

Data Analysis Through Modeling:
Thinking and Writing in Context
Supplement: Computer Guide for Excel and StatPro

Kris Green and Allen Emerson

Fall 2014 Edition¹

¹©2014 Kris H. Green and W. Allen Emerson

Contents

1	Format of computer information in this text	1
2	Basic Excel Information	3
2.1	Advice on computers and doing work electronically	3
2.2	Using the help system in Microsoft Office	4
2.3	Excel desktop	5
2.4	Saving and Opening Files	6
2.5	A Note About Naming Files	7
2.6	Folders and Organization	7
2.7	Variable names in Excel and StatPro	7
2.8	Good Data Entry Practice	8
2.9	Comments in Excel	8
2.10	Copying and pasting between programs	9
2.11	Renaming a worksheet in Excel	9
2.12	Sizing columns to make data fit	9
2.13	Pasting graphs and charts from Excel into a Word document to make a report	9
3	Cell References in Excel	13
3.1	Cell References in Excel	13
3.2	Absolute Cell References in Excel	14
3.3	Three dimensional cell references in Excel	14
3.4	Naming Cell Ranges	14
4	Copying Formulas in Excel	17
4.1	Using Copy and Paste Commands	17
4.2	Dragging the Fill Handle	17
4.3	Double-clicking the fill handle	17
5	Computing Summary Statistics in Excel	19
5.1	Computing standard deviation in Excel with the built-in formula	19
5.2	Computing Means in Excel	19
5.3	Adding up a list of values	20
5.4	Computing deviations in Excel	20
5.5	Computing standard deviation in Excel with a user-created formula	20
5.6	Computing Medians in Excel (Method #1)	21

5.7	Computing Medians in Excel (Method #2)	21
5.8	Computing the Mode in Excel	21
5.9	Computing z-scores	21
5.10	Percentiles in StatPro	22
5.11	Percentiles in Excel	22
5.12	Correlation Tables using StatPro	22
6	Summary statistics in StatPro	23
6.1	Correlation with Excel	25
7	Using the Excel 2007 status bar	27
8	Making and Using Pivot Tables	29
8.1	Advice on Making Pivot Tables	29
8.2	Making a simple Pivot Table	29
8.3	Advanced Pivot Table Options	30
8.4	The Pivot Table Ribbon	31
8.5	Grouping items in the table	31
8.6	Making a more complex pivot table	33
9	Sorting data	35
9.1	Using the Data Ribbon	35
9.2	Using Autofilter to Sort and View Data	36
9.3	How to Stack and Unstack Data Using StatPro	37
10	Making Charts and Graphs with Excel and StatPro	39
10.1	Making Boxplots with StatPro	39
10.2	Histograms in StatPro	41
10.2.1	Making histograms of z-scores with StatPro	43
10.2.2	Adding Information to Histograms	43
10.3	Histograms in Excel	44
10.3.1	Using array formulas to make a frequency table	44
10.3.2	Creating a Histogram from a Frequency Table	45
10.4	Scatterplots with StatPro	46
10.5	Scatter Plots in EXCEL without StatPro	47
10.6	Moving a Graph To Another Worksheet	47
10.7	Adding Trend Lines to a Scatter Plot	48
10.8	Using the drawing tools in Excel (or Word)	49
10.9	Adding Trendlines for Non-proportional Models	50
10.10	Making a 3D surface plot	50
10.11	Logarithmic and Log-Log plots	50
10.12	Create a scatter plot with more than one y -variable on the same axes	52
10.13	Scatterplots with multiple curves	52
10.13.1	Finding the Intersection Point of Two Curves	52
10.13.2	Graphing More Than One Scatterplot and Trendline on the Same Axes	53

10.13.3 Graphing More Than One Equation on the Same Axes	55
10.14 Drawing on a chart	55
11 Estimating Statistics from Summary Data	59
11.1 Estimating the mean in Excel	59
11.2 Estimating the standard deviation of the data in Excel	60
12 Cumulative Distributions in Excel (with help from StatPro)	61
13 Checking for Normality with StatPro and Histograms	63
14 Entering an Algebraic Model into EXCEL	65
15 Using Goal Seek	69
16 Constructing Regression Models in Excel and StatPro	71
16.1 Simple Regression with StatPro	71
16.2 Linear Regression in Excel (without StatPro)	72
16.3 How the Fill Handle Works to Complete a sequence of numbers	73
16.4 Multiple regression in StatPro	74
16.5 Stepwise regression in StatPro	75
16.6 Easily predicting values from a multiple linear model	76
16.7 Multiple regression with dummy variables	77
17 Creating New Variables From Old	79
17.1 Generating dummy variables in StatPro	79
17.1.1 One dummy based on a numerical variable	79
17.1.2 Several dummy variables from a categorical	80
17.2 Dummy variables with IF functions in Excel	80
17.3 Creating Interaction Variables	81
17.3.1 Creating interaction terms with StatPro	81
17.3.2 Interacting a variable with itself	82
17.3.3 Pitfalls of numerical-numerical interaction variables	82
17.4 Transforming Data	83
17.4.1 Using StatPro's Data Transformation Utility	83
17.4.2 What transformations to use	83
17.4.3 Transformations without StatPro	84
18 Using a VLOOKUP table	85
19 Computing Values of Exponentials and Logarithms	87
20 Setting up functions in Excel for shifting and Scaling	89
21 Calculating R^2 for your fit of the data	91
22 Using Goal Seek to find the best values for the shifts and scales	93

23 Computing S_e and R^2 for nonlinear models	95
24 Creating a column-oriented, one-variable data table	101
25 Using mixed cell references to compute a table of function values	103
26 Adjusting security settings for macros in Excel	105
27 Typesetting and formatting equations in MS Word	109
28 The PMT Function	111
29 Using SOLVER	113
29.1 Introduction to using SOLVER to minimize and maximize a function.	113
29.2 Setting up constraints in Excel	113
29.3 Installing the SOLVER add-in	115
29.4 Adding constraints in Solver	116
29.5 Changing constraints in solver	117
29.6 Deleting constraints	117
29.7 Options in solver	117
29.8 Errors in Solver	118
29.9 More tricks solver will play on you	119
30 A Basic Integration Tool in Excel	121
30.1 Cell References and Formulas in the Basic Tool	122
30.2 How to Modify the Basic Integration Tool for Other Functions	122
A Excel Cursor Shapes	123
B Excel Errors	125
C Other Features of Excel	127

Chapter 1

Format of computer information in this text

This textbook contains three kinds of computer information to help you. Each will be formatted a little differently, so we have included a brief overview of each to help you.

Chapter 2

Basic Excel Information

Excel information will be presented with a brief set up of how the spreadsheet should be organized and what the new tool will do. Step-by-step instructions will follow. Any formulas in the format below show the syntax of formulas to be typed into Excel.

`=FormulaName(argument 1, argument 2, etc.)`

Any formulas shown in this format, with specific cells or cell ranges in the formula, should be typed exactly as shown, assuming that your spreadsheet is set up as described in the information or as shown in the accompanying screenshots and images.

`=FormulaName(A1:A10, 3)`

2.1 Advice on computers and doing work electronically

There is nothing so tragic as bad things happening to good students.

Unknown Instructor

If you want to avoid being one of those good students to whom bad things happen, take heed of the following advice. It should become a mantra, repeated to yourself over and over until it is a part of your psyche:

SAVE EARLY, SAVE OFTEN.

Anytime you make a substantial change to your work like pasting a graphic in, typing a whole sentence or paragraph, adding a table, or reformatting, you should save your file. Save as soon as possible after starting a file. There is also a keyboard shortcut for saving files: CTRL+S. Use this frequently to avoid losing a substantial part of your work.

2.2 Using the help system in Microsoft Office

The help system for Microsoft Office is fairly extensive. In Word or Excel (or Power Point), there is a button in the upper right corner of the screen that is labeled with a question mark. Clicking this button will activate the help system, shown in figure 2.1

The basic help dialog box has a search feature and some organized links about Excel. For starters, try searching for help on "help". Simply type the word "help" into the search box. If you are offline (not connected to the internet) you will have access to the basic help features; if you are online the help center can search a huge database of information to help you with your questions.

You should get used to using the help features. Very often, your first try will not turn up anything, but always check the "see also" line that appears with most help information. This will link you to other information that is related to the topic you originally searched for.

Within the information portion of the help window, most of the phrases and sentences are hyperlinked to allow you to navigate through the information to locate what you need.



Figure 2.1: Microsoft Excel 2007 help interface

2.3 Excel desktop

This lesson will introduce you to the Excel desktop. To begin this lesson, start Microsoft Excel. The Microsoft Excel window appears and your screen will look similar to the one shown here.

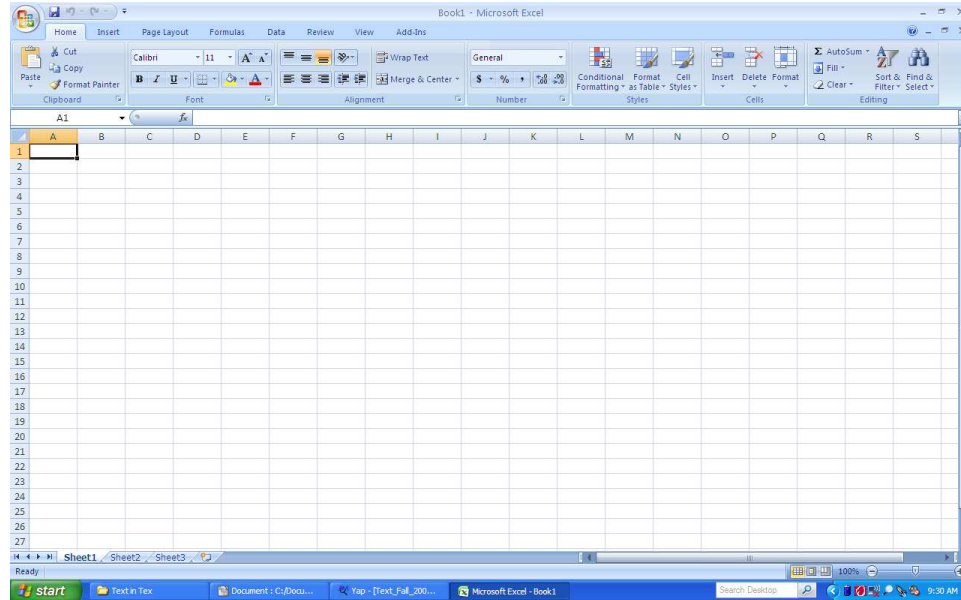


Figure 2.2: Microsoft Excel 2007 Desktop

Starting at the top of the Excel desktop, you will see the following items:

- The program control bar: This displays the name of the current workbook and provides buttons for minimizing, maximizing, and closing the program on the right-hand side. The left-hand side has a quick button for saving the workbook (the disk icon), undoing and redo-ing actions (the looping arrows), and a pull-down menu to customize the interface. The extreme left of the program control bar has a Microsoft Office icon; clicking this accesses features that let you save, open, or print workbooks.
- Menu tabs: These tabs, labeled "Home", "Insert", "Page Layout", "Formulas", "Data", "Review", "View", and "Add-ins", control the ribbon below. Unless you have installed an Add-in (like StatPro) you may not see the last of these options.
- Ribbon: The ribbon (formerly the toolbar) has icons for most of the actions you might need to perform in Excel. Selecting different menu tabs changes the icons on the ribbon to the ones associated with that tab. Usually, these are grouped together. Hovering the cursor over any of the icons provides additional information about the tool.
- Formula Bar and Name Box: The formula bar has two regions. The left area (the Name Box) allows you to select, view, or name cell ranges in the current worksheet. The right area (the Formula Bar) displays the formula entered into the current active cell, or allows you to type in a formula.

- **Workspace:** The main area of the screen is a grid of cells into which you enter information, data, and formulas. Each of these cells has a name, identified by first the column (A, B, C, etc.) and then the row (1, 2, 3, etc.) So cell D6 is in the fourth column (labeled D) and the sixth row.
- **Worksheet Control:** This area, just below the workspace, has tabs to select different worksheets in the workbook.
- **Status Bar:** The status bar provides quick statistics for the region of data that is currently selected in the worksheet along the right side. Along the left side is where you will see error messages and notifications.

2.4 Saving and Opening Files

Remember, you should save your work early and often to avoid loss of data. Office 2007 saves files in a new format compared to old versions of Office. When saving, make sure you know where you are saving the file and what format you are using. For example, using the "Save As..." feature allows you to change the format of the file. To do this, use the pull-down menu below the file name to select a different file type.

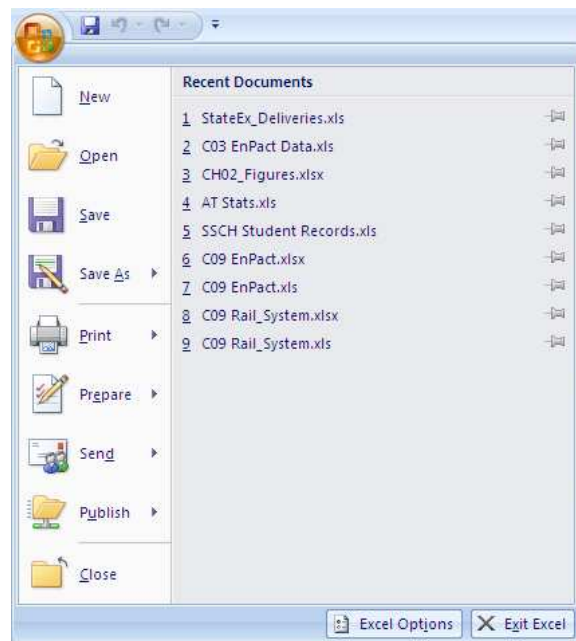


Figure 2.3: Microsoft Excel 2007 file menu

To open a file, you can either double-click the file icon in the browser or the Windows explorer, or you can open the file from within Excel. Simply open the file menu and choose "Open"; then browse through the folders on the computer to locate the file you want.

2.5 A Note About Naming Files

When you save your files ("Early and often", remember) be sure to save them with a meaningful name. If the file includes your solution to homework 2, then include "homework 2" in the title. You may also want to save all the files for each course you are taking into a separate folder, named for the course. Finally, if the file is going to be sent electronically to your instructor (through email or some course management system) it's a good idea to make sure that your name appears on the file in some way. After all, unless you are the only student in the class, the file name "homework 2" could belong to anyone. Your instructor may even establish guidelines for naming files in order to make file management for the entire course easier on him/herself and the teaching assistants (if any). Be sure to check whether your instructor has a preferred file-naming system. It is also helpful when saving files to name them meaningfully. If you name the first Excel workbook for every course you take "File1" you will have a lot of files with the same name. It is helpful to come up with a naming convention that clearly helps you locate the files you want. For example, you could name all your files with the following convention: Class+First Initial+Last Name+Assignment. Then, a file named

MSTI130 KGreen Memo01.xlsx

clearly is the first memo assignment written by K. Green in the MSTI 130 class. If your instructor collects work electronically, she may have a particular file-naming convention that she expects you to use in order to make tracking everyone's work easier. Be sure to check your syllabus or ask your instructor.

2.6 Folders and Organization

When saving your files, it also helps to have some sort of plan for organizing the files. In Windows, the way to do this is to use **folders**. These can be named anything you want, and you can have as many folders inside a folder as you want. You can also put folders inside other folders. Just be careful: it's easy to create such a complex nest of storage folders that you cannot remember where your files are.

2.7 Variable names in Excel and StatPro

In order to use the features of StatPro, you will need to make sure that whenever you enter data into a spreadsheet, the top cell of each column of data contains a variable name. Variable names should be typed as simple text, without an "=" sign in front of the name. (All Excel formulas must start with either an equal sign or a plus sign; text is not a formula, so do not start text with an equal sign or a plus sign.)

When naming the variables, do not use spaces. If your variable name has two or more words in it (like "Annual Salary"), you can either

- Enter the name with an underscore character to separate the words: `Annual_Salary`,
or

- Enter the name as a single word, with the first letters of each sub-word capitalized: `AnnualSalary`.

It is also very important that no special characters appear in variable names. In other words, your variable names should be made up of only letters (upper and lower case), numbers, and the underscore character ("_"). Other software may have other requirements; be sure you check the help guide or user manual.

2.8 Good Data Entry Practice

Organize your spreadsheets so that the data is stored with the variables in columns and observations are stored in rows. Make sure that each variable has a heading at the top of the column of data to identify it. It's a good idea to add comments to each variable name in order to explain the coding and the units of the data. Make sure each observation has a unique identifier.

It is also very important that each cell in the data contain information from only one variable. For example, if you are coding information about homes and you want to record data on the garage, you have two things to deal with: whether the garage is attached to the house or not, and the number of cars that the garage can hold. You would not want to have the cells coded as "Detached 2" and "Attached 1" and so forth. That is mixing two variables, type of garage and size of garage, into a single variable. It would be better to either

- create two variables, one for "Type," coded as "attached" or "detached" and a separate variable for number of cars, or
- code a single variable (nominal categorical) to include the information, perhaps using the codes below
 - 1 = attached, 1 car garage
 - 2 = attached, 2 car garage
 - 3 = detached, 1 car garage
 - 4 = detached, 2 car garage
 - 5 = other type of garage
 - 6 = no garage

2.9 Comments in Excel

Excel allows you to add notes, called "comments" to any cell. These comments are not part of the data or formulas in the cell, and they do not normally appear in the worksheet. Instead, any cell with an attached comment will have a small red triangle in the upper right corner. If you place the mouse pointer over a commented cell, the comment will appear. Comments are used to include such information as the way in which a variable is coded, the units of numerical data, and references to the source of the data.

To add a comment to a cell, right click on the cell. In about the middle of the context menu, the option "Add comment..." should appear. Select this option, and an editable comment box will appear. Type your comment in the box. When you are done, select another cell with the mouse. Your comment will be entered into the spreadsheet. To make changes to an existing comment, right click on the commented cell and select "Edit comment..." To delete a comment from a cell, right click on the cell and select "Delete comment..."

2.10 Copying and pasting between programs

Microsoft Office is designed so that you can select information in one program, copy it (using either the keyboard shortcut CTRL + C or the menu command "Edit/ Copy") and then paste it into another program. When you copy selections, they are placed in an area called the "clip board". To take these selections from the clipboard and place them into the document (either another location in the same document, or in another document altogether) simply place the cursor where you want the information to go and either use the keyboard shortcut "CTRL + V" or the menu "Edit/ Paste" to paste the object in the location you have selected.

2.11 Renaming a worksheet in Excel

An Excel workbook can contain many different worksheets. By default, these are named "Sheet1", "Sheet2", etc. Double-click on the name of the sheet and it the sheet name will be highlighted. Type in the new name for the sheet and hit ENTER. (You can also right click on the sheet name and select "Rename" from the menu.)

2.12 Sizing columns to make data fit

You may also run into the problem that information you enter into cells in a spreadsheet might not fit. You have two options to get information to fit: You can either resize the columns or you can enter the text on multiple lines. To resize the columns, you can go to the column header and either clip-and-drag the width of the column to the desired size or you can double-click on it so that it automatically resizes to be wide enough for the widest entry in that column.

To enter information in multiple lines within a single cell in Excel, first type the information on the first line and then hit ALT+ENTER to move to the second line. You can use as many lines as you want.

2.13 Pasting graphs and charts from Excel into a Word document to make a report

Microsoft Windows-based computers have the powerful ability to share information between programs. This is especially useful when creating technical reports that involve data analysis.

You can use the powerful data analysis routines of Excel and StatPro, and then share that information with Microsoft Word by copying and pasting the charts and computations. There are several ways to do this, however, and each is useful for certain types of information. When copying a chart from Excel into Word, the important thing to remember is that the chart is full-screen/full-page in Excel, but will only be about one-third as big in your report in Word. This means that all the text on your chart - the labels for the axes and the title of the chart, for example - will shrink to almost unreadable size when copying from Excel to Word. This can be avoided by simply preparing the chart before you copy it.

1. Prepare the chart for copying.

The first thing to do is to make the font size of the chart larger. To do this easily, right-click on the chart area and select "Font" from the context menu that appears (see figure 2.4.) Change the "size" setting to something larger. We recommend either 16- or 18-point font. That seems to carry over into Word very nicely. When you hit "OK" all the text on the chart should change to the size you selected.

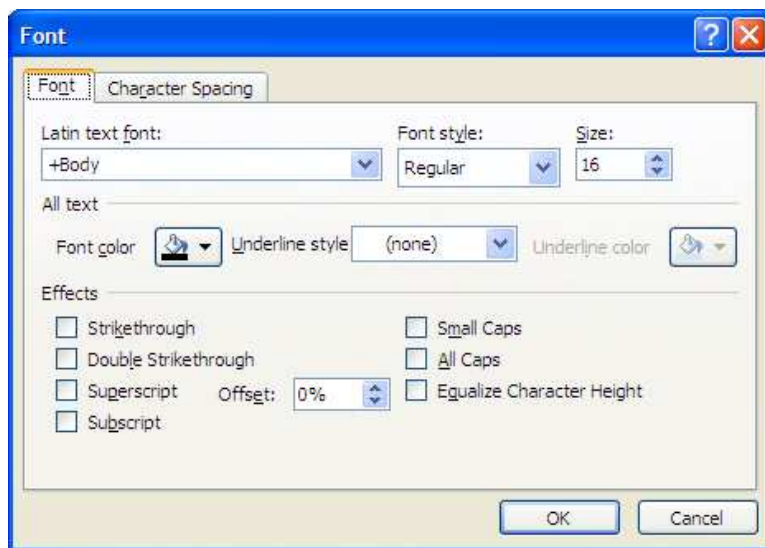


Figure 2.4: Modifying the chart for pasting into Word.

The other change you may want to make is to be sure the background color is turned off. Sometimes this doesn't come through well in Word, especially if you plan to print the report. To change the background, simply right-click on the chart and select "Format plot area..." from the context menu. Then, select "no fill" for the Fill setting.

2. Copy the chart.

As you probably know, there are many ways to copy and paste in Windows-based programs. We'll show you all of them here. Each begins with the same step: select the object to be copied. For a chart, you should click in the white region so that the entire chart is selected. This region is the same region where you double-clicked to change the font and is described under "step 1" above.

