the original data and **Paste**. **Move** the columns  $x$  and *Residuals* together. Now, run the 2<sup>nd</sup> regression, using  $y$  as the **Y-Range** and  $x$  and *residuals* as the **X-Range**. Place the results on a worksheet named **Hausman step 2**. No need to store residuals for this step.

The regression output is:

|    | A                            | B                   | C                     | D             |
|----|------------------------------|---------------------|-----------------------|---------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |
| 2  |                              |                     |                       |               |
| 3  | <i>Regression Statistics</i> |                     |                       |               |
| 4  | Multiple R                   | 0.919262166         |                       |               |
| 5  | R Square                     | 0.84504293          |                       |               |
| 6  | Adjusted R Square            | 0.841847939         |                       |               |
| 7  | Standard Error               | 0.731267595         |                       |               |
| 8  | Observations                 | 100                 |                       |               |
| 9  |                              |                     |                       |               |
| 10 | ANOVA                        |                     |                       |               |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression                   | 2                   | 282.873178            | 141.436589    |
| 13 | Residual                     | 97                  | 51.87097263           | 0.534752296   |
| 14 | Total                        | 99                  | 334.7441507           |               |
| 15 |                              |                     |                       |               |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept                    | 1.137590589         | 0.079746437           | 14.26509615   |
| 18 | Residuals                    | 0.995728288         | 0.162938901           | 6.111053161   |
| 19 | x                            | 1.039871983         | 0.133013055           | 7.817818964   |

Based on the  $t$ -test of the coefficient on *Residuals*, we reject the null hypothesis of no correlation between  $x$  and  $e$ , and conclude that instrumental variable estimation is the procedure in this case.

## 10.4 TESTING FOR WEAK INSTRUMENTS: SIMULATED DATA

Instrumental variables must be as strongly correlated with the endogenous variable as possible. A standard rule of thumb is that they will at least have a  $t$ -statistic of 3.3 or  $F$ -statistic of 10 in the reduced form regression.

|    | A                     | B                   | C                     | D             |
|----|-----------------------|---------------------|-----------------------|---------------|
| 1  | SUMMARY OUTPUT        |                     |                       |               |
| 2  |                       |                     |                       |               |
| 3  | Regression Statistics |                     |                       |               |
| 4  | Multiple R            | 0.2246167           |                       |               |
| 5  | R Square              | 0.050452662         |                       |               |
| 6  | Adjusted R Square     | 0.040763404         |                       |               |
| 7  | Standard Error        | 0.936954039         |                       |               |
| 8  | Observations          | 100                 |                       |               |
| 9  |                       |                     |                       |               |
| 10 | ANOVA                 |                     |                       |               |
| 11 |                       | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression            | 1                   | 4.571198885           | 4.571198885   |
| 13 | Residual              | 98                  | 86.03252141           | 0.877882871   |
| 14 | Total                 | 99                  | 90.60372029           |               |
| 15 |                       |                     |                       |               |
| 16 |                       | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept             | 0.213950552         | 0.094344497           | 2.267758666   |
| 18 | z2                    | 0.209097976         | 0.091633243           | 2.281900853   |

From the above output,  $t$ -statistic for the second instrumental variable is 2.28. Since the  $t$ -statistics less than the rule of thumb of 3.3, usage of this instrumental variable may not yield satisfactory results.

## **10.5 TESTING THE VALIDITY OF SURPLUS INSTRUMENTS: SIMULATED DATA**

In addition to being strongly correlated with the endogenous variable, a “good” instrument needs to be uncorrelated with the error term. We can test the validity of the instrument by an  $LM$  test.  $LM$  test uses 2SLS residual  $\hat{e}_{2SLS}$ , as the dependent variable, and all the available exogenous and instrumental variables as independent variables.  $LM$  statistic is calculated from this regression by  $LM = NR^2$  which has  $\chi^2_{(k)}$  where  $k$  is the number of surplus instrumental variables. Estimate an auxiliary regression using the  $\hat{e}_{2SLS}$  from the 2<sup>nd</sup> step of 2SLS estimation,  $y_i = \beta_1 + \beta_2 \hat{x}_i + e_i$  as the dependent variables and use instruments as the independent variable. The output is:

## **10.6 ESTIMATION USING THE MROZ DATA**

We will use the Mroz data to illustrate the endogenous variables with real data. Open file *mroz.xls*. This is a log-linear wage model on working woman.

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

### **10.6.1 Least squares regression**

The data set includes observations for non-working women. To eliminate the non-working woman, use the variable *lfp* (labor force participation). Delete the rows where *lfp* = 0. Next, create

$lwage = \ln(wage)$  and  $exper2 = exper^2$ . Next, run the least squares regression model using the  $lwage$  as **Y-Range** and  $educ$ ,  $exper$ ,  $exper2$  as **X-Range**. Include the labels, name the worksheet **LSE** and check the residual box. The results are:

|    | A                     | B                   | C                     | D             |
|----|-----------------------|---------------------|-----------------------|---------------|
| 1  | SUMMARY OUTPUT        |                     |                       |               |
| 2  |                       |                     |                       |               |
| 3  | Regression Statistics |                     |                       |               |
| 4  | Multiple R            | 0.396005544         |                       |               |
| 5  | R Square              | 0.156820391         |                       |               |
| 6  | Adjusted R Square     | 0.150854498         |                       |               |
| 7  | Standard Error        | 0.666420217         |                       |               |
| 8  | Observations          | 428                 |                       |               |
| 9  |                       |                     |                       |               |
| 10 | ANOVA                 |                     |                       |               |
| 11 |                       | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression            | 3                   | 35.02229655           | 11.67409885   |
| 13 | Residual              | 424                 | 188.3051439           | 0.444115905   |
| 14 | Total                 | 427                 | 223.3274405           |               |
| 15 |                       |                     |                       |               |
| 16 |                       | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept             | -0.522040561        | 0.198632066           | -2.628178678  |
| 18 | educ                  | 0.107489639         | 0.014143478           | 7.598332013   |
| 19 | exper                 | 0.041566511         | 0.013175198           | 3.154906015   |
| 20 | exper2                | -0.000811193        | 0.000393242           | -2.062833684  |

### 10.6.2 Two-stage least squares

In this example,  $educ$  is considered the endogenous variable as it could be correlated with factors in the regression error term such as ability. In the 1<sup>st</sup> stage of the 2SLS estimation, we will estimate the first one by using the  $educ$  as **Y-Range** and  $exper$ ,  $exper2$  and  $mothereduc$  (the instrumental variable) as the **X-Range**. Keep the predicted values from the first stage to be used in the 2<sup>nd</sup> stage.

|    | A                     | B                   | C                     | D             |
|----|-----------------------|---------------------|-----------------------|---------------|
| 3  | Regression Statistics |                     |                       |               |
| 4  | Multiple R            | 0.390760937         |                       |               |
| 5  | R Square              | 0.15269411          |                       |               |
| 6  | Adjusted R Square     | 0.146699021         |                       |               |
| 7  | Standard Error        | 2.111099613         |                       |               |
| 8  | Observations          | 428                 |                       |               |
| 9  |                       |                     |                       |               |
| 10 | ANOVA                 |                     |                       |               |
| 11 |                       | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression            | 3                   | 340.5378336           | 113.5126112   |
| 13 | Residual              | 424                 | 1889.658428           | 4.456741576   |
| 14 | Total                 | 427                 | 2230.196262           |               |
| 15 |                       |                     |                       |               |
| 16 |                       | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept             | 9.77510269          | 0.423888615           | 23.06054547   |
| 18 | exper                 | 0.0488615           | 0.04166926            | 1.172603007   |
| 19 | exper2                | -0.001281065        | 0.001244906           | -1.029045855  |
| 20 | mothereduc            | 0.267690809         | 0.031129797           | 8.599182717   |

Observe that the  $mothereduc$  is a good instrument with a  $t$ -statistic of approximately 8.6 which is greater than the rule of thumb of 3.3

Estimate the 2<sup>nd</sup> stage regression by using the *lwage* as **Y-Range** and *exper*, *exper2* and predicted values of *educ* (from the previous regression) as **X-Range**.

|    | A                            | B                   | C                     | D             |
|----|------------------------------|---------------------|-----------------------|---------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |
| 2  |                              |                     |                       |               |
| 3  | <i>Regression Statistics</i> |                     |                       |               |
| 4  | Multiple R                   | 0.21351507          |                       |               |
| 5  | R Square                     | 0.045588685         |                       |               |
| 6  | Adjusted R Square            | 0.038835775         |                       |               |
| 7  | Standard Error               | 0.709015788         |                       |               |
| 8  | Observations                 | 428                 |                       |               |
| 9  |                              |                     |                       |               |
| 10 | ANOVA                        |                     |                       |               |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     |
| 12 | Regression                   | 3                   | 10.18120435           | 3.393734782   |
| 13 | Residual                     | 424                 | 213.1462361           | 0.502703387   |
| 14 | Total                        | 427                 | 223.3274405           |               |
| 15 |                              |                     |                       |               |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> |
| 17 | Intercept                    | 0.198186076         | 0.493342657           | 0.401720941   |
| 18 | exper                        | 0.044855849         | 0.014164402           | 3.166801513   |
| 19 | exper2                       | -0.000922076        | 0.000423969           | -2.174867605  |
| 20 | Predicted educ               | 0.049262951         | 0.039056201           | 1.261334918   |

Do not forget to modify the standard errors, *t*-statistics and the *p*-values are described in the previous section. First we need to modify the model standard error.

| <i>lwage</i> | <i>wage</i> | <i>educ</i> | <i>mothereduc</i> | <i>exper</i> | <i>exper2</i> | <i>fathereduc</i> | <i>residuals= y-b1+b2exper+b3exper2+b4educ</i> | <i>sighat_2sls</i>       | 0.678803546 |
|--------------|-------------|-------------|-------------------|--------------|---------------|-------------------|------------------------------------------------|--------------------------|-------------|
| 1.210154     | 3.354       | 12          | 12                | 14           | 196           | 7                 | -0.026442777                                   | <i>sighat_wrong</i>      | 0.709015515 |
| 0.328512     | 1.389       | 12          | 7                 | 5            | 25            | 7                 | -0.66205676                                    | <i>correction factor</i> | 0.957388847 |
| 1.514138     | 4.546       | 12          | 12                | 15           | 225           | 7                 | 0.259425652                                    |                          |             |
| 0.092123     | 1.097       | 12          | 7                 | 6            | 36            | 7                 | -0.933158549                                   |                          |             |
| 1.524272     | 4.592       | 14          | 12                | 7            | 49            | 14                | 0.367595506                                    |                          |             |

Then, apply the new standard error to the coefficient standard errors, *t*-statistics and *p*-values.

|                          | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
|--------------------------|---------------------|-----------------------|---------------|----------------|
| Intercept                | 0.198186076         | 0.493342657           | 0.401720941   | 0.688091817    |
| exper                    | 0.044855849         | 0.014164402           | 3.166801513   | 0.001652599    |
| exper2                   | -0.000922076        | 0.000423969           | -2.174867605  | 0.030192088    |
| Predicted educ           | 0.049262951         | 0.039056201           | 1.261334918   | 0.207881726    |
| <b>correction factor</b> | <b>0.957388847</b>  |                       |               |                |
|                          |                     |                       |               |                |
|                          |                     |                       |               |                |
|                          | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept                | 0.198186076         | 0.472320758           | 0.419600606   | 0.688091817    |
| exper                    | 0.044855849         | 0.01356084            | 3.307748489   | 0.001652599    |
| exper2                   | -0.000922076        | 0.000405903           | -2.271665909  | 0.030192088    |
| Predicted educ           | 0.049262951         | 0.037391972           | 1.317474004   | 0.207881726    |

## 10.7 TESTING THE ENDOGENEITY OF EDUCATION

We will use Hausman test to test whether *educ* is endogenous and correlated to the regression error term. First run the 1<sup>st</sup> stage regression using *educ* as the **Y-Range** and all other explanatory variables as the **X-Range**. We only need the residuals from the 1<sup>st</sup> step to use as an additional explanatory variable in the 2<sup>nd</sup> stage. For the second stage, use the *lwage* as the **Y-Range** and *educ*, *exper*, *exper2*, and residuals as **X-Range**.

|    | A                     | B            | C              | D           |
|----|-----------------------|--------------|----------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |             |
| 2  |                       |              |                |             |
| 3  | Regression Statistics |              |                |             |
| 4  | Multiple R            | 0.402927346  |                |             |
| 5  | R Square              | 0.162350446  |                |             |
| 6  | Adjusted R Square     | 0.15442941   |                |             |
| 7  | Standard Error        | 0.665015919  |                |             |
| 8  | Observations          | 428          |                |             |
| 9  |                       |              |                |             |
| 10 | ANOVA                 |              |                |             |
| 11 |                       | df           | SS             | MS          |
| 12 | Regression            | 4            | 36.25730963    | 9.064327408 |
| 13 | Residual              | 423          | 187.0701308    | 0.442246172 |
| 14 | Total                 | 427          | 223.3274405    |             |
| 15 |                       |              |                |             |
| 16 |                       | Coefficients | Standard Error | t Stat      |
| 17 | Intercept             | 0.048100303  | 0.394575257    | 0.121904001 |
| 18 | educ                  | 0.061396628  | 0.030984942    | 1.981498884 |
| 19 | exper                 | 0.044170394  | 0.013239447    | 3.336271785 |
| 20 | exper2                | -0.00089897  | 0.000395913    | -2.27062255 |
| 21 | Residuals -hausman    | 0.058166612  | 0.034807276    | 1.671104998 |

Based on the *t*-test of the coefficient on *Residuals*, we fail to reject the null hypothesis at 5% level of no correlation between *x* and *e*, and conclude that OLS estimation is appropriate. However, *educ* is endogenous at 10% level of significance.

## 10.8 TESTING FOR WEAK INSTRUMENTS

To test whether the relationship between the instruments and *educ* is strong enough, estimate the restricted and unrestricted reduced form equations and test the joint significance of the instruments. The unrestricted and restricted models have the following ANOVA tables:

|    |            |              |                |              |             |
|----|------------|--------------|----------------|--------------|-------------|
| 10 | ANOVA      |              |                |              |             |
| 11 |            | df           | SS             | MS           | F           |
| 12 | Regression | 4            | 471.6203982    | 117.9052496  | 28.36041288 |
| 13 | Residual   | 423          | 1758.575263    | 4.15738833   |             |
| 14 | Total      | 427          | 2230.196262    |              |             |
| 15 |            |              |                |              |             |
| 16 |            | Coefficients | Standard Error | t Stat       | P-value     |
| 17 | Intercept  | 9.10264011   | 0.426561367    | 21.33957927  | 4.09847E-69 |
| 18 | mothereduc | 0.157597033  | 0.035894116    | 4.390609167  | 1.42964E-05 |
| 19 | exper      | 0.045225423  | 0.040250712    | 1.123593117  | 0.261822938 |
| 20 | exper2     | -0.001009091 | 0.001202445    | -0.838571743 | 0.402183285 |
| 21 | fathereduc | 0.18954841   | 0.033756467    | 5.615173276  | 3.56151E-08 |

| ANOVA      |                     |                       |               |                |
|------------|---------------------|-----------------------|---------------|----------------|
|            | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| Regression | 2                   | 10.97967328           | 5.489936639   | 1.051372495    |
| Residual   | 425                 | 2219.216368           | 5.22168562    |                |
| Total      | 427                 | 2230.196262           |               |                |
|            |                     |                       |               |                |
|            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept  | 12.369356           | 0.322313045           | 38.37684003   | 3.5283E-140    |
| exper      | 0.056491946         | 0.045093499           | 1.252773616   | 0.210976963    |
| exper2     | -0.001904334        | 0.001345228           | -1.415622479  | 0.157618444    |

Using the MSE from the above models, to test the instruments jointly, we can conduct an  $F$ -test where the unrestricted model has all exogenous variables and the restricted model has *exper* and *exper2* as the explanatory variables. Recall that the rule-of-thumb threshold value for adequate instruments is an  $F$ -value of 10.0.

|    | A                                  | B           |
|----|------------------------------------|-------------|
| 1  | <b>Hypothesis Testing - F-Test</b> |             |
| 2  |                                    |             |
| 3  | <b>Data Input</b>                  |             |
| 4  | J                                  | 2           |
| 5  | N                                  | 428         |
| 6  | K                                  | 5           |
| 7  | SSE-RESTRICTED                     | 5.22168562  |
| 8  | SSE-UNRESTRICTED                   | 4.15738833  |
| 9  | ALPHA                              | 0.05        |
| 10 |                                    |             |
| 11 | <b>Computed Values</b>             |             |
| 12 | df-numerator                       | 2           |
| 13 | df-denominator                     | 423         |
| 14 | F                                  | 54.1443     |
| 15 | Right Critical value               | 3.017048903 |
| 16 | Decision                           | Reject Null |
| 17 | p-value                            | 1.15768E-21 |

## 10.9 TESTING THE VALIDITY OF SURPLUS INSTRUMENTS

We can check the validity of the surplus instruments using the  $LM$  test. For this purpose, we need to run an auxiliary regression where the residuals are the **Y-Range** and all the exogenous variables are the **X-Range**. We calculate the  $LM$  test statistic as  $N \cdot R^2$  which is 0.3780714 for this example.



# CHAPTER 11

## Simultaneous Equations Models

### CHAPTER OUTLINE

- 11.1 Truffle Supply and Demand
- 11.2 Estimating the Reduced Form Equations
- 11.3 2SLS Estimates of Truffle Demand and Supply
  - 11.3.1 Correction of 2SLS standard errors
  - 11.3.2 Corrected standard errors in truffle demand and supply
- 11.4 Supply and Demand of Fish
- 11.5 Reduced Forms for Fish Price and Quantity

In this chapter, we estimate **simultaneous equation** models where there are two or more dependent variables that need to be estimated jointly. Ordinary least squares estimation is not possible when we are dealing with more than one equation. For example to explain both price and quantity of a good, we need both supply and demand equations which work together to determine price and quantity *jointly*.

### 11.1 TRUFFLE SUPPLY AND DEMAND

Consider the supply and demand for truffles:

$$\text{Demand: } Q_i = \alpha_1 + \alpha_2 P_i + \alpha_3 PS_i + \alpha_4 DI_i + e_i^d$$

$$\text{Supply: } Q_i = \beta_1 + \beta_2 P_i + \beta_3 PF_i + e_i^s$$

In the demand equation,  $Q$  is the quantity of truffles traded at time (in ounces)  $P$  is the market price for truffles in dollars per ounce,  $PS$  is the market price for substitutes for truffles in dollars per pound, and  $DI$  is the per capita disposable income, in thousands of dollars. In the supply equation,  $PF$  is the price of the factor of production.  $P$  and  $Q$  are endogenous variables meaning their values are determined within the system of equations. The equilibrium levels of price and

quantity,  $P^*$  and  $Q^*$  are determined by both of these equations.  $PS$ ,  $DI$ ,  $PF$  are exogenous meaning that we take the values as given.

Open *truffles.xls* and obtain the summary statistics. Part of the print out is given below.

|   | A                  | B           | C           | D           | E           | F           |
|---|--------------------|-------------|-------------|-------------|-------------|-------------|
| 1 |                    | $p$         | $q$         | $ps$        | $di$        | $pf$        |
| 2 |                    |             |             |             |             |             |
| 3 | Mean               | 62.72400013 | 18.45833333 | 22.0220001  | 3.526966667 | 22.7533334  |
| 4 | Standard Deviation | 18.72346189 | 4.613087857 | 4.077237313 | 1.040803254 | 5.329653589 |
| 5 | Minimum            | 29.639999   | 6.37        | 15.21       | 1.525       | 10.52       |
| 6 | Maximum            | 105.449997  | 26.27       | 28.98       | 5.125       | 34.009998   |
| 7 | Count              | 30          | 30          | 30          | 30          | 30          |

## 11.2 ESTIMATING THE REDUCED FORM EQUATIONS

The reduced form equation expresses each endogenous variable,  $P$  and  $Q$ , in terms of the exogenous variables  $PS$ ,  $DI$ ,  $PF$ . This can be accomplished by setting the structural equation equal to each other and solving for the endogenous variables.

$$Q_i = \pi_{11} + \pi_{21}PS_i + \pi_{31}DI_i + \pi_{41}PF_i + v_{i1}$$

$$P_i = \pi_{12} + \pi_{22}PS_i + \pi_{32}DI_i + \pi_{42}PF_i + v_{i2}$$

These equations can be estimated by least squares since all independent variables are exogenous and uncorrelated with the error terms. We will estimate two regressions for the reduced form equations. Once we estimate these reduced form equations, we will obtain and store the predicted values of price,  $\hat{P}$ , and then estimate the structural equations using  $\hat{P}$  and the other exogenous variables using the 2SLS technique which was first introduced in Chapter 10.