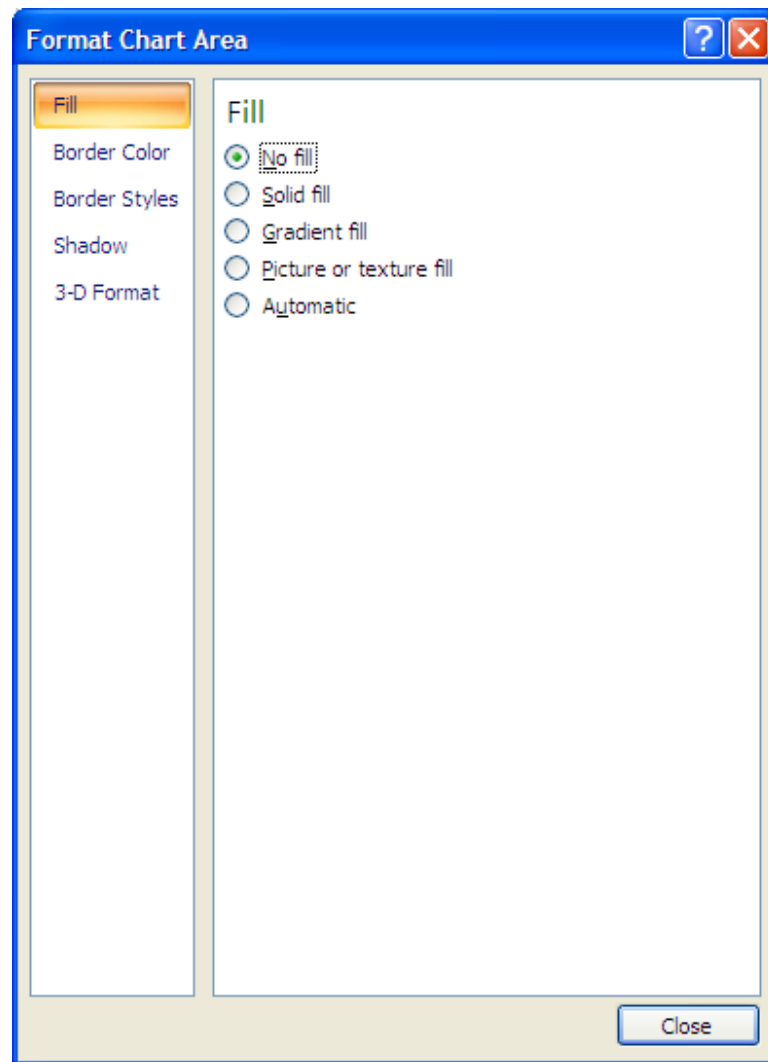


Figure 2.5: Changing the background fill of a chart before pasting into Word.

- (a) Using the ribbon. Now that the chart is selected, go to the Home menu and select "Copy".
- (b) Using the keyboard. After selecting the chart, hit CTRL + C. This is the keyboard shortcut for the "Edit/ Copy" menu option above. You'll also notice that this keyboard shortcut appears on the edit menu next to the copy command.
- (c) Using the mouse. Right-click on the chart and select "Copy".

Before we complete the process, it's important to understand a little about how Windows works with copying and pasting. Anytime you copy an object (a word, a chart, a paragraph, a file) it is placed in a temporary storage space called the Clipboard. Starting with Microsoft Office 2000, this clipboard can hold several objects at once. In Microsoft Office 2007, you can access the entire clipboard by clicking on the small

arrow next to the word "clipboard" along the left-hand side of the Home ribbon. Once you have the entire clipboard displayed, you can select any item to paste.

3. Paste the chart into your report.

Now here is where it gets a little confusing. The easiest way to paste a chart is to place the cursor where you want the chart and either use the "Edit/ Paste" menu command, the keyboard shortcut CTRL + V, or right-click with the mouse and select "paste" from the context menu that appears. If you are pasting a chart from Excel, this will automatically select the format in which the object is pasted to be a Microsoft graphic object.

You can, however, choose other formats for pasting objects. By clicking on the small down arrow below "Paste" on the Home ribbon, you can use the "Paste Special.." option, which allows you to select many different formats. Usually, this is not necessary. In older versions of Word, the paste special feature could help by selecting a format that used less memory, making it easier to transfer and store files. The newer version is smarter about its default selection process, and modern memory storage and data transfer rates make saving a few kilobytes of space unnecessary.

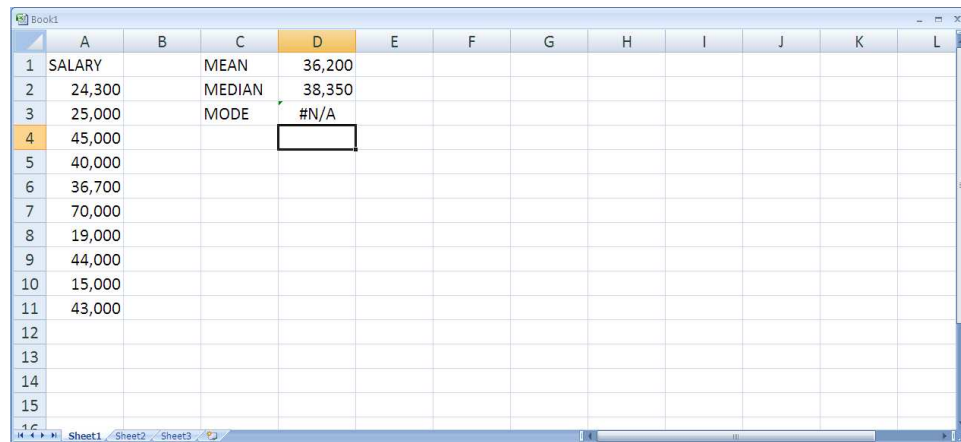
The following comments are based on the data file "C05 Homes.xls" which has data in cells A3:M278. The data records information about homes that sold in a three-month period in Rochester, NY. The variables are address (identifier), location (categorical), taxes (numerical), style (categorical), bath (numerical), bed (numerical), rooms (numerical), garage (numerical), year (numerical), acres (numerical), size (numerical), value (numerical), price (numerical).

Chapter 3

Cell References in Excel

3.1 Cell References in Excel

Excel organizes information into sheets. Each worksheet is then organized by columns (labeled by letters) and rows (labeled by numbers). Thus, every cell (rectangle on the worksheet that contains information) has a name, called a cell reference. This cell reference is usually given the way you called out locations on the game Battleship: as a column and a row. For example, in the worksheet shown at the right, the word "SALARY" is in cell A1. The mean of the salary data is in cell D1. Such a reference is called a relative cell reference.



The screenshot shows an Excel worksheet with columns A through L and rows 1 through 15. Column A contains the word "SALARY" in row 1 and a list of salary values from row 2 to row 11. Column C contains statistical labels: "MEAN" in row 1, "MEDIAN" in row 2, and "MODE" in row 3. Column D contains the corresponding statistical results: "36,200" in row 1, "38,350" in row 2, and "#N/A" in row 3. The cell D4 is currently selected and is empty.

	A	B	C	D	E	F	G	H	I	J	K	L
1	SALARY		MEAN	36,200								
2	24,300		MEDIAN	38,350								
3	25,000		MODE	#N/A								
4	45,000											
5	40,000											
6	36,700											
7	70,000											
8	19,000											
9	44,000											
10	15,000											
11	43,000											
12												
13												
14												
15												

Figure 3.1: Excel worksheet

There are other types of cell references (called absolute cell references) that you will learn about below. The other important thing to know about cell references is that you can easily refer to a block of cells, as long as the block of cells forms some sort of rectangle. For example, to refer to the salaries themselves on the worksheet above, you would refer to all the cells in the rectangle starting in cell A2 and continuing through cell A11. In Excel notation, this entire range of cells is identified by A2:A11.

3.2 Absolute Cell References in Excel

Above, you learned how to refer to any cell or range of cells using the grid system in Excel. If there is data in the cell in column D in row 2, this cell is referred to as D2. However, this type of cell reference (a relative reference) will change if the formula is copied to another cell. Many times (as in the example below of computing deviations) a particular cell reference will need to be absolute. This means that it will not change if the formula is copied. To make a cell reference absolute, place a dollar sign (\$) in front of both the column and row. Thus, an absolute reference to cell D2 would look like \$D\$2.

As you may have guessed, you can have mixed references also, where either the column or row is absolute. In general, if you don't want part of the reference (either the row or column) to change as you copy the formula, be sure to place a dollar sign in front of it.

When you are typing a cell reference into a formula, you do not have to type the dollar signs to convert them to absolute references. After you type a cell reference in a formula (say you type A2), hit the F4 button along the top row of the keyboard. This converts the current cell reference into an absolute reference (so now you would have \$A\$2). If you hit the F4 button again, it is converted to a mixed reference with the row fixed (A\$2), hitting it again will convert it to a mixed reference with the column fixed (\$A2). Finally, hitting F4 a fourth time will cycle back to a relative reference (A2).

3.3 Three dimensional cell references in Excel

In addition to referring to cells by the column and row, Excel allows you to build formulas that include references to cells on other worksheets in the current workbook. Suppose you are entering a formula in 'Sheet 1' of a workbook and there is a number in cell D4 of 'Sheet 2' that you want the formula to look up. Simply typing D4 in the current formula will not work; Excel will simply look up the value in cell D4 of the workbook containing the formula. To get around this, you must use a 3D cell reference. All this involves is including the name of the worksheet in single quotes, followed by the "bang" or exclamation mark symbol (!) and then the normal cell reference. So, in our example, to get a formula in 'Sheet 1' to use the value in cell D4 from 'Sheet 2', you need to type the cell reference exactly in the form

`'Sheet 2'!D4`

3.4 Naming Cell Ranges

There is another way to refer to cell ranges (or individual cells) besides a cell reference. You can give the cells or cell ranges their own names and then use these names in formulas for computations. Figure 3.2 shows the C05 Homes.xls data with the data in the Price variable (column M) selected. To give this range of cells a name, we simply click on the "Name Box" to the left of the formula bar and type a name; in this case, we'll call the range of cells "Price". Note: there are no spaces allowed in the name box.

Now, in any formula in the worksheet we can use "Price" instead of the range M4:M278. Instead of typing the formula =AVERAGE(M4:M278), you could just type the formula

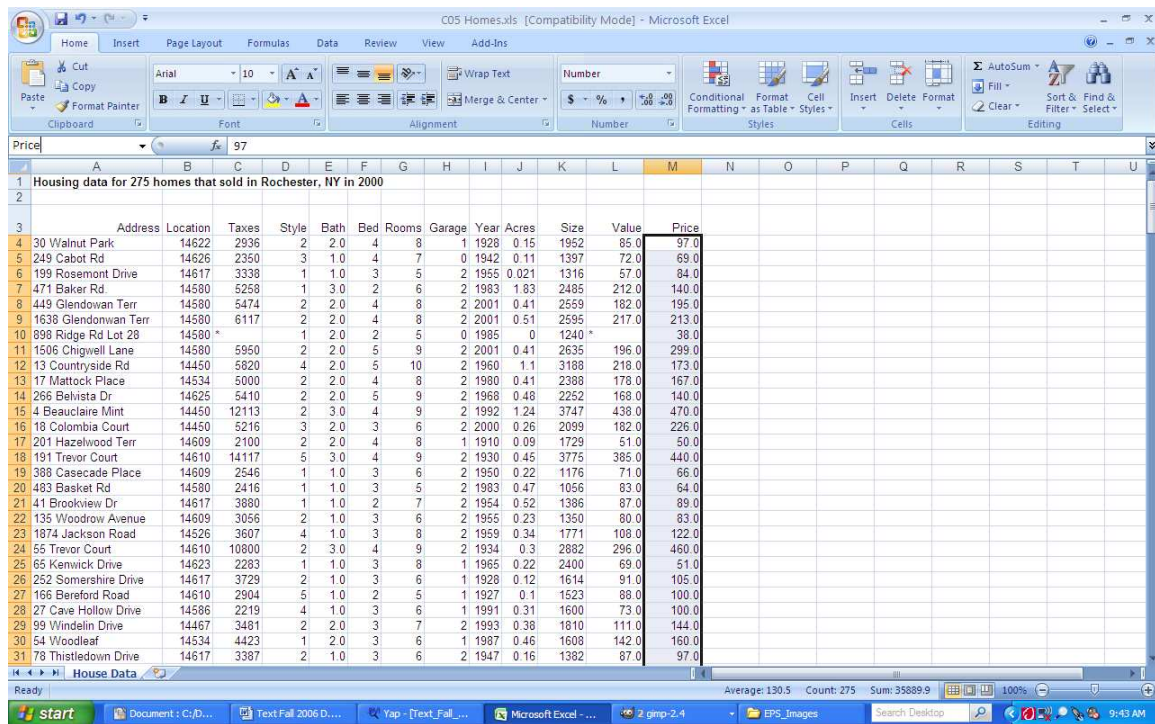


Figure 3.2: Homes data showing the price variable selected and being assigned the name "Price" using the name box to the left of the formula bar, below the ribbon.

=AVERAGE(Price). Notice that such references are always absolute. This has the benefit of making all the formulas more readable. You can see a list of all the named ranges in the workbook by clicking on the downward pointing triangle next to the name box. Notice that if you use any of StatPro's commands to manipulate your work, there will be a lot of ranges named for you, because that's what StatPro does first in order to simplify its formulas. If you want to get rid of these, StatPro has a command for this: go to "StatPro/Clean Up..." and select "Delete range names". StatPro will then wipe out all of these range names. This may free up some space and make the Excel file smaller.

Chapter 4

Copying Formulas in Excel

There are three different ways to copy formulas in Excel from one cell to another cell or to a group of cells (like a whole column): standard copy and paste commands, dragging the fill handle, or double-clicking the fill handle.

4.1 Using Copy and Paste Commands

This method is the most obvious. First select the cell with the formula you want to copy. Copy this using either CTRL+C, the copy button on the toolbar, or the "Edit/Copy" menu command. Now highlight the cell or cells where you want the formula to be placed and paste it in using either CTRL+V, the paste button on the toolbar, or the "Edit/Paste" menu command.

4.2 Dragging the Fill Handle

If you want to copy the formula to the column of cells beneath it, or to the row of cells beside it, you can use the fill handle. The fill handle is a tiny black square that appears in the lower right corner of a cell you have selected. If you click on this fill handle and drag down (or right), then, when you release the mouse button, the formula from the first cell (or group of cells!) is copied to all the cells in the area you highlighted by dragging. Be sure that you are clicking on the fill handle, though. You'll know for certain that you are on the fill handle because the cursor will change from a fat plus sign to a skinny plus sign.

4.3 Double-clicking the fill handle

In certain circumstances, you can double-click the fill handle and Excel will automatically copy and paste the formula all the way down the column until it reaches the end of the column to the left of the one in which you are pasting the formula. So, if you have a row of salaries in cells A2:A20, and you enter a formula to compute a raise in cell B2, you can copy this formula into cells B2:B20 by double-clicking the fill handle.

N.B. If the column to the left is empty or has a break in the data (an empty cell), this trick will not fill the column out all the way.

Chapter 5

Computing Summary Statistics in Excel

5.1 Computing standard deviation in Excel with the built-in formula

There are two different standard deviations in Excel, depending on whether the data is from a sample or a population. To compute the standard deviation of a sample (this is the most commonly used version), use the formula

`=stdev(range of cells)`

For the standard deviation of a population, use the formula

`=stdevp(range of cells)`

5.2 Computing Means in Excel

Excel uses the function AVERAGE for the mean. To compute the mean of the data in cells A2:A11, we enter the formula

`= AVERAGE(A2:A11)`

into any cell on the spreadsheet. If you later move or copy the cell containing this formula, the cell references will be changed since we used relative cell references. This means that the formula will probably not point to the right cells anymore. Also remember that if you change any of the data in cells A2:A11 the mean will be re-calculated instantly. If, however, you add data outside this range, you will need to change the formula.

For the purposes of the remainder of this guide, assume that there is data in cells A2:A20 with a variable name in cell A1.

5.3 Adding up a list of values

If you have a list of values, you can quickly add them together using the SUM command in Excel. For example, if your values to be added are in cells A2:A26, entering the command

$$=SUM(A2:A26)$$

into cell B2 (or any other cell) will add the values together.

5.4 Computing deviations in Excel

In order to compute the deviations in Excel, we first need the mean of all the data. Let's calculate this with Excel by typing the formula below into cell F1.

$$=average(A2:A20)$$

Now, we will create a new column for the deviations. In cell B1, type "Deviation" so that the column has a label. Now, in cell B2, we want to enter a formula to compute difference between the first data point (in cell A2) and the average (an absolute reference to cell F1). Thus, we enter the formula

$$=A2 - \$F\$1$$

Now we simply copy this formula (see below) down to the other cells in column B.

5.5 Computing standard deviation in Excel with a user-created formula

One could also create the formula for standard deviation by hand, simply by constructing each piece of the formula separately and then combining them together. In Excel, the square root of a number is computed with the sqrt(number) function. For the standard deviation, we want the square root of the sum of the squared deviations divided by the number of observations minus one. First, we create a column of deviations (see above). Next, create a column of deviations squared (If the deviations are in column B, starting at row 2, then the deviation squared would be calculated with the formula $=B2^2$ which can then be copied to the rest of the column to compute the other deviations squared. Now, we should have the deviations in cells B2:B20, and the deviations squared in cells C2:C20.) Finally, we construct the entire formula:

$$=sqrt(sum(C2:C20)/(count(C2:C20)-1))$$

As an alternate version, we could have Excel square the deviations and add them together using the "sumproduct" function (mathematically, it is equivalent to a vector dot-product). This formula works directly with the deviations in column B with no need to construct a column for the deviations squared.

$$=sqrt(sumproduct(B2:B20,B2:B20)/(count(B2:B20)-1))$$

5.6 Computing Medians in Excel (Method #1)

To compute the median of the data in cells A2:A11, we enter the formula

= MEDIAN(A2:A11)

into any cell on the spreadsheet. Remember, though, that if you later move or copy the cell, the cell references will be changed since we used relative cell references. Also remember that if you change any of the data in cells A2:A11 the median will be re-calculated instantly. If, however, you add data outside this range, you will need to change the formula.

5.7 Computing Medians in Excel (Method #2)

Since the median is also the second quartile (see chapter ?? (page ??)), we could use the formula below to compute the median.

= QUARTILE(A2:A11,2)

Notice that this function uses two inputs, a range of data (A2:A11) and a number indicating which quartile is desired, 1=first quartile, 3=third quartile.

5.8 Computing the Mode in Excel

The mode is computed with the formula

=MODE(A2:A11)

You may get the result #N/A if there is no mode. If there is more than one mode, Excel just guesses and gives one of them. The fact that there may be more than one mode, or no mode at all, is why this statistic is rarely used except for categorical data.

For all of the information below, assume that the spreadsheet shown above in the "How To Guide" for chapter ?? (page ??) is being used. It contains data on sample sales figures from Toys-For-Tots (from Exploration ?? (page ??)) in cells A2:B20. For a picture of this spreadsheet, see figure ?? (page ??).

5.9 Computing z-scores

To compute z-scores for the variable Price (cells M3:M278), we first need to compute the mean and standard deviation. In cell P1 enter =AVERAGE(M3:M278) to compute the average and in cell P2 enter =STDEV(M3:M278) to get the standard deviation. Now, in column N, enter "Z Score" in N3 and enter the formula below into N4

=(M4 - \$P\$1)/\$P\$2

All that is left is to copy the formula to the rest of column N (N5:N278).

5.10 Percentiles in StatPro

Using StatPro, it is easy to calculate the 1st, 5th, 10th, 90th, 95th and 99th percentiles. Start by generating the "One variable summary statistics" (see chapter 3). At the bottom of the dialog box with the statistic options, select the percentiles that you want calculated.

5.11 Percentiles in Excel

To calculate percentiles in Excel, use the formula

`=PERCENTILE(array of cells, percentile)`

Note that percentile should be entered as a decimal number. Thus, for the 80% percentile, you should enter 0.80. For the 35th percentile, enter 0.35.

5.12 Correlation Tables using StatPro

To create a table of correlations, follow the usual steps in activating a StatPro routine. In step 2, select "StatPro/Summary Stats/Correlations and Covariances". In step 5, you will see a screen like the in figure 5.1. Usually, you can just click "OK" at this dialog box to move on, because none of these settings will need to be changed. We suggest placing the resulting calculations on a new worksheet and naming the worksheet something like "Correlations."

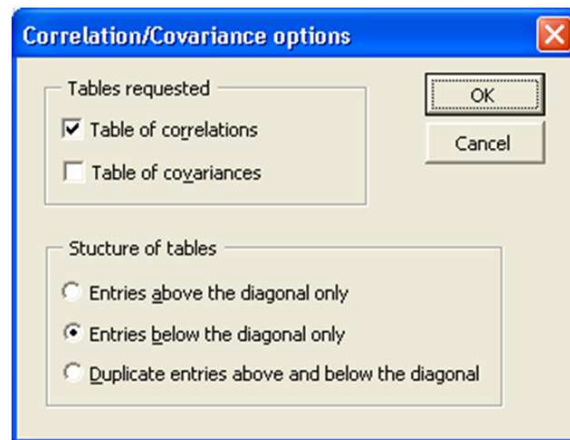


Figure 5.1: Creating a correlation matrix in StatPro.

Covariances are related to correlations, but are much harder to interpret since they have units and may range in size from negative infinity to positive infinity. The second set of options, "Structure of tables" doesn't really matter; it simply allows you to calculate the correlations above the diagonal, below the diagonal, or both. But remember, the values are the same above and below the diagonal.

Chapter 6

Summary statistics in StatPro

Statistical add-ins like StatPro often have a more convenient way to compute summary statistics than to enter formulas for the mean, median, and so forth, separately. Often, they include a routine that will compute all of the possible statistics, or a selection of them, at one time. Each routine in StatPro is structured the same way. The six steps below will take you through any of StatPro's useful routines. The only differences among the routines occur in step 5.

1. Select the region of the worksheet that contains the data.

When using any StatPro routine, you must first select the data to which the routine will be applied. To do this, simply click on any cell inside the region of the worksheet that contains the data. For example, if the data is in the region A1:B20 (including the variables names), clicking on any cell or collection of cells in that region will work. Be careful, though! StatPro assumes that the region is rectangular, and it will include any rows or columns that have any data that touches the region you are interested in. This is why it is always a good idea to make sure there is at least one blank row and one blank column between the data region and any other information on the worksheet.

2. Select the StatPro routine to apply to the data.

In this case, we want to use the statistics routines. Select the Add-Ins ribbon. Click on the "StatPro" menu, then select "Summary Statistics..." and choose "One variable summary statistics..."

3. Verify that the data region is correct.

Check to be sure the highlighted chapter of the worksheet includes your data (and only your data), then click "OK". StatPro will then create a list of variables that are included in your data so that each variable can be referred to by name rather than by row and column references. If there is an error in the data region that has been selected by StatPro you can either hit "Cancel" and start the procedure over, being certain to select the data region correctly this time at step 1, or you can type the correct region into the dialog box in the blank next to the prompt "Data range:" (see figure 6.1.)

4. Select the variables to which the routine will apply.

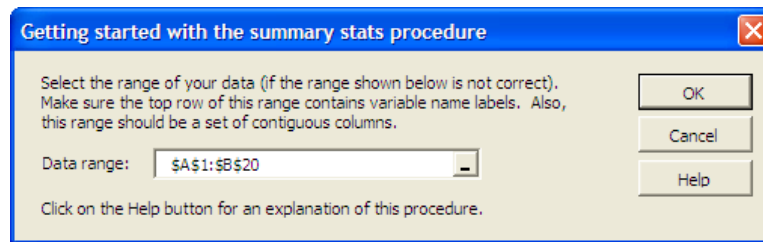


Figure 6.1: Dialog box in StatPro to verify that the correct data region is selected.

From the list that StatPro generates, choose the variable or variables you wish to get statistics for, and hit "OK" (see figure 6.2.) To select multiple variables that are listed next to each other, either click and drag with the mouse, or select the first variable in the list that you want and hold down the SHIFT key while selecting the last variable that you want. All the variables in between these two will be highlighted. If the variables that you want are not listed together, simply hold down the CONTROL (CTRL) key while selecting each variable with the mouse.

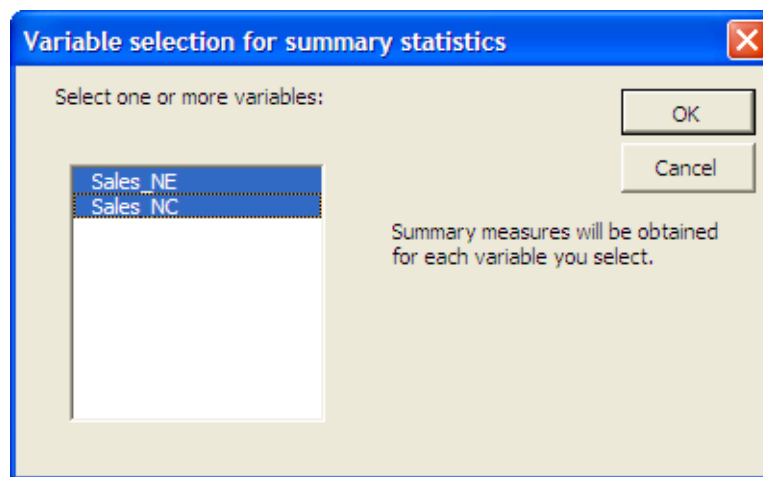


Figure 6.2: Dialog box in StatPro to select variables for the summary statistics procedure.

5. Fill in the details of the routine.

Each StatPro routine has different information at this point, depending upon what the routine is designed to do with the data. For the one variable statistics you should see a dialog box like figure 6.3. Check off the statistics that you want computed (some are already checked by default) and hit "OK".

6. Select the placement for the output of the routine.

Next, you will be asked where the statistics should be placed. You have three options, shown in figure 6.4. We recommend that you always place the computations, graphs, and statistics on separate worksheets in the workbook. Each worksheet should have

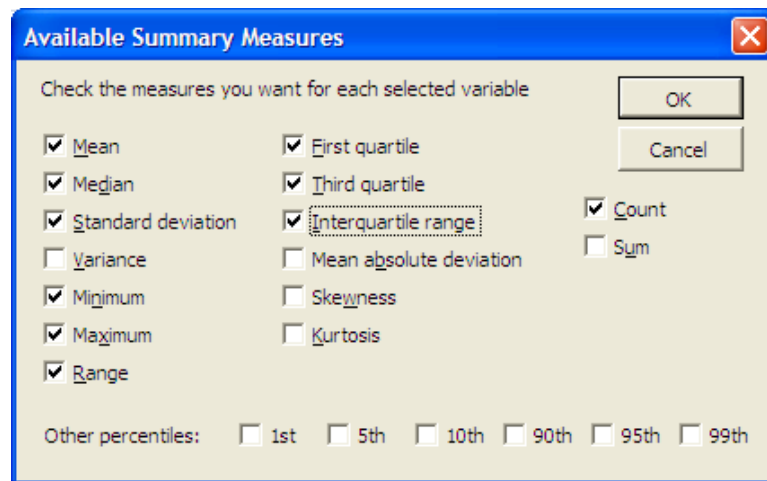


Figure 6.3: Dialog box in StatPro to select which summary statistics are computed.

a descriptive name, something like "Statistics," so that you know which worksheet contains the original data, and which contains the statistics on that data. Sometimes we will ask you to place the results of a routine on the same worksheet (to the right of the data) so that you can see everything at once. We'll let you know when we want you to do this; otherwise, put everything on separate worksheets.

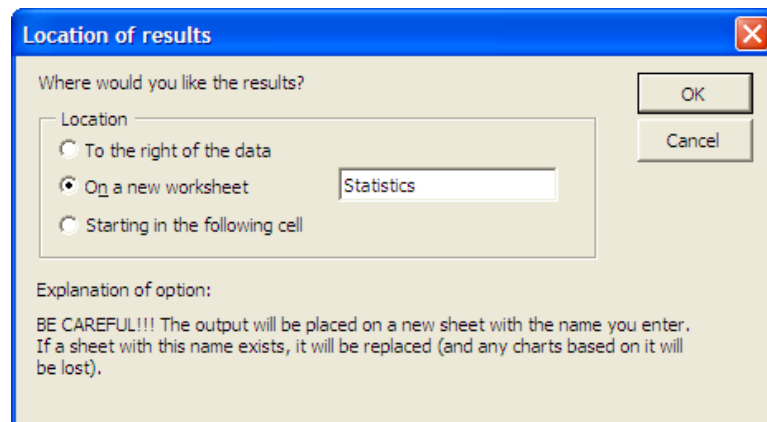


Figure 6.4: Dialog box in StatPro to select where to place the summary statistics.

6.1 Correlation with Excel

To calculate the correlation between two variables (with the same number of observations) using Excel, type

`=CORREL(X values, Y values)`

With this formula, it is critically important that you have the same number of observations of both variables, or you will get an error message.

Chapter 7

Using the Excel 2007 status bar

One nice feature of Excel is called the "status bar" (in Excel XP, this was the Autosum menu, but had far fewer features and information than Excel 2007.) If you highlight a set of numbers, any set of numbers, in a workbook, Excel automatically computes some information about those cells and displays the result in the lower right portion of the screen, just below the horizontal scroll bar. This makes it easy to see what is going on with your data without having to enter formulas. By default, the status bar shows information about the average, count, and sum of the numerical values highlighted.

This feature is good for other things as well. If you right-click on the status bar area you will get a menu of other options to have Excel "auto-display". You can select any of the following:

Count	Displays the number of cells highlighted that have data in them (blank cells are ignored)
Average	Computes the mean of the cells highlighted, ignoring blank cells
Minimum	Computes the smallest element of the highlighted cells
Maximum	Computes the largest element of the highlighted cells
Sum	Computes the sum of the highlighted cells
Numerical Count	Counts only the cells containing numerically stored data in the highlighted region

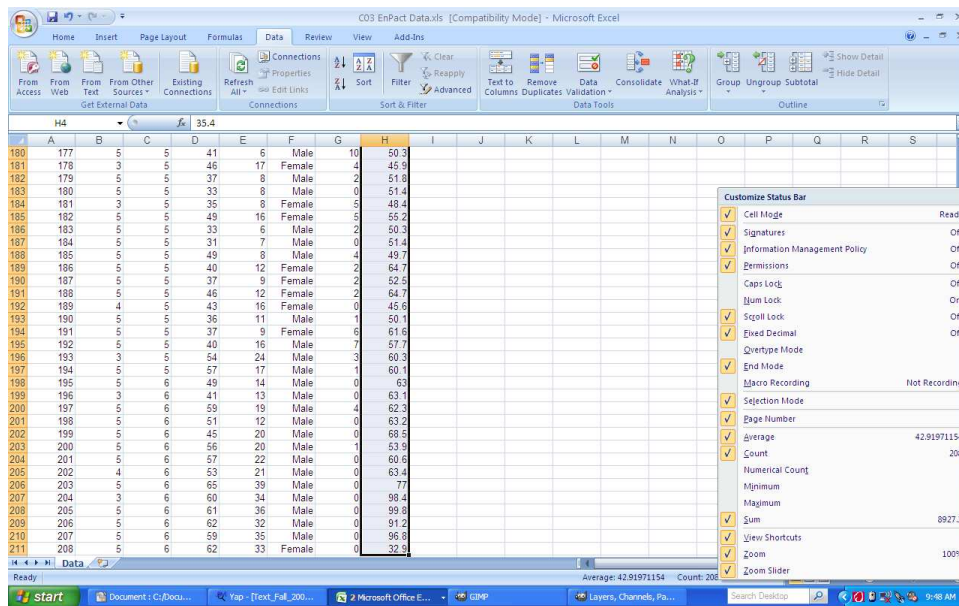


Figure 7.1: The Excel 2007 status bar, showing the options available for display by right-clicking in the status bar region (where the average, sum and count are displayed in the figure)

Chapter 8

Making and Using Pivot Tables

1. Start by selecting a cell inside the range of the data. For example, the file C03 EnPact Data.xls has data in cells A1:G251 (with row 1 containing headings or variable names), so you might select cell A1 (or any other cell in the range of data).
2. Go to the "Insert" menu and select "Pivot Table". This will bring up the Pivot Table dialog box shown below in figure 8.1 (page 30).
3. Normally, you will not need to change anything else at this stage. Verify that the data range is correct, make sure it has "New worksheet" selected for where to create the table, and click "OK" and Excel will create a worksheet like the one shown in figure 8.2 (page 31).
4. What you see on this are (a) the pivot table area, (b) the pivot table ribbon (along the top) and (c) the pivot table field list (along the right-hand side; these fields should match the variable names/headings in the spreadsheet; remember: field = variable in a database).

8.1 Advice on Making Pivot Tables

While you can place any variable in any field, and in particular you can use either row or column variables interchangeably, if your variable has many values (for example, "Years Experience" in the C03 EnPact.xls file) you are best off making it a row variable, rather than a column variable in order to make it easier to read the resulting table, and avoid having to scroll horizontally to get information.

8.2 Making a simple Pivot Table

Starting from the screen shown above, you now simply drag fields from the field list into the various regions of the pivot table. No data will be displayed until there is one field in the area marked "Drop Data Items Here". For example, to look at the average salaries of the employees, broken down by gender, complete the following.

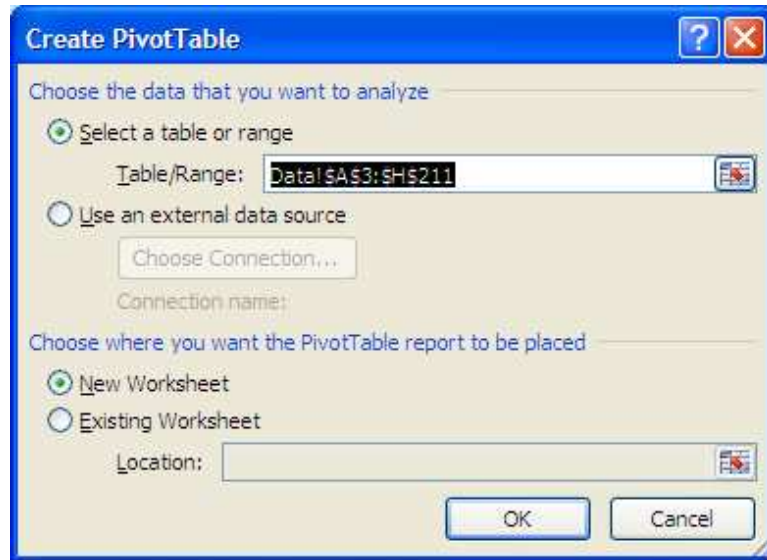


Figure 8.1: The basic Pivot Table dialog box.

1. Drag *Gender* to the area marked *Drop Row Fields Here* or drag it into the area in the lower right marked *row labels*.
2. Drag *Salary* to the area marked *Drop Data Items Here* or drag it into the area in the lower right marked *values*.
3. By default, the pivot table will either show the sum of the data variable (in this case, the total of all salaries for males and females) or the count of the data variable (the number of males and females). We would rather see the averages. To display the averages, double-click where it says *Sum of Salary*. You will see the dialog box shown at the right of figure 8.3 (page 32).
4. To summarize the data by averages, select *Average* from the list on the left. You can also format the data by clicking the *Number Format* button on the left side of the dialog box in figure 8.3 (page 32).

8.3 Advanced Pivot Table Options

For more sophisticated tables, you can display the data as percentages, etc. To access this feature, simply click the tab marked *Show values as* in the pivot table field display dialog box shown in the right half of figure 8.3 (page 32). To select a different way of presenting the data, select one of the options from the pull down menu under *Show values as*. This gives you the various options for displaying the data. The most useful items from the list are probably *Normal*, *% of row* and *% of column*.

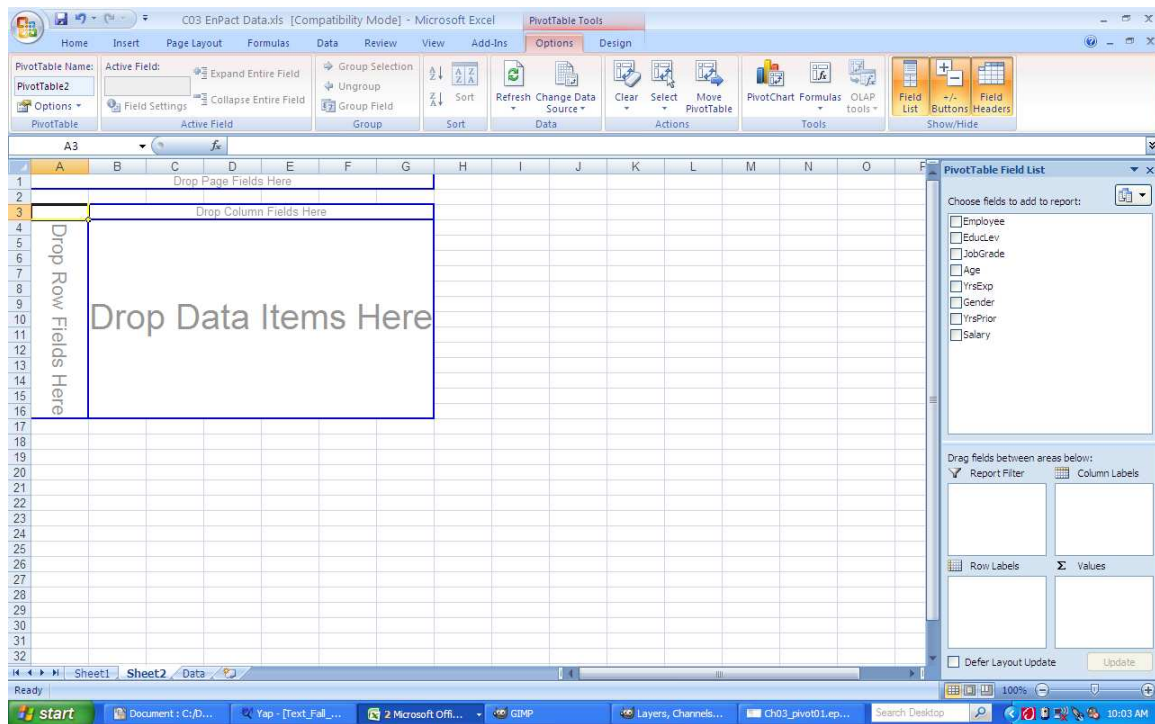


Figure 8.2: A blank pivot table.

8.4 The Pivot Table Ribbon

The pivot table ribbon in figure 8.4 (page 32) provides a lot of flexibility for working with the pivot table. Two of the most important items on the ribbon are the *Refresh* button and the *Change Data Source* button. Refresh forces Excel to re-check the original data and re-build the current pivot table. This is useful if you change or add data to the original database. This makes it easy to update information, without having to create the pivot table again. If you have more data - that is, data outside the original range of the pivot table - you can use the *Change Data Source* button to modify the data range.

8.5 Grouping items in the table

This feature allows you to take a variable that has many values (like a numerical variable) and group it together in the pivot table. For example, one could easily use "Year Experience" as a row variable in the pivot table above, but the wide variety of values makes it hard to see any details or compare results. However, if you group some years experience together (like 0-9, 10-19, etc.) you can see more interesting results as illustrated below.

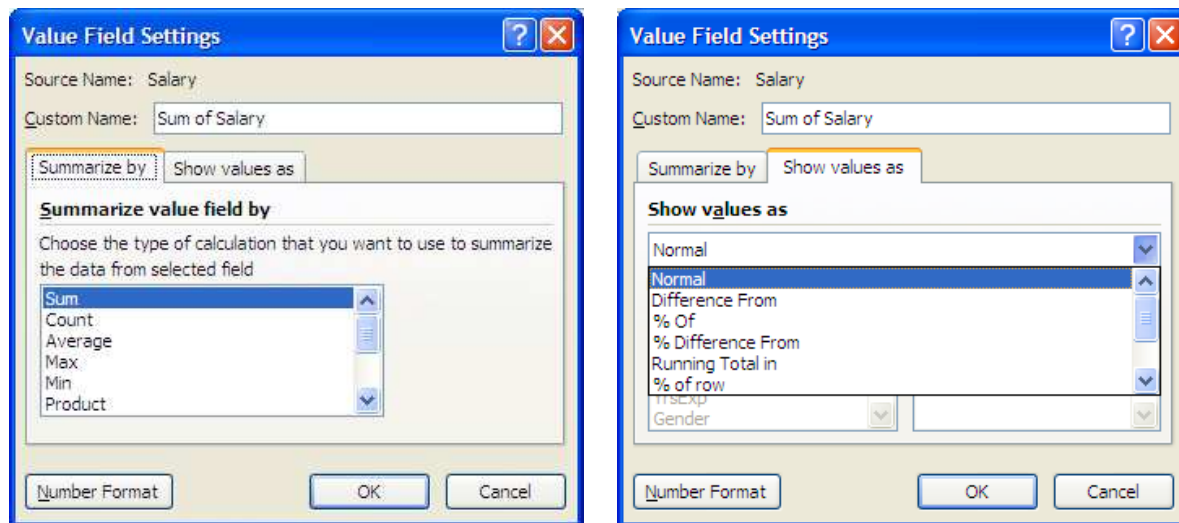


Figure 8.3: Pivot table field display options. These are accessed by double-clicking on the pivot table portion that explains what the data field is, in this case, *sum of salary*.



Figure 8.4: The pivot table ribbon.

Average of Salary	Gender		
YrsExp	Female	Male	Grand Total
0-9	38.77	41.27	39.67
10-19	42.45	55.46	44.47
20-29	44.12	59.22	52.35
30-39	32.90	92.64	82.68
Grand Total	40.21	48.51	42.92

To group the salary categories in order to see the data more easily, select a range of years experience (like all the rows with fewer than ten (10) years experience) and click the *Group Selection* button on the ribbon. Repeat this with the other ranges of experience. Then you can collapse or expand the individual groups of experience to look at the data more easily.

You can also right click on the *YrsExp* field in the pivot table. Select *Group...* from the context-sensitive menu that appears. You should see a dialog box like the one in figure 8.5 (page 33). Here you can select the starting value, ending value and space between groupings. The settings below, for example, produce the groupings shown in the table above.



Figure 8.5: The pivot table grouping tool.

8.6 Making a more complex pivot table

Pivot tables can be used to cross-chapter the data on up to three variables at once. You can select one variable as the Row Field, one as the Column Field and one as the Page Field. For example, if we take the "StateEx_Deliveries.xls" file and create a pivot table like the one shown in example ?? (page ??), we can further explore the data by adding "Shift" as the page variable. Notice that next to each variable (field) there is a pull-down menu. This pull-down menu allows you to select specific values of the variable (field) to display in the table. For example, pulling down the menu next to the page variable "Shift" in this example allows you to display data from just one of the shifts (Day, Evening, or Night) as shown in figure 8.6 (page 34). One can then explore how each of the variables relates to the data.

One can also create more complex pivot tables by placing several variables in one location. For example, by dragging first the "Truck" and then the "Shift" variable into the Column Field area, we produce a pivot table like the one shown below. This pivot table breaks down the data even further than before, showing us how the "Truck" and "Shift" variables relate to the distribution of unloading times.

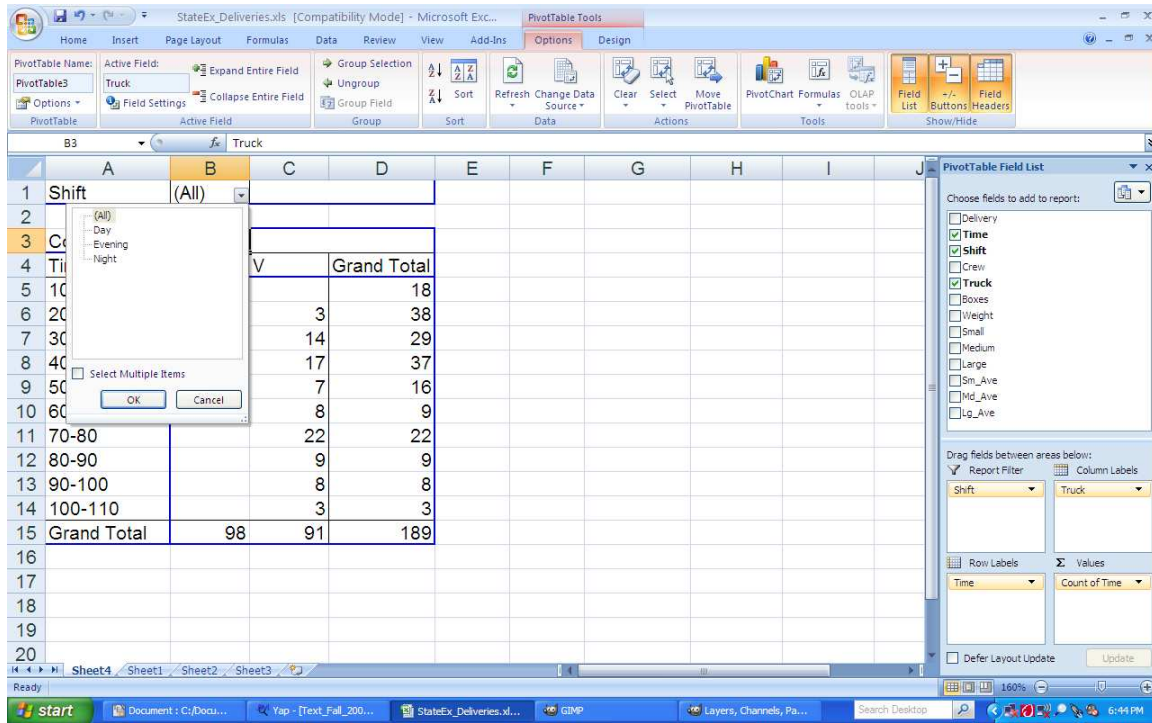


Figure 8.6: Selecting values of the "Page Field" variable for display in the pivot table.