First, estimate the quantity equation using  $Q$  as the **Y-Range**, and  $PS$ ,  $DI$  and  $PF$  as the **X-Range**. Include labels and place results on the worksheet named **reduced eq quant** and Click **OK**.

**Regression**

Input

Input Y Range:

Input X Range:

☒ Labels ☐ Constant is Zero

☐ Confidence Level:  %

Output options

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

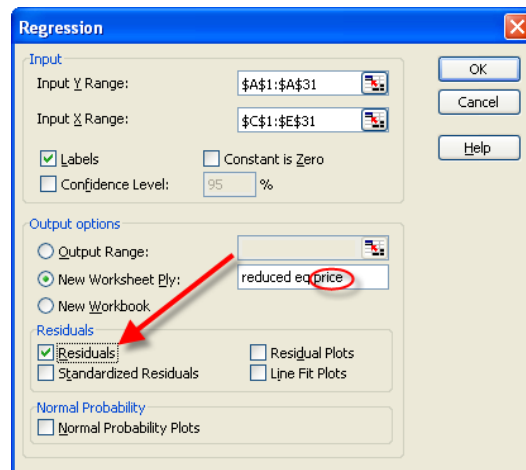
☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

OK Cancel Help

Next, estimate the regression for the price equation using  $P$  as the **Y-Range** and  $PS$ ,  $DI$  and  $PF$  as the **X-Range**. Include labels and place results on a worksheet named **reduced eq price**. Make sure to include **Residuals** option to obtain the  $\hat{P}$  and click **OK**.



The Quantity ( $Q$ ) reduced form estimates are:

|    |           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
|----|-----------|---------------------|-----------------------|---------------|----------------|
| 16 |           |                     |                       |               |                |
| 17 | Intercept | 7.895100328         | 3.243421325           | 2.434188944   | 0.022099332    |
| 18 | ps        | 0.656402011         | 0.142537596           | 4.605114937   | 9.53266E-05    |
| 19 | di        | 2.167156078         | 0.700473729           | 3.093843477   | 0.004680802    |
| 20 | pf        | -0.506982392        | 0.121261645           | -4.180896549  | 0.000291281    |

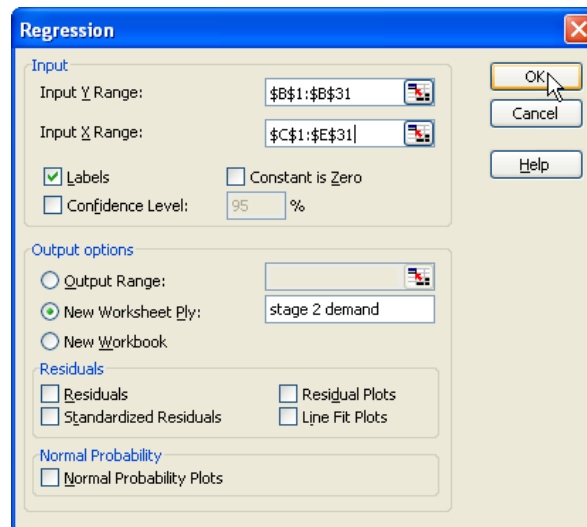
The Price ( $P$ ) reduced form estimates are:

|    |           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
|----|-----------|---------------------|-----------------------|---------------|----------------|
| 16 |           |                     |                       |               |                |
| 17 | Intercept | -32.51242016        | 7.984235283           | -4.072076912  | 0.000387308    |
| 18 | ps        | 1.708147148         | 0.350880625           | 4.868171757   | 4.75902E-05    |
| 19 | di        | 7.602492026         | 1.724335664           | 4.408939736   | 0.000159932    |
| 20 | pf        | 1.353905695         | 0.298506239           | 4.535602665   | 0.000114523    |

All explanatory variables in both models are significant.

### 11.3 2SLS ESTIMATES OF TRUFFLE DEMAND AND SUPPLY

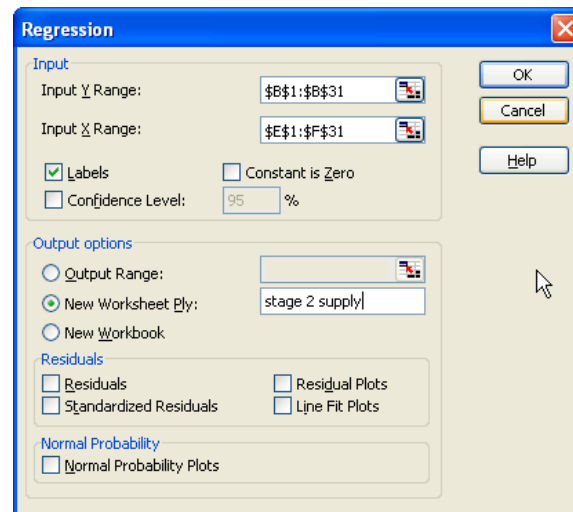
To obtain 2SLS estimates, we replace  $P$  in the structural equation with  $\hat{P}$  from the reduced form equation. To accomplish that, return to the worksheet containing the original data. Create an empty column next to the other explanatory variables as Excel requires all explanatory variables to be next to each other. Return to the **reduced eq price** worksheet and **Copy** cells B26 through B56 (*Predicted p*) to the original data worksheet. Estimate a regression of the structural demand equation using  $Q$  as the **Y-Range** and *Predicted p*,  $PS$  and  $PF$  as the **X-Range**. Include labels and place results on the worksheet named **stage 2 demand**.



The Regression dialog box is shown with the following settings:

- Input:**
  - Input Y Range:  $\$B\$1:\$B\$31$
  - Input X Range:  $\$C\$1:\$E\$31$
  - ☒ Labels
  - ☐ Constant is Zero
  - ☐ Confidence Level: 95 %
- Output options:**
  - ☐ Output Range:
  - ☒ New Worksheet Ply: stage 2 demand
  - ☐ New Workbook
- Residuals:**
  - ☐ Residuals
  - ☐ Standardized Residuals
  - ☐ Residual Plots
  - ☐ Line Fit Plots
- Normal Probability:**
  - ☐ Normal Probability Plots

Return to the worksheet containing the original data. Move the columns around so that *phat* and *PF* are next to each other to estimate the structural supply equation since Excel requires the **X-Range** data to be contiguous. Estimate the regression, using *Q* as the **Y-Range** and *Predicted p*, and *PF* as the **X-Range**. Include labels and place results on a worksheet names **stage 2 supply**.



The Regression dialog box is shown with the following settings:

- Input:**
  - Input Y Range:  $\$B\$1:\$B\$31$
  - Input X Range:  $\$E\$1:\$F\$31$
  - ☒ Labels
  - ☐ Constant is Zero
  - ☐ Confidence Level: 95 %
- Output options:**
  - ☐ Output Range:
  - ☒ New Worksheet Ply: stage 2 supply
  - ☐ New Workbook
- Residuals:**
  - ☐ Residuals
  - ☐ Standardized Residuals
  - ☐ Residual Plots
  - ☐ Line Fit Plots
- Normal Probability:**
  - ☐ Normal Probability Plots

The 2<sup>nd</sup> stage regression results for demand equation are:

|    |           | Coefficients | Standard Error | t Stat       | P-value     |
|----|-----------|--------------|----------------|--------------|-------------|
| 16 |           |              |                |              |             |
| 17 | Intercept | -4.279473279 | 3.013833748    | -1.41994338  | 0.167504529 |
| 18 | ps        | 1.296033361  | 0.193094429    | 6.711914817  | 4.02707E-07 |
| 19 | di        | 5.013978871  | 1.241414409    | 4.038924337  | 0.000422352 |
| 20 | phat      | -0.374459162 | 0.089564321    | -4.180896549 | 0.000291281 |

The results for supply equation are:

| 16 |           | Coefficients | Standard Error | t Stat       | P-value     |
|----|-----------|--------------|----------------|--------------|-------------|
| 17 | Intercept | 20.03280279  | 2.165698376    | 9.250042857  | 7.35757E-10 |
| 18 | phat      | 0.337981554  | 0.04412361     | 7.659879835  | 3.07433E-08 |
| 19 | pf        | -1.000909364 | 0.146127429    | -6.849565273 | 2.33394E-07 |

Since the model standard error is based on the LS residuals, the standard errors,  $t$ -statistics and confidence intervals are incorrect. Since Excel doesn't have a built in function for these errors, we will need to do some calculations. The calculation is explained in the next section.

### 11.3.1 Correction of 2SLS standard errors

In the simple linear regression model  $y_i = \beta_1 + \beta_2 x_i + e_i$  the 2SLS estimator is the least squares estimator applied to  $y_i = \beta_1 + \beta_2 \hat{x}_i + e_i$  where  $\hat{x}_i$  is the predicted value from a reduced form equation. So, the 2SLS estimators are

$$\hat{\beta}_2 = \frac{\sum(\hat{x}_i - \bar{x})(y_i - \bar{y})}{\sum(\hat{x}_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

In large samples the 2SLS estimators have approximate normal distributions. In the simple regression model

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(\hat{x}_i - \bar{x})^2}\right)$$

The error variance  $\sigma^2$  should be estimated using the estimator

$$\hat{\sigma}_{2SLS}^2 = \frac{\sum(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2}{N - 2}$$

with the quantity in the numerator being the sum of squared 2SLS residuals, or  $SSE_{2SLS}$ . The problem with doing 2SLS with two least squares regressions is that in the second estimation the estimated variance is

$$\hat{\sigma}_{wrong}^2 = \frac{\sum(y_i - \hat{\beta}_1 - \hat{\beta}_2 \hat{x}_i)^2}{N - 2}$$

The numerator is the  $SSE$  from the regression of  $y_i$  on  $\hat{x}_i$ , which is  $SSE_{wrong}$ .

Thus correct 2SLS standard error is

$$se(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}_{2SLS}^2}{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}_{2SLS}^2}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{2SLS}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}}$$

and the “wrong” standard error, calculated in the 2<sup>nd</sup> least squares estimation, is

$$se_{wrong}(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}_{wrong}^2}{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\sqrt{\hat{\sigma}_{wrong}^2}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}} = \frac{\hat{\sigma}_{wrong}}{\sqrt{\sum(\hat{x}_i - \bar{x})^2}}$$

Given that we have the “wrong” standard error in the 2<sup>nd</sup> regression, we can adjust it using a correction factor

$$se(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}_{2SLS}^2}{\hat{\sigma}_{wrong}^2}} se_{wrong}(\hat{\beta}_2) = \frac{\hat{\sigma}_{2SLS}}{\hat{\sigma}_{wrong}} se_{wrong}(\hat{\beta}_2)$$

### 11.3.2 Corrected standard errors in truffle demand and supply

The first step to correcting the standard errors is to calculate the sigma-hats of the 2SLS for both the supply and demand equations. As in the previous chapter, go back to the original data and create two columns.

- Label them *residualdemand* and *residualsupply*, respectively.
- Calculate the residuals for both columns using the 2SLS estimates.

|   | A     | B     | C     | D     | E        | F     | G                                  | H                             |
|---|-------|-------|-------|-------|----------|-------|------------------------------------|-------------------------------|
| 1 | p     | q     | ps    | di    | phat     | pf    | residualdemand= y-b1-b2ps-b3di-b4p | residualsupply= y-b1-b2p-b3pf |
| 2 | 29.64 | 19.89 | 19.97 | 2.103 | 31.83041 | 10.52 | -1.157741013                       | -20.69014321                  |
| 3 | 40.23 | 13.04 | 18.04 | 2.043 | 40.46577 | 19.67 | -1.240036587                       | -40.27768788                  |
| 4 | 34.71 | 19.61 | 22.36 | 1.87  | 38.50108 | 13.74 | -1.468496308                       | -25.90663584                  |
| 5 | 41.43 | 17.13 | 20.87 | 1.525 | 39.03302 | 17.95 | 2.228780053                        | -34.87170364                  |

- Label column I as shown below to calculate the correction factors for both the demand and the supply functions.
- In cells J1 and J5, we calculate the correct sigma-hat\_2sls for the demand and supply functions, respectively.
- In cells J2 and J6, we copy and paste the *sigmat\_wrong* from the demand and supply outputs and
- calculate the correction factors in cells J3 and J7 for demand and supply, respectively.

| G                                  | H                             | I                        | J                       |
|------------------------------------|-------------------------------|--------------------------|-------------------------|
| residualdemand= y-b1-b2ps-b3di-b4p | residualsupply= y-b1-b2p-b3pf | SighatD_2sls             | =SQRT(SUMSQ(G2:G31)/26) |
| -1.157741013                       | -20.69014321                  | SighatD_wrong            | 2.68008449800055        |
| -1.240036587                       | -40.27768788                  | Correction factor Demand | =+J1/J2                 |
| -1.468496308                       | -25.90663584                  |                          |                         |
| 2.228780053                        | -34.87170364                  | SighatS_2sls             | =SQRT(SUMSQ(H2:H31)/27) |
| 7.582987119                        | -29.24334634                  | SighatS_wrong            | 2.65168723436824        |
| -8.116766311                       | -51.65454187                  | Correction factor Supply | =+J5/J6                 |
| 4.890932986                        | -47.5673206                   |                          |                         |
| -3.472370467                       | -47.15280603                  |                          |                         |

The resulting sighat\_2sls figures and the correction factor for demand are

| I                        | J           |
|--------------------------|-------------|
| SighatD_2sls             | 4.929960233 |
| SighatD_wrong            | 2.680084498 |
| Correction factor Demand | 1.839479403 |

Now, we can plug in the correction factor of 1.839479403 into the demand model coefficient page and calculate the corrected standard errors, *t*-test, and the *p*-values. From the results, we have a demand function is downward sloping, and shifts to the right as the price of substitutes rises, and shifts to the right as income rises (truffles are a normal good). All of our coefficients are significant at the 5% level.

|    |                          | Coefficients | Standard Error           | t Stat           | P-value           |
|----|--------------------------|--------------|--------------------------|------------------|-------------------|
| 16 |                          |              |                          |                  |                   |
| 17 | Intercept                | -4.279473279 | 3.013833748              | -1.41994338      | 0.167504529       |
| 18 | ps                       | 1.296033361  | 0.193094429              | 6.711914817      | 4.02707E-07       |
| 19 | di                       | 5.013978871  | 1.241414409              | 4.038924337      | 0.000422352       |
| 20 | phat                     | -0.374459162 | 0.089564321              | -4.180896549     | 0.000291281       |
| 21 |                          |              |                          |                  |                   |
| 22 |                          | Coefficients | Corrected Standard Error | Corrected t Stat | Corrected P-value |
| 23 | Intercept                | -4.279473279 | 5.543885104              | -0.771926762     | 0.447117761       |
| 24 | ps                       | 1.296033361  | 0.355193225              | 3.648812161      | 0.001160083       |
| 25 | di                       | 5.013978871  | 2.283556237              | 2.195688807      | 0.037235204       |
| 26 | phat                     | -0.374459162 | 0.164751724              | -2.272869455     | 0.03153503        |
| 27 |                          |              |                          |                  |                   |
| 28 | Correction factor Demand | 1.839479403  |                          |                  |                   |

## 11.4 SUPPLY AND DEMAND OF FISH

The second example is from the Fulton Fish market. The demand equation is:

$$\ln(QUAN_t) = \alpha_1 + \alpha_2 \ln(PRICE_t) + \alpha_3 MON_t + \alpha_4 TUE_t + \alpha_5 WED_t + \alpha_6 THU_t + e_t^d$$

Where  $\ln(QUAN_t)$  the quantity is sold in pounds, and  $PRICE_t$  is the average price per pound. The remaining are the dummy variables for the days of the week which capture the demand shifts.  $\alpha_2$  is the price elasticity of demand, which is should be negative. The supply equation is:

$$\ln(QUAN_t) = \beta_1 + \beta_2 \ln(PRICE_t) + \beta_3 STORMY_t + e_t^s$$

$\beta_2$  is the price elasticity of supply. The variable *STORMY* is a dummy variable indicating stormy weather during the previous three days. Below are the partial summary statistics on the data set *fultonfish.xls*.

|   | A                  | B             | C            | D          | E          | F          | G          | H             |
|---|--------------------|---------------|--------------|------------|------------|------------|------------|---------------|
| 1 |                    | <i>lprice</i> | <i>lquan</i> | <i>mon</i> | <i>tue</i> | <i>wed</i> | <i>thu</i> | <i>stormy</i> |
| 2 |                    |               |              |            |            |            |            |               |
| 3 | Mean               | -0.19368      | 8.52343      | 0.189189   | 0.207207   | 0.189189   | 0.207207   | 0.288288      |
| 4 | Standard Deviation | 0.381935      | 0.741672     | 0.393435   | 0.407143   | 0.393435   | 0.407143   | 0.45502       |
| 5 | Minimum            | -1.10775      | 6.194406     | 0          | 0          | 0          | 0          | 0             |
| 6 | Maximum            | 0.664327      | 9.981374     | 1          | 1          | 1          | 1          | 1             |
| 7 | Count              | 111           | 111          | 111        | 111        | 111        | 111        | 111           |

### 11.5 REDUCED FORMS FOR FISH PRICE AND QUANTITY

The reduced form equations are estimated using the LSE. We can get them by estimating the price and quantity equations as a function of all the exogenous variables. In this case,

$$Q_t = \pi_{11} + \pi_{21}MON_t + \pi_{31}TUE_t + \pi_{41}WED_t + \pi_{51}THU_t + \pi_{61}STORMY_t + v_{t1}$$

$$P_t = \pi_{12} + \pi_{22}MON_t + \pi_{32}TUE_t + \pi_{42}WED_t + \pi_{52}THU_t + \pi_{62}STORMY_t + v_{t2}$$

The output for the reduced form equations are:

For the quantity:

|    |           |                     |                       |               |                |
|----|-----------|---------------------|-----------------------|---------------|----------------|
| 16 |           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept | 8.810062929         | 0.147024398           | 59.92245519   | 5.06767E-83    |
| 18 | mon       | 0.101005158         | 0.206502523           | 0.489123117   | 0.625774713    |
| 19 | tue       | -0.484706164        | 0.201145868           | -2.409724684  | 0.017704406    |
| 20 | wed       | -0.553113922        | 0.205806011           | -2.687549888  | 0.008371018    |
| 21 | thu       | 0.053693174         | 0.201048771           | 0.267065417   | 0.789942623    |
| 22 | stormy    | -0.387772227        | 0.143730552           | -2.697910927  | 0.008131513    |

For the price:

|    |           |                     |                       |               |                |
|----|-----------|---------------------|-----------------------|---------------|----------------|
| 16 |           | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept | -0.271705457        | 0.076388974           | -3.55686747   | 0.000564633    |
| 18 | mon       | -0.11292248         | 0.107291824           | -1.052479825  | 0.294995532    |
| 19 | tue       | -0.04114933         | 0.104508685           | -0.393740767  | 0.694570871    |
| 20 | wed       | -0.011824968        | 0.10692994            | -0.110586129  | 0.912155647    |
| 21 | thu       | 0.049645652         | 0.104458236           | 0.475267953   | 0.635583179    |
| 22 | stormy    | 0.346405584         | 0.074677601           | 4.638681183   | 1.01527E-05    |

Now, using the predicted price, *phat*, we can estimate the structural equations. The output for the structural equations is:



Supply equation:

|    |                  |                     |                       |               |                |
|----|------------------|---------------------|-----------------------|---------------|----------------|
| 16 |                  | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept        | 8.62835419          | 0.388927755           | 22.18497927   | 5.0627E-42     |
| 18 | Predicted lprice | 0.001058583         | 1.309403799           | 0.000808446   | 0.999356445    |
| 19 | stormy           | -0.363245814        | 0.464861724           | -0.781406157  | 0.436273259    |

Demand equation:

|    |                  |                     |                       |               |                |
|----|------------------|---------------------|-----------------------|---------------|----------------|
| 16 |                  | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept        | 8.505911279         | 0.160846266           | 52.88224267   | 1.66981E-77    |
| 18 | mon              | -0.025402162        | 0.20789713            | -0.122186211  | 0.902985096    |
| 19 | tue              | -0.530769415        | 0.201339958           | -2.636185187  | 0.009655531    |
| 20 | wed              | -0.56635099         | 0.205942487           | -2.750044429  | 0.007018196    |
| 21 | thu              | 0.109267351         | 0.202101276           | 0.54065641    | 0.589889505    |
| 22 | Predicted lprice | -1.11941679         | 0.414919848           | -2.697910927  | 0.008131513    |

Recall that, although the point estimates are correct, the standard errors are NOT. You need to adjust the standard errors the way illustrated in Section 11.3.2 above. The discussion in the text explains why the estimated supply equation is not credible.