Count of Time	Truck	Shift							Grand Total
		Semi		Semi Total		Van		Van Total	
Time	Day	Evening	Night		Day	Evening	Night		
10-20		7	11	18					18
20-30	8	10	17	35			3	3	38
30-40	5	4	6	15	6	6	2	14	29
40-50	3	6	11	20	10	6	1	17	37
50-60		5	4	9	1	2	4	7	16
60-70		1		1	4	2	2	8	9
70-80					9	7	6	22	22
80-90					4	4	1	9	9
90-100					6	2		8	8
100-110					2	1		3	3
Grand Total	16	33	49	98	42	30	19	91	189

Chapter 9

Sorting data

9.1 Using the Data Ribbon

Excel makes it relatively easy to sort your data on many variables simultaneously. In order to use this effectively, though, you need to have your data organized as we have discussed in chapter two: your variables (fields) should be the columns and the observations (records) should be the rows. It is also a lot easier if you make sure the first row of the data contains headers (variables names).

To start sorting the data, first select (click on) any cell in the data range. Then go to the data ribbon and select "Sort". This will bring up a dialog box like the one shown. The example shown (figure 9.1) uses the file "C03 EnPact Data.xls". The sort feature automatically assumes that you want to sort on the first column of the data; this is indicated by the line that says "sort by". You can change this, however, by using the pull down menus along that row of information. These let you select a different variable to sort on and different sort orders. You can sort on several variables, by adding more sort conditions using the "Add Level" button. You can delete conditions or add as many as you like. (Excel XP only allowed three sort conditions at a time.) In the upper right-hand side of the dialog box make sure the "My data has headers" is checked, so that Excel knows what the variable names for each column are.

Say you wanted to sort the data so that employees of the highest job grade are at the top of the data. Then simply select "Job Grade" and "Descending" in the top of the dialog box, then hit "OK".

If you wanted to sort by job grade and then by gender, you might select "Job Grade" in the first sort condition (and "largest to smallest" for the sort order) and then add another level to the sort and select "Gender" for that level. This will sort the list on two variables. Excel will collect all the employees with JobGrade = 6 at the top of the list, and within that group, the Female employees will be at the top of the list and the Male employees at the bottom. Sorting on three variables is similar. For example, we could take the list produced above and sort first on JobGrade, then on Gender and finally on Education.

Sorting data can be very useful for identifying outliers in the data or other anomalies. For example, if you have data on the diameter of parts being produced by one of your factory machines, and you determine that the mean size of these is 0.45" with a standard deviation

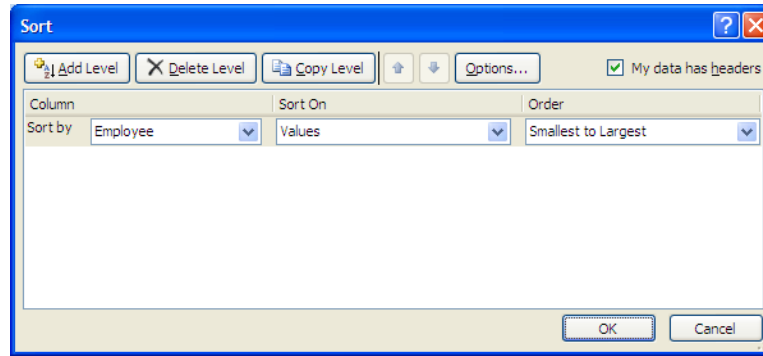


Figure 9.1: The sorting wizard in Excel 2007

of 0.03", sorting the data on the diameter variable would let you quickly find any parts produced that are too far above or below the standards of your company.

9.2 Using Autofilter to Sort and View Data

Often, our data contains so much information that it can be difficult to work with it and locate specific pieces of information. For example, in the pivot table above, we see that there is one semi load that took a little longer than the others. This occurred on the night shift. So it might be useful to arrange the data to focus on just the night shift. One tool for this is called "Autofilter". To activate the Autofilter, first click on any cell in the data, use the menu option "Data/ Filter/ Autofilter". (When you are finished, you select "Autofilter" again to turn it off; it works like a toggle switch.)

When Autofilter is active, you will see a little pull-down arrow next to each of the fields in your data. Excel assumes that you have arranged the data with columns as variables (fields) and rows as observations (records). Pulling down the arrow next to a field will allow you to select any of the values of that field; the table will then display only the records with that value. So, we could select "Evening" from the "Shift" field and "S" (for Semi) from the "Truck" field to display only the data for unloading times during the evening for a semi. While there is still a lot of data to explore, this simplifies the problem considerably. You could further reduce the data by filtering based on crew size or some other variable.

Notice that whatever fields have been filtered have a blue pull-down arrow; fields that have not been used to filter have a black pull-down arrow. You can unfilter on one variable by simply selecting "all" from its pull-down menu. The pull-down menus also allows you to "Sort ascending" or "Sort descending". Finally, if your variable has many values (like a numerical variable) you can select "Custom" from its menu and entering the information in dialog box. For the data in the example, if you display only semis unloaded on the evening shift and then select "Custom" for the "Time" field and use this to select only times greater than 48 minutes, we see that all the long unloading times were 2 or 3 person crews, giving us greater insight into the situation at StateEx.

For all of the information below, assume that the spreadsheet shown in figure 3.1 (page 13) is being used. It contains data on sample salaries (from Exploration ?? (page ??)).

9.3 How to Stack and Unstack Data Using StatPro

This is a handy feature if you have a lot of data with several categorical variables. You may want to analyze each group of data separately. For example, you may want to compute the average salary of males and females separately in order to compare them. If the data is not sorted to make this easy, the best approach is to unstack the data. This is a standard StatPro procedure, so the six steps discussed in chapter 3 apply:

1. Select the region of the worksheet that contains the data. This step is the same as it is described in chapter 3 for "Summary Statistics in StatPro".
2. Select the StatPro routine to apply to the data. The routine for this is under "Data Utilities..." then select "Unstack Variables..."
3. Verify that the data region is correct. This does not happen for this routine.
4. Select the variables to which the routine will apply. First we select the code variable. This should be a categorical variable with a small number of categories. For example "Gender" might be a good choice, since there are only two options, Male and Female.
5. Fill in the details of the routine. Next select all the variables that you want unstacked. In this example, if you only wanted the salary data for males and females, just select salary. The routine will create new variables called "Salary male" and "Salary female". If you want several variables unstacked, the routine will create a new variable called "Old Variable Category #1" and then "Old variable Category #2", etc.
6. Select the placement for the output of the routine. Select whether you want the results (the new variables) to be placed in a cell next to the data, on a new worksheet, or in a particular cell.

The following guide uses the file C05 FamilyIncome.xls which contains the following variables, observed for many different families in a particular city: Family Size, Location (where in the city they live), Ownership (whether they own their home), First income, Second income, Monthly payment, Utilities, and Debt. This data occupies cells A1:I503 of the worksheet "Data".

Chapter 10

Making Charts and Graphs with Excel and StatPro

10.1 Making Boxplots with StatPro

Microsoft Excel has the built-in ability to make many useful graphs; the boxplot, however, is not one of these. Fortunately, StatPro does include a powerful boxplot routine that has options for making one of three types of boxplots: a single boxplot, a side-by-side boxplot for several unstacked variables, or a side-by-side boxplot from a single variable that is stacked based on a second variable.

1. Select the region of the worksheet that contains the data. This step is the same as it is described above under "Summary Statistics in StatPro".
2. Select the StatPro routine to apply to the data. To make a boxplot, we want to select "StatPro", then "Charts" and then "BoxPlot(s)..."
3. Verify that the data region is correct. This is the same as above under "Summary Statistics in StatPro".
4. Select the variables to which the routine will apply. For boxplots, this step actually comes later, as part of step 5 because there are two types of boxplots that you can make. Each requires slightly different information.
5. Fill in the details of the routine. The next screen you see will ask you whether you want to make a single boxplot or side-by-side boxplots. Single boxplots are for displaying a boxplot of a single variable, while side-by-side boxplots are excellent for comparing two or more variables that are in the same units with about the same range.
6. (Option 1.) Making a single boxplot. After you select "single boxplot" from the list and hit "OK" you will be asked to select the variable (as in the usual step 4). After you hit "OK", StatPro will automatically create a new worksheet called "BOX - (your variable name)" and take you to it. Unfortunately, in Excel 2007, the boxplot procedure has a slight error. When it completes the graph, it selects the wrong line type for the box plot, producing something like the graph shown in figure 10.2.

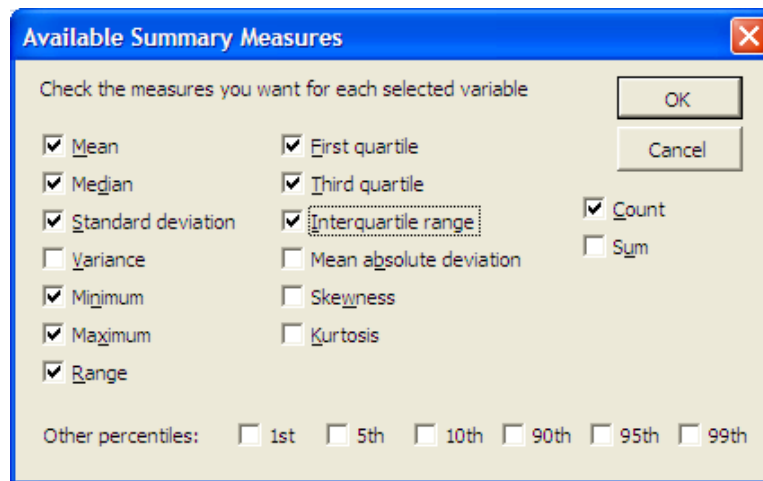


Figure 10.1: StatPro dialog box for selecting the statistics you want to compute.

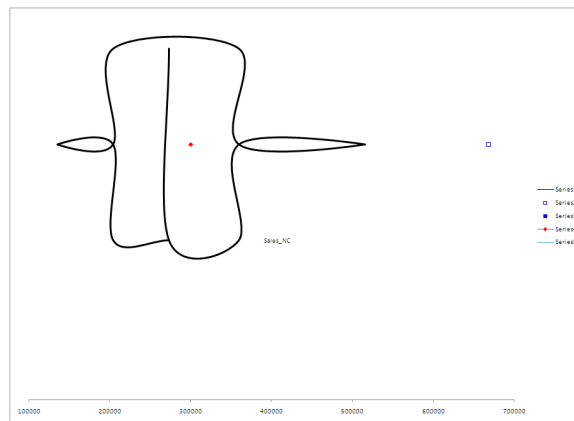


Figure 10.2: The default (incorrect) boxplot made by StatPro in Excel 2007.

Fortunately, there is an easy way to fix this. First click on your newly created chart to activate the design ribbon for chart tools (figure 10.3.) On the left-hand side is the option to change the chart type. The only change you need to make is to select a different type of XY (scatter) chart. Instead of the one with the smooth lines connecting the points, boxplots should use the straight lines (no data points shown) option. See figure 10.4 for the chart type dialog box showing the correct option highlighted. One other change you may want to make for boxplots is to remove the legend (placed along the right-hand side of the graph) because it is totally meaningless.

7. (Option 2.) Making side-by-side boxplots. If you select "side-by-side" boxplot, StatPro will ask you whether the data is stacked or unstacked. If you want to make several boxplots to compare different variables, select "unstacked", then choose your variables, and hit OK. The option to make boxplots of stacked data is only applicable if you have data in which one variable is a "Code" for another variable. For example, if you



Figure 10.3: The Excel 2007 design ribbon for chart tools.

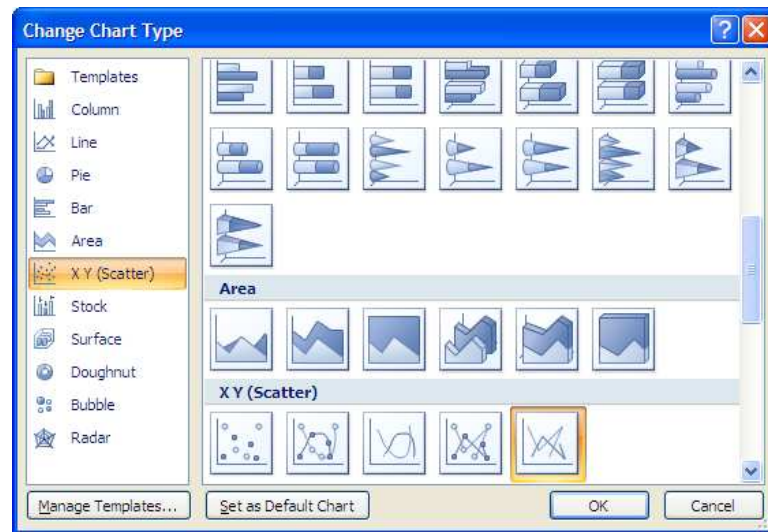


Figure 10.4: Selecting the correct chart type for boxplots in Excel 2007.

have one variable that lists house prices in a neighborhood, and another variable that describes the house type as "Ranch", "Cape Cod", or "Colonial", the stacked option will allow you to create three boxplots for house price: one for each type of house.

8. Select the placement for the output of the routine. This is handled automatically by the "Chart" routines in StatPro. All charts are placed on new worksheets.

10.2 Histograms in StatPro

Making histograms in StatPro is easy. The process follows the same basic structure as for all StatPro procedures.

1. Select a cell in the region of the worksheet that contains the data.
2. Select the StatPro routine to apply to the data. In this case, you should select "Charts/Histograms".
3. Verify that the data region is correct.
4. Select the variables to which the routine will apply. Choose any of the variables in the list. Notice that only numerical variables and categorical variables that are coded as

numbers (like Likert data) are available. In this example, we have selected the Debt variable.

5. Fill in the details of the routine. For this routine, you will see a dialog box like the one below. It includes some information about the variable you selected, and asks for you to fill in three pieces of information: minimum, number of categories, and category width. We have selected the values shown in the figure; in reality, there is an art to selecting these. Usually you will need to try several different combinations to get a reasonably good representation of the data.

Defining histogram categories

Notes on inputs below
 Left-most category extends up to 'Minimum value'.
 'Number of categories' includes open-ended categories on left and right.
 'Category length' applies to all other (non-open-ended) categories.

Information of variable being plotted
 Here is some useful information about Debt
 Minimum: 227.000 5th percentile: 1654.650
 Maximum: 9104.000 95th percentile: 7020.250

Category settings for variable Debt
 Minimum value:
 Number of categories:
 Category length:

Example of settings
 Entering 500, 6, and 100 would produce categories: ≤ 500 , 500 to 600, 600 to 700, 700 to 800, 800 to 900, and >900 .

☐ Test normal fit

OK Cancel

Figure 10.5: Histogram settings dialog box.

6. Select the placement for the output of the routine. As with all charts made through StatPro, you have no choice on this one. StatPro will automatically place the chart on a new worksheet called "Hist - (your variable name)". Notice that another worksheet is also created during this process. StatPro generates a frequency table of the data on a worksheet called "Hist - (your variable) Data". What StatPro does is to create a frequency table of the data using your settings for minimum, number of categories and category width, and then it creates a bar chart from this. By changing some of the features of the bar chart, the final graph looks like a histogram.

Notice that the first category in the histogram is actually a category less than what you enter as the minimum value in the dialog box in figure 10.5. So you need to think about this when choosing values for the minimum, number of categories and category length.

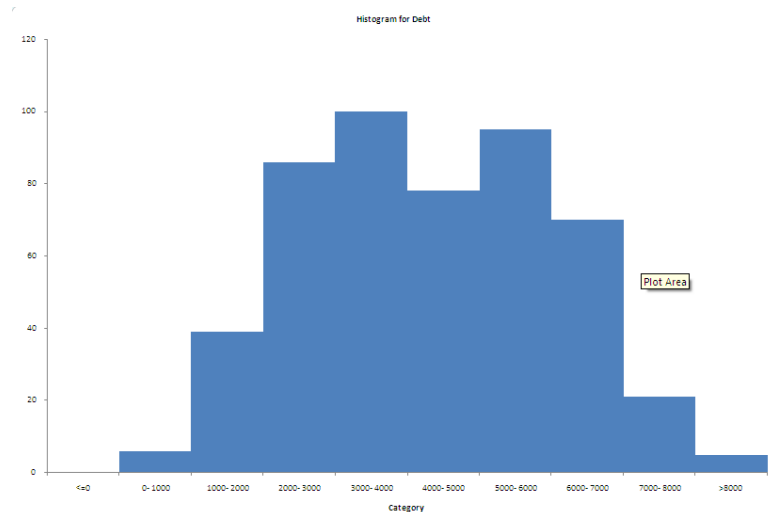


Figure 10.6: Histogram produced using the settings in figure 10.5 for the Debt variable in the file C05 FamilyIncome.xls.

10.2.1 Making histograms of z-scores with StatPro

Sometimes it can be difficult to select reasonable values for the settings to produce an informative histogram. One way to simplify this process is to make a histogram of z-scores, rather than a histogram of the actual data.

1. Compute the mean, the standard deviation.
2. Compute the z-scores for each piece of data.
3. Generate the histogram. When you create the histogram, use the -3 as the minimum, make 8 bins (explained below), and make the width 1 unit wide. For the z-scores of family debt in C05 FamilyIncome.xls, this produces the following histogram, to which we have added the frequency information (see below, "Adding Information to Histograms"). If you have done this correctly, the mean of the data should fall exactly in the center of the horizontal axis, at the zero point between the fourth and fifth bins. Each bin will be one standard deviation wide, easily showing you the number of observations that are within k standard deviations of the mean.

10.2.2 Adding Information to Histograms

Rather than include a frequency table with every histogram, it is usually best to combine the two types of information into one graph. This will make it easier to read the graph and interpret the information that is presented. Adding this information is easy in Excel 2007.

1. Start with any histogram on the screen in Excel.
2. Click on the "Layout" ribbon to access the chart features.

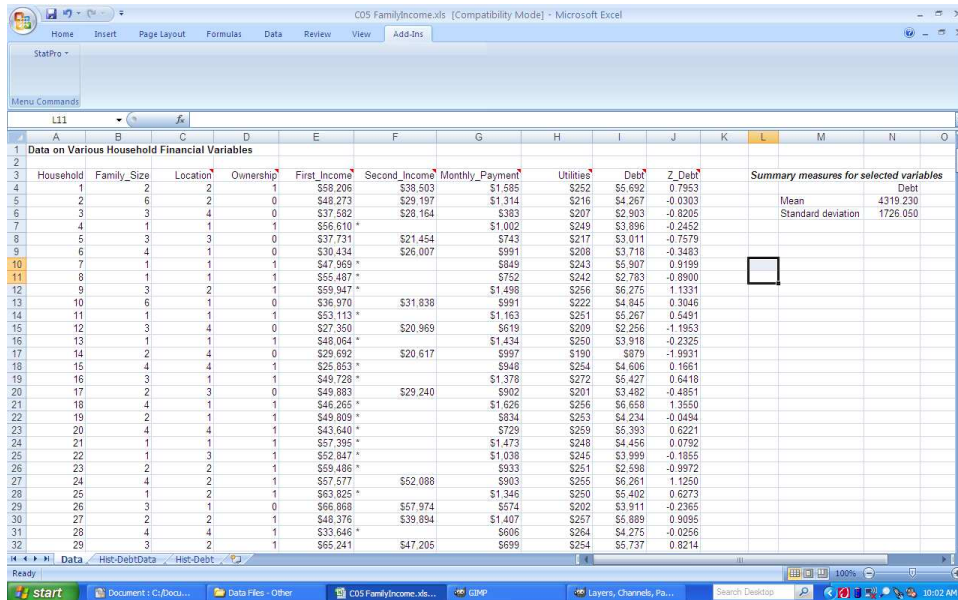


Figure 10.7: Family Income data, with z-scores for debt computed in column J.

- You can then add the data table, data labels, or change just about any aspect of the chart. For the graph shown in figure 10.8 (page 45), we added data labels using the "outside end" option for their placement.

10.3 Histograms in Excel

10.3.1 Using array formulas to make a frequency table

Some of Excel's formulas (such as the "Frequency" formula) are array formulas. Array formulas work with an entire group of cells at once and produce multiple outputs at one time, each of which is placed in its own cell.

To enter an array formula, highlight all of the cells in which the formula should be calculated (this is the "array"). Type your formula (for example, you might want the frequencies for a set of data in a named range called "return" and the frequency ranges are in "bins" so `=FREQUENCY(return, bins)` would be the formula). Next, hit Control + Shift + Enter to enter the array formula in all the cells you highlighted at one time.

If you hit ENTER instead of CTRL+SHIFT+ENTER, the formula will only be entered in the first cell of the array, and you will need to start over.

If you make a mistake entering an array formula you must start over completely (if you try to type something else it gives you an error, saying "you cannot change part of an array formula"). To start over, highlight all the cells in the array. Hit "Delete" on the keyboard. Then begin the process over at the beginning (as above).

If you get trapped and cannot get out of editing a cell in an array formula or any other formula, hit the ESCAPE button on the keyboard. This will take you out of "edit mode" and restore the original contents of the cell that you were trying to change.

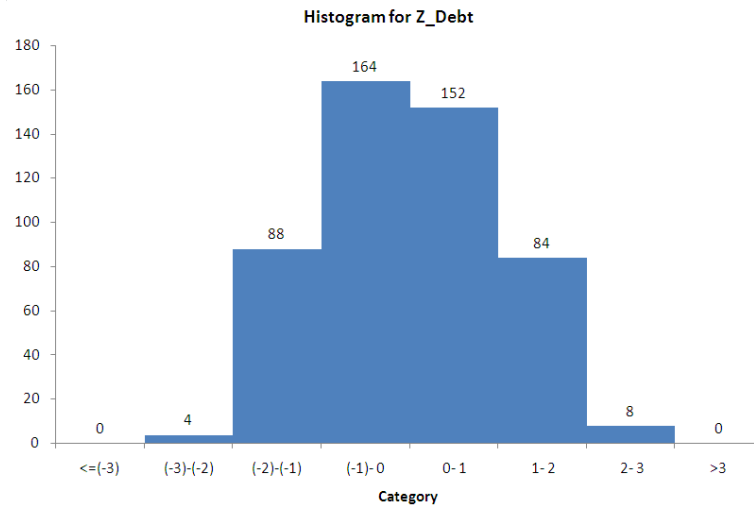


Figure 10.8: Histogram of z-scores for debt in the family income data.



Figure 10.9: The chart layout ribbon for a histogram (or a bar/column chart).

10.3.2 Creating a Histogram from a Frequency Table

1. First follow the steps above to make a frequency table. (But you don't have to use the mean and the standard deviation; you can pick any bin width and any starting point.)
2. Next, highlight the frequency table and click to activate the Insert ribbon.
3. Select the "column" type of chart and use the first subtype "clustered column".
4. Make any changes you want to the legend, data tables, and so forth. When you click "Next", you will have the option of making this a graph in the current sheet, or making it an entirely new sheet in the workbook. (StatPro's histogram feature automatically puts the graph on a new page.)
5. To make your graph look more like the ones that StatPro produces, place the cursor on one of the bars of the graph and RIGHT-CLICK. Select "Format Data Series..." from the bottom of the pop-up menu. Click on the "Series Options" tab and set the "Gap Width" to be "no gap". You can also outline the columns on the chart to stand out better by clicking on the "Border Color" tab, selecting "Solid Line" for the outline type and setting the color of the outline to black using the pull-down menu.

10.4 Scatterplots with StatPro

StatPro makes scatterplots very easy. The procedure follows the same basic steps as all other StatPro routines:

1. Select the region of the worksheet that contains the data.
2. Select the StatPro routine to apply to the data. This is located under "Charts/ Scatterplots".
3. Verify that the data region is correct.
4. Select the variables to which to apply the routine. For scatterplots, you can select as many variables as you like (hold down the control key "CTRL" and use the mouse to select several variables). Every possible combination of the variables selected will be graphed. This means that if you select two variables, you get one graph. Three variables: three graphs. Four variables: six graphs. If you selected ten variables, you would get forty-five graphs!
5. Fill in the details of the routine. For each possible combination of variables, StatPro will ask you which one you want to be on the vertical axis. Be sure to pick the proper variable so that the graph will show the relationship you are interested in seeing. The dialog box for this is shown in figure 10.10.

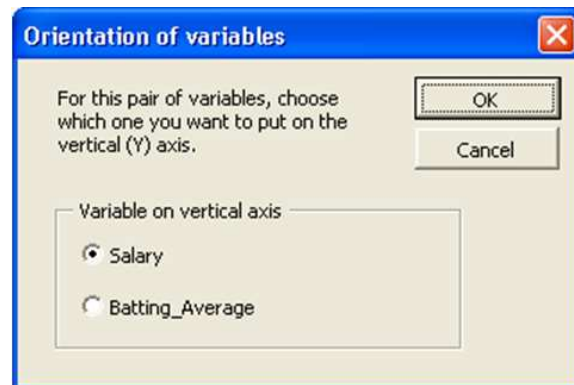


Figure 10.10: Selecting variables for making a scatterplot in StatPro.

6. Select the placement for the output of the routine. StatPro automatically puts each graph on a separate worksheet in the workbook. The sheet will be named in the form "Y-variable VS X-variable" with the names of your two variables filled in. The best part about using StatPro to make the scatterplots is that you can select any variables for the independent and dependent variables. If you use Excel's normal graphing tools, you are limited in your options as to how the graph can be made (unless you go to a lot of trouble).

10.5 Scatter Plots in EXCEL without StatPro

First, select the data you want. For Excel, this means that you must highlight all the data (and the variable names at the tops of the columns) that you want to graph. If the two variables are not right next to each other, highlight the first column of data, then hold down the control key (CTRL) and highlight the second column of data. Click the "Insert" ribbon and select *scatter* from the list of plot types. Then select the subtype of graph that you want to create. See figure 10.11.

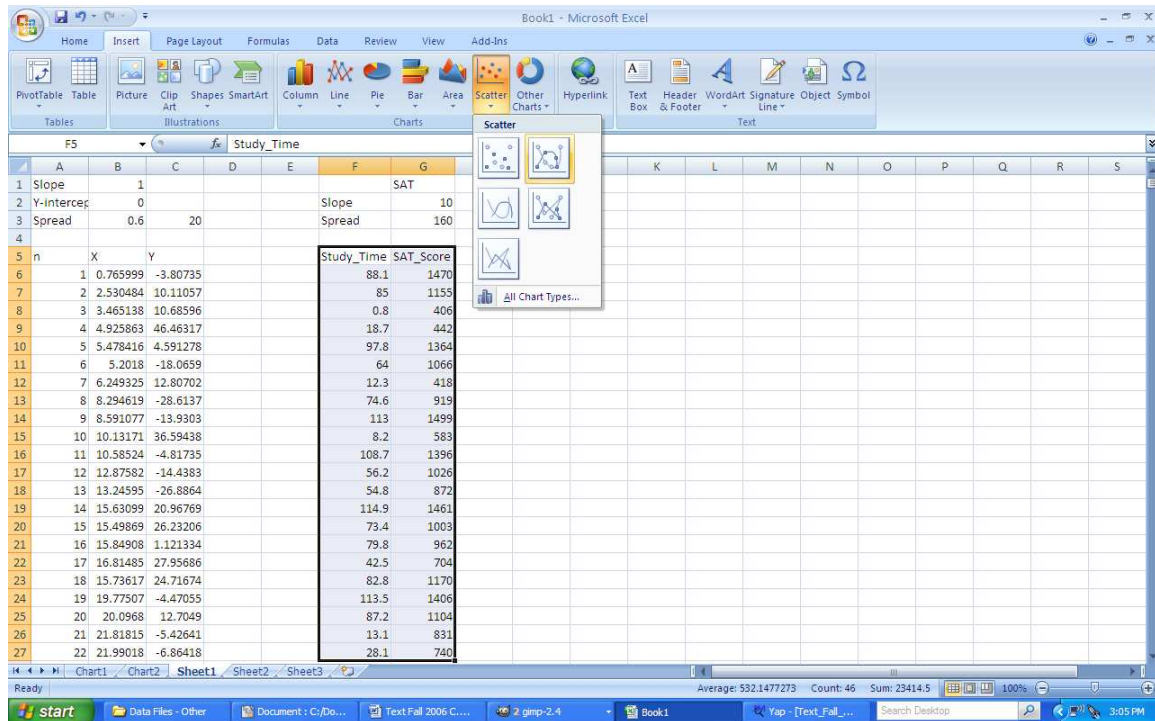


Figure 10.11: Screen image showing the selection of data (highlighted) and inserted a scatter plot.

Note that when making scatterplots in Excel without the use of StatPro, Excel will assume that the left column variable is the independent variable and the right-column variable is the dependent variable. To change this, you will have to first make the graph, then select the graph, and then click "Select Data" from the "Graph/Design" ribbon.

10.6 Moving a Graph To Another Worksheet

Any chart or graph, whether created in Excel or with StatPro, can be moved to either be a chart in an existing worksheet or a separate worksheet page by itself. To do this, click on the chart and select "Move Chart" from the "Graph/Design" ribbon. Then select the option you want from the dialog box shown in figure 10.12

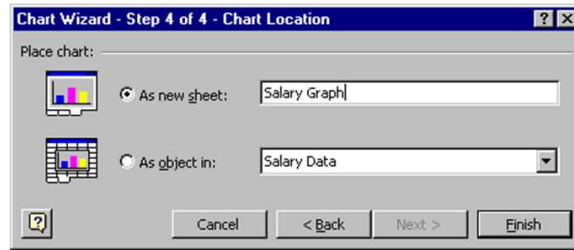


Figure 10.12: Dialog box for moving a chart to a new location.

10.7 Adding Trend Lines to a Scatter Plot

Now we will use EXCEL's capabilities to explore the relationship between the two variables by creating a "Trend line".

1. Position your pointer over one of the points on the scatter plot and right-click your mouse. Select "Add Trendline..." from the menu that appears.
2. You will now have a window (see figure 10.13) that shows several different types of functions that EXCEL can graph on top of your data. Let's select "Linear", which is the default choice. Don't click on "OK" yet, as we have some options to set in order to really take advantage of EXCEL.
3. Make sure you select "Display Equation on Chart" and "Display R-squared value on chart". This will help us in the future.
4. When you have finished setting the options, click on "Close". You should now see your scatter plot with two new things added. One of these will be a solid line, the other will be a little text box that displays the equation of the line and the R-squared value. For right now, think of R-squared as a measure of how closely the line resembles the data. The closer this number is to "1", the better the line describes the data. We can also get R^2 values and equations for the other types of trend lines that EXCEL will graph.
5. Try out some other trend lines with this graph. EXCEL can put as many onto the same graph as you want. Simply repeat all the steps above for making a trend line, except choose a different function in step 2 each time.

A note about the other shapes for trendlines: In later chapters, we'll explore the other types of trendlines and what they are good for. For right now, just be aware that straight lines aren't the only option. Also, as we'll find out, some trendlines simply can't be used with certain data. If this is the case, Excel will automatically "grey out" those choices from the list.

A note about the Polynomial choice for trend lines: Polynomials come in different degrees. You can control the degree of the polynomial that Excel uses by adjusting the number in the box next to the polynomial trendline. Excel allows degree 2 through 6 polynomials.

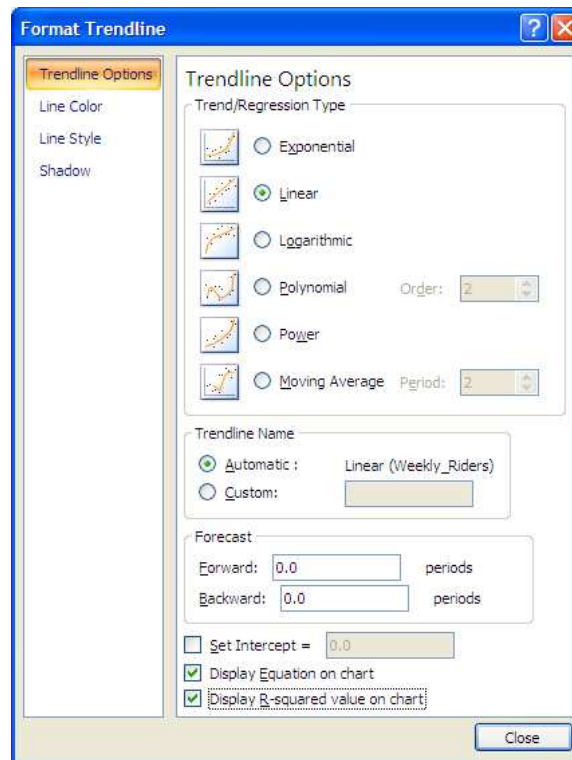


Figure 10.13: Trendline tool in Excel (access by right-clicking on a data point, and selecting "Add trendline...").

10.8 Using the drawing tools in Excel (or Word)

Microsoft Office has many tools designed to help you add graphics to your work in order to enhance its appearance and improve productivity. The Insert ribbon has most of these drawing tools easily accessible. Most of the items that you might wish to draw - lines, arrows, circles, rectangles, etc. - are available from the "Shapes" menu. To draw lines, simply select the line tool. Once the line tool is selected simply click on the page where you want the first end-point of the line to be. Then drag the mouse across the screen and release it where you want the line to stop. To draw perfect horizontal or vertical lines, hold the shift key down while you sketch out the line; this will constrain the line to draw at forty-five degree angles.

The other drawing tools work very similarly to the line tool. For rectangles, holding the shift key while drawing will force the rectangle to be a perfect square. For ovals, holding the shift key will make them into perfect circles.

You can also add a textbox using that menu option. If you select this, and drag out an area on the drawing, you will create a region in which you can type text. This is useful for labeling your drawings and for pointing out important features of the drawings.

Once a drawing object is added to the page (or "canvas") you can make changes to it. Select the object and right click to see available options. You can layer objects, placing one in front of the other, by controlling the "order" of the objects. You can group objects together so that they can be moved as a whole. If you double-click (or right click and select

"format drawing object") you can control the color of the lines, whether the object is filled in (and with what color) and its layout on the screen.

The best way to learn about the drawing tools is to experiment with using them.

10.9 Adding Trendlines for Non-proportional Models

Excel can add trendlines for some non-proportional models to graphs. The process is virtually identical to the process used to add linear trendlines in Excel. The only difference is that in step 2 you should select the following options:

- To get an exponential fit, choose "exponential"
- To get a logarithmic fit, choose a "logarithmic"
- To get the square fit, use "Polynomial" and select "Order 2"
- It is not possible to force Excel to generate trendlines for reciprocals or square roots directly. As it turns out, these are specific cases of the more general "Power" models. However, if you add a "power" trendline to a graph, the power is one of the parameters in the model (like slope or y -intercept) so you probably will not get a power of 0.5 ($= \frac{1}{2}$ which is a square root model) or a power of -1 (for a reciprocal model). We'll talk about how to force Excel to fit these types of models with regression in the next chapter.

10.10 Making a 3D surface plot

Once you have a table of values for a function of two variables, whether generated from actual data or from a formula (see above), it is relatively easy to create a surface plot. First highlight all of the table (in file "C13 HowTo.xls", we would highlight B2:M18). Notice that the first cell, the empty one, is included in this. Then from the insert ribbon, select "Other charts" and choose the first of the surface chart types.

10.11 Logarithmic and Log-Log plots

When you have data that spans many order of magnitude (like 1, 10, 100, 1000, 10000...) taking the logarithm of the data reduces it to a much more manageable set of numbers. For example, if we take the base-10 logarithm of each number in the preceding list, we get the numbers (0, 1, 2, 3, 4, 5...) which are much easier to use. This is the essence of many commonly used scales of measurement (the Richter scale for measuring earthquake energy and the unit of measuring sound, the decibel, are both logarithmic). This is also useful in dealing with models in which the variability in the residuals increases.

An alternate approach to actually computing the logarithm of each data point is to simply graph the data on a logarithmic scale. This is easy to do in Excel. For example, if you enter the pairs of (x, y) data points shown below and generate a standard XY (scatter) plot of the data, the graph is obviously curved, indicating a nonlinear relationship between the variables.

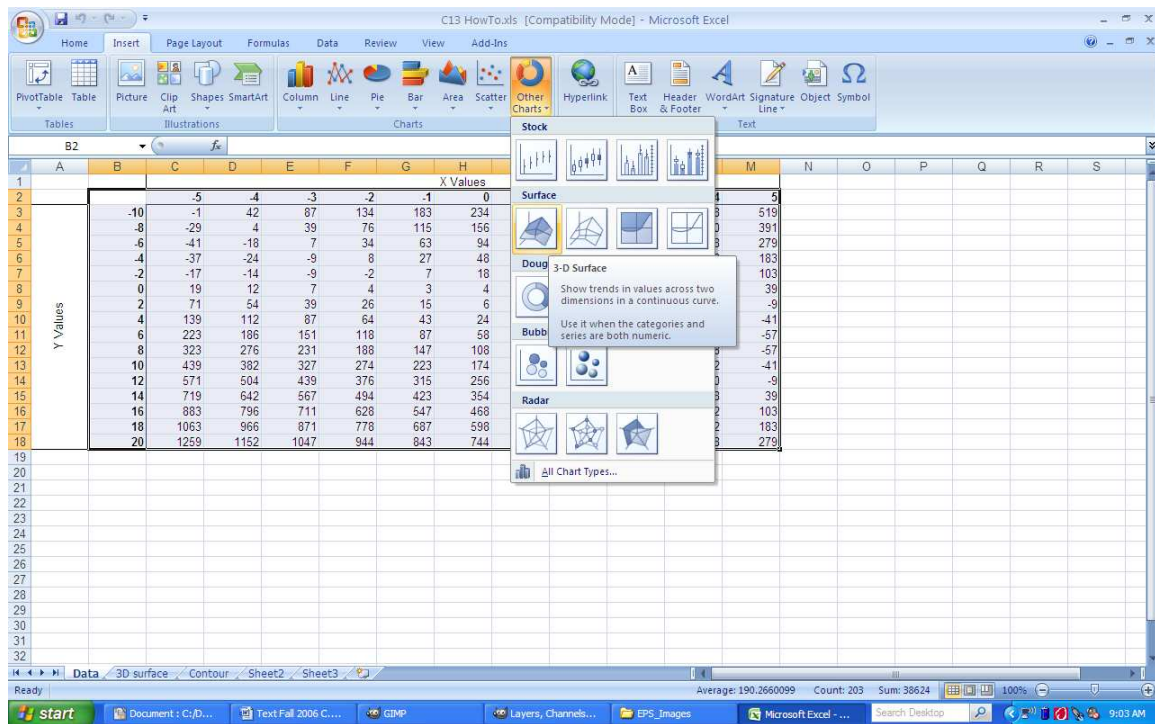


Figure 10.14: Inserting a 3D surface graph in Excel.

x	y
1	2
2	7
3	20
4	54
5	148
6	400
7	1100
8	2900
9	8100

Click on the graph, then select the layout ribbon from the toolbar. Click on "axes" (not "axis titles") and select primary vertical axis. From the pull down list, choose "show axis with log scale". The graph will still display exactly the same data, but will appear to represent an almost linear relationship. This is shown on in figure 10.17. Notice that the vertical axis now looks very different. In the original graph, the evenly spaced gridlines represented an increase in the y variable of 1,000, regardless of whether you were at the top of the axis or the bottom. The spacing on the logarithmic scale, though, increases by a factor of 10 for each gridline (from 1 to 10, 10 to 100, 100 to 1000, etc.)

You can change the scale on the horizontal axis as well, letting you create log-linear, linear-log and log-log type graphs.

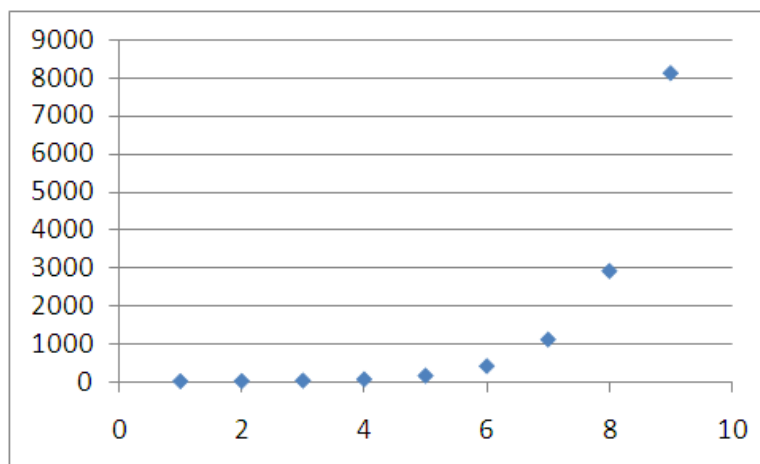


Figure 10.15: XY data showing a nonlinear relationship.

10.12 Create a scatter plot with more than one y -variable on the same axes

StatPro will not allow you to create a graph with more than one dependent variable plotted against the same independent variable. For example, in the graph below you see two y -values plotted for each x -value. To generate such graphs, you will have to use the chart wizard in Excel. The steps are outlined below.

1. Highlight all the data you wish to graph, including all the y -variables. Make sure that your data is organized with the x -variable (the independent variable) in the left-hand column.
2. Click on the Chart icon in Excel or go under "Insert" and click on "Chart"
3. Click on xy -scatter
4. Click through all the "Next"'s and click "Finish"
5. Move/resize your graph

10.13 Scatterplots with multiple curves

10.13.1 Finding the Intersection Point of Two Curves

Suppose you want to find the point of intersection of the two curves above given by the equations:

$$\begin{aligned} D(x) &= -0.0029x^2 - 0.0139x + 118.26 \\ S(x) &= 0.0015x^2 + 0.0806x + 30.596 \end{aligned}$$

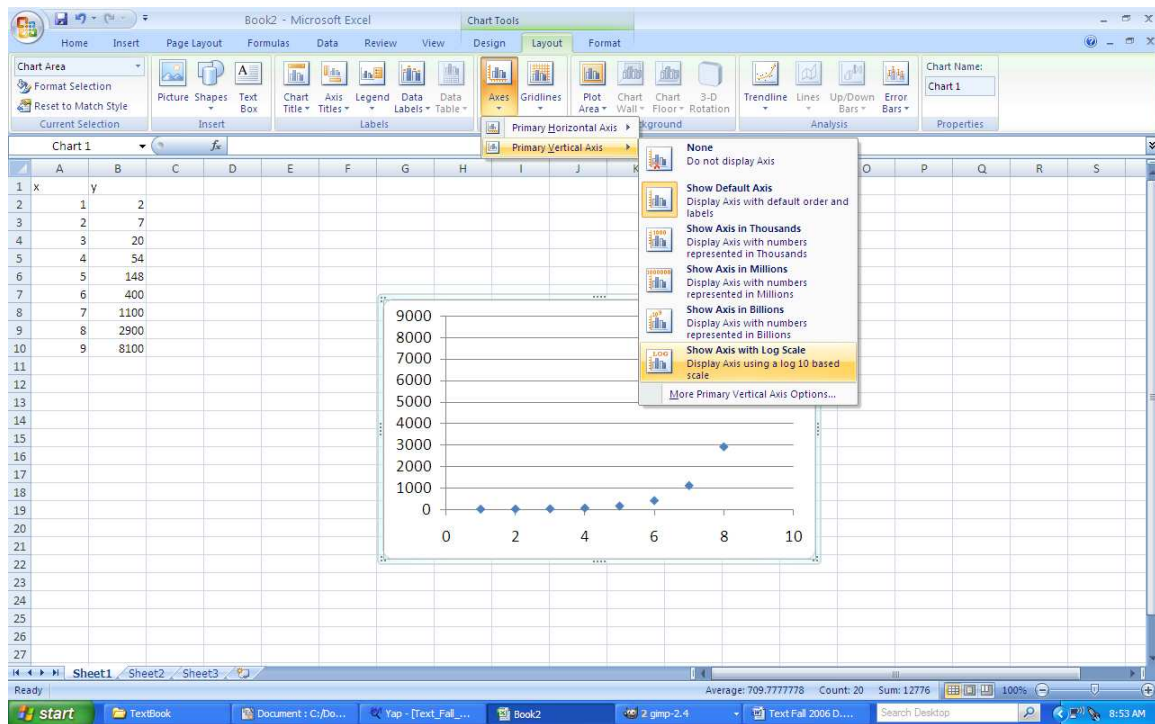


Figure 10.16: Converting the vertical axis of a graph to a logarithmic scale.

The point of interchapter occurs when the difference of the two functions, $D(x) - S(x)$, is zero. We will use Goal Seek to find the x that will make this difference zero. If you set up your spreadsheet as below, you will also find the y that goes with this x . To use Goal Seek, see the How to Guide for Chapter 7.B.

In A1, we have a "guess" value that Goal Seek needs to get going. We entered 120 because $x = 120$ is pretty close to the interchapter of the two curves representing the graphs of our two functions. Cell C2 contains the formula for $D(x)$, cell D2 contains the formula for $S(x)$, and cell B2 contains $D(x) - S(x)$. B2 is the cell we want to set to zero by changing B1. You should get the following screens:

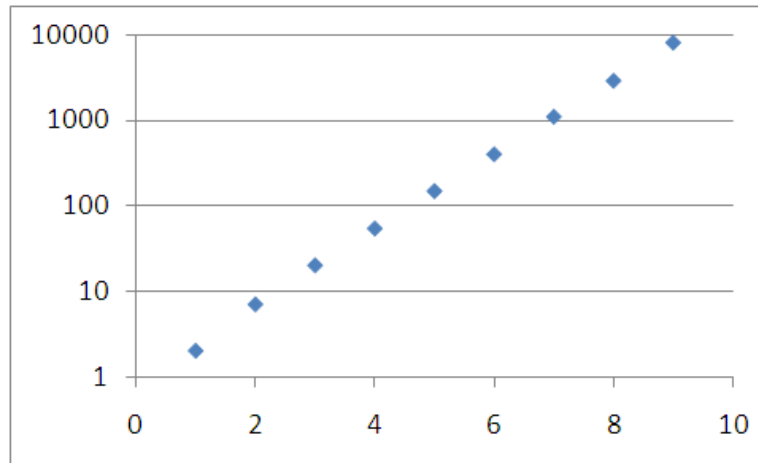
So the x -coordinate of the point of interchapter is $x = 131$ in cell A2. The y -coordinate of the point of interchapter is $y = 67$, as we can see in cells C2 and D2. In our example, the point of equilibrium is (131, 67).