# CHAPTER 12

## Nonstationary Time-Series Data and Cointegration

### CHAPTER OUTLINE

12.1 Stationary and Nonstationary Data  
12.2 Spurious Regression  
12.3 Unit Root Tests for Stationarity

12.4 Integration and Cointegration  
12.5 Engle-Granger Test

### 12.1 STATIONARY AND NONSTATIONARY DATA

One of the fundamental principles of econometrics is that the statistical properties of estimators depend on how the data behaves. When time series data is concerned, the data need to be stationary for the usual econometric procedures to have desired statistical properties. Time series data is stationary when the means, variances, and covariances are constant and don't depend on the period in which they are measured.

To illustrate stationarity, we will use *usa.xls* and look at the mean and variance of GDP in different sub period using time series plots and summary statistics. Open your file *usa.xls*. **Insert** a column to the left of GDP and name it *t* for time.

- Type 1 in cell A2.
- Type =A2+1 in cell A3 and copy the cell down the entire column.
- After creating a time counter, highlight the *t* and *GDP* columns and go to **Insert>Chart>Scatter Plot**. After some formatting, your time series plot will look like:



We can also create summary statistics to compare the means in different periods. Below are the summary statistics for the first half, second half and overall data.

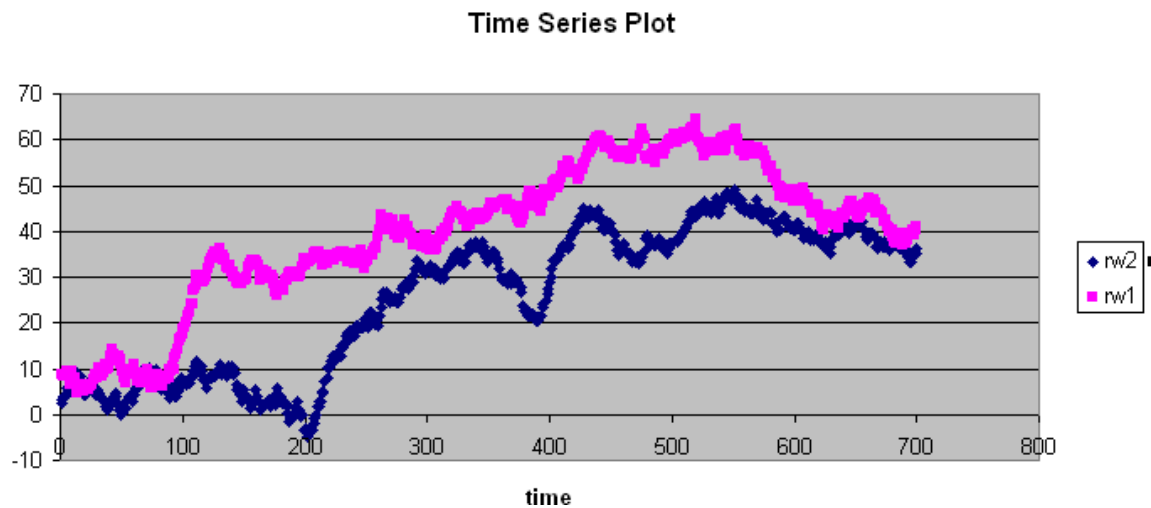
Both plot and the descriptive statistics indicate the GDP data is non-stationary.

It is very possible to estimate a regression and find statistically significant relationship where no such relationship exists. In time series analysis this is a common phenomenon when data is non-stationary. Our example will use two time series, *rw1* and *rw2* that are generated by random walks where

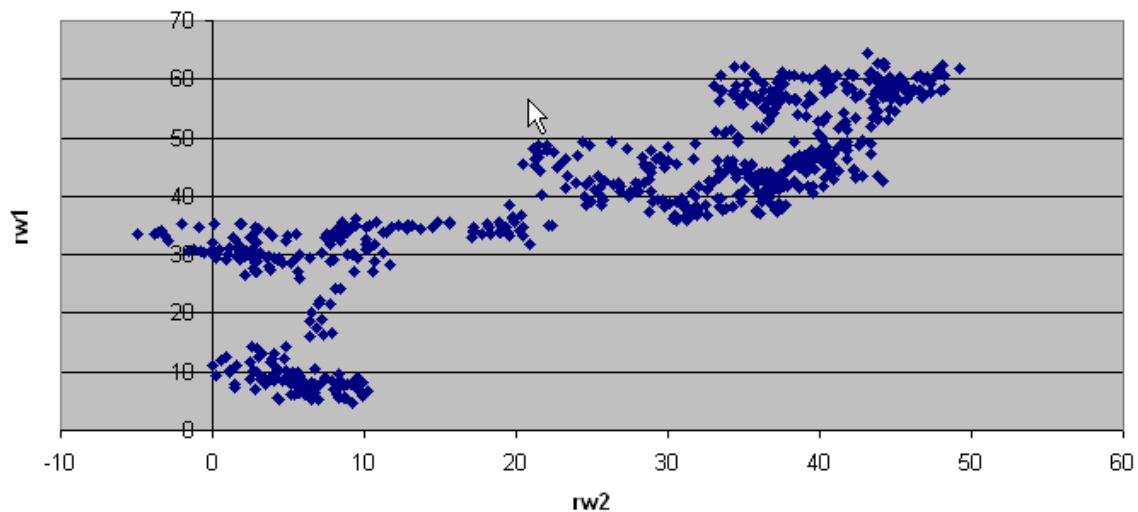
$$rw_2 : x_{t+1} + v_{2t}$$

The errors are independent standard normal generated by a pseudo random generator. By construction,  $rw_1$  and  $rw_2$  are not related in any way. Now let's look at the empirical relationship by opening the file *spurious.xls*.

Insert a column at Column A and create the time counter as described in the previous section. Then plot both of the series against the  $t$  variable.



Next, scatter plot  $rw1$  against  $rw2$  and observe the potential spurious relationship between the two variables.



You can also estimate a linear regression that will confirm the suspicion. The coefficient on  $rw2$  is positive and significant. However, these variables are not related in any way. The cause of the spurious relationship is the nonstationarity of the two series.

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.83960906          |                       |               |                |
| 5  | R Square                     | 0.704943374         |                       |               |                |
| 6  | Adjusted R Square            | 0.704520657         |                       |               |                |
| 7  | Standard Error               | 8.557267989         |                       |               |                |
| 8  | Observations                 | 700                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 1                   | 122116.5568           | 122116.5568   | 1667.647606    |
| 13 | Residual                     | 698                 | 51112.33113           | 73.22683543   |                |
| 14 | Total                        | 699                 | 173228.8879           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 17.81804111         | 0.620477603           | 28.71665477   | 2.4603E-120    |
| 18 | rw2                          | 0.84204116          | 0.020619645           | 40.83684128   | 3.5686E-187    |

Therefore, it is very important to check for stationarity first, whenever you are dealing with time series data.

### 12.3 UNIT ROOT TEST FOR STATIONARITY

The Dickey Fuller (DF) and augmented Dickey-Fuller (ADF) are unit root tests that are used to test stationarity. Before conducting the test, there are certain decisions that need to be made about the nature of the data in order to implement the correct regression model. These choices can be made by visual inspection of the data various plots. For example, if the data has nonzero mean, a constant term in the regression is appropriate. Below are the different possible regression models for series with different characteristics:

$$\Delta y_t = \gamma y_{t-1} + v_t \quad \text{No constant and no trend}$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + v_t \quad \text{Constant but no trend}$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \lambda t + v_t \quad \text{Constant and trend}$$

Recall the GDP plot, which is slightly quadratic in time, so you would choose the regression model that included a constant and a trend to conduct the unit root test. The test is conducted by estimating the regression and implementing a  $t$ -test for the following hypothesis:

$$H_0 : \gamma = 0$$

$$H_1 : \gamma < 0$$

The augmented version of the DF test (ADF) adds lagged differences to the model and the models become:

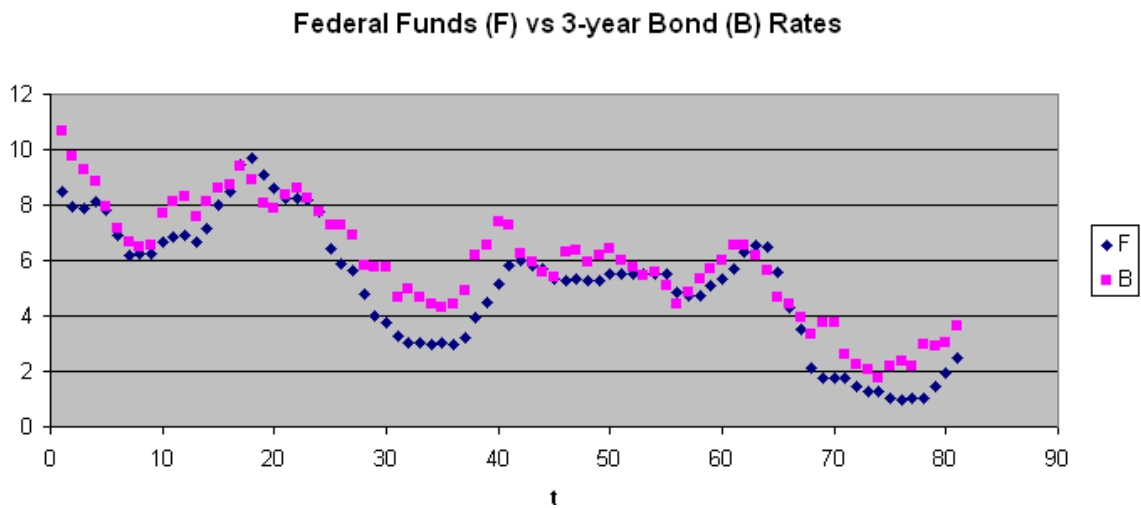
$$\Delta y_t = \gamma y_{t-1} + \sum_{s=1}^m a_s \Delta y_{t-s} + v_t$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{s=1}^m a_s \Delta y_{t-s} + v_t$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \lambda t + \sum_{s=1}^m a_s \Delta y_{t-s} + v_t$$

You have to pick a lag length to implement this test. The lag length should be enough to ensure that the residuals are not autocorrelated.

Let's consider Federal Funds rate ( $F_t$ ) and the 3-year Bond rate ( $B_t$ ) for this test. The plots indicated both series are nonstationary.



Since the series fluctuated from a non-zero mean and didn't seem to have trend, we will use the model with a constant but no trend. We will use one lag for the ADF test. So we will be estimating the following regression model for both the  $F_t$  and the  $B_t$ .

$$\Delta y_t = \alpha + \gamma y_{t-1} + a_1 \Delta y_{t-1} + v_t$$

Open the file *usa.xls* and estimate two regressions.

- Create 3 columns to the right of column F.
- Label them *diffF*, *lagF* and *difflagF* for differenced F, lagged F and differenced lagged F, respectively.

|   | A        | B          | C          | D       | E        | F       | G        | H       | I        | J       | K        |
|---|----------|------------|------------|---------|----------|---------|----------|---------|----------|---------|----------|
| 1 | <i>t</i> | <i>gdp</i> | <i>inf</i> | F       | diffF    | lagF    | diffF    | B       | diffB    | lagB    | diffB    |
| 2 | 1        | 4119.5     | 3.548619   | 8.47667 |          |         |          | 10.6767 |          |         |          |
| 3 | 2        | 4178.4     | 3.645685   | 7.92333 | =D3-D2   | =D2     |          | 9.76333 | =E3-E2   | =E2     |          |
| 4 | 3        | 4261.3     | 3.309139   | 7.9     | -0.02333 | 7.92333 | =F4-F3   | 9.28667 | -0.47666 | 9.76333 | =G4-G3   |
| 5 | 4        | 4321.8     | 3.469453   | 8.10333 | 0.20333  | 7.9     | -0.02333 | 8.84333 | -0.44334 | 9.28667 | -0.47666 |
| 6 | 5        | 4385.6     | 3.06095    | 7.82667 | -0.27666 | 8.10333 | 0.20333  | 7.93667 | -0.90666 | 8.84333 | -0.44334 |
| 7 | 6        | 4425.7     | 1.626267   | 6.92    | -0.90667 | 7.82667 | -0.27666 | 7.18    | -0.75667 | 7.93667 | -0.90666 |

- In cell E3, type **=D3-D2**.
- **Copy** this formula down the column.
- In cell F3, type **=D2**, and **Copy** the formula down the column.
- In cell G4, type **=F4-F3**, and **Copy** the formula down the column.
- Repeat the same for series F. The formulas have been highlighted in the above sheet.

Next, estimate two regressions; one for B and the other for F. Use the *diffF* as the **Y-Range**, and *lagF* and *diffF* as the **X-Range**. Notice that the first observations appear in Row 3. Place results on a worksheet named **ADF-f**.

Repeat for the B series and store the output in **ADF-b**.

**Regression**

**Input**

Input Y Range:

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level:  %

**Output options**

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

**Residuals**

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

The results for the B series are:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.371685373         |                       |               |                |
| 5  | R Square                     | 0.138150016         |                       |               |                |
| 6  | Adjusted R Square            | 0.115469754         |                       |               |                |
| 7  | Standard Error               | 0.49540856          |                       |               |                |
| 8  | Observations                 | 79                  |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 2                   | 2.989922068           | 1.494961034   | 6.091200007    |
| 13 | Residual                     | 76                  | 18.65265275           | 0.245429642   |                |
| 14 | Total                        | 78                  | 21.64257482           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.285224314         | 0.178082305           | 1.601643207   | 0.113383195    |
| 18 | lagB                         | -0.056219537        | 0.028444983           | -1.976430699  | 0.051734625    |
| 19 | diffB                        | 0.31506378          | 0.105991584           | 2.972535827   | 0.00395437     |

The results for the F series are:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.675282422         |                       |               |                |
| 5  | R Square                     | 0.456006349         |                       |               |                |
| 6  | Adjusted R Square            | 0.441690727         |                       |               |                |
| 7  | Standard Error               | 0.354361732         |                       |               |                |
| 8  | Observations                 | 79                  |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 2                   | 7.9998949             | 3.99994745    | 31.85375643    |
| 13 | Residual                     | 76                  | 9.543490008           | 0.125572237   |                |
| 14 | Total                        | 78                  | 17.54338491           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.177861681         | 0.100751146           | 1.765356406   | 0.081520684    |
| 18 | lagF                         | -0.037066841        | 0.01773275            | -2.09030419   | 0.039934653    |
| 19 | diffF                        | 0.672477694         | 0.08536637            | 7.877548199   | 1.88803E-11    |

The  $t$ -statistics in this case is also called the **tau-statistics** within the context of unit root testing. These values are compared to the ADF critical values. For both of the series, when compared the test statistics to the critical values for the Dickey-Fuller test (which are shown in Table 12.2 of



*POE*), we do not reject our null hypothesis, and conclude that the levels of the series are nonstationary.

## 12.4 INTEGRATION AND COINTEGRATION

Two non-stationary series are said to be cointegrated if their differences are stationary. For example, Fed Funds rate and the 3-year bond are non-stationary. Both series are “integrated of order 1” or  $I(1)$ . If the two series move together through time, they are said to be cointegrated. We can test cointegration by running a regression of one of the  $I(1)$  series on the other and testing the residuals for stationarity using the augmented Dickey-Fuller test where the null hypothesis is that the residuals are non-stationary. Rejecting the hypothesis leads to the conclusion that the residuals are stationary and therefore, the series are cointegrated.

## 12.5 ENGLE-GRANGER TEST

The test described above is commonly referred as the Engle-Granger test. We will illustrate the test by regressing  $B$  on  $F$ , save the residuals then use these in an augmented Dickey-Fuller regression on the residuals.

Return to *usa.xls* file. Estimate the regression  $B$  as the **Y-Range**, and  $F$  as the **X-Range**. Include labels and choose the **Residuals** option. Place the results on a **cointegration** worksheet.

**Regression**

**Input**

Input Y Range: \$H\$1:\$H\$82

Input X Range: \$D\$1:\$D\$82

☒ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

**Output options**

☐ Output Range:

☒ New Worksheet Ply: Cointegration

☐ New Workbook

**Residuals**

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

Copy the residuals from the output section to the data section and paste it in the next available column. Create *differ*, *lage* and *difflage* and write the relevant formulas as shown below.

| J       | K        | L         | M        | N        | O        |
|---------|----------|-----------|----------|----------|----------|
| lagB    | difflagB | Residuals | diffe    | lage     | difflage |
|         |          | 1.976095  |          |          |          |
| 10.6767 |          | 1.523383  | =L3-L2   | =L2      |          |
| 9.76333 | -0.91337 | 1.066146  | -0.45724 | 1.523383 | =N4-N3   |
| 9.28667 | -0.47666 | 0.453532  | -0.61261 | 1.066146 | -0.45724 |
| 8.84333 | -0.44334 | -0.22281  | -0.67634 | 0.453532 | -0.61261 |
| 7.93667 | -0.90666 | -0.22467  | -0.00186 | -0.22281 | -0.67634 |
| 7.18    | -0.75667 | -0.15415  | 0.070521 | -0.22467 | -0.00186 |

Estimate a regression, using *diffe* as the **Y-Range** and *lage* and *difflage* as the **X-Range**.

**Regression**

**Input**

Input **Y** Range:

Input **X** Range:

☐ Labels ☐ Constant is **Z**ero

☐ Confidence Level:  %

**Output options**

☐ Output Range:

☒ New Worksheet **P**ly:

☐ New **W**orkbook

**Residuals**

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

The output is:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.491775216         |                       |               |                |
| 5  | R Square                     | 0.241842863         |                       |               |                |
| 6  | Adjusted R Square            | 0.221891359         |                       |               |                |
| 7  | Standard Error               | 0.401454738         |                       |               |                |
| 8  | Observations                 | 79                  |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 2                   | 3.907156588           | 1.953578294   | 12.1215357     |
| 13 | Residual                     | 76                  | 12.24860893           | 0.161165907   |                |
| 14 | Total                        | 78                  | 16.15576551           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | -0.019239664        | 0.045260927           | -0.425083293  | 0.671977115    |
| 18 | lage                         | -0.314927566        | 0.069576341           | -4.526359986  | 2.182E-05      |
| 19 | diffIage                     | 0.312402954         | 0.10285232            | 3.037393364   | 0.003269178    |

Compare the results to the 1%, 5%, 10% ADF critical values. If the values are less than the *t-value*, reject the null hypothesis and conclude that the residuals are stationary. You can find the critical values for this test in Table 12.3 of *POE*.

# **CHAPTER 13**

## **An Introduction to Macroeconometrics: VEC and VAR Models**

### **CHAPTER OUTLINE**

13.1 VEC and VAR models

13.2 Estimating VECM

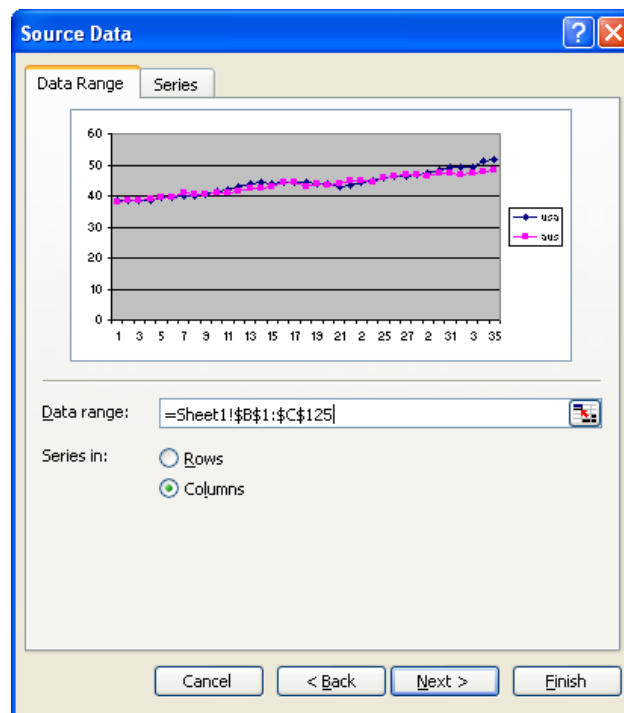
13.3 Estimating VAR

### **13.1 VEC AND VAR MODELS**

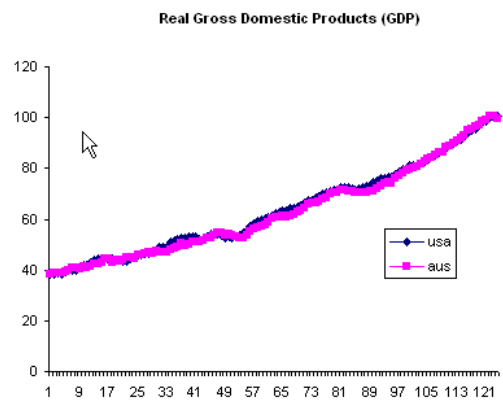
The vector autoregressive (VAR) model is a general framework used to describe dynamic interrelationship among stationary variables. So first step of any VAR analysis should be to determine whether the levels of your data is stationary or not. If not, take the first difference and try again. Usually, if the levels of the time series are not stationary, the differences will be. If the time series is stationary, then we need to modify the VAR framework to allow consistent estimation of the relationship among series. The vector error correction (VEC) model is a special case of the VAR for variables that are  $I(1)$ . The VEC can also take into account any cointegrating relationship among the series.