10.13.2 Graphing More Than One Scatterplot and Trendline on the Same Axes

There are two different ways to go about this.

Method 1. Creating the graph from scratch.

Bring up the data set, which will contain an independent variable and two or more dependent variables. Make sure the independent variable is in the left column. You cannot use StatPro to create a graph with more than one dependent variable. You must use Excel's chart wizard.

Figure 10.17: XY data showing a nonlinear relationship on a $\log(Y)$ scale.

	A	B	C	D
1	x value	$D(x) - S(x)$	$D(x)$	$S(x)$
2	120	$=C2-D2$	$=-0.0029*A2*A2-0.0139*A2+118.26$	$=0.0015*A2*A2+0.0806*A2+30.596$
3				

Figure 10.18: Setting up Excel to use goal seek to find the point of interchapter of the supply and demand curves.

1. Highlight the independent and dependent data you wish to graph.
2. Go to "Insert" and click on "Scatter"
3. Select the sub-type you want
4. Add the appropriate information about your graph in the Chart Title and the x- and y-axes.
5. Add the appropriate trend line(s) displaying the equation and the R^2 .
6. Repeat step 5 for the each data series (set of y values) on the graph.

Method 2. Adding a scatterplot to an existing graph.

If your chart already exists and you want to add a new series of y -values to the graph, you do not have to start over and make a new chart (although you could, using Method 1 above to get all the y -data on the graph). The first step is to right-click in the white space around the existing graph. Choose "Select data..." from the menu that appears to bring up the dialog box in figure 10.21. For example, suppose we had the function D from the example above plotted by itself (see figure 10.20) and we wanted to add a graph of function S to the same axes.

The easiest way to add the new column of data is to simply highlight all of the data (the new and the old) that you now want to appear in the graph. If the columns of data you

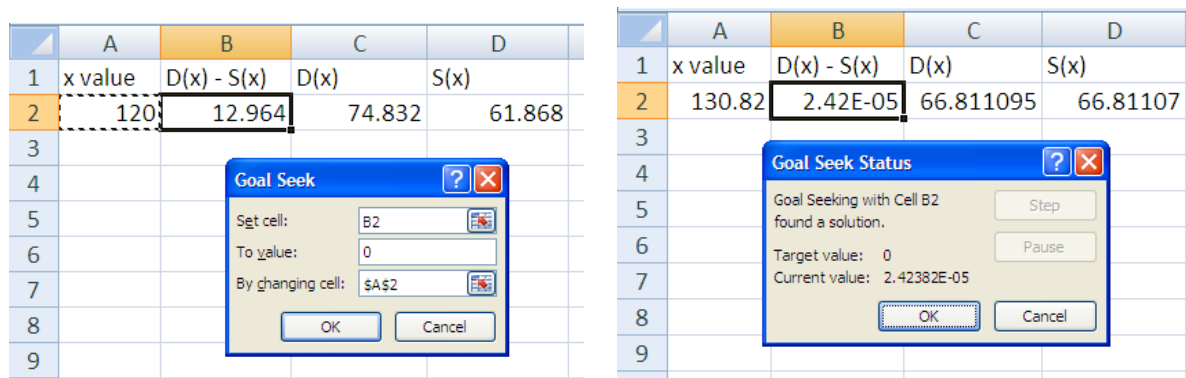


Figure 10.19: Goal seek finds the point where the curves intersect.

want graphed are not adjacent to each other, remember to hold down the control key while you select the regions. Alternatively, you can select "Add" to add a new data series, but that takes a little more work.

When you are finished, hit "OK" and the new data should appear on the chart.

10.13.3 Graphing More Than One Equation on the Same Axes

Suppose you know the equations you wish to plot. Perhaps you have graphed two scatterplots and their trendlines as above. You wish, however, to graph just the trendlines by themselves without the scatterplots. Bring up an Excel spreadsheet. You need to determine how many x -values you wish to use. If you have already done scatterplots and trendlines of the data, you can use it as a guide. To illustrate, suppose you want x to range from 1 to 160 and suppose you wish to graph two functions:

$$\begin{aligned}
 D(x) &= -0.0029x^2 - 0.0139x + 118.26 \\
 S(x) &= 0.0015x^2 + 0.0806x + 30.596
 \end{aligned}$$

Set up your spreadsheet as you see below. Enter 1-160 under the x column (highlight 1, 2, 3 and drag the fill handle down to A161) and enter the two equations in B2 and C2, respectively. Double click the fill handle on B2 and C2 to generate the remaining values. Next, highlight all the data (the column of x data and the columns of y data) and insert a new scatter plot. Select the sub-type of chart on the left side of the second row (see figure 10.22).

10.14 Drawing on a chart

Occasionally, you will want to add details to a chart in order to call attention to certain features or to fill in missing information. The drawing tools can help you do this. To access the drawing tools, go to "View/Toolbars/Drawing". One example of this would be adding in the line which marks the equilibrium level in a graph showing the supply and demand curves. You can do this one of two ways.

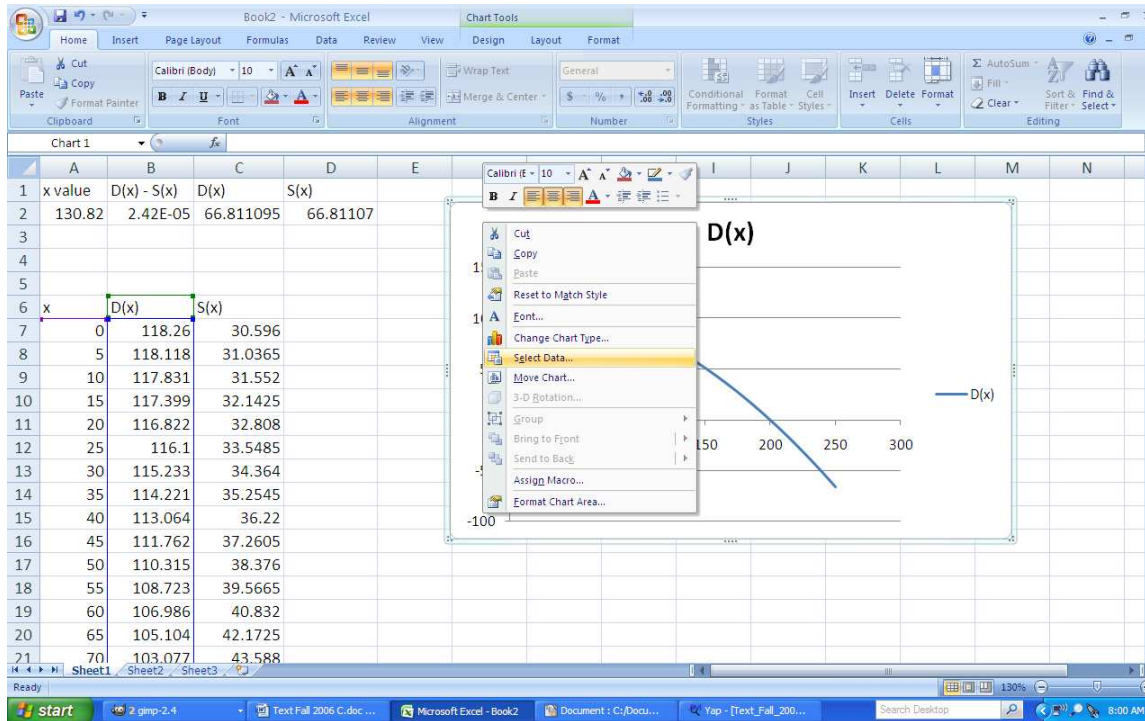


Figure 10.20: Adding a second set of y values to a graph by right-clicking on the graph.

Method 1. Draw the line yourself Select the line tool from the drawing menu; it looks exactly like a straight line. Position the mouse where you want the line to start and then LEFT-CLICK and DRAG the other end of the line until it is where you want it. Release the mouse and your line is there. Double-clicking on the line will allow you to change features like its color, thickness, and so forth.

TIP: To force the line to be exactly horizontal, hold down the SHIFT key while dragging and do not release the shift key until after you release the mouse button.

Method 2. Add the line as the graph of a horizontal line You could also add a new column of y -values to the data table. Fill in any y -value you wish in the first cell in the column (such as cell D2 in figure 17.24). Then, in the second cell (D3) enter the formula " $=D2$ " and copy this formula down the column. Add this series to the graph (or create a new graph with all the series included). Now you can adjust the value in cell D2 and the entire line will remain horizontal. Keep adjusting the position of the line until it is exactly where you want it.

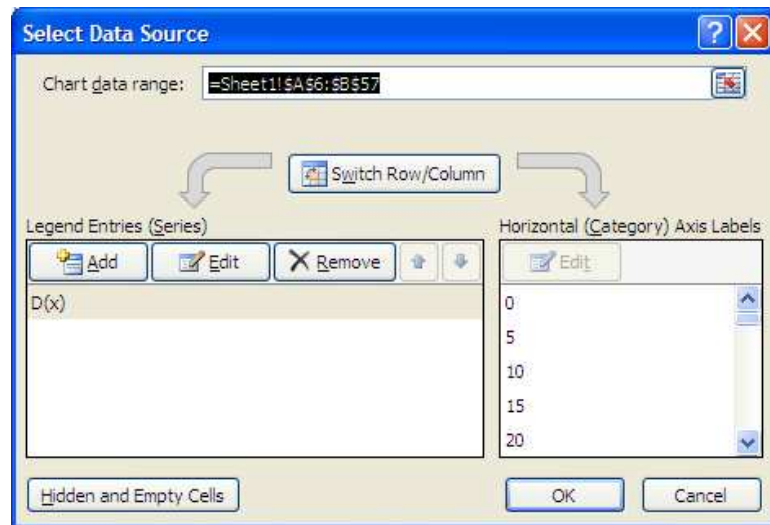


Figure 10.21: Selecting new data for a graph.

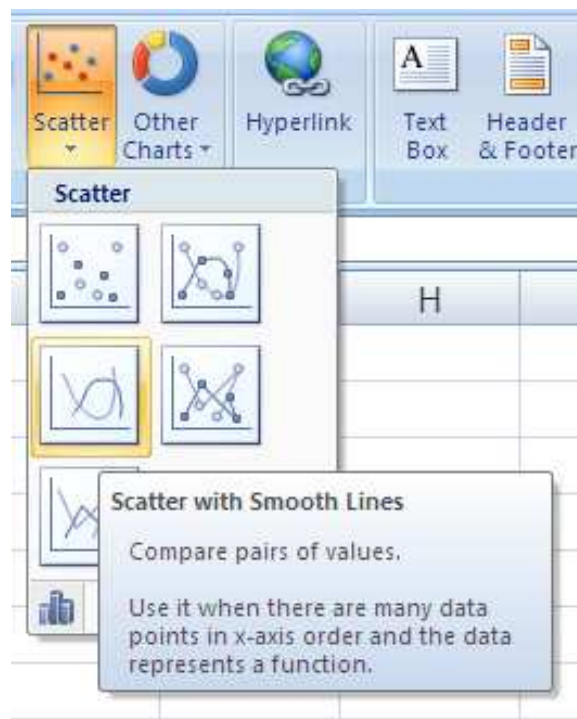


Figure 10.22: Adding a scatterplot without the data points.

Chapter 11

Estimating Statistics from Summary Data

11.1 Estimating the mean in Excel

We will start by assuming that we want to enter the data from example ?? (page ??) and estimate its mean. The easiest way to do this is to structure the data table in Excel as shown below in cells A1:B6, with the minimum value of each salary bin (or whatever variable is used to summarize the data) in one column (marked "Low" in the example) and enter the maximum values of each bin (marked "High") in a separate column. This will allow us to enter these as numbers and to have Excel to calculate the middle of each bin for use in the calculations.

	A	B	C	D	E
	Low	High	Number of Employees	Midpoint	Count*Mid
1	200000	250000	1	\$225,000.00	225000
2	150000	199999	2	\$174,999.50	349999
3	100000	149999	5	\$124,999.50	624997.5
4	50000	99999	13	\$ 74,999.50	974993.5
5	0	49999	38	\$ 24,999.50	949981
6					
7					
8		SUM	59		3124971
9					
10		Mean (estimated)			\$ 52,965.61

Figure 11.1: Estimating the mean of a set of data given only by frequency counts.

Now, column D contains your estimate of the midpoint for each bin. In cell D2, enter

$$= (A2 + B2)/2$$

Now copy this formula down the column. Column E is just the individual frequency counts multiplied by the midpoint for the bin, so in E2, enter

$$= D2*C2$$

After copying this formula down, you will need to compute the sum of the frequency counts (in cell C8 in the example) and the sum of the frequency count*midpoint data (in cell E8 in the example). The estimate for the mean is then simply **=E8/C8**.

11.2 Estimating the standard deviation of the data in Excel

To estimate the standard deviation for the data shown in the previous example, we need to add two columns to the data. Column F will contain the deviation of each data bin from the estimated average. Column G will contain the frequency data times the square of the deviations. This is shown in the figure below.

	A	B	C	D	E	F	G
1	Low	High	Number of Employees	Midpoint	Count*Mid	Deviation	Count*Dev^2
2	200000	250000	1	\$225,000.00	225000	\$172,034.39	\$29,595,831,284.36
3	150000	199999	2	\$174,999.50	349999	\$122,033.89	\$29,784,540,534.33
4	100000	149999	5	\$124,999.50	624997.5	\$72,033.89	\$25,944,406,420.57
5	50000	99999	13	\$74,999.50	974993.5	\$22,033.89	\$6,311,399,913.82
6	0	49999	38	\$24,999.50	949981	\$(27,966.11)	\$29,719,926,084.46
7							
8		SUM	59		3124971		1.21356E+11
9							
10		Mean (estimated)			\$52,965.61		
11							
12					Std Dev (estimated)		\$45,742.18

Figure 11.2: Estimating the standard deviation of a set of data given only by frequency counts.

To compute the deviations, enter the formula below in cell F2 and copy it down.

$$= D2 - \$E\$10$$

To compute the frequency times the square of the deviations, enter the formula below in cell G2 and copy it down.

$$= C2*F2*F2$$

Now, sum up the weighted squares of the deviations by entering

$$= \text{SUM}(G2:G6)$$

in cell G8. To complete the calculation that estimates the standard deviation, we need to divide the sum of the squared deviations by the total number of observations (minus 1 for the loss of one degree of freedom) and take the square root. So we enter the following in cell G12 (Note that, in this formula, the parentheses are all necessary in order to enforce the proper order of operations):

$$= \text{SQRT}(G8/(C8-1))$$

Chapter 12

Cumulative Distributions in Excel (with help from StatPro)

First make a histogram of the actual data, using whatever minimum, number of categories, and category length makes sense. Notice that StatPro generates two worksheets. One will be called "Hist-(variable name)" and contains the actual histogram. The other sheet is called "Hist-(variable name)Data" and contains a frequency table of the data which is used to generate the histogram. To make the cumulative distribution graph, we will use the frequency table and add two columns to it. We will then make a column chart of the data under the "Category" column and the data in our last column that we create.

Start by adding headings for each column into cells D3 ("Total") and E3 ("Percent").

To calculate the values for the "Total" column, enter the following formulas: In cell D4, enter " $=C4$ ". This looks up the first bin (C4). In cell D5, enter " $=D4 + C5$ ". This adds the next bin (C5) to the previous total (D4). Now, select cell D5 and double-click on the fill handle. This should complete the column. The total in the last row should be equal to the total number of observations in the data.

To calculate the values for the "Percent" column, enter the formula " $=D4/(\text{last total})$ " where (last total) is an absolute cell reference to the last cell in the "Total" column. If you have 12 categories, this cell will be cell \$D\$15. If you have a different number of categories, just enter the appropriate cell reference, but make sure it is an absolute reference.

Now we have all the data we need; all that's left is to make a bar graph of the data in columns B ("Category") and E ("Percent"). To do this, select all the cells with data in column B. Now, hold down the CONTROL (CTRL) key and select the data cells in row E. When you release the mouse button, all the data in both columns should be highlighted, but nothing else should be highlighted. Be sure to include the headings for each column as well, if you want the graph to automatically set up labels for the axes.

Now click on the "chart wizard" button on the toolbar. Select the "Column" type and choose the subtype in the upper left corner, the "clustered column" graph. Click "Next". At step 2 click "Next". At step 3, enter the titles you want and adjust the legend (probably you can remove the legend; it's not very helpful and clutters up the graph). To make the chart look like StatPro's histograms, click "As new sheet..." in step 4.

Now you can clean the chart up to make it look better. Right click on the grey area and select "Format plot area". Under "Area", select "None" and hit "OK". This will remove

the background, making it easier to print and easier to read when you copy the chart into a Word document. Next, click on one of the columns in the graph and select "Format Data Series". Go to the last tab, marked "Options". Set the gap width to 0 (either type "0" in the box or use the arrow buttons) and hit "OK". You will be left with a graph that looks a lot like StatPro's!

Chapter 13

Checking for Normality with StatPro and Histograms

When making histograms using StatPro, you might have noticed the "test normal fit" option in the lower left corner of the dialog box where you set up the categories.

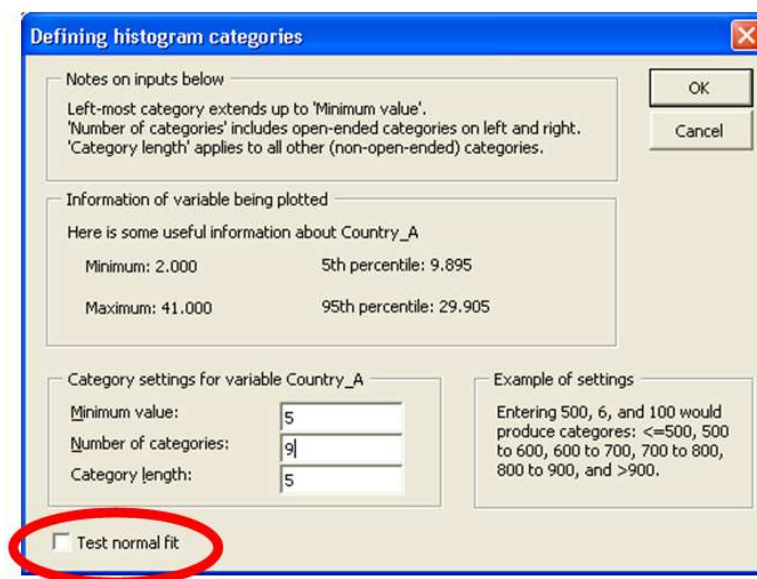


Figure 13.1: Testing a distribution for normalcy with StatPro's histogram routine.

If you click on this box, StatPro will add some information to your histogram and to the frequency table of the data. First off, you will notice that a dialog box appears over the histogram telling you whether the data could be from a normal distribution. To determine this, StatPro uses a p-value. We won't get into the way p-values are calculated, but the main idea is that if the p-value is close to 0, then the data is probably not from a normal distribution. Clicking "OK" will remove the dialog box and let you see the graph.

Notice that the graph has your data shown in blue, but there are other data outlined in between each of your columns. This extra information shows you what the heights of the

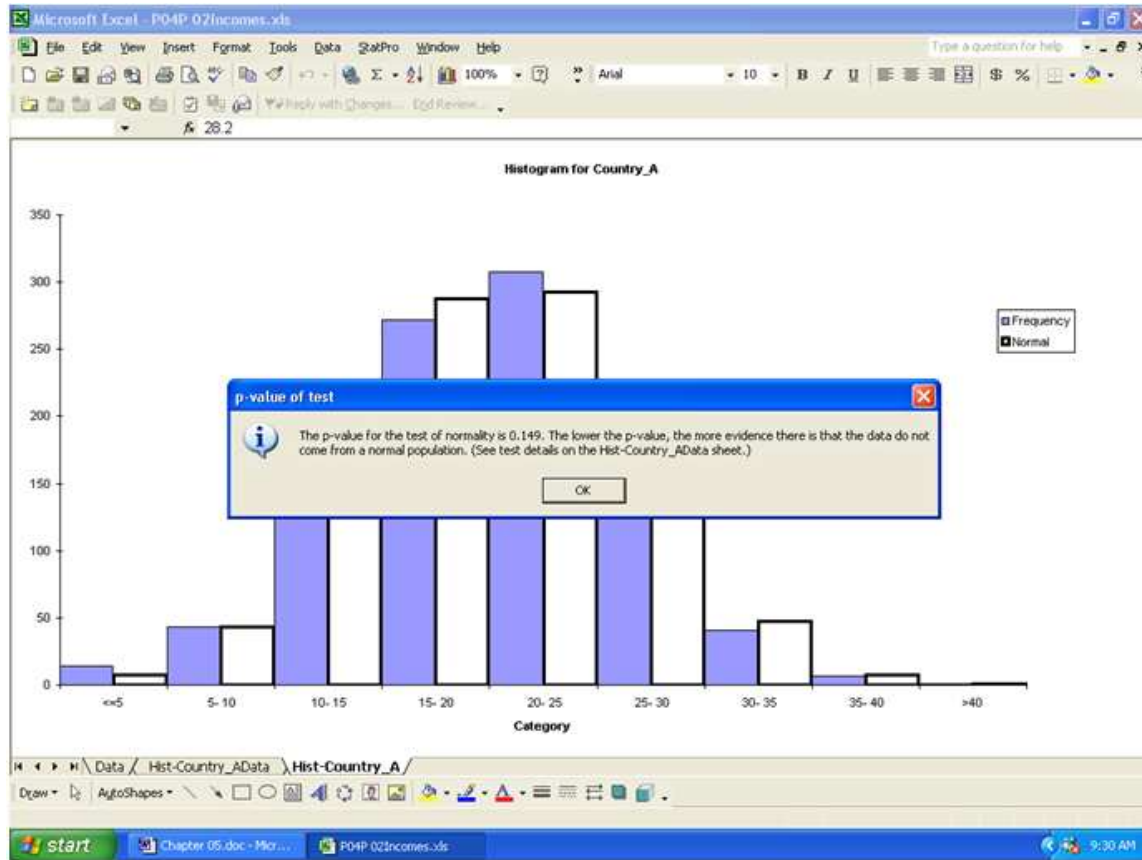


Figure 13.2: Results of testing for normalcy using StatPro.

bins should be if the data came from a perfect normal distribution. The bigger the difference between your data and the theoretical data, the lower the p-value.