### **13.2 ESTIMATING A VEC MODEL**

We will use gdp.xls data on Gross Domestic Product of U.S. and Australia to estimate a VEC model. We are using a VEC model because (1) both series are nonstationary in levels and stationary in differences (2) the variables are integrated. The following is the time series plot for the two series. Go to **Insert>Chart>** select **line graph** and highlight both columns of data.



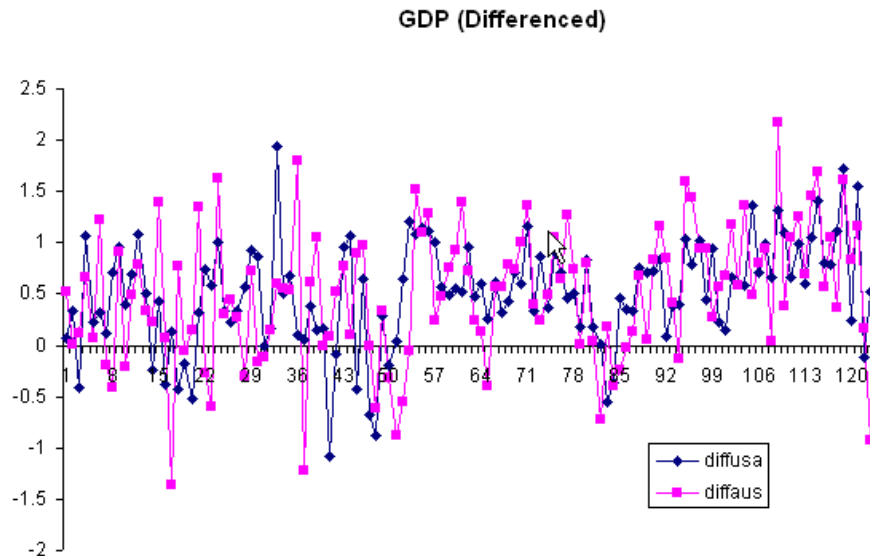
After some formatting your plot should look like



Neither of these series is stationary in its levels and they appear to have a common trend. Next, we will take the difference of both series. To accomplish that, label cells D1 and E1, *diffusa* and *diffaus*, for differenced U.S. GDP and differenced Australian GDP, respectively.

|   | A       | B       | C        | D        |
|---|---------|---------|----------|----------|
| 1 | usa     | aus     | diffusa  | diffaus  |
| 2 | 38.3011 | 38.2355 |          |          |
| 3 | 38.3734 | 38.7551 | =+B3-B2  | =+C3-C2  |
| 4 | 38.7137 | 38.7706 | 0.340297 | 0.015499 |
| 5 | 38.2991 | 38.8948 | -0.4146  | 0.124199 |
| 6 | 39.3615 | 39.5621 | 1.062401 | 0.667301 |

- In cell C3, type **=B3-B2** to difference the series *usa* and **Copy** this formula down the column.
- Repeat it for *aus*; in cell D3, type **=C3-C2**, and **Copy** the formula down the column.
- Highlight both C and D columns and go to **Insert>Chart>** select **line graph**.
- After some formatting the plot for the differenced data will look like



The differences of both series seem to be stationary, and possible they may be cointegrated.

Based on the graphical information, we will estimate Dickey-Fuller regressions for levels with intercept, time trend and 2 lags for US equation and 3 lags for Australian equation. Recall that the model with intercept and trend is

$$\Delta y_t = \alpha + \gamma y_{t-1} + \lambda t + \sum_{s=1}^m a_s \Delta y_{t-s} + v_t$$

Before we estimate the model, we need get the data ready. Return to *gdp.xls*.

- Insert 5 columns after *usa* and label those new columns *diffusa*, *lag usa*, *diff11usa*, *diff12usa*, and *trend*.
- In cell B3, type **=A3-A2** and **Copy** this formula down the column (You already have this column from the previous section).
- In cell C3, type **=A2** to get the lagged values of *usa*, and **Copy** the formula down the column.
- In cell D4, type **=C3** to get the lagged values of the differenced data for *usa*, and **Copy** the formula down the column.
- In cell E5, type **=C3** to get the double lagged values of the differenced data for *usa*, and **Copy** the formula down the column.
- In cell F1, type 1, in cell F2 type 2. Highlight the two cells and drag it down to create a time trend.

- Repeat the same process for series *aus* except add another lag for the difference since we are using 3 lags for *aus*. The formulas have been highlighted in the below sheet.

|    | A         | B        | C         | D        | E        | F     | G         | H         | I         | J         | K         | L        | M     |
|----|-----------|----------|-----------|----------|----------|-------|-----------|-----------|-----------|-----------|-----------|----------|-------|
| 1  | usa       | diffusa  | lag usa   | diff1usa | diff2usa | trend | aus       | diffaus   | lag aus   | diff1aus  | diff2aus  | diff3aus | trend |
| 2  | 38.301102 |          |           |          |          | 1     | 38.2355   |           |           |           |           |          | 1     |
| 3  | 38.373402 | 0.0723   | 38.301102 |          |          | 2     | 38.7551   | 0.5196    | 38.2355   |           |           |          | 2     |
| 4  | 38.713699 | 0.340297 | 38.373402 | 0.0723   |          | 3     | 38.770599 | 0.015499  | 38.7551   | 0.5196    |           |          | 3     |
| 5  | 38.299099 | -0.4146  | 38.713699 | 0.340297 | 0.0723   | 4     | 38.894798 | 0.124199  | 38.770599 | 0.015499  | 0.5196    |          | 4     |
| 6  | 39.3615   | 1.062401 | 38.299099 | -0.4146  | 0.340297 | 5     | 39.562099 | 0.667301  | 38.894798 | 0.124199  | 0.015499  | 0.5196   | 5     |
| 7  | 39.583599 | 0.222099 | 39.3615   | 1.062401 | -0.4146  | 6     | 39.640202 | 0.078103  | 39.562099 | 0.667301  | 0.124199  | 0.015499 | 6     |
| 8  | 39.897301 | 0.313702 | 39.583599 | 0.222099 | 1.062401 | 7     | 40.861401 | 1.221199  | 39.640202 | 0.078103  | 0.667301  | 0.124199 | 7     |
| 9  | 40.011398 | 0.114097 | 39.897301 | 0.313702 | 0.222099 | 8     | 40.674099 | -0.187302 | 40.861401 | 1.221199  | 0.078103  | 0.667301 | 8     |
| 10 | 40.722401 | 0.711003 | 40.011398 | 0.114097 | 0.313702 | 9     | 40.264198 | -0.409901 | 40.674099 | -0.187302 | 1.221199  | 0.078103 | 9     |
| 11 | 41.683998 | 0.961597 | 40.722401 | 0.711003 | 0.114097 | 10    | 41.1712   | 0.907002  | 40.264198 | -0.409901 | -0.187302 | 1.221199 | 10    |

Next, estimate two regressions; one for *usa* and the other for *aus*. Use the *diffusa* as the **Y-Range**, and *lag usa*, *diff1usa*, *diff2usa*, and *trend* as the **X-Range**. Notice that the first 4 rows will not be used due to missing values. Place results on a worksheet named **ADF for USA**

**Regression**

**Input**

Input Y Range:

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level:  %

**Output options**

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

**Residuals**

☐ Residuals ☐ Residual Plots

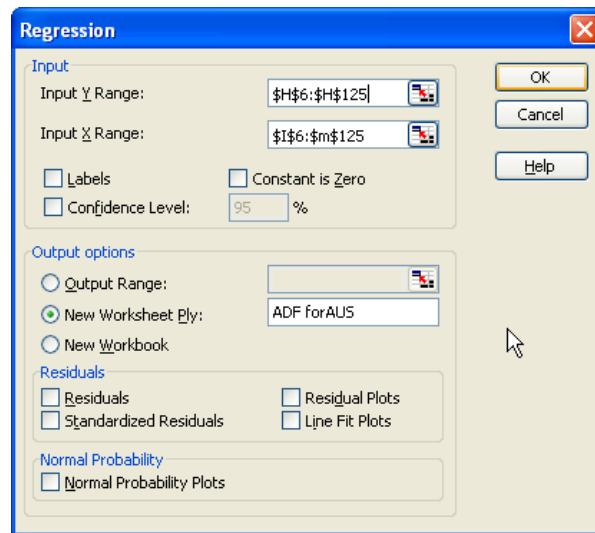
☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

Repeat for the *aus* series and store the output in **ADF for AUS**



The image shows the 'Regression' dialog box in Microsoft Excel. The 'Input' section has 'Input Y Range' set to '\$H\$6:\$H\$125' and 'Input X Range' set to '\$I\$6:\$m\$125'. The 'Labels' checkbox is unchecked, 'Constant is Zero' is unchecked, and 'Confidence Level' is set to '95 %'. The 'Output options' section has 'Output Range' empty, 'New Worksheet Ply:' selected with 'ADF for AUS' entered, and 'New Workbook' unselected. The 'Residuals' section has 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' all unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. Buttons for 'OK', 'Cancel', and 'Help' are on the right.

The ADF results for USA are:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.421979542         |                       |               |                |
| 5  | R Square                     | 0.178066734         |                       |               |                |
| 6  | Adjusted R Square            | 0.149724208         |                       |               |                |
| 7  | Standard Error               | 0.477910919         |                       |               |                |
| 8  | Observations                 | 121                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | ANOVA                        |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 4                   | 5.739818131           | 1.434954533   | 6.282669784    |
| 13 | Residual                     | 116                 | 26.49426622           | 0.228398847   |                |
| 14 | Total                        | 120                 | 32.23408435           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.606481834         | 0.54685214            | 1.109041712   | 0.26970618     |
| 18 | lag usa                      | -0.01426451         | 0.016603754           | -0.859113561  | 0.39204974     |
| 19 | diff1 usa                    | 0.202463125         | 0.092160702           | 2.196848768   | 0.030019379    |
| 20 | diff2 usa                    | 0.204936241         | 0.095625644           | 2.143109646   | 0.034193279    |
| 21 | trend                        | 0.00950405          | 0.007947703           | 1.195823466   | 0.234204198    |

and for Australia:



|    | A                     | B            | C              | D            | E           |
|----|-----------------------|--------------|----------------|--------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |              |             |
| 2  |                       |              |                |              |             |
| 3  | Regression Statistics |              |                |              |             |
| 4  | Multiple R            | 0.323637007  |                |              |             |
| 5  | R Square              | 0.104740912  |                |              |             |
| 6  | Adjusted R Square     | 0.065475163  |                |              |             |
| 7  | Standard Error        | 0.6334085    |                |              |             |
| 8  | Observations          | 120          |                |              |             |
| 9  |                       |              |                |              |             |
| 10 | ANOVA                 |              |                |              |             |
| 11 |                       | df           | SS             | MS           | F           |
| 12 | Regression            | 5            | 5.351065163    | 1.070213033  | 2.667487923 |
| 13 | Residual              | 114          | 45.73752131    | 0.401206327  |             |
| 14 | Total                 | 119          | 51.08858647    |              |             |
| 15 |                       |              |                |              |             |
| 16 |                       | Coefficients | Standard Error | t Stat       | P-value     |
| 17 | Intercept             | 0.724289943  | 0.595013684    | 1.217266027  | 0.226016839 |
| 18 | lag aus               | -0.018175522 | 0.018602333    | -0.977056081 | 0.330610216 |
| 19 | diff1 aus             | 0.063265454  | 0.097241016    | 0.650604623  | 0.516610867 |
| 20 | diff2 aus             | 0.014950618  | 0.098350356    | 0.152013869  | 0.8794447   |
| 21 | diff3 aus             | 0.1919537    | 0.098280123    | 1.953128397  | 0.053255    |
| 22 | trend                 | 0.012244468  | 0.008839102    | 1.385261502  | 0.168677111 |

In each case, the null hypothesis of nonstationarity can not be rejected at any significance. Notice that both lagged differences are significant in the US equation and the 3<sup>rd</sup> lag in the Australian equation are significant. Next, we will estimate the cointegration equation. Notice that the cointegration equation does not include a constant. Return to *gdp.xls* file. Estimate the regression *aus* as the **Y-Range**, and *usa* as the **X-Range**. Include labels and choose the **Residuals** option. Place the results on **LS for cointegration** worksheet.

**Regression**

**Input**

Input Y Range: \$G\$1:\$G\$125

Input X Range: \$A\$1:\$A\$125

☒ Labels ☒ Constant is Zero

☐ Confidence Level: 95 %

**Output options**

☐ Output Range:

☒ New Worksheet Ply: LS for Cointegration

☐ New Workbook

**Residuals**

☒ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

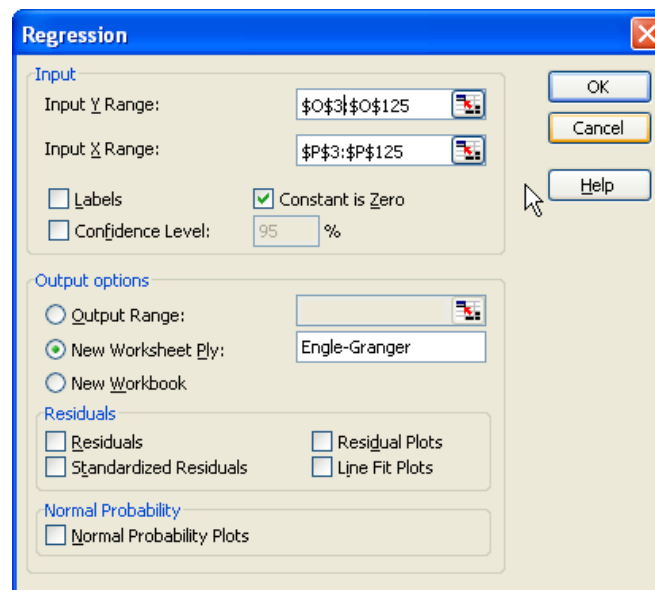
The output is:

|    | A                     | B                   | C                     | D             | E              |
|----|-----------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT        |                     |                       |               |                |
| 2  |                       |                     |                       |               |                |
| 3  | Regression Statistics |                     |                       |               |                |
| 4  | Multiple R            | 0.999826204         |                       |               |                |
| 5  | R Square              | 0.999652439         |                       |               |                |
| 6  | Adjusted R Square     | 0.991522358         |                       |               |                |
| 7  | Standard Error        | 1.219374742         |                       |               |                |
| 8  | Observations          | 124                 |                       |               |                |
| 9  |                       |                     |                       |               |                |
| 10 | ANOVA                 |                     |                       |               |                |
| 11 |                       | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression            | 1                   | 526014.2115           | 526014.2115   | 353771.6996    |
| 13 | Residual              | 123                 | 182.8855956           | 1.486874761   |                |
| 14 | Total                 | 124                 | 526197.0971           |               |                |
| 15 |                       |                     |                       |               |                |
| 16 |                       | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept             | 0                   | #N/A                  | #N/A          | #N/A           |
| 18 | usa                   | 0.985349542         | 0.001656642           | 594.7871045   | 1.3477E-214    |

We saved the residual to for the Engle-Granger test of cointegration as described in Section 12.5 of this manual. Copy the residuals from the output section to the data section and paste it in the next available column and rename the column *e*. Create *diffe*, *lag e* and *diff1e* and write the relevant formulas as shown below.

| N        | O            | P            | Q             |
|----------|--------------|--------------|---------------|
| <b>e</b> | <b>diffe</b> | <b>lag e</b> | <b>diff1e</b> |
| 0.495527 |              |              |               |
| 0.943655 | =N3-N2       | =+N2         |               |
| 0.624073 | -0.31981     | 0.943886     | =O3           |
| 1.156798 | 0.532725     | 0.624073     | -0.31981      |
| 0.777263 | -0.37954     | 1.156798     | 0.532725      |
| 0.636521 | -0.14074     | 0.777263     | -0.37954      |

We will estimate the relationship  $\Delta \hat{e}_t = \phi \hat{e}_{t-1} + v_t$ .



which produces the following result:

|    | A                     | B                   | C                     | D             | E              |
|----|-----------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT        |                     |                       |               |                |
| 2  |                       |                     |                       |               |                |
| 3  | Regression Statistics |                     |                       |               |                |
| 4  | Multiple R            | 0.253071206         |                       |               |                |
| 5  | R Square              | 0.064045035         |                       |               |                |
| 6  | Adjusted R Square     | 0.055848314         |                       |               |                |
| 7  | Standard Error        | 0.598499617         |                       |               |                |
| 8  | Observations          | 123                 |                       |               |                |
| 9  |                       |                     |                       |               |                |
| 10 | ANOVA                 |                     |                       |               |                |
| 11 |                       | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression            | 1                   | 2.990323011           | 2.990323011   | 8.34815202     |
| 13 | Residual              | 122                 | 43.70061859           | 0.358201792   |                |
| 14 | Total                 | 123                 | 46.6909416            |               |                |
| 15 |                       |                     |                       |               |                |
| 16 |                       | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept             | 0                   | #N/A                  | #N/A          | #N/A           |
| 18 | lag e                 | -0.127936484        | 0.044279146           | -2.889316878  | 0.004570296    |

The  $t$ -ratio is  $-2.89$ . The 5% critical value for the model with no intercept is  $-2.76$ . The  $t$ -ratio falls within the rejection region and the null hypothesis of no cointegration is rejected and we conclude that the two real GDP series are cointegrated.

To get the VEC model results for Australia, we will estimate the following regression.

**Regression**

**Input**

Input Y Range: \$H\$3:\$H\$125

Input X Range: \$P\$3:\$P\$125

☐ Labels ☒ Constant is Zero

☐ Confidence Level: 95 %

**Output options**

☐ Output Range:

☒ New Worksheet Ply: VEC for AUS

☐ New Workbook

**Residuals**

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

The results are:

|    | A                            | B                   | C                     | D             | E              |
|----|------------------------------|---------------------|-----------------------|---------------|----------------|
| 1  | SUMMARY OUTPUT               |                     |                       |               |                |
| 2  |                              |                     |                       |               |                |
| 3  | <i>Regression Statistics</i> |                     |                       |               |                |
| 4  | Multiple R                   | 0.18556239          |                       |               |                |
| 5  | R Square                     | 0.0344334           |                       |               |                |
| 6  | Adjusted R Square            | 0.026453511         |                       |               |                |
| 7  | Standard Error               | 0.640878604         |                       |               |                |
| 8  | Observations                 | 123                 |                       |               |                |
| 9  |                              |                     |                       |               |                |
| 10 | <i>ANOVA</i>                 |                     |                       |               |                |
| 11 |                              | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression                   | 1                   | 1.772289218           | 1.772289218   | 4.315022349    |
| 13 | Residual                     | 121                 | 49.69777165           | 0.410725386   |                |
| 14 | Total                        | 122                 | 51.47006087           |               |                |
| 15 |                              |                     |                       |               |                |
| 16 |                              | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept                    | 0.491705874         | 0.057909469           | 8.490940752   | 6.12439E-14    |
| 18 | lag e                        | -0.098702699        | 0.047515741           | -2.077263187  | 0.039892949    |

The significant negative coefficient on *lag e* indicates that the Australian GDP responds to disequilibrium between the US and Australia.