On the frequency table, you will also notice additional information. Two new columns ("Normal" and "Distance measure") are in the table, and there is a "Test of Normal Fit" off to the side. The "Normal" column is used to graph the theoretical normal data. The other information is used to generate the p-value. Notice that there is no definite answer as to whether the data is or is not normal. This is because the data is a sample from a distribution; we don't have all the population information, so it's possible that the underlying population characteristic is normal, but your sample is slightly skewed.

Chapter 14

Entering an Algebraic Model into EXCEL

In order to take a mathematical model and input it into Excel to make predictions, we need to complete three steps: Enter values for the parameters, create the data table for the x and y variables, and plot the data. These steps are outlined below for the linear model

$$y = A + Bx.$$

1. Enter the parameters. Notice that our model has two parameters, A and B . (It also has two variables, X and Y .) We need to tell EXCEL what numbers we want to use for A and B . Let's try $A = 0.5$ and $B = 3$. Now we will enter this into EXCEL, along with some labels so that we can read the spreadsheet when we are done. Enter the labels "A" into cell A3 and "B" into cell A4, and enter the values of these two parameters into cells B3 and B4, respectively. This is shown in the screen illustration in figure 14.1.
2. Create the data table. Remember: EXCEL works best with data, and we have been working with data that is organized with variables listed across in columns and observations of those variables listed as rows. Our variables in the linear model are X and Y . So we need two columns, labeled X and Y . I'll enter these labels in cells D1 and E1. Next, in the column under the "X" we need to enter some values for the independent variable, X . We can pick anything we like, but it's easiest if we pick a nice pattern like 1, 2, 3, 4, 5... or 0.1, 0.2, 0.3, 0.4, 0.5... We'll need a lot of values, so we'll let EXCEL get the pattern going. I'd like to use 1, 2, 3, 4, 5... so I'll enter 1 in cell D2 and 2 in cell D3. Then I'll highlight those two cells. Notice that if you position the cursor over the little box in the lower right corner of the highlighted box, it turns into a "+". Click the left mouse button, hold it down, and drag straight down the screen as far as you want to go. Notice what happens: all the cells you dragged through have a dashed box around them and there is a little yellow box with a number in it. As soon as you release the mouse button, EXCEL will fill in the values according to the pattern that you started: 1, 2, 3, 4, 5... as far as you dragged down to. This setup is illustrated in figure 14.1 with the results shown in figure 14.2

We have now created half of the data table, the X values. We need to use a formula to get the Y values. Notice that our model says we can get Y by computing the value

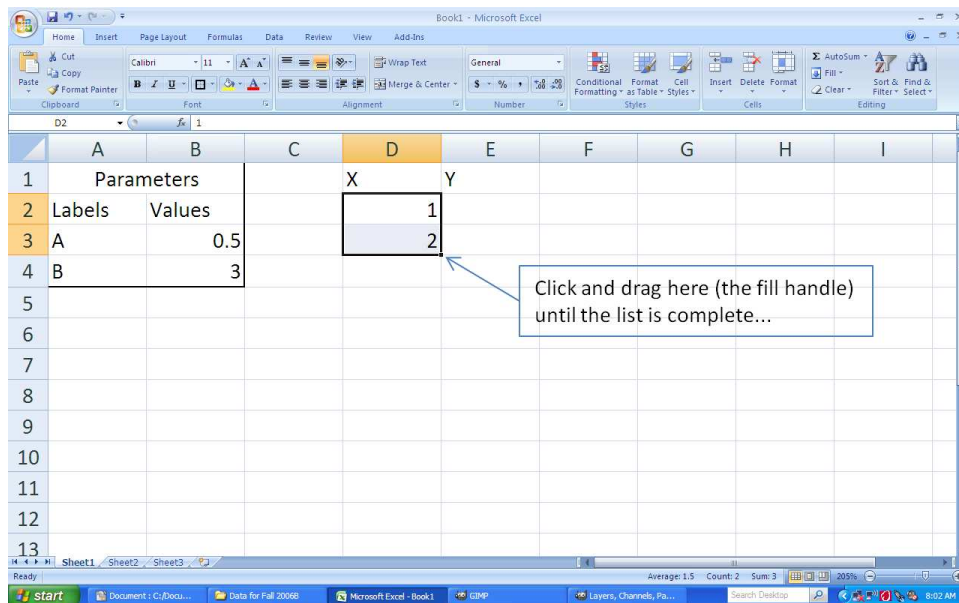


Figure 14.1: Setting up parameters and data table for creating a linear function.

of $A * X + B$. A and B will always be the numbers we put in cells B1 and B2, but there are lots of X values that we just created.

To enter the formula for the Y variable, we need to first click on cell E2. All formulas in EXCEL begin with an equal sign, so type $=$. Now we need to tell EXCEL to look up the value of the parameter A . We've put this in cell B3 so we'll enter $\$B\1 . Now we multiply this by X ; the first value of X is in cell D2, so type D2. (The $"*"$ stands for multiplication.) Finally, we need to add the value of B to this. So type $+\$B\2 since the parameter B is stored in cell B4. You should have typed the formula below into cell E2:

$$= \$B\$3 * D2 + \$B\$4$$

Why the dollar signs for the parameter cells (B3 and B4) but not the variable (D2)? Remember, there is only one value for A and one for B . We need to make sure that EXCEL always uses cell B3 for the value of A . The dollar signs tell EXCEL "No matter what, do not change the cell reference from B3." This is the way we force Excel to use an absolute cell reference.

Now, we just need to copy this formula to all the other cells in column E so that we get one Y value for each and every X value. Click on cell E2, position the cursor over the little box in the lower right corner of the cell, click the left mouse button, hold the button down and drag the cursor so that you highlight each cell in column E that has an X value next to it in column A. Now release the mouse button and EXCEL will fill the formula in.

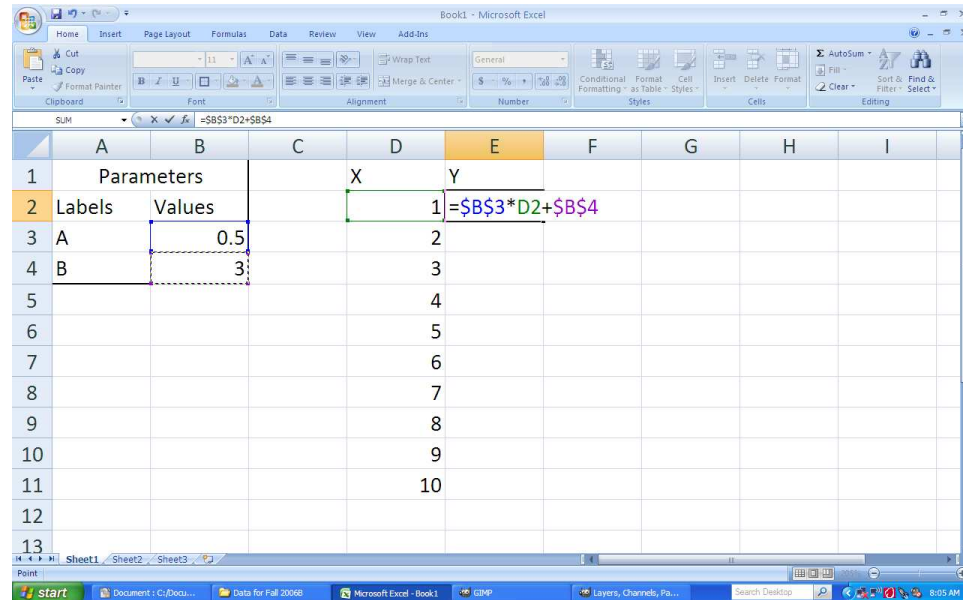


Figure 14.2: Entering the formula for a line.

To see more about the dollar signs for B3 and B4, click on cell E7 and then look up at the formula bar. Notice that when EXCEL copied the formula, the parameter references (B3 and B4) didn't change, but it is looking for the X variable in cell D7 instead of D2! This is because X is a variable - each value of X corresponds to a different value of Y - the A and the B are parameters; they never change once we set them at the top. Each value of Y uses the same values for A and B , those that we typed into the top of the spreadsheet.

3. Create the plot of the model. This is the easy part; it's just like making a scatter plot, which we've discussed in previous chapters. First highlight all the data; in this example, it's cells D1 to E11. Now click on the chart wizard icon on the tool bar. Select "XY (scatter)" and pick a subtype; we've chosen the one in the second row and second column. After you're done, it should look like the figure.

Now you can try changing the parameters and observing what happens to the graph. You could also try different models by changing the formula that you type in step 2. Explore! That's the best way to learn.

Chapter 15

Using Goal Seek

Goal seek is a way to have Excel find approximate solutions to equations. To set it up, you need to set up your spread sheet so that there are two cells with information. The first cell contains a guess for the solution to the equation. The second cell contains a formula to calculate the actual result, based on the cell containing your guess. So, in the backpack example above, you might guess that the backpack will hold 40 books. Place this information in cell B1 and label it with "Books" in cell A1. Then, in A2, put the label "Price" and in B2 enter the formula to calculate the price: $= -30.68 + 1.46*B1$. Select cell B2 (containing the price calculation) and activate the Data Ribbon. From there, select "What if analysis" and choose Goal Seek from menu. This is shown in figure 15.1.

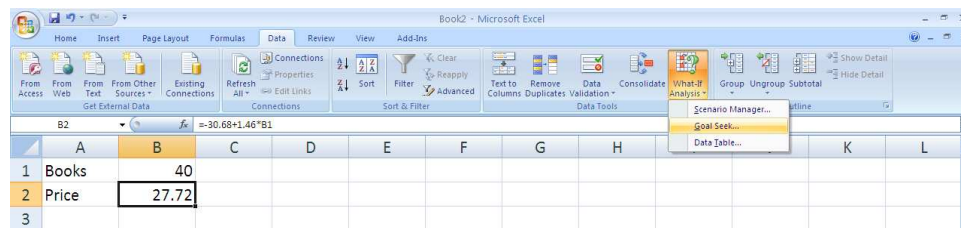


Figure 15.1: Setting up for using Goal Seek.

Fill in the values shown in figure 15.2 and hit "OK". Excel will place the results in the cells on the spreadsheet. (Note: Goal Seek needs a good guess in order to work!)

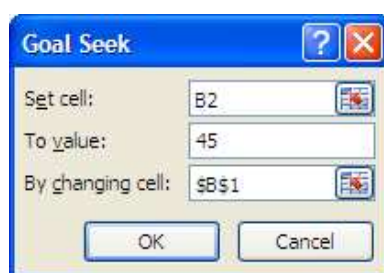


Figure 15.2: Entering values to determine how many books a \$45 backpack will hold.

Chapter 16

Constructing Regression Models in Excel and StatPro

16.1 Simple Regression with StatPro

StatPro makes regression analysis fairly easy. Follow the steps below with any set of data to develop a regression model along with its associated measures and graphs.

1. Select the region of the worksheet that contains the data
2. Select the StatPro routine to apply to the data.

The routine for simple regression is under "Regression/ Simple..." Note that there are other types of regression. In this text, we'll focus on "simple" and "multiple". "Stepwise" and "block" regression are similar to each other and are a modification of multiple regression. "Forward" and "backward" regression are related to time series analysis.

3. Verify that the data region is correct.

At this point, the simple regression routine will warn you that it has not been designed to deal with certain problems in the data. Just click "OK" and move on.

4. Select the variables to which the routine will be applied.

StatPro will give you two screens on which to select variables. The first screen allows you to select the dependent (response) variable. The second allows you to select the explanatory (independent) variable.

5. Fill in the details of the routine.

For simple regression, the routine will ask you which diagnostic graphs you want produced (the dialog box is shown below). Each of these graphs will be discussed later in the text. Usually, though, you will only need the first two options ("Fitted versus actual" and "Residuals versus fitted"). These two scatterplots allow you to determine whether the model, in this case a linear model, is a good choice for your data. Explanations of these graphs appear in a later chapter; for now, we'll rely on some other tools.

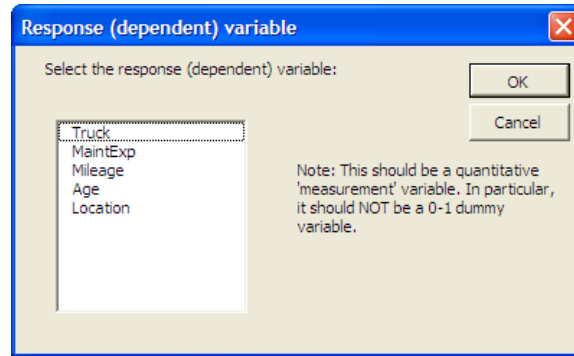


Figure 16.1: Selecting the variables in the regression routine of StatPro.

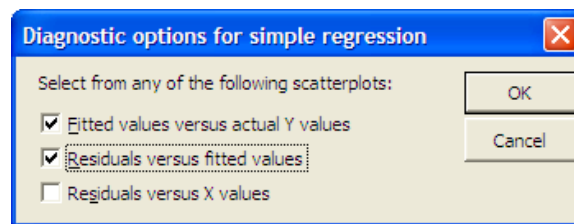


Figure 16.2: Diagnostic options that are useful when running regression.

6. Select the placement for the output of the routine.

For regression output, we definitely suggest putting the output on its own sheet. Name the sheet so that you can tell which model it contains (in case there are several different explanatory variables to choose from).

You will also notice that the simple regression routine will add two new columns of data to your worksheet. One column contains the predicted y -values (referred to as "fitted values") and the other column contains the residuals for each observation.

Now that you have the regression coefficients, you can easily create Excel formulas to compute the predicted y -values, based on given x -values. This will let you explore the model's predictive power.

16.2 Linear Regression in Excel (without StatPro)

If you are ever trying to do linear regression and you do not have StatPro available, there are several options available. One would be to use the actual formulas for the regression coefficients and the summary measures to compute the slope, y -intercept, coefficient of determination and standard error of estimate. This would be a little tough. Fortunately, there are formulas already built in to Excel. The LINEST formula is used to ESTimate LINear trendlines. Let's use the backpack data from the examples and perform a linear regression on it to compute "Price" as a function of "Number of Books." This file has the x -data

(number of books) in cells C2:C31 and the y-data in cells A2:A31. The LINEST formula has the following syntax:

`=LINEST(known y values, known x values, const, stats)`

Const refers to whether you want to calculate the y-intercept (the constant) from the regression (make it TRUE) or whether to force it to be zero (FALSE). We'll usually use TRUE.

Stats is another true/false variable. It indicates whether to calculate and output the summary measures. We'll almost always want it to be TRUE.

However, before you type in the formula, you should know that the output of it will have ten (10) pieces of information. Obviously, we can't put ten different numbers in a single cell, so we have to enter the formula as an array calculation.

First, highlight a block of cells that is two columns wide by five columns high.

Now, type the formula

`=LINEST(A2:A31, C2:C31, TRUE, TRUE)`

and hit CTRL+SHIFT+ENTER. (If you hit enter, you will only get the first of the ten numbers; then you have to start over!) The output will then appear in a 5 row by 2 column grid with the information shown below. The most important information is shown in bold.

Slope	Y-intercept	Regression coefficients
S_e for slope	S_e for y-intercept	
R^2 for model	S_e for model	Summary measures
F	Df	ANOVA information
SS (reg)	SS (resid)	

For more information about the LINEST function, type "regression" into the help system. If you check the "See also" portion of the help information, you will find out about the TREND function which helps you calculate other values, based on a set of known x and y values. There is also a separate SLOPE function which computes just the slope of the regression line. It has the syntax SLOPE(known y values, known x values). Used with INTERCEPT(known y values, known x values), you can get both coefficients.

16.3 How the Fill Handle Works to Complete a sequence of numbers

In Excel, you may have used the fill handle to copy a formula down a column or across a row. Remember, the fill handle is the little dot in the lower right corner of the active cell or active cell region. The fill handle can also be used to fill in patterns in a sequence of numbers that you enter.

For example, suppose you want to generate a column of numbers 10, 20, 30, 40, on up to 300. It would be tedious to type these by hand. Excel can help! Start by typing 10 in cell A1, 20 in cell A2 and 30 in cell A3. Now highlight the cells (A1:A3). Click and drag the fill

handle all the way down the column until the little floating box that follows the cursor says "300". Release the mouse button and your list of numbers is filled in!

How does Excel know what you want? It uses simple linear regression to get the answer! When you highlight the first part of the pattern, it takes those numbers and treats them as the "known y values". Then, it assumes that the x -values are 1, 2, 3, 4... and fills in the linear regression.

16.4 Multiple regression in StatPro

StatPro makes multiple regression analysis fairly easy. Follow the steps below with any set of data to develop a regression model along with its associated measures and graphs.

1. Select the region of the worksheet that contains the data.
2. Select the StatPro routine to apply to the data.

The routine for simple regression is under "Regression/ Multiple..."

3. Verify that the data region is correct.
4. Select the variables to which the routine will apply.

StatPro will give you two screens on which to select variables. The first screen allows you to select the dependent (response) variable. The second allows you to select the explanatory (independent) variables. To select several explanatory variables, hold down the control (CTRL) key while you select them with the mouse.

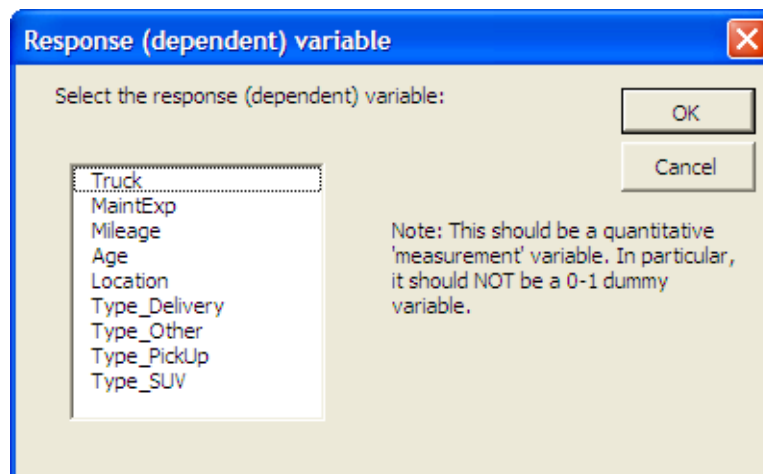


Figure 16.3: Selecting variables in multiple regression.

5. Fill in the details of the routine.

For multiple regression, the routine will ask you which diagnostic graphs you want produced (the dialog box is shown below). Usually, though, you will only need the

first and third graphs ("Fitted versus actual" and "Residuals versus fitted"). These two scatterplots allow you to determine whether the model, in this case a linear model, is a good choice for your data. You may also want to have the regression routine calculate the predicted ("fitted") values and the residuals. If so, be sure to check the box at the bottom of the dialog box. Unlike simple regression, these are not calculated automatically. If you select the second and fourth options, StatPro will produce quite a few graphs; each choice creates one graph for each explanatory variable.

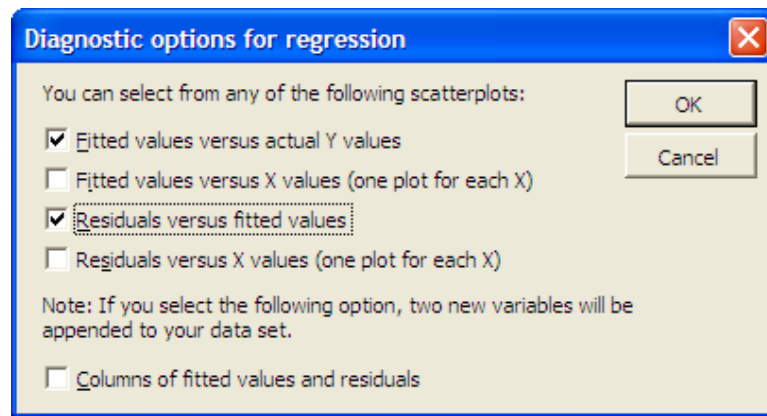


Figure 16.4: Selecting diagnostic graphs in multiple regression.

6. Select the placement for the output of the routine.

For regression output, we definitely suggest putting the output on its own sheet. Name the sheet so that you can tell which model it contains (in case there are several different explanatory variables to choose from).

Now that you have the regression coefficients, you can easily create Excel formula to compute the predicted y-values, based on given x-values. This will let you explore the model's predictive power. See below for how to do this most easily (to avoid re-typing the coefficients).

16.5 Stepwise regression in StatPro

Stepwise regression is very similar to performing multiple regression. In steps 1 - 4, nothing should be changed, except to use "StatPro/ Regression Analysis... /Stepwise..." Also, in step 4, select all of the explanatory variables that you think might be important. Usually, all of the non-identifier variables are selected. In step 5, there will be two more screens to fill in. For right now, leave everything as is and just click "OK".

For this routine, definitely have the output placed on a new worksheet. We advise that you call the new sheet "Step" so that you know it is the stepwise regression model. When the routine is finished (it may take a little longer than the multiple regression routine) you will have a lot more information. Basically, the routine will make several multiple variable models, carefully adding one explanatory variable (from the list you chose) at a time and

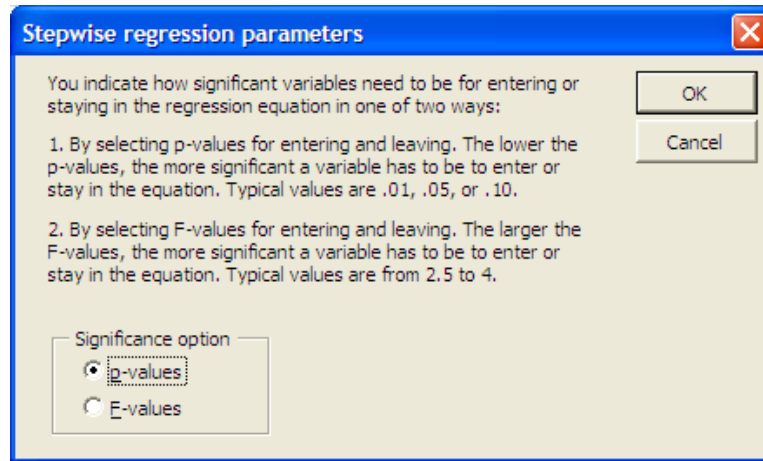


Figure 16.5: Selecting the p-value option for stepwise regression in StatPro.

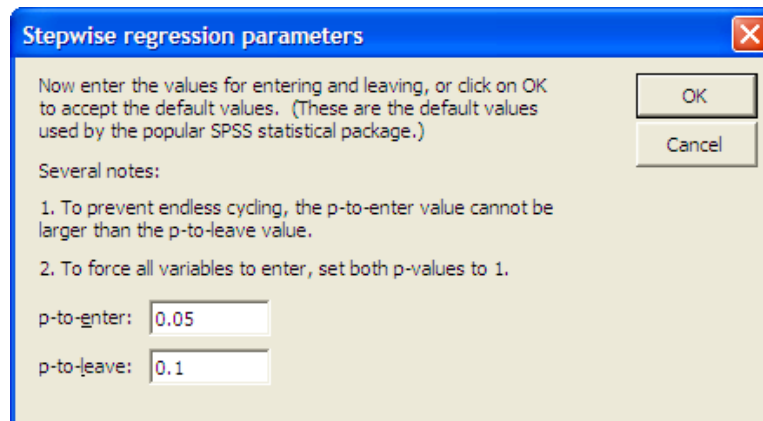


Figure 16.6: Setting significance for p-values in stepwise regression.

seeing how it affects the summary measures and the p-values. Scroll down the output screen to see how each variable was added. The last step will show you the best model the routine could develop, based on the variables you gave the routine and the parameters you set in step 5 to determine when to throw variables out.