If we estimate the VEC model for USA,

**Regression**

**Input**

Input Y Range:

Input X Range:

☐ Labels ☐ Constant is Zero

☐ Confidence Level:  %

**Output options**

☐ Output Range:

☒ New Worksheet Ply:

☐ New Workbook

**Residuals**

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

**Normal Probability**

☐ Normal Probability Plots

OK Cancel Help

We obtain:

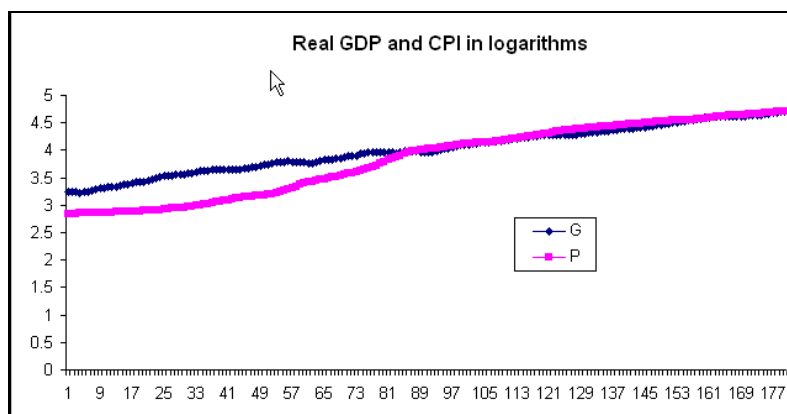
|    | A                     | B            | C              | D           | E           |
|----|-----------------------|--------------|----------------|-------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |             |             |
| 2  |                       |              |                |             |             |
| 3  | Regression Statistics |              |                |             |             |
| 4  | Multiple R            | 0.071619325  |                |             |             |
| 5  | R Square              | 0.005129328  |                |             |             |
| 6  | Adjusted R Square     | -0.003092744 |                |             |             |
| 7  | Standard Error        | 0.516568014  |                |             |             |
| 8  | Observations          | 123          |                |             |             |
| 9  |                       |              |                |             |             |
| 10 | ANOVA                 |              |                |             |             |
| 11 |                       | df           | SS             | MS          | F           |
| 12 | Regression            | 1            | 0.166469321    | 0.166469321 | 0.623848573 |
| 13 | Residual              | 121          | 32.2879441     | 0.266842513 |             |
| 14 | Total                 | 122          | 32.45441342    |             |             |
| 15 |                       |              |                |             |             |
| 16 |                       | Coefficients | Standard Error | t Stat      | P-value     |
| 17 | Intercept             | 0.509884284  | 0.046676827    | 10.92371351 | 9.50767E-20 |
| 18 | lag e                 | 0.030250241  | 0.038299159    | 0.789840853 | 0.431165838 |

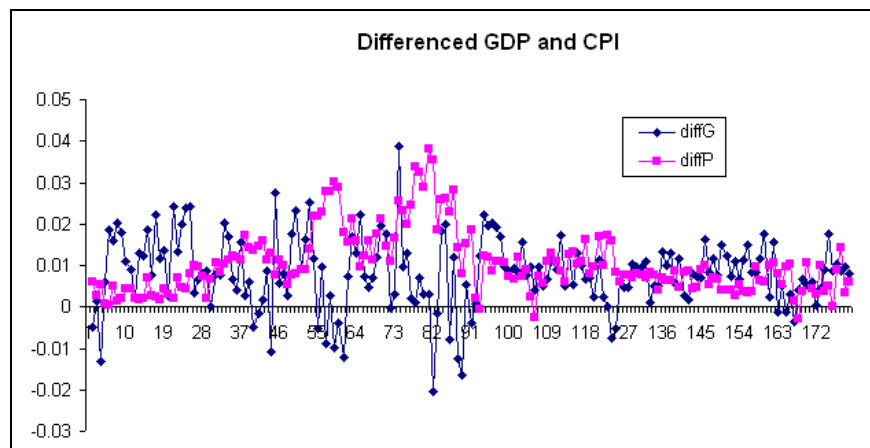
The US, does not appear to respond to the disequilibrium between the two economies. These finding support the idea that the economic conditions in Australia depend on those in the US but not vice verse. In this simple model of two economy trade, the US acts as a large closed economy and Australia as a small open economy.

### 13.3 ESTIMATING A VAR

The VEC model incorporates cointegration equation. It is relevant when two variables are I(1) and integrated. Vector autoregressive model (VAR) is used when there is no cointegration among variables and is estimated using time series that have been transformed to their stationary values. To illustrate VAR model, we will use *growth.xls*. As earlier, the first step is to determine whether the variables are stationary. If they are not, we will difference them and make sure the difference is stationary. Next, we will test for cointegration, if they are cointegrated, we will estimate the VEC model. If the series are not cointegrated, then using the differences and the lagged differences, we will estimate a VAR model.

Let's first plot the series and the differenced series. Plotting will help us determine the existence of intercept and/or trend components.





The fitted least squares regression results of  $G_t$  on  $P_t$  are:

|    | A                     | B            | C              | D           | E           |
|----|-----------------------|--------------|----------------|-------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |             |             |
| 2  |                       |              |                |             |             |
| 3  | Regression Statistics |              |                |             |             |
| 4  | Multiple R            | 0.977255712  |                |             |             |
| 5  | R Square              | 0.955028727  |                |             |             |
| 6  | Adjusted R Square     | 0.954776079  |                |             |             |
| 7  | Standard Error        | 0.08828569   |                |             |             |
| 8  | Observations          | 179          |                |             |             |
| 9  |                       |              |                |             |             |
| 10 | ANOVA                 |              |                |             |             |
| 11 |                       | df           | SS             | MS          | F           |
| 12 | Regression            | 1            | 29.4633339     | 29.4633339  | 3780.082302 |
| 13 | Residual              | 178          | 1.387396627    | 0.007794363 |             |
| 14 | Total                 | 179          | 30.85073053    |             |             |
| 15 |                       |              |                |             |             |
| 16 |                       | Coefficients | Standard Error | t Stat      | P-value     |
| 17 | Intercept             | 1.631500506  | 0.039323334    | 41.48937383 | 1.94269E-93 |
| 18 | P                     | 0.623453779  | 0.010140366    | 61.48237391 | 7.8928E-122 |

Observe the very large  $R^2$ , and  $t$ -ratio. However, based on the plots, the series appear to be nonstationary. To test for cointegration, we estimate the Dickey-Fuller regression. The following are the results from that regression:

|            |              |                |              |             |
|------------|--------------|----------------|--------------|-------------|
| ANOVA      |              |                |              |             |
|            | df           | SS             | MS           | F           |
| Regression | 1            | 0.000111625    | 0.000111625  | 0.954408746 |
| Residual   | 178          | 0.020818365    | 0.000116957  |             |
| Total      | 179          | 0.02092999     |              |             |
|            |              |                |              |             |
|            | Coefficients | Standard Error | t Stat       | P-value     |
| Intercept  | 0            | #N/A           | #N/A         | #N/A        |
| lag e      | -0.009036793 | 0.009250115    | -0.976938456 | 0.329925471 |

Since the  $\tau$  (unit root  $t$ -value) of  $-0.977$  is greater than the 5% critical value of the test for stationarity of  $-3.37$ , we conclude the residuals are nonstationary. Hence the relationship between  $G$  and  $P$  is spurious. Therefore instead of a VEC model, we need to apply a VAR model to these series.

For illustrative purposes, the order of lag in this example will be restricted to 1. In general, one should test for the significance of lag terms greater than 1. The models we will be running are:

$$\Delta P_t = \beta_{10} + \beta_{11}\Delta P_{t-1} + \beta_{12}\Delta G_{t-1}$$

$$\Delta G_t = \beta_{20} + \beta_{21}\Delta P_{t-1} + \beta_{22}\Delta G_{t-1}$$

We need to modify the *growth.xls* worksheet to include the above variables as shown below:

|    | A        | B        | C        | D        | E        | F        | G        | H        | I        |
|----|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1  | G        | P        | e        | lag e    | diffe    | diffG    | diffP    | diff1G   | diff1P   |
| 2  | 3.244286 | 2.836813 | -0.15584 |          |          |          |          |          |          |
| 3  | 3.23927  | 2.842983 | -0.1647  | -0.15584 | -0.00886 | =A3-A2   | =B3-B2   |          |          |
| 4  | 3.240825 | 2.845566 | -0.16475 | -0.1647  | -5.5E-05 | 0.001555 | 0.002583 | =F3      | =G3      |
| 5  | 3.227784 | 2.851042 | -0.18121 | -0.16475 | -0.01646 | -0.01304 | 0.005476 | 0.001555 | 0.002583 |
| 6  | 3.233824 | 2.851683 | -0.17557 | -0.18121 | 0.00564  | 0.00604  | 0.000641 | -0.01304 | 0.005476 |
| 7  | 3.252434 | 2.852006 | -0.15716 | -0.17557 | 0.018409 | 0.01861  | 0.000323 | 0.00604  | 0.000641 |
| 8  | 3.2685   | 2.857125 | -0.14429 | -0.15716 | 0.012875 | 0.016066 | 0.005119 | 0.01861  | 0.000323 |
| 9  | 3.288727 | 2.858399 | -0.12485 | -0.14429 | 0.019433 | 0.020227 | 0.001274 | 0.016066 | 0.005119 |
| 10 | 3.306542 | 2.860313 | -0.10823 | -0.12485 | 0.016622 | 0.017815 | 0.001914 | 0.020227 | 0.001274 |
| 11 | 3.317417 | 2.864758 | -0.10013 | -0.10823 | 0.008104 | 0.010875 | 0.004445 | 0.017815 | 0.001914 |

The least square results for CPI equation are:

| ANOVA      |              |                |             |             |
|------------|--------------|----------------|-------------|-------------|
|            | df           | SS             | MS          | F           |
| Regression | 2            | 0.006972498    | 0.003486249 | 175.4642224 |
| Residual   | 175          | 0.003477025    | 1.98687E-05 |             |
| Total      | 177          | 0.010449523    |             |             |
|            | Coefficients | Standard Error | t Stat      | P-value     |
| Intercept  | 0.001432711  | 0.000710429    | 2.016685391 | 0.045257125 |
| diff1G     | 0.046446771  | 0.03985809     | 1.165303502 | 0.245482057 |
| diff1P     | 0.826820668  | 0.044706493    | 18.4944203  | 5.01742E-43 |

The results indicate that the growth in price ( $\Delta P_t$ ) is significantly related to its own past value ( $\Delta P_{t-1}$ ) but insignificantly related to growth rate in GDP in the previous period ( $\Delta G_{t-1}$ ).

The least square results for  $G$  are:

| ANOVA      |              |                |              |             |
|------------|--------------|----------------|--------------|-------------|
|            | df           | SS             | MS           | F           |
| Regression | 2            | 0.002188739    | 0.001094369  | 17.76577384 |
| Residual   | 175          | 0.010779979    | 6.15999E-05  |             |
| Total      | 177          | 0.012968718    |              |             |
|            | Coefficients | Standard Error | t Stat       | P-value     |
| Intercept  | 0.009814317  | 0.001250908    | 7.845753633  | 4.09303E-13 |
| diff1G     | 0.228509949  | 0.070181314    | 3.255994191  | 0.001356833 |
| diff1P     | -0.326946958 | 0.078718283    | -4.153380182 | 5.11284E-05 |

These results indicate that  $\Delta G_t$  is significantly and positively related to its own past value and significantly and negatively related to last period's inflation ( $\Delta P_{t-1}$ ).

# CHAPTER 14

## An Introduction to Financial Econometrics: Time-Varying Volatility and ARCH Models

### CHAPTER OUTLINE

14.1 ARCH Model and Time Varying Volatility

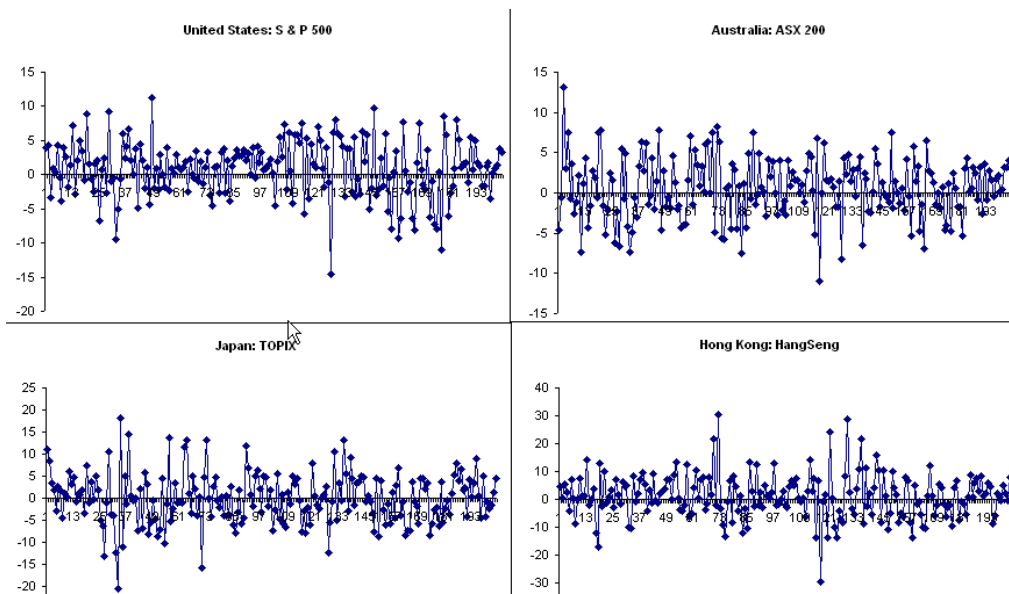
14.2 Testing, Estimating, and Forecasting

### 14.1 ARCH MODEL AND TIME VARYING VOLATILITY

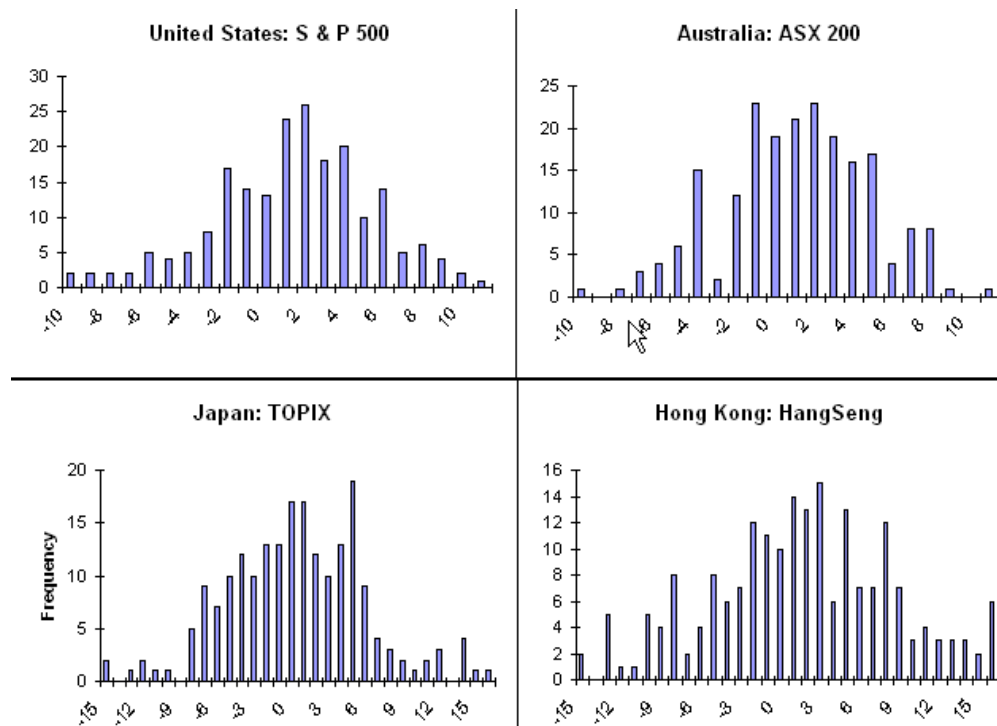
In this chapter, we will estimate several models in which the variance of the dependent variable changes over time. These types of models are broadly referred to as autoregressive conditional heteroskedasticity (ARCH) models. These models have become popular because the variance specification can capture commonly observed features of the time series of financial variables. They are especially useful in modeling volatility and changes in volatility over time. To understand volatility and time-varying volatility, we will look at the *returns.xls* which shows the time series of the monthly returns to four different stock market indices.

Open *returns.xls* and plot each series using **Insert>Chart** and choosing **Line** plot. After some editing, the plots look like this:





The series are characterized by random, rapid changes and are said to be volatile. The volatility seems to change over time for all 4 series. Next, to generate histograms of the series, go to **Tools>Data Analysis>Histogram**. After some editing the histograms look like:



Based on the histograms, we can say the series are leptokurtic. In other words, they have a lot of observations around the average and a relatively large number of observations that are far from average.

## 14.2 TESTING, ESTIMATING, AND FORECASTING

The basic ARCH model has two equations; one equation to describe the behavior of the mean, and another to describe the behavior of the variance.

The mean equation is a linear regression function that contains a constant and possibly some explanatory variables. The example below contains only an intercept.

$$y_t = \beta_0 + e_t \text{ where } e_t \sim N(0, \sigma_t^2)$$

In this case, the series is expected to vary random about its mean. The error of the regression is normally distributed and heteroskedastic. The variance of the current period depends on the information revealed in the proceeding period.

$$h_t = \alpha + \alpha_1 e_{t-1}^2$$

where the variance of  $e_t$  is given by  $h_t$ . Notice that  $h_t$  depends on the squared error in the preceding period. The parameters in this equation have to be positive to ensure a positive variance.

To test for the presence of the ARCH effects, we can use a Lagrange Multiplier (*LM*) test. To conduct this test, we first estimate the mean equation, store least squares residuals from the mean equation, square them and run an auxiliary regression to observe the impact of previous error term for a 1<sup>st</sup> order ARCH process. In other words, testing for ARCH(1), requires to regress  $\hat{e}_t^2$  on the lagged residual  $\hat{e}_{t-1}^2$ .

$$\hat{e}_t^2 = \gamma_0 + \gamma_1 \hat{e}_{t-1}^2 + v_t$$

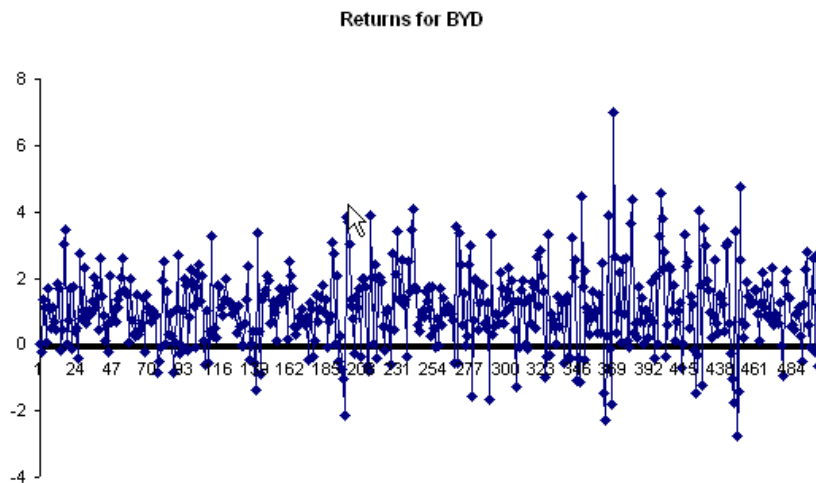
where  $v_t$  is the random error of the auxiliary relationship. The null and alternative hypotheses to be tested are:

$$H_0 : \gamma_1 = 0$$

$$H_1 : \gamma_1 \neq 0$$

The *LM* test statistic is  $NR^2$  with a  $\chi_{(1)}$  distribution under null hypothesis, where  $N$  is the number of observations in the auxiliary regression. The general version of the *LM* test is to conduct an ARCH( $q$ ) test and include  $q$  lags of  $\hat{e}_t^2$  as regressor. This version will have  $\chi_{(q)}$  distribution if the null hypothesis is true and  $N$  in this case will be  $T - q$ .

We will use *byd.xls* data to as the first ARCH example. Open the Excel file named *byd.xls* and go to **Insert>Chart** and plot the **line plot** for visual inspection of the time series.



As can be seen from the plot, there is visual evidence of volatility in the series. We will propose and test an ARCH(1) model against the null hypothesis of no ARCH using the LM test discussed earlier. First, estimate the mean equation. Recall that the mean equation we are running does not have any explanatory variables. Excel will not estimate the model without defining the **X-Range** so we will have to trick Excel. In cell B1, type *intercept* and create a column of ones. Next, estimate the regression using the *r* as the **Y-Range**, and *intercept* as the **X-Range**. Make sure to suppress the intercept, include the labels and keep the residuals.

In order to test the ARCH(1) model, we now need to estimate the auxiliary regression. Copy the residuals from the output section to the data section and paste it in the next available column. Rename it *e*, and create *e square*, and *lag e square* and write the relevant formulas as shown below.

|   | A        | B                | C        | D               | E                   |
|---|----------|------------------|----------|-----------------|---------------------|
| 1 | <b>r</b> | <b>intercept</b> | <b>e</b> | <b>e square</b> | <b>lag e square</b> |
| 2 | 0        | 1                | -1.07829 | =C2^2           |                     |
| 3 | -0.22719 | 1                | -1.30549 | 1.704297        | =+D2                |
| 4 | 1.350843 | 1                | 0.272549 | 0.074283        | 1.704296882         |
| 5 | 1.105559 | 1                | 0.027265 | 0.000743        | 0.074282837         |
| 6 | 0.065499 | 1                | -1.0128  | 1.025754        | 0.000743368         |
| 7 | 1.681189 | 1                | 0.602895 | 0.363482        | 1.025754158         |
| 8 | 1.051365 | 1                | -0.02693 | 0.000725        | 0.363482116         |
| 9 | 0.481928 | 1                | -0.59637 | 0.355653        | 0.000725183         |

Now, to test for first-order ARCH, regress  $\hat{e}_t^2$  on  $\hat{e}_{t-1}^2$  by estimating the regression using the *e square* as the **Y-Range**, and *lag e square* as the **X-Range**.

|    | A                     | B            | C              | D           | E           |
|----|-----------------------|--------------|----------------|-------------|-------------|
| 1  | SUMMARY OUTPUT        |              |                |             |             |
| 2  |                       |              |                |             |             |
| 3  | Regression Statistics |              |                |             |             |
| 4  | Multiple R            | 0.352942118  |                |             |             |
| 5  | R Square              | 0.124568139  |                |             |             |
| 6  | Adjusted R Square     | 0.122806707  |                |             |             |
| 7  | Standard Error        | 2.450017971  |                |             |             |
| 8  | Observations          | 499          |                |             |             |
| 9  |                       |              |                |             |             |
| 10 | ANOVA                 |              |                |             |             |
| 11 |                       | df           | SS             | MS          | F           |
| 12 | Regression            | 1            | 424.501819     | 424.501819  | 70.71979869 |
| 13 | Residual              | 497          | 2983.286264    | 6.002588057 |             |
| 14 | Total                 | 498          | 3407.788083    |             |             |
| 15 |                       |              |                |             |             |
| 16 |                       | Coefficients | Standard Error | t Stat      | P-value     |
| 17 | Intercept             | 0.908261837  | 0.124401233    | 7.301067805 | 1.14072E-12 |
| 18 | lag e square          | 0.35307145   | 0.0419848      | 8.409506448 | 4.3871E-16  |

$$LM = (T-q)R \text{ Square} \\ = (500-1) 0.12457 \\ = 62.16$$

Given the  $LM=62.16$ , we reject the null hypothesis, there is evidence of presence of ARCH(1) process.

Unfortunately, ARCH models and all its extensions are estimated using by the maximum likelihood method and beyond the capabilities of Excel.

# CHAPTER 15

## Panel Data Models

### CHAPTER OUTLINE

- 15.1 Sets of Regression Equations
- 15.2 Seemingly Unrelated Regressions
  - 15.2.1 Breusch-Pagan test of independence
- 15.3 The Fixed Effects Model
  - 15.3.1 A dummy variable model
- 15.4 Random Effects Estimation

### 15.1 SETS OF REGRESSION EQUATIONS

Times series data are observations on the same unit taken over time. For example, annual GDP over a ten-year period would be time-series data. Cross-sectional data are observations at one point in time, over different units, such as 1990 per capita income for each of the 50 U.S. states. In this section we will examine investment data from two firms (cross section) and for 20 periods (time series). Open *grunfeld2.xls*. The descriptive statistics for the two firms can be obtained by choosing **Tools>Data Analysis>Descriptive Statistics**. Below are some of the descriptive statistics after some editing:

|                           | inv_ge      | v_ge        | k_ge        | inv_we     | v_we        | k_we        |
|---------------------------|-------------|-------------|-------------|------------|-------------|-------------|
| <b>Mean</b>               | 102.29      | 1941.325    | 400.16      | 42.8915    | 670.91      | 85.64       |
| <b>Standard Deviation</b> | 48.58449937 | 413.8432895 | 250.6188475 | 19.1101886 | 222.3919274 | 62.26493818 |
| <b>Minimum</b>            | 33.1        | 1170.6      | 97.8        | 12.93      | 191.5       | 0.8         |
| <b>Maximum</b>            | 189.6       | 2803.3      | 888.9       | 90.08      | 1193.5      | 213.5       |
| <b>Count</b>              | 20          | 20          | 20          | 20         | 20          | 20          |

The equations we consider first are the individual investment models.

$$INV_{GE,t} = \beta_1 + \beta_2 V_{GE,t} + \beta_3 K_{GE,t} + e_{GE,t} \quad t=1, \dots, 20$$

$$INV_{WE,t} = \beta_1 + \beta_2 V_{WE,t} + \beta_3 K_{WE,t} + e_{WE,t} \quad t=1, \dots, 20$$

If the models have the same parameters, we can estimate a pooled regression model using all 40 observations. However, if the parameters are not identical the models will be:

$$INV_{GE,t} = \beta_{1,GE} + \beta_{2,GE} V_{GE,t} + \beta_{3,GE} K_{GE,t} + e_{GE,t} \quad t=1,\dots,20$$

$$INV_{WE,t} = \beta_{1,WE} + \beta_{2,WE} V_{WE,t} + \beta_{3,WE} K_{WE,t} + e_{WE,t} \quad t=1,\dots,20$$

We will first estimate two separate regressions using **Tools>Data Analysis>Regression**. Check the **Residuals** option for later use. Click **OK**.

The screenshot shows the 'Regression' dialog box with the following settings:

- Input**
  - Input Y Range: \$A\$1:\$A\$21
  - Input X Range: \$B\$1:\$C\$21
  - ☒ Labels
  - ☐ Confidence Level: 95 %
  - ☐ Constant is Zero
- Output options**
  - ☐ Output Range:
  - ☒ New Worksheet Ply: GE
  - ☐ New Workbook
- Residuals**
  - ☒ Residuals
  - ☐ Standardized Residuals
  - ☐ Residual Plots
  - ☐ Line Fit Plots
- Normal Probability**
  - ☐ Normal Probability Plots

Buttons: OK, Cancel, Help.

Estimate another regression for *WE* using D1 through D21 as the **Y-Range** and E1 through F21 as the **X-Range**, include labels and check the **Residuals** option for later use. Click **OK**.

The screenshot shows the 'Regression' dialog box with the following settings:

- Input**
  - Input Y Range: \$D\$1:\$D\$21
  - Input X Range: \$E\$1:\$F\$21
  - ☒ Labels
  - ☐ Confidence Level: 95 %
  - ☐ Constant is Zero
- Output options**
  - ☐ Output Range:
  - ☒ New Worksheet Ply: WE
  - ☐ New Workbook
- Residuals**
  - ☒ Residuals
  - ☐ Standardized Residuals
  - ☐ Residual Plots
  - ☐ Line Fit Plots
- Normal Probability**
  - ☐ Normal Probability Plots

Buttons: OK, Cancel, Help.

The least squares results for *GE* are:

|            | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
|------------|---------------------|-----------------------|---------------|----------------|
| Regression | 2                   | 31632.03023           | 15816.01511   | 20.34354567    |
| Residual   | 17                  | 13216.58777           | 777.4463394   |                |
| Total      | 19                  | 44848.618             |               |                |
|            |                     |                       |               |                |
|            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept  | -9.956306455        | 31.37424914           | -0.317340071  | 0.754849936    |
| v_ge       | 0.026551189         | 0.015566104           | 1.705705484   | 0.106265098    |
| k_ge       | 0.15169387          | 0.025704083           | 5.901547565   | 1.74209E-05    |

The least squares results for *WE* are:

|            | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
|------------|---------------------|-----------------------|---------------|----------------|
| Regression | 2                   | 5165.552925           | 2582.776462   | 24.76108713    |
| Residual   | 17                  | 1773.23393            | 104.3078783   |                |
| Total      | 19                  | 6938.786855           |               |                |
|            |                     |                       |               |                |
|            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept  | -0.509390184        | 8.015288941           | -0.063552317  | 0.950067995    |
| v_we       | 0.052894126         | 0.015706501           | 3.367658052   | 0.003654762    |
| k_we       | 0.092406492         | 0.056098974           | 1.647204673   | 0.117874323    |

To test whether we should pool the data or not, create the below dummy variable model:

$$INV_{i,t} = \beta_{1,GE} + \delta_1 D_i + \beta_{2,GE} V_{i,t} + \delta_2 D_i + \beta_{3,GE} K_{i,t} + \delta_3 D_i + e_{i,t}$$

where  $D = 1$  for Westinghouse observations and 0 otherwise.

To estimate the dummy variable model in Excel, create three new columns and name them *D*, *DV* and *DK* respectively. Column *D* will take the value 1 for *WE* and 0 for *GE*, Column *DV* will take the respective *v* values for Westinghouse and 0 for General Electric, and Column *DK* will take the respective *k* values for Westinghouse and 0 for General Electric. The worksheet will look like:

|    | A      | B      | C     | D | E     | F    |
|----|--------|--------|-------|---|-------|------|
| 1  | inv_ge | v_ge   | k_ge  | d | dv    | dk   |
| 2  | 33.1   | 1170.6 | 97.8  | 0 | 0     | 0    |
| 3  | 45     | 2015.8 | 104.4 | 0 | 0     | 0    |
| 4  | 77.2   | 2803.3 | 118   | 0 | 0     | 0    |
| 5  | 44.6   | 2039.7 | 156.2 | 0 | 0     | 0    |
| 6  | 48.1   | 2256.2 | 172.6 | 0 | 0     | 0    |
| 7  | 74.4   | 2132.2 | 186.6 | 0 | 0     | 0    |
| 8  | 113    | 1834.1 | 220.9 | 0 | 0     | 0    |
| 9  | 91.9   | 1588   | 287.8 | 0 | 0     | 0    |
| 10 | 61.3   | 1749.4 | 319.9 | 0 | 0     | 0    |
| 11 | 56.8   | 1687.2 | 321.3 | 0 | 0     | 0    |
| 12 | 93.6   | 2007.7 | 319.6 | 0 | 0     | 0    |
| 13 | 159.9  | 2208.3 | 346   | 0 | 0     | 0    |
| 14 | 147.2  | 1656.7 | 456.4 | 0 | 0     | 0    |
| 15 | 146.3  | 1604.4 | 543.4 | 0 | 0     | 0    |
| 16 | 98.3   | 1431.8 | 618.3 | 0 | 0     | 0    |
| 17 | 93.5   | 1610.5 | 647.4 | 0 | 0     | 0    |
| 18 | 135.2  | 1819.4 | 671.3 | 0 | 0     | 0    |
| 19 | 157.3  | 2079.7 | 726.1 | 0 | 0     | 0    |
| 20 | 179.5  | 2371.6 | 800.3 | 0 | 0     | 0    |
| 21 | 189.6  | 2759.9 | 888.9 | 0 | 0     | 0    |
| 22 | 12.93  | 191.5  | 1.8   | 1 | 191.5 | 1.8  |
| 23 | 25.9   | 516    | 0.8   | 1 | 516   | 0.8  |
| 24 | 35.05  | 729    | 7.4   | 1 | 729   | 7.4  |
| 25 | 22.89  | 560.4  | 18.1  | 1 | 560.4 | 18.1 |
| 26 | 18.84  | 519.9  | 23.5  | 1 | 519.9 | 23.5 |
| 27 | 28.57  | 628.5  | 26.5  | 1 | 628.5 | 26.5 |
| 28 | 48.51  | 537.1  | 36.2  | 1 | 537.1 | 36.2 |

Estimate a regression, using A1 through A41 for the **Y-Range**, and cells B1 through F41 as the **X-Range**. Place the results on a worksheet named *OLS with dummies*.

The results are

|            | df | SS          | MS          | F           |
|------------|----|-------------|-------------|-------------|
| Regression | 5  | 72079.40118 | 14415.88024 | 32.69818266 |
| Residual   | 34 | 14989.8217  | 440.8771088 |             |
| Total      | 39 | 87069.22288 |             |             |

|           | Coefficients | Standard Error | t Stat       | P-value     |
|-----------|--------------|----------------|--------------|-------------|
| Intercept | -9.956306455 | 23.62636475    | -0.421406618 | 0.676110524 |
| v_ge      | 0.026551189  | 0.011722048    | 2.265064025  | 0.02999627  |
| k_ge      | 0.15169387   | 0.019356449    | 7.836864688  | 4.01579E-09 |
| d         | 9.446916271  | 28.80535074    | 0.327956995  | 0.744955281 |
| v_we      | 0.026342937  | 0.034352676    | 0.766837983  | 0.448470065 |
| k_we      | -0.059287378 | 0.116946431    | -0.506961843 | 0.61545399  |

From these results, if we plug in 1 for  $D$ , we obtain two estimated regressions, one for  $GE$  and one for Westinghouse. Since  $D = 1$  for  $WE$  and  $D = 0$  for  $GE$ , the respective equations are:

$$GE (D = 0): \quad INV_{i,t} = -9.9563 + 0.02655V_t + 0.15169K_{i,t}$$

$$\text{Westinghouse } (D = 1): \quad INV_{i,t} = -0.5094 + 0.0529V_t + 0.0924K_t$$

The parameter estimates here are exactly as those from the separate regression equations above, but the standard errors are different. This new model treats the coefficients in the same way as before, but now assumes constant variance across all the observations.

To test for the significant differences in the variances, we use the Goldfeld-Quandt test, where -

$$GQ = SSE_{GE} / SSE_{WE}$$

We can get the  $SSE$ s for the separate regressions from the output and calculate  $GQ$  as:



$$GQ = 13216.5871/1773.23405 = 7.45338$$

The  $F(17,17,95)$  critical value is 2.2718 with a  $p$ -value of 0.3284. We reject the null hypothesis of equal variance and find that there is strong evidence that the error variance for the two equations are different.

## 15.2 SEEMINGLY UNRELATED REGRESSIONS

We will now assume that the error terms in the two equations, at the same point in time, are correlated. This is called **contemporaneous correlation**. Adding this assumption of contemporaneous correlation provides additional information to our model and it should be incorporated. Seemingly unrelated regression (SUR) permits equation coefficients and variances to differ and also allows for contemporaneous correlation between the errors,

$$\text{cov}(e_{GE,t}, e_{WE,t}) = \sigma_{GE,WE}$$

While most statistical packages perform SUR estimation automatically, Excel does not. The basic procedures used in SUR are (1) estimate the two equations separately via least squares, (2) use the residuals in step (1) to obtain estimates of the variances and covariances in order to transform the data, and (3) estimate the equations jointly via generalized least squares. Unfortunately, the transformation necessary to perform the generalized least squares is beyond the scope of text.

### 15.2.1 Breusch-Pagan test of independence

In order to determine when to pool the data and use SUR or when to estimate the equations separately, we need to test for the independence of errors. If the errors are not correlated, separate estimation is fine. The test for correlation between the errors is called the Breusch-Pagan test. If the null hypothesis of zero correlation is not rejected, again, separate estimation is fine.

$$BP = Tr_{GE,WE}^2$$

$$\text{where } r_{GE,WE}^2 = \frac{\sigma_{GE,WE}^2}{\sigma_{GE}^2 \sigma_{WE}^2} \text{ and } \sigma_{GE,WE} = \frac{1}{T} = \sum_{i=1}^{20} \hat{e}_{GE,t} \hat{e}_{WE,t}.$$

The test statistic is distributed as  $\chi_{(1)}^2$  under the null hypothesis of no contemporaneous correlation. To calculate the estimated correlation between the equations, we will use the residuals from the separate OLS results.

- Return to the worksheet *GE* and highlight cells **C25 to C45**.
- Choose **Edit>Copy** or click the **Copy** icon.
- Go to the *WE* worksheet and **Paste** in cells D25 through D45.
- In cell E26, type **=C26\*D26** and **copy** this formula down to cell E45.

| 25 | Observation | Predicted i we | Residuals    | Residuals    | e*e         |
|----|-------------|----------------|--------------|--------------|-------------|
| 26 | 1           | 9.786166672    | 3.143833328  | -2.860176107 | -8.99191697 |
| 27 | 2           | 26.85790414    | -0.957904138 | -14.40242074 | 13.79613842 |
| 28 | 3           | 38.73423587    | -3.684235868 | -5.174518855 | 19.06414797 |
| 29 | 4           | 30.80503565    | -7.915035651 | -23.29473664 | 184.378671  |
| 30 | 5           | 29.1618186     | -10.3218186  | -28.03084857 | 289.329334  |
| 31 | 6           | 35.18334018    | -6.613340178 | -0.562215299 | 3.718121027 |
| 32 | 7           | 31.24516001    | 17.26483999  | 40.74959444  | 703.5352276 |
| 33 | 8           | 34.79310815    | 8.546891845  | 16.03552218  | 137.0538737 |
| 34 | 9           | 39.93597243    | -2.915972431 | -23.71921299 | 69.16457116 |
| 35 | 10          | 41.06683077    | -3.256830775 | -26.78010044 | 87.21825527 |
| 36 | 11          | 47.02251951    | -7.752519512 | 1.768123007  | -13.7074081 |
| 37 | 12          | 47.6635511     | 5.796448895  | 58.73723628  | 340.4673884 |
| 38 | 13          | 40.50961605    | 15.05038395  | 43.93586896  | 661.2516972 |
| 39 | 14          | 46.59067745    | 2.969322552  | 31.22712944  | 92.72341968 |
| 40 | 15          | 43.47344125    | -11.43344125 | -23.5520062  | 269.2804791 |
| 41 | 16          | 45.72092623    | -13.48092623 | -37.51099533 | 505.6829607 |
| 42 | 17          | 49.76050037    | 4.619499633  | -4.983022245 | -23.0190694 |
| 43 | 18          | 58.64156885    | 13.13843115  | 1.892879122  | 24.86946203 |
| 44 | 19          | 78.77240423    | 11.30759577  | 5.086901827  | 57.52062956 |
| 45 | 20          | 82.10522249    | -13.50522249 | -2.563001836 | 115.645245  |
| 46 |             |                |              |              | 3528.981227 |
| 47 |             |                |              |              | 176.4490614 |

- In cell E46, type **=SUM(E26:E45)**.
- In cell E47, type **=E46/20**. The result is 176.4490614 and represents  $\sigma_{GE,WE}$ .

Next, we need a degrees of freedom adjustment for the calculation. Excel calculates the estimated model variance by dividing by  $T-K$ . Because the Breusch-Pagan test is only asymptotically justified, we divide only by  $T$ . Since the estimated model variance is the  $SSE$  divided by  $T-K$ , we can simply recalculate using information from the ANOVA tables of our separate regressions.

- In cell D9 of the *GE* worksheet, type **=C13/20**. The result is 660.8293885.
- In cell D9 of the *WE* worksheet, type **=C13/20**. The result is 88.66169652.
- In a blank cell, say E7 of *WE*, type **=(176.4490614^2)/(660.8293885\*88.66169652)**. The result is 0.531389929.
- In cell F7, type **=E7\*20** and
- in cell G7, type **=CHIDIST(F7,1)**.

| BP test statistic | p-value for BP        |
|-------------------|-----------------------|
| <b>=E7*20</b>     | <b>=CHIDIST(F7,1)</b> |
| 0.531389929       | 10.62779858           |
|                   | 0.001114003           |

Based on the *BP test*, we reject the null hypothesis and find evidence of contemporaneous correlation between the error terms. *SUR* is the appropriate analysis.

### 15.3 THE FIXED EFFECTS MODEL

The fixed effects model is a model for pooling data where the intercept is allowed to change across firms but not across time and slope coefficients are assumed to be the same across firms.

Since all behavioral differences between firms and over time will be captured by the intercept, this model is called a **fixed effects model**.

### 15.3.1 A Dummy Variable Specification

The fixed effects model permits cross-section heterogeneity by allowing only the intercept to vary across individuals using dummy variables. We will also extend our model to include data on all ten firms in *grunfeld.xls*. This data set has the investment data for ten companies for 20 periods each. Below is an edited portion of the descriptive statistics of the dataset.

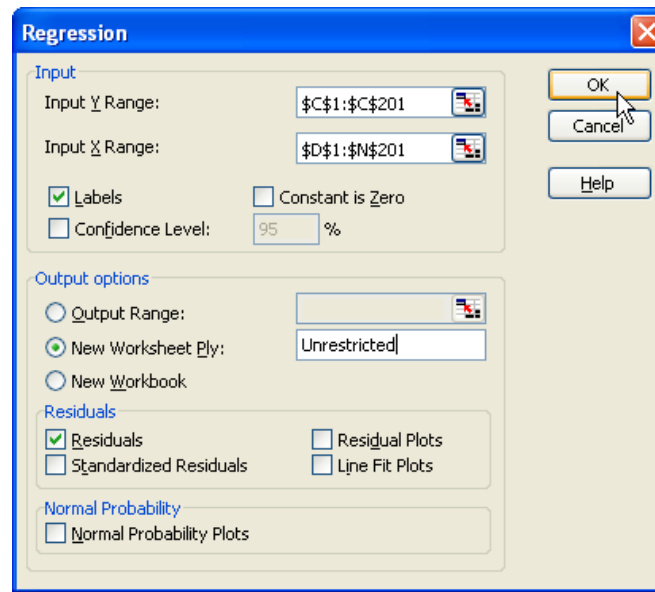
|   | A                  | B           | C           | D           | E           | F           |
|---|--------------------|-------------|-------------|-------------|-------------|-------------|
| 1 |                    | <i>i</i>    | <i>t</i>    | <i>inv</i>  | <i>v</i>    | <i>k</i>    |
| 2 |                    |             |             |             |             |             |
| 3 | Mean               | 5.5         | 10.5        | 145.9067502 | 1081.681102 | 276.5171506 |
| 4 | Standard Deviation | 2.879489067 | 5.780751286 | 216.8855026 | 1314.469701 | 300.9475476 |
| 5 | Minimum            | 1           | 1           | 0.93        | 58.119999   | 0.8         |
| 6 | Maximum            | 10          | 20          | 1486.699951 | 6241.700195 | 2226.300049 |
| 7 | Count              | 200         | 200         | 200         | 200         | 200         |

We will need nine dummy variables for ten firms, one firm will be the base firm and captured by the intercept term. And the coefficient for each firm will be the difference between the intercept for its firm and the intercept for the "base" firm (the "variable of omission"). Recall that if we include all ten dummy variables, we will have perfect multicollinearity and estimation would not be possible.

- First, we will create the dummy variables. Open the file named *grunfeld.xls*.
- Label cells F1 through N1, *D1* through *D9*.
- In cells F2 through F21, type "1" and type "0" in the remaining cells of the column.
- In cells G22 through G41, type "1" and type "0" in all other cells of the column.
- Continue in this fashion, typing "1" appropriately for each firm, and "0" otherwise using the **Copy** and **Paste** functions.

|   | A        | B        | C          | D        | E        | F         | G         | H         | I         | J         | K         | L         | M         | N         |
|---|----------|----------|------------|----------|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1 | <i>i</i> | <i>t</i> | <i>inv</i> | <i>v</i> | <i>k</i> | <i>D1</i> | <i>D2</i> | <i>D3</i> | <i>D4</i> | <i>D5</i> | <i>D6</i> | <i>D7</i> | <i>D8</i> | <i>D9</i> |
| 2 | 1        | 1        | 317.6      | 3078.5   | 2.8      | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| 3 | 1        | 2        | 391.8      | 4661.7   | 52.6     | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| 4 | 1        | 3        | 410.6      | 5387.1   | 156.9    | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| 5 | 1        | 4        | 257.7      | 2792.2   | 209.2    | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| 6 | 1        | 5        | 330.8      | 4313.2   | 203.4    | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| 7 | 1        | 6        | 461.2      | 4643.9   | 207.2    | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |
| 8 | 1        | 7        | 512        | 4551.2   | 255.2    | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 0         | 0         |

- Estimate a regression, using column C as the **Y-Range**, and columns D through N as the **X-Range**.
- Include labels and place results on a worksheet named *Unrestricted*.



The image shows the 'Regression' dialog box in Microsoft Excel. The 'Input' section has 'Input Y Range' set to '\$C\$1:\$C\$201' and 'Input X Range' set to '\$D\$1:\$N\$201'. The 'Labels' checkbox is checked, and 'Constant is Zero' is unchecked. The 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected with the name 'Unrestricted'. The 'Residuals' section has 'Residuals' checked, while 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' are unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. The 'OK' button is highlighted with a mouse cursor.

The results are

|            | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
|------------|---------------------|-----------------------|---------------|----------------|
| Regression | 11                  | 8837969.789           | 803451.799    | 288.8925207    |
| Residual   | 188                 | 522855.1361           | 2781.144341   |                |
| Total      | 199                 | 9360824.926           |               |                |
|            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept  | -6.546268194        | 11.81986618           | -0.553836067  | 0.580349165    |
| v          | 0.109771099         | 0.011854899           | 9.259555738   | 4.70566E-17    |
| k          | 0.310644119         | 0.01737039            | 17.88354297   | 1.98326E-42    |
| D1         | -62.59718141        | 50.30696922           | -1.244304365  | 0.214936634    |
| D2         | 107.408677          | 26.92852115           | 3.988658581   | 9.50954E-05    |
| D3         | -228.5724613        | 26.49588482           | -8.62671554   | 2.61477E-15    |
| D4         | -21.08870132        | 18.0386602            | -1.169083573  | 0.243850136    |
| D5         | -108.770633         | 18.42532863           | -5.903321197  | 1.63464E-08    |
| D6         | -16.52729642        | 17.11168235           | -0.965848716  | 0.33536068     |
| D7         | -60.13665761        | 17.43587875           | -3.449017882  | 0.000694446    |
| D8         | -50.81232281        | 17.97745009           | -2.826447719  | 0.005215892    |
| D9         | -80.7307421         | 17.36690092           | -4.648540491  | 6.28015E-06    |

To obtain the results in Table 17.2, remember that we did not include a dummy variable for the tenth firm, so its intercept is  $-6.546$ . And for all the other firms, the intercept is the difference between the coefficient of the firm dummy and the intercept. For example, for Firm 1, the intercept is the difference  $-6.546 - 62.60 = -69.146$ . Based on the  $p$ -values for the dummy coefficients, it appears that Firms 1, 4, and 6 do not have intercepts that differ significantly from  $-6.546$ .

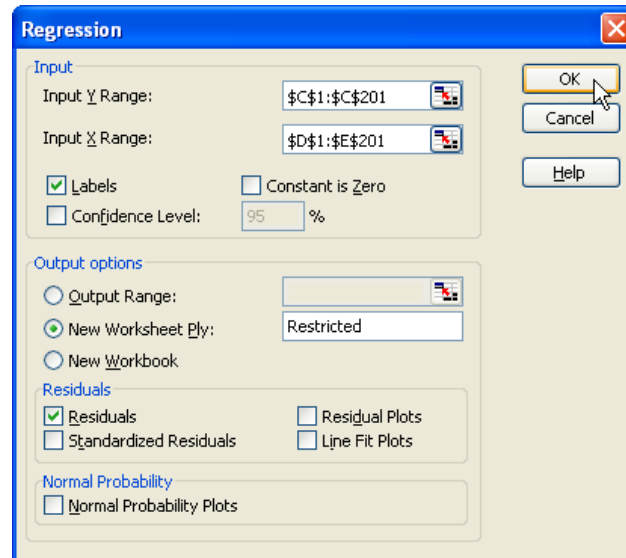
To test the equality of the intercepts, the null and alternative hypotheses are:

$$H_0 : \beta_{1,D1} = \beta_{1,D2} = \beta_{1,D3} = \beta_{1,D4} = \beta_{1,D5} = \beta_{1,D6} = \beta_{1,D7} = \beta_{1,D8} = \beta_{1,D9}$$

$$H_1 : \beta_{1,i} \text{ are not all equal}$$

This is a regular  $F$ -test where we can use the *template* we previously created. The model we estimated above is the unrestricted model, so now we just need to run the restricted model, where we force all nine intercept coefficients to be equal. We can run the restricted model by returning

to the worksheet containing the data and running a regression, including only  $V$  and  $K$  as the **X-Range**. Place the results on a worksheet named *Restricted*.



The image shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' set to '\$C\$1:\$C\$201' and 'Input X Range' set to '\$D\$1:\$E\$201'. The 'Labels' checkbox is checked, and 'Constant is Zero' is unchecked. The 'Confidence Level' is set to 95%. The 'Output options' section has 'New Worksheet Ply' selected with the name 'Restricted'. The 'Residuals' section has 'Residuals' checked, while 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' are unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. The 'OK' button is highlighted with a mouse cursor.

The results are:

|    |            |                     |                       |               |                |
|----|------------|---------------------|-----------------------|---------------|----------------|
| 11 |            | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| 12 | Regression | 2                   | 7611697.286           | 3805848.643   | 428.6434938    |
| 13 | Residual   | 197                 | 1749127.64            | 8878.820507   |                |
| 14 | Total      | 199                 | 9360824.926           |               |                |
| 15 |            |                     |                       |               |                |
| 16 |            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| 17 | Intercept  | -43.02447743        | 9.497895769           | -4.529895724  | 1.02136E-05    |
| 18 | v          | 0.115374318         | 0.005830177           | 19.78916102   | 1.04325E-48    |
| 19 | k          | 0.231931394         | 0.025464874           | 9.107894735   | 9.5536E-17     |

- Open the *F-template* created earlier.
- **Copy** the SSE from cell C13 of the *Unrestricted* model and
- **Paste** it to cell B7 of the template.
- From the *Restricted* model, **Copy** the SSE in cell C13 and
- **Paste** it to cell B6 of template.
- We are testing for 9 restrictions and sample size is 200, so set the  $J = 9$ ,  $T = 200$ , and  $K = 12$ .

|                             |             |
|-----------------------------|-------------|
| Hypothesis Testing - F-Test |             |
| <b>Data Input</b>           |             |
| J                           | 9           |
| T                           | 200         |
| K                           | 12          |
| SSE-RESTRICTED              | 1749127.683 |
| SSE-UNRESTRICTED            | 522855.1655 |
| ALPHA                       | 0.05        |
| <b>Computed Values</b>      |             |
| df-numerator                | 9           |
| df-denominator              | 188         |
| F                           | 48.99152204 |
| Right Critical Value        | 1.92995686  |
| Decision                    | Reject Ho   |
| p-value                     | 1.11131E-44 |

We reject the null hypothesis and find evidence that at least one firm has an intercept different from the rest.

## **15.4 RANDOM EFFECTS ESTIMATION**

We only make inferences about the firms on which we have data. The error components model assumes the intercepts are random variables, drawn from a population distribution of firm intercepts. One result of this is that the error terms from the same firm in different time periods are correlated. The error components model, therefore, is sometimes called a **random effects model**.

We know that generalized least squares estimation is appropriate when we have correlated error terms. While it is possible to carry out the calculations with Excel, you are best advised to use an econometric software package with a specialized routine for estimation of the random effects model.

# CHAPTER 16

## Qualitative and Limited Dependent Variable Models

### CHAPTER OUTLINE

16.1 Models with Binary Dependent Variables  
16.1.1 The linear probability model

16.1.2 Least squares estimation of the linear probability model

### 16.1 MODELS WITH BINARY DEPENDENT VARIABLES

So far, our focus has been on models in which the dependent variables is continuous; prices, revenues, quantities etc. However, as a general theory of choice, microeconomics deals with many choices where the outcomes are qualitative. In this chapter, we will consider choices that are of the "either-or" type. That is, we choose to buy a particular car or not, we choose one job over another, we vote either for or against a particular issue. In trying to explain these types of choices, the dependent variable is *dichotomous*, or *binary*, since we quantify the choices by assigning values zero or one. We then construct a statistical model that explains why particular choices are made and what factors influence those choices.

To illustrate a model with a dichotomous dependent variable, consider a problem from transportation economics. Workers can either drive to work (private transportation) or take a bus (public transportation). For simplicity, we will assume these are the only two alternatives. The individual's choice will be represented by a dummy variable  $y$ , where

$$y = \begin{cases} 1 & \text{if the individual drives to work} \\ 0 & \text{if the individual takes a bus to work} \end{cases}$$

If we collect 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 this 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  $\text{var}(y) = p(1-p)$ .

If we assume that the only factor that determines the probability that an individual chooses one mode of transportation over the other is the difference in time to get to work between the two modes, then we define the explanatory variable  $x$  as

$$x = (\text{commuting time by bus} - \text{commuting time by car})$$

While there are other factors that affect this choice, we will focus on this single explanatory variable. *A priori* we expect a positive relationship between  $x$  and  $p$  that is as  $x$  increases, the individual will be more inclined to drive.

### 16.1.1 The linear probability model

In regression analysis, the dependent variable is broken into two parts; fixed (systematic) and random (stochastic). If we apply this to random variable  $y$ , we have

$$y = E(y) + e = p + e$$

We then relate the systematic portion of  $y$  to the explanatory variables that we believe will help explain the expected value. We are assuming that the probability of driving is positively related to the difference in driving times,  $x$ , in this example. If we assume a linear relationship, then we will have the following *linear probability model*.

$$E(y) = p = \beta_1 + \beta_2 x$$

In this chapter, we will examine the problems with least squares estimation in the context of binary choice models. However, for these models least squares estimation methods are not the best choice. Instead, maximum likelihood estimation (see Appendix C.8 of your book) is the method to use. Excel does not have the capabilities to perform maximum likelihood estimation. Other statistical packages such as EViews or SAS should be used when dealing with binary choice models.

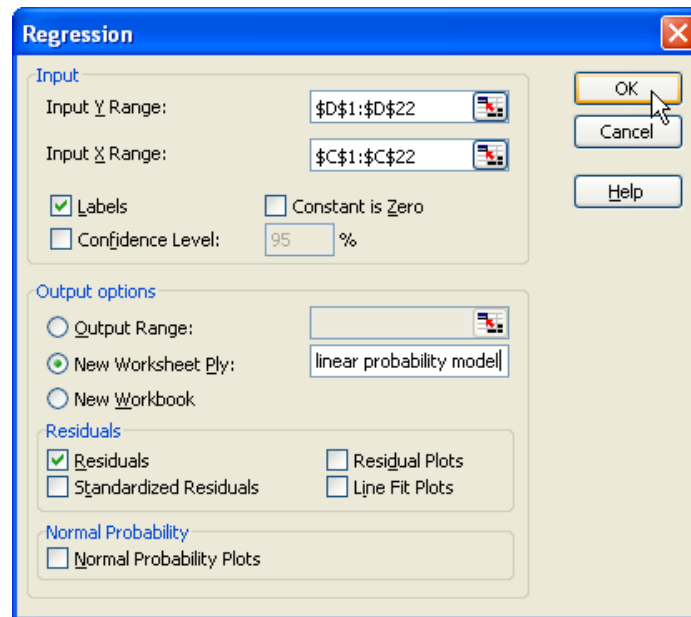
## 16.2 Least squares estimation of the linear probability model

The linear regression model for explaining the choice variable  $y$  is called the *linear probability model* and is given by

$$y = E(y) + e = \beta_1 + \beta_2 x + e$$

To see how to apply the *linear probability model* in Excel, open the file **transport.xls**. Estimate the regression, using *auto* as the **Y-Range** and *dtime* as the **X-Range**. Include labels and check the **Residuals** option. label the worksheet “**linear probability model**” and click **OK**.





The least squares results are

| ANOVA      |                     |                       |               |                |
|------------|---------------------|-----------------------|---------------|----------------|
|            | <i>df</i>           | <i>SS</i>             | <i>MS</i>     | <i>F</i>       |
| Regression | 1                   | 3.202181455           | 3.202181455   | 29.8840983     |
| Residual   | 19                  | 2.035913784           | 0.107153357   |                |
| Total      | 20                  | 5.238095238           |               |                |
|            | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
| Intercept  | 0.484795068         | 0.071449411           | 6.785151347   | 1.76499E-06    |
| dtime      | 0.007030992         | 0.001286164           | 5.466635007   | 2.8342E-05     |

The explanatory variable is significant, suggesting that an increase of one minute in the difference between the time it takes to get to work by bus versus by car increases the probability of driving to work. However, the linear probability model has a very serious problem. Let's look at the fitted model, using least squares estimation:

$$E(y) = \hat{p} = 0.4848 + 0.007dtime$$

For certain values of dtime, the estimated probability might turn out less than zero or greater than one which is NOT possible for any valid probability function. If we look at the residual output we can observe multiple occurrences of this problem.

| RESIDUAL OUTPUT    |                       |                  |
|--------------------|-----------------------|------------------|
| <i>Observation</i> | <i>Predicted auto</i> | <i>Residuals</i> |
| 1                  | 0.143791971           | -0.143791971     |
| 2                  | 0.656351265           | -0.656351265     |
| 3                  | 1.066961201           | -0.066961201     |
| 4                  | 0.311832673           | -0.311832673     |
| 5                  | 0.262615731           | -0.262615731     |
| 6                  | 1.124615311           | -0.124615311     |
| 7                  | 0.851109721           | 0.148890279      |
| 8                  | -0.131822882          | 0.131822882      |
| 9                  | 0.365268209           | -0.365268209     |
| 10                 | 0.122698996           | -0.122698996     |
| 11                 | -0.152915857          | 0.152915857      |
| 12                 | 0.945325023           | 0.054674977      |

The problem arises because, the linear probability model is an increases function in  $x$  and the increase is constant. However, given the requirement for a valid probability function of  $0 \leq p \leq 1$ , a constant rate of increase is not possible. Unfortunately, more appropriate models such as the Logit and Probit model can not be estimated using the standard version of Excel.

# CHAPTER 17

## Importing Internet Data

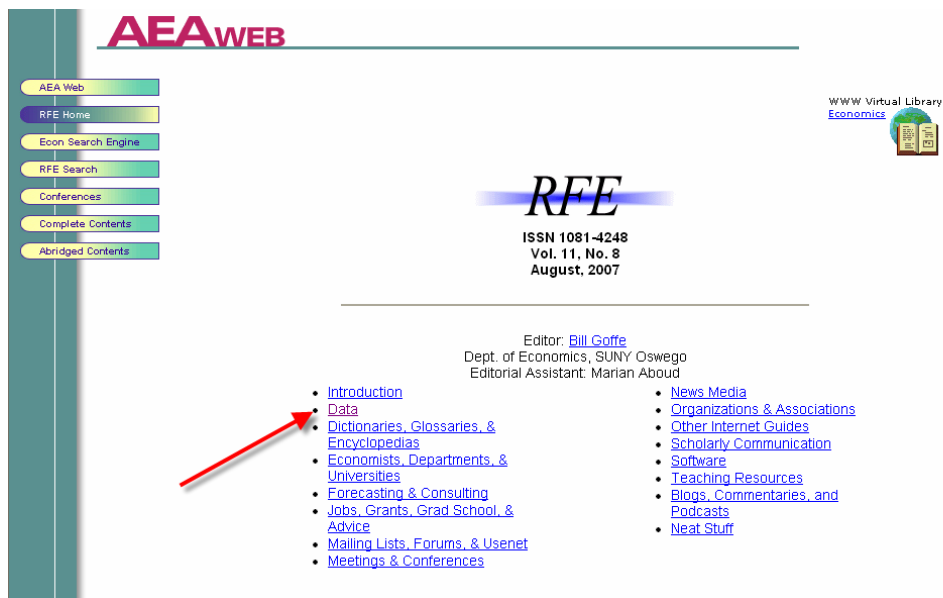
Up to now, we have taken you through various econometric methodologies and applications using already prepared Excel workfiles. In this chapter, we show you how to import data into an Excel spreadsheet.

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 *POE* 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.

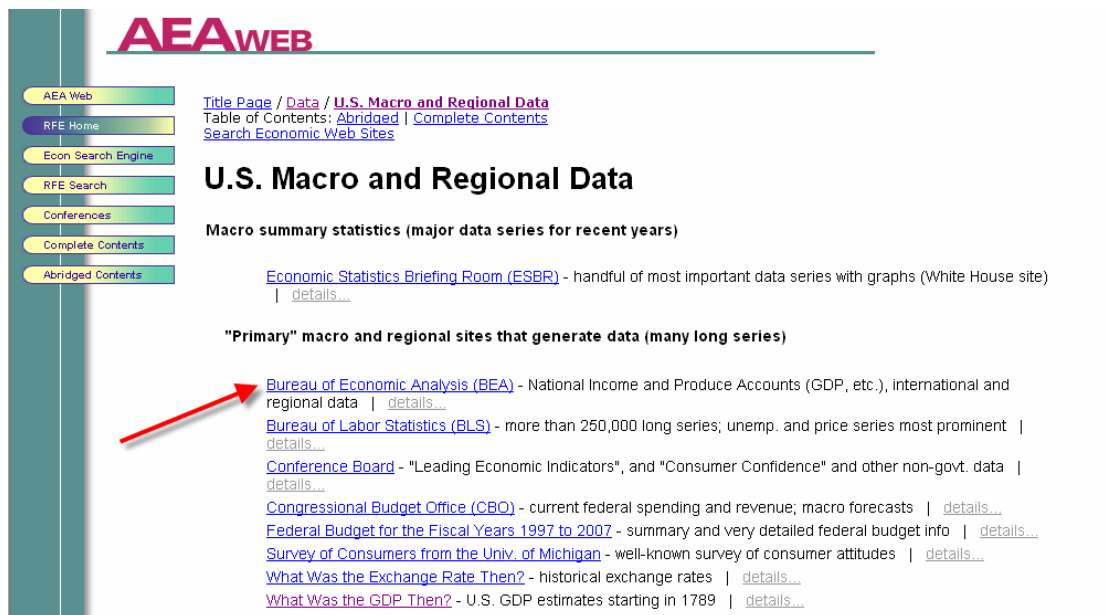
First, open up the website: [www.rfe.org](http://www.rfe.org) :



Select the **Data** option and then select **U.S. Macro and Regional Data**.



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.



From the screen below, select the Gross Domestic Product (GDP) option.

| U.S. Economic Accounts                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>National</b><br><hr/> Access National Economic Accounts Data <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>• Research and Development</li> </ul> </li> <li>▶ View all <a href="#">National Accounts</a> Information...</li> </ul> | <b>International</b><br><hr/> Access International Economic Accounts Data <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>                                                               |
| <b>Regional</b><br><hr/> Access Regional Economic Accounts Data <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>                    | <b>Industry</b><br><hr/> Access Industry Economic Accounts Data <ul style="list-style-type: none"> <li>▶ <a href="#">Annual Industry Accounts</a> <ul style="list-style-type: none"> <li>• GDP by Industry</li> <li>• Input-Output Accounts</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>• Research and Development</li> <li>• Travel and Tourism</li> </ul> </li> <li>▶ <a href="#">Supplemental Estimates</a></li> <li>▶ View all <a href="#">Industry Accounts</a> Information...</li> </ul> |

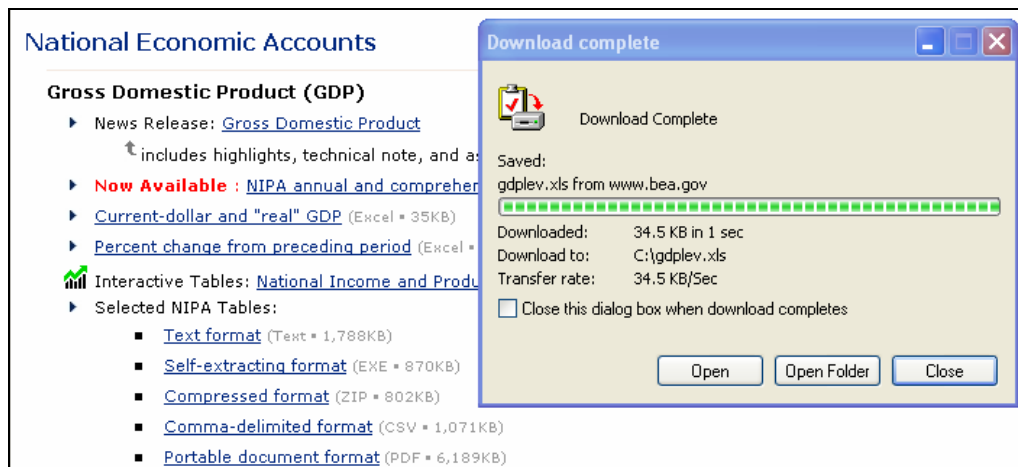
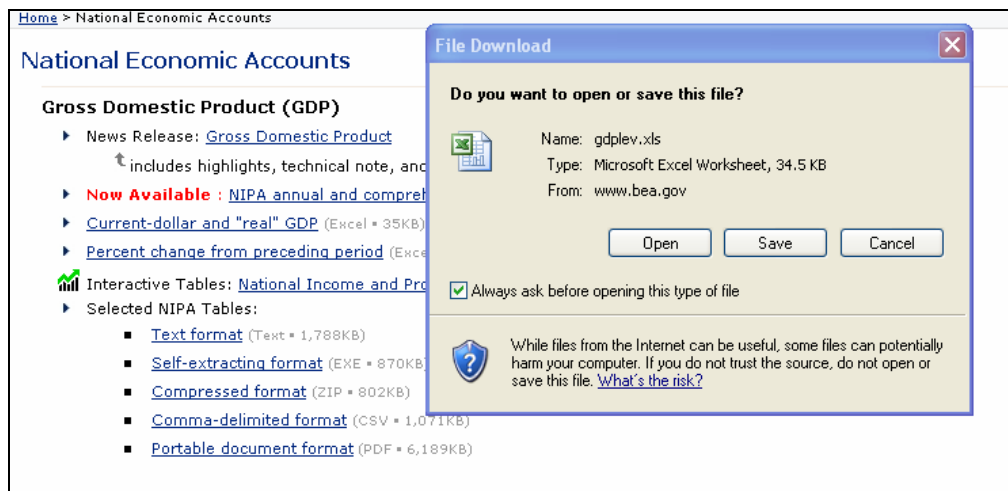
Most websites allow you to download data conveniently in an Excel format.

## 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*.



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  |                                                         |             |              |   |                                    |              |         |
|    |                                                         | GDP in      | GDP in       |   | GDP in                             | GDP in       |         |
|    |                                                         | billions of | billions of  |   | billions of                        | billions of  |         |
|    |                                                         | current     | chained      |   | current                            | chained      |         |
|    |                                                         | dollars     | 2000 dollars |   | dollars                            | 2000 dollars |         |
| 6  |                                                         |             |              |   |                                    |              |         |
| 7  |                                                         |             |              |   |                                    |              |         |
| 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 |

# **APPENDIX B**

## Review of Probability Concepts

### **CHAPTER OUTLINE**

#### **B.1 Binomial Probabilities**

##### **B.1.1 Computing binomial probabilities directly**

##### **B.1.2 Computing binomial probabilities using BINOMDIST**

#### **B.2 The Normal Distribution**

Excel has a number of functions for computing probabilities. In this chapter we will show you how to work with the probability function of a binomial random variable, how to compute probabilities involving normal random variables.

### **B.1 BINOMIAL PROBABILITIES**

A binomial experiment consists of a fixed number of trials,  $n$ . On each independent trial the outcome is success or failure, with the probability of success,  $p$ , being the same for each trial. The random variable  $X$  is the number of successes in  $n$  trials, so  $x = 0, 1, \dots, n$ . For this discrete random variable, the probability that  $X = x$  is given by the probability function

$$P(X = x) = f(x) = \left( \frac{n!}{x!(n-x)!} \right) p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

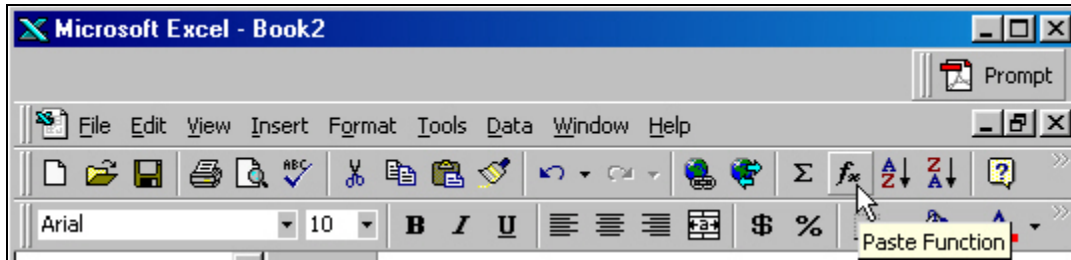
We can compute these probabilities two ways: the hard way and the easy way.

#### **B.1.1 Computing binomial probabilities directly**

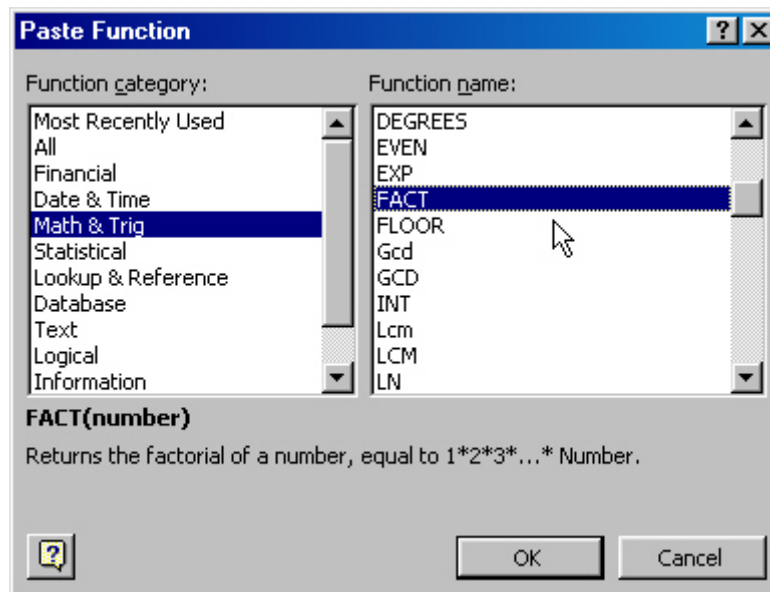
Excel has a number of mathematical functions that make computation of formulas straightforward. Assume there are  $n = 5$  trials, that the probability of success is  $p = 0.3$ , and that we want the probability of  $x = 3$  successes. What we must compute is

$$P(X = 3) = f(3) = \left( \frac{5!}{3!(5-3)!} \right) .3^3 (1-.3)^{5-3}$$

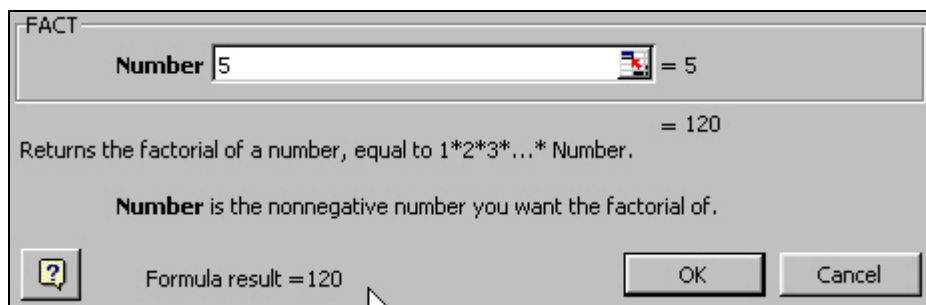
Eventually you will learn many shortcuts in Excel, but should you forget how to compute some mathematical or statistical quantity, there is a **Paste Function** ( $f_*$ ) button on the Excel toolbar,



Click on the **Paste Function** button, select **Math & Trig** in the first column, and scroll down the list of functions in the right-hand column. When you reach **Fact** you see that this function returns the factorial of a number.

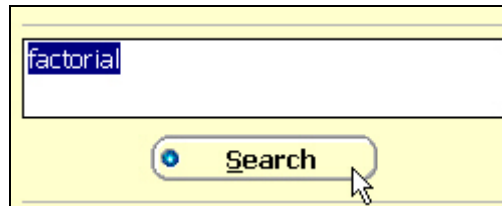


Click **OK**. In the resulting dialog box, enter 5 and Excel determines that  $5! = 120$ .

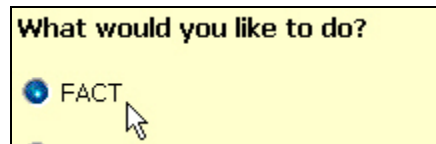




Alternatively, click on **Help**. In the resulting dialog box, enter **factorial** and click **Search**



Click on **FACT**.



You are presented with an Excel function, **FACT(number)**, a definition and some examples.

The other mathematical operations we need to compute the binomial probability are multiplication (\*), division (/) and power (^).

In cell A1 type “f(3)”, and in B1 type the formula

$$=(\text{FACT}(5)/(\text{FACT}(3)*\text{FACT}(2)))*(0.3^3)*(0.7^2)$$

It will look like

|    |      |   |                                              |
|----|------|---|----------------------------------------------|
| B1 |      | = | =(FACT(5)/(FACT(3)*FACT(2)))*(0.3^3)*(0.7^2) |
|    | A    |   | B                                            |
| 1  | f(3) |   | =(FACT(5)/(FACT(3)*FACT(2)))*(0.3^3)*(0.7^2) |

Note that we have used parentheses to group operations.

Hit **<enter>**, and the result is 0.1323.

### B.1.2 Computing binomial probabilities using BINOMDIST

The Excel function **BINOMDIST** can be used to find either cumulative probability,  $P(X \leq x)$  or the probability function,  $P(X = x)$  for a Binomial random variable. Syntax for the function is

**BINOMDIST(number\_s, trials, probability\_s, cumulative)**

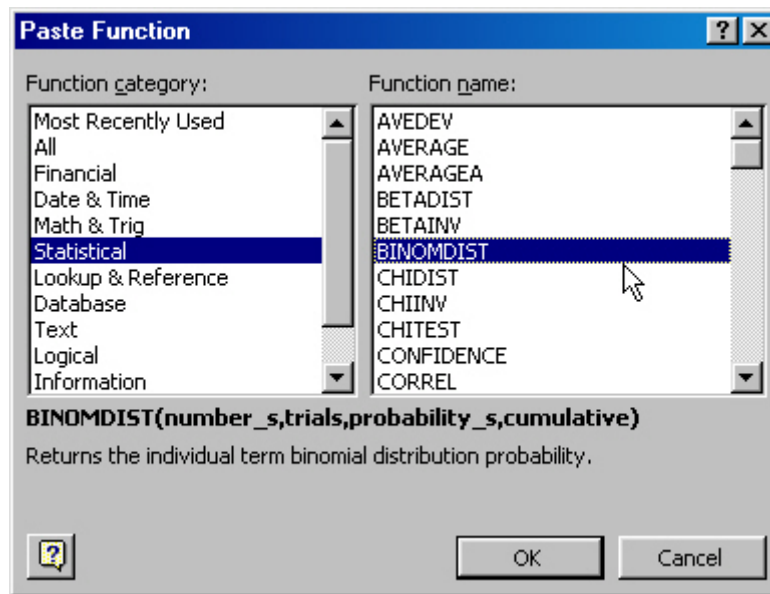
where **number\_s** is the number of successes in  $n$  trials

**trials** is the number of independent trials ( $n$ )

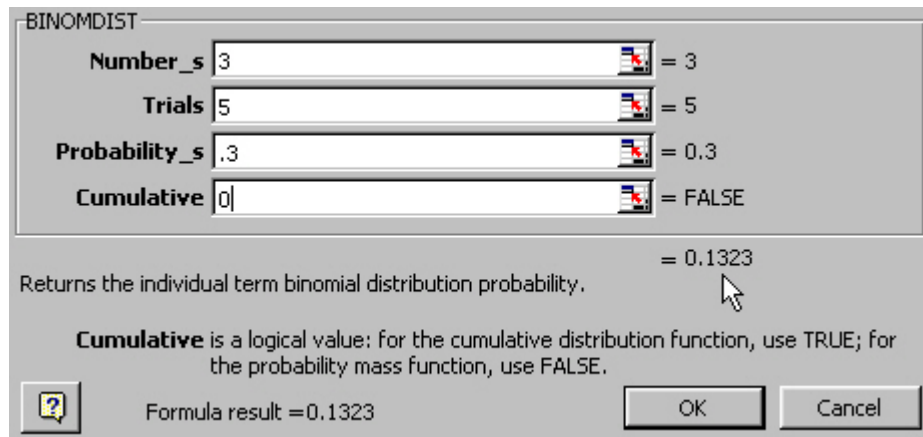
**probability** is  $p$ , the probability of success on any one trial

**cumulative** is a logical value. If set equal to 1 (true), the cumulative probability is returned; if set to 0 (false), the probability mass function is returned.

Access this function by clicking the **Paste Function** button. Select **Statistical** in the Function category and **BINOMDIST** in the Function name.



Using the values  $n = 5$ ,  $p = .3$ , and  $x = 3$  we obtain the probability 0.1323, as above.



Alternatively, we can type the function equation directly into a cell. For example, if  $p = .2$  and  $n = 10$ , to find the probability that  $X = 4$  and  $X \leq 4$ , the worksheet would appear as follows:

|                               |                 |
|-------------------------------|-----------------|
| <b>=BINOMDIST(4,10,0.2,0)</b> | <b>0.08808</b>  |
| <b>=BINOMDIST(4,10,0.2,1)</b> | <b>0.967207</b> |

The formulas in the first column produce the results reported in the second column.

## **B.2 THE NORMAL DISTRIBUTION**

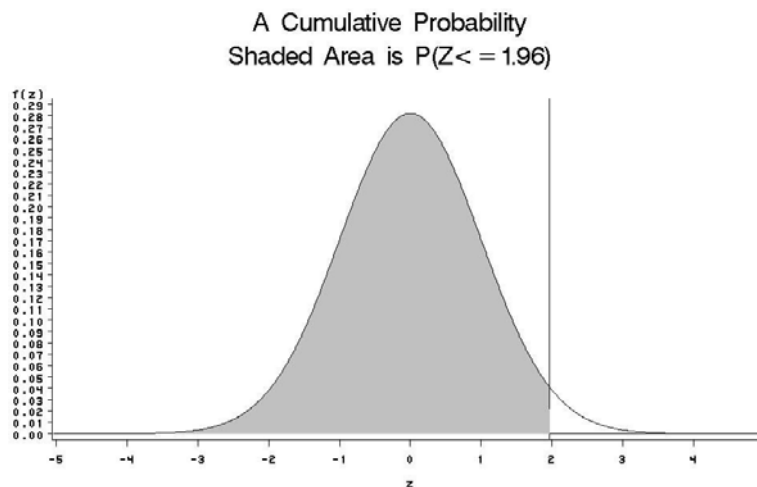
Excel provides several functions related to the Normal and Standard Normal Distributions.

1. The **STANDARDIZE** function computes the  $Z$  value for given values of  $X$ ,  $\mu$ , and  $\sigma$ . The format of this function is

**STANDARDIZE( $X$ ,  $\mu$ ,  $\sigma$ )**

Referring to the example in *POE* Section B.5.1 in which  $\mu = 3$  and  $\sigma = 3$ , if we wanted to find the  $Z$  value corresponding to  $X = 6$ , we would enter **=STANDARDIZE(6,3,3)** in a cell, and the value computed would be 1.0.

2. The **NORMSDIST** function computes the area, or cumulative probability, less than a given  $Z$  value. Geometrically, the cumulative probability is



The format of this function is

**NORMSDIST( $Z$ )**

If we wanted to find the area below a  $Z$  value of 1.0, we would enter **=NORMSDIST( 1.0)** in a cell, and the value computed would be .8413.

3. The **NORMSINV** function computes the  $Z$  value corresponding to a given cumulative area under the normal curve. The format of this function is

**NORMSINV(prob)**

where **prob** is the area under the standard normal curve less than  $z$ . That is,  $prob = P(Z < z)$ . If we wanted to find the  $z$  value corresponding to a cumulative area of .10, we would enter **=NORMSINV(.10)** in a cell and the value computed would be  $-1.2815$ .

4. The **NORMDIST** function computes the area or probability less than a given  $X$  value. The format of this function is

**NORMDIST( $X, \mu, \sigma, \text{TRUE}$ )**

TRUE is a logical value, which can be replaced by 1. If we wanted to find the area below an  $X$  value of 6, we would enter **=NORMDIST(6,3,3,1)** in a cell, and the value computed would be .8413.

5. The **NORMINV** function computes the  $x$  value corresponding to a cumulative area under the normal curve. The format of this function is

**NORMINV(prob,  $\mu, \sigma$ )**

where **prob** is the area under the normal curve less than  $x$ . That is,  $prob = P(X < x)$ . To compute the value of  $x$  such that .10 of the probability is to the left, enter **=NORMINV(.10,3,3)** in a cell, yielding  $-0.8446$ .

For the example in Section 2.6, a template can be built in Excel to compute probabilities and values of  $X$  corresponding to particular probabilities. The highlighted cells require user input. The formulas in the other cells do the computations. Set up a spreadsheet that looks like the following

|    | <b>A</b>                  | <b>B</b>                                     |
|----|---------------------------|----------------------------------------------|
| 1  | Normal Probabilities      |                                              |
| 2  | mean                      |                                              |
| 3  | standard_dev              |                                              |
| 4  |                           |                                              |
| 5  | Left-tail Probability     |                                              |
| 6  | a                         |                                              |
| 7  | $P(X \leq a)$             | =NORMDIST(B6,B2,B3,1)                        |
| 8  |                           |                                              |
| 9  | Right-tail Probability    |                                              |
| 10 | a                         |                                              |
| 11 | $P(X \geq a)$             | =1-NORMDIST(B10,B2,B3,1)                     |
| 12 |                           |                                              |
| 13 | Interval Probability      |                                              |
| 14 | a                         |                                              |
| 15 | b                         |                                              |
| 16 | $P(a \leq X \leq b)$      | =NORMDIST(B15,B2,B3,1)-NORMDIST(B14,B2,B3,1) |
| 17 |                           |                                              |
| 18 | <b>Inverse cumulative</b> |                                              |
| 19 | Left-tail probability     |                                              |
| 20 | Quantile                  | =NORMINV(B19,B2,B3)                          |

Using  $X \sim N(3,9)$ , the above template would produce the following results:

### Normal Probabilities

|              |                                |
|--------------|--------------------------------|
| mean         | <input type="text" value="3"/> |
| standard_dev | <input type="text" value="3"/> |

### Left-tail Probability

|               |                                |
|---------------|--------------------------------|
| a             | <input type="text" value="6"/> |
| $P(X \leq a)$ | 0.84134474                     |

### Right-tail Probability

|               |                                |
|---------------|--------------------------------|
| a             | <input type="text" value="6"/> |
| $P(X \geq a)$ | 0.15865526                     |

### Interval Probability

|                      |                                |
|----------------------|--------------------------------|
| a                    | <input type="text" value="4"/> |
| b                    | <input type="text" value="6"/> |
| $P(a \leq X \leq b)$ | 0.210786144                    |

### Inverse cumulative

|                       |                                   |
|-----------------------|-----------------------------------|
| Left-tail probability | <input type="text" value="0.95"/> |
| Quantile              | 7.934559001                       |

Note that the Quantile = 7.93 gives the top 5% "cut off" value.

Once again, if you forget these formulas, use the **Paste Function** ( $f_*$ ) button on the Menu Bar.