16.6 Easily predicting values from a multiple linear model

For this example, suppose you have values of the variables on the data sheet in cells C42:F42 and suppose the regression output is on a worksheet called "FullModel" with the coefficients in cells C16:C20 (the constant is in C16). The following formula will calculate the value of the response variable that corresponds to the values of the explanatory variables in C42:F42:

$$= \text{Constant Cell} + \text{MMULT}(\text{X values}, \text{Coefficients})$$

So for our setup, use "=FullModel!C16 + MMULT(C42:F42, FullModel!C17:C20)".

Be sure that you have the variables in the same order in both places. Also be sure that the X-values are listed first in the model. The x-values must also be listed in a row (not a column) for the formula to work. If the x-values are in a column (like the coefficients) use the formula

$$= \text{Constant Cell} + \text{SUMPRODUCT}(\text{X values}, \text{Coefficients})$$

Also remember that if you want to calculate values for several sets of x-values by copying the formula to other cells, be sure that the cell references to the coefficients are absolute and not relative (the reference to the constant term must also be absolute).

16.7 Multiple regression with dummy variables

This is just like normal multiple regression with the following modifications:

1. First create all the dummy variables that you want to use in the model.
2. When selecting explanatory variables, leave one dummy variable out from each categorical variable. Otherwise, you will get the "exact multicollinearity" error.
3. When selecting explanatory variables, be sure to not use both the original categorical variable and the dummy variables.

Chapter 17

Creating New Variables From Old

17.1 Generating dummy variables in StatPro

StatPro has a variety of tools available to manipulate your data. Under "Data utilities/ Create dummy variables..." you can convert either a categorical variable into several dummy variables (one for each category) or a numerical variable into a single dummy variable, based on a cutoff value.

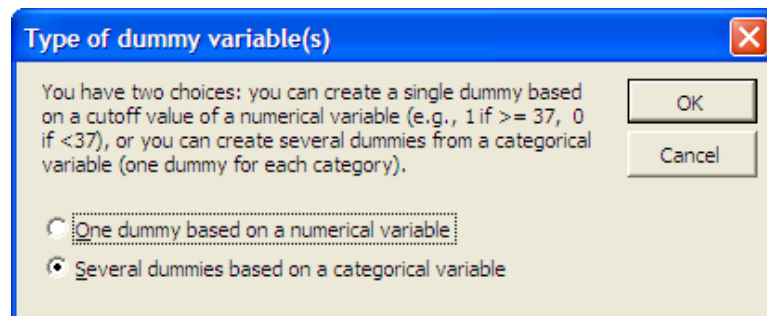


Figure 17.1: Two options for making dummy variables in StatPro. Usually you will want to choose to make several dummy variables from a single categorical variable (the second choice, shown selected.)

17.1.1 One dummy based on a numerical variable

This option will not be used too often. After you select a variable (pick a numerical one or a categorical variable that is coded with numbers, like "Education level" or "Job Grade"), you will need to determine the definition of the dummy variable. Simply enter a cutoff value and select a criterion. The name of the new variable will be a description of the cutoff criterion added to the end of the original variable name. The new variable will be added to the data set at the far right in a new column.

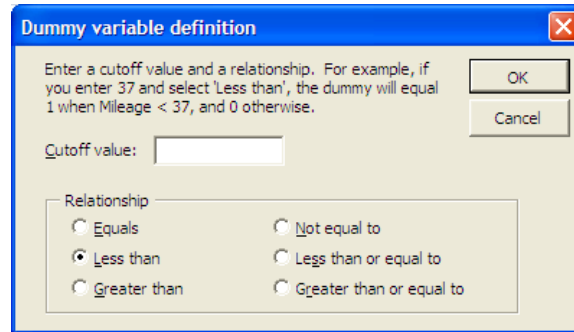


Figure 17.2: Dialog box for choosing how to construct a single dummy variable from a numerical variable.

17.1.2 Several dummy variables from a categorical

This will be the option you use most often. Once you select a categorical variable, the routine automatically creates several dummy variables from it at the far right of your data set. Each is named by adding a category to the end of the name of the original categorical variable. Thus, a "Gender" variable which was originally coded "Male" or "Female" would be turned into two new variables "GenderMale" and "GenderFemale".

17.2 Dummy variables with IF functions in Excel

If you do not have StatPro available, it is easy to create a formula to make a dummy variable. Suppose you have a variable "Gender" coded as "Male" or "Female" in column B2:B40, with the variable name "Gender" in cell B1. Let's say you want to create a "GenderMale" variable in column F. First type the variable name "GenderMale" in cell F1. Then, in cell F2, enter the formula

```
=IF(B2='Male', 1, 0)
```

After the formula is entered, select the cell and double-click the fill handle to copy the formula down the column. You can enter whatever category name you want in the formula in place of "Male", but be certain that you type it exactly as it is coded in the original categorical variable.

The IF function has the following syntax.

```
IF(condition, value if true, value if false)
```

The condition can be any sort of logical condition and can include checking for whether a cell is equal to a particular value, greater than a particular value, or whatever. See the help files on "IF" for more information.

17.3 Creating Interaction Variables

17.3.1 Creating interaction terms with StatPro

1. Go to StatPro/Data utilities/Create Interaction Variable(s)
2. Click past the next window
3. Select the two variables you wish to interact. Remember to hold the control key in order to select variables that are not listed next to each other on the list. Click OK.

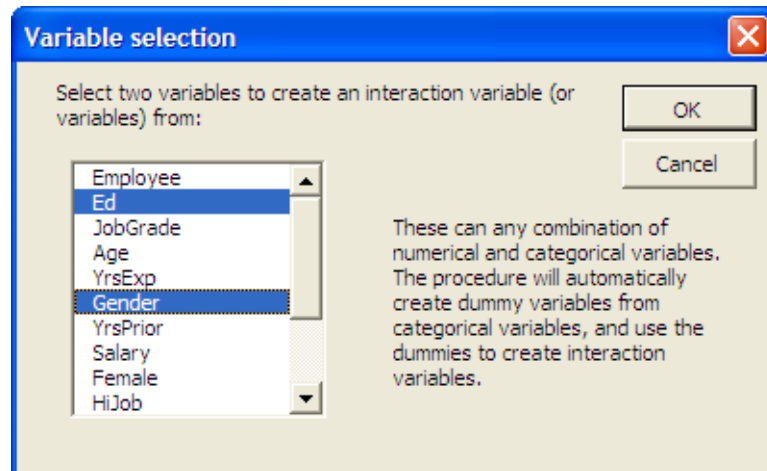


Figure 17.3: Selecting variables to create interaction terms with StatPro.

4. Check any of the two variables that happen to be categorical variables. Do NOT check numerical variables, even if they were originally created from categorical variables. Click OK.

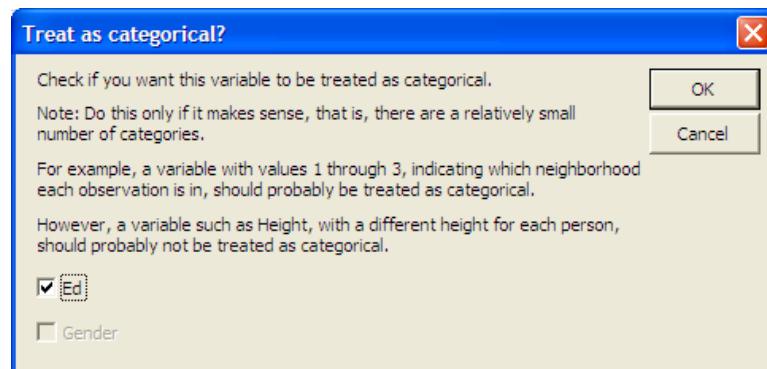


Figure 17.4: Ensuring that StatPro treats the Education Level (Ed) as a categorical variable.

You can see a portion below of the interaction terms that StatPro has created.

	P	Q	R	S	T	U	V
3	Ed_1*Gender_Female	Ed_1*Gender_Male	Ed_2*Gender_Female	Ed_2*Gender_Male	Ed_3*Gender_Female	Ed_3*Gender_Male	Ed_4*Gender_Female
4	0	0	0	0	0	1	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	0	1	0	0	0	0
8	0	0	0	0	0	1	0
9	0	0	0	0	1	0	0
10	0	0	0	0	1	0	0

Figure 17.5: A portion of the new interaction terms created by StatPro after the above steps.

17.3.2 Interacting a variable with itself

StatPro will let you create a new interaction variable from two different variables; however, it will not allow you to interact a variable with itself. This is easy to do manually with a small formula, though. First, you create a new column for the interaction variable. To remind yourself what it represents, you might call it "Variable2" where "variable" is replaced with the name of the base variable. Second, you enter a formula in the first cell of the new column that computes the product of the variable with itself. Finally, you copy this formula down the column.

In the spreadsheet shown below (the original file without the interaction terms is "C13 Rail System.xls") you see that we have added a column for the new variable "Park2" which will be the interaction of the variable ParkingRate with itself (column G). Next, in cell G4 we entered $=F4*F4$ and then copied this down column G by double-clicking on the fill handle for cell G4.

	A	B	C	D	E	F	G
3	Year	Weekly_Riders	Price_per_Ride	Population	Income	Parking_Rate	Park2
4	1966	1,190	\$0.15	1,800	\$2,900	\$0.50	0.25
5	1967	1,135	\$0.15	1,790	\$3,100	\$0.50	0.25
6	1968	1,161	\$0.15	1,780	\$3,200	\$0.60	0.36
7	1969	1,165	\$0.25	1,778	\$3,250	\$0.60	0.36
8	1970	1,171	\$0.25	1,750	\$3,275	\$0.60	0.36
9	1971	1,136	\$0.25	1,740	\$3,290	\$0.70	0.49

Figure 17.6: Numerical-on-numerical interaction terms in the rail system data.

17.3.3 Pitfalls of numerical-numerical interaction variables

When creating models with numerical-numerical interaction variables, you are much more likely to encounter an error. The most common is the "multiple colinearity" error. In well-designed and collected data sets with several numerical variables, it is quite likely that there are many hidden relationships among the different variables. For example, if you attempt to predict car maintenance costs based on the two variables of age and mileage, we expect that older cars also have more mileage. Thus, the two independent variables are not truly independent.

17.4 Transforming Data

17.4.1 Using StatPro's Data Transformation Utility

Often it will be necessary to create models of the data that are not linear models. In order to do this using regression analysis, we must first transform the data with a nonlinear function (using StatPro, or typing the formulas into Excel ourselves) and then perform regression using these new, transformed variables as the data.

To have StatPro transform the data, follow these steps:

- Select the data (click anywhere in the data)
- Open the "StatPro" menu on the toolbar
- Select "Data Utilities"
- Select "Transform"
- Choose an appropriate transformation, based on the model you need (see below)

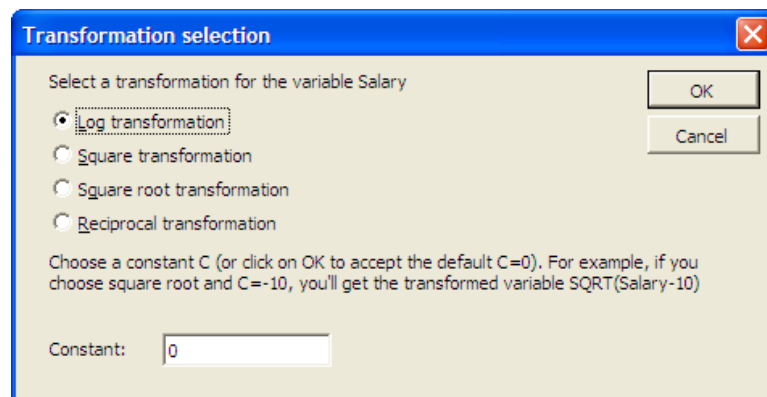


Figure 17.7: Dialog box for transforming a variable in StatPro.

There are four basic transformations available in StatPro (shown in the screen shot above). These can be combined to produce a wide variety of models, shown in the chart below, under "What transformations to use". Also note that the data transformation tool will allow you to shift the data by a known constant, opening up the number of different types of models you can produce to be much larger than your patience would allow you to investigate.

17.4.2 What transformations to use

Okay, now that you know how to use the transformation tool in StatPro, you need to know what variables to transform for each of the nonlinear models that you might encounter. Consult the table below for the type of model you want, and what variables you need to

create in order to produce a regression model of that type. We assume that the original data contains two variables, the response variable and the explanatory variable, and we assume that you are trying to develop a model equation to predict the response as a function of the explanatory.

To create this type of model	Response Variable	Explanatory Variable(s)
Linear model	Response	Explanatory
Square model	Response	Sqr(Explanatory)
Quadratic model	Response	BOTH Explanatory and Sqr(Explanatory)
Logarithmic model	Response	Log(Explanatory)
Exponential model	Log(Response)	Explanatory
General Power model	Log(Response)	Log(Explanatory)
Multivariate Power Model (multiplicative model)	Log(Response)	Log(Explanatory 1), Log(Explanatory 2), etc.
Reciprocal model	Response	Reciprocal(Explanatory)
Square Root model	Response	SqRt(Explanatory)

17.4.3 Transformations without StatPro

It is easily possible to construct these transformations without StatPro. Suppose you have a variable "X" extending from cell B1 to cell B51, with a label for the variable name in B1. In column C we will construct the natural log of this data, using the built-in Excel formula for the natural log: LN. Thus, the natural log of the observation in B2 is "=LN(B2)". Then fill down formula down to all the cells in column C.

To compute the square of a number, use the caret symbol (^) and the power 2 in the formula. Thus, in column D we can construct the square of the data "=B2^2".

To compute the square root, use the formula "SQRT". To compute the reciprocal, use the fact that the reciprocal of X is just 1/X.

Chapter 18

Using a VLOOKUP table

In doing some tasks, we find that we need some way to use different information depending on the result of some number. For example, in calculating employee pay, different job types might have different, standardized pay rates at our company. Wouldn't it be nice if Excel could figure it out from the information given and calculate the pay rate correctly? Using a lookup table, in this case a VLOOKUP table, Excel can.

If you open the file "C10 HowTo.xls" you'll see an example. Shown below is an image of the screen illustrating a sample employee database. This database contains information on each employee: hours worked that week, job type, and years of experience.

	A	B	C	D	E	F	G	H	I
1	Employee	Hours	JobType	YrsExp	GrossPay				
2	1	36	5	9	\$ 1,184.40		JobType	BasePay	Raise
3	2	44	2	12	\$ 741.40		1	\$ 6.00	\$ 0.55
4	3	20	4	13	\$ 562.00		2	\$ 7.25	\$ 0.80
5	4	16	4	5	\$ 296.00		3	\$ 8.25	\$ 0.95
6	5	40	5	11	\$ 1,484.00		4	\$ 12.50	\$ 1.20
7	6	36	3	13	\$ 741.60		5	\$ 14.00	\$ 2.10
8	7	38	1	12	\$ 478.80				
9	8	42	3	2	\$ 426.30				
10	9	35	3	11	\$ 654.50				
11	10	27	2	11	\$ 433.35				
12	11	5	2	1	\$ 40.25				
13	12	18	1	9	\$ 197.10				
14	13	40	4	3	\$ 644.00				
15	14	37	1	13	\$ 486.55				
16	15	41	3	11	\$ 766.70				

Figure 18.1: Employee database illustrating use of VLOOKUP tables.

Off to the right of database, in cells G2:I7 is the lookup table. (Normally, one would put this on a different sheet of the workbook and name the entire range of cells to make it easier to reference, but for this example, we wanted to keep it easy to visualize.) Now we want Excel to take the employees hours and multiply it by the correct hourly rate, based on the job type and the years of experience. This hourly pay rate will be something like

$$(\text{Base Pay Rate}) + (\text{Years Experience}) * (\text{Annual Raise})$$

But Excel must use the Job Type to determine both the base pay rate and annual raise. To do this, we use VLOOKUP:

`=VLOOKUP(Lookup Value, Lookup range, Column, [range lookup])`

So, we can find the base hourly rate for employee 1 by looking up his/her job type (cell C2) in the lookup table (\$G\$3:\$I\$7 - the absolute reference is a MUST here!) and using the information in column 2 of the table. To find the annual raise, we perform the same lookup, but instead of returning the information in column 2, we want the information in column 3. Thus, we can compute employee 1's pay by the following formula (shown in text and Excel notation to make it easier to read).

$\text{Pay} = (\text{Hours Worked}) * ((\text{Base Pay Rate}) + (\text{Years Experience}) * (\text{Annual Raise}))$

`E2 = B2*(VLOOKUP(C2,$G$3:$I$7,2) + D2*VLOOKUP(C2,$G$3:$I$7,3))`

Copying this formula to the cells in E3:E16 will compute each employee's pay, using the correct job type to calculate the pay rate. One could also use this to calculate the taxes based on the number of dependents declared on W4 forms, or practically anything.

IMPORTANT TIP: Lookup tables must be organized a certain way. Excel always uses the leftmost column of the table to match with the LookupValue in the formula, so be sure this is the way it is organized. It is also vital that the table be sorted in ascending order by the first column. If it is not sorted, Excel cannot find the proper match, and you will see an error in the calculation.

NICE FEATURE: Lookup tables don't have to return numbers; they can return any type of data. And, they don't require an exact match. If you have a range of possible values that should return a certain result, then just put the lower end of each range in the left column.

Chapter 19

Computing Values of Exponentials and Logarithms

Excel uses a standard notation to compute the exponential or logarithm of a number. The notation looks a lot like the notation we have been using above:

- To compute the value of e^3 , type `"=EXP(3)"` in a cell and hit enter.
- To get the value of e raised to whatever is in cell B2, type `"=EXP(B2)"`
- To compute the natural logarithm of 3, type `"=LN(3)"`
- To compute the natural log of the number in cell B2, type `"=LN(B2)"`

Note that Excel (and most calculating tools) have another logarithm function. This is the LOG(x) function. There is a slight difference between LOG(x) and LN(x). For our purposes, we will always use LN(x) when we talk about the logarithm of x .

Technical details: LOG(x) stands for the base-10 logarithm of x . LN(x) stands for the base- e logarithm of x . Essentially, when we compute a base- b logarithm of the number x we are finding the value of a so that the following equation is true: $b^a = x$. For example, since $10^2 = 100$, we know that the base-10 logarithm of 100 is 2 (i.e., $\log_{10}(100) = 2$.) Since $2^5 = 32$ we know that $\log_2(32) = 5$. Excel really only has options for base- e logarithms (LN) and base-10 logarithms (LOG). There are many other useful logarithm bases, but these are the most common, and there is a mathematical technique that relates logarithms of any two bases: $\log_b x = \frac{\ln(x)}{\ln(b)}$.

Chapter 20

Setting up functions in Excel for shifting and Scaling

Previously, we introduced the idea of setting up an Excel spreadsheet to calculate a table of function values. We can use this same idea for calculating the values of a function with arbitrary shifts (horizontal and vertical) and scalings. For example, suppose we want to fit a shifted curve to a set of data that has x -values in cells A6:A25 and y -values in cells B6:B25. Let's add in some parameter values. Enter labels for each shift in A1:A3 and sample values for these shifts in B1:B3. You now have a worksheet that looks something like the one at the right.

Now we need to add in a column of values for the predicted y data, according to our formula, using the shifts and scales. Suppose that we want to use a logarithmic function to try and fit the data. So, we want to try to use the formula

$$y = (\text{Vertical shift}) + (\text{Vertical Scale}) * \ln(X + \text{Horizontal shift})$$

To do this, we enter the following formula into cell C6 and copy it down the column (Note: This is a formula for the logarithmic model we are currently working with; for other models, you will have to develop a different formula):

$$= \$B\$1 + \$B\$3 * \ln(A6 + \$B\$2)$$

Notice that we are using absolute cell references to look up the values of the parameters and compute the predicted y -values. This way, the constants will remain correct as we copy the formula down, but the x -values will change, based on which row we are in. This format will easily allow us to change the shifts and scales to try and match the actual data (in column B). A visual representation (a scatterplot) would also help, since the graph would help you see the shifts and try to move the graph of the predicted y -values closer to the actual data.

	A	B
1	Vertical Shift	1
2	Horizontal Shift	2
3	Vertical Scaling	-3
4		
5	X	Y
6	0	5.841117
7	1	3.408326
8	2	1.682234
9	3	0.343373
10	4	-0.75056
11	5	-1.67546
12	6	-2.47665
13	7	-3.18335
14	8	-3.81551
15	9	-4.38737
16	10	-4.90944
17	11	-5.3897
18	12	-5.83434
19	13	-6.2483
20	14	-6.63553
21	15	-6.99928
22	16	-7.34223
23	17	-7.66663
24	18	-7.97439
25	19	-8.26713
26	20	-8.54625

Figure 20.1: Setting up a function with vertical and horizontal shifts, and a scaling factor.

Chapter 21

Calculating R^2 for your fit of the data

Now, it can be very hard to see if your shifts and scales have done a good job at fitting the data, so it is helpful to add in computations to determine the value of R^2 the coefficient of determination. For this, we'll need to compute the mean of the actual y -data, the variation in the actual y -data, the residuals, and finally the R^2 .

1. Compute the mean of the actual y -data.

A good place to put the mean of the actual y -data is to put it at the bottom of the column of y -data, separated from the data by a blank row. So, in cell A28 we place a label "Y Mean" and in cell B28, we place the formula "`=average(B6:B26)`" to compute the mean of the y -data.

2. Compute the variation in the actual y -data.

The variation in the y -data is the sum of the squares of the differences between each y -value and the mean of the y -data. So, we add a column to the right of the data to compute the square of these differences. In cell D5, place the label " $(y-\text{mean})^2$ ". In cell D6, we place a formula to calculate the squared difference: "`=(B6-$B$28)^2`" and then copy this formula to the other cells in the column. At the bottom of the column, in cell D28, we compute the sum of these squared differences: "`=sum(D6:D26)`". This sum of the squared differences is the total variation in the y -data.

3. Compute the residuals (and the sum of their squares).

We now do a similar calculation in column E. We need the squares of the differences between the actual y -data and the predicted y -data. In cell E5 we place the label " Residual^2 ". In cell E6, we put "`=(B6-C6)^2`" and copy this formula to the rest of column E. At the bottom of this column, in cell E28, we sum up these squared residuals: "`=sum(E6:E26)`".

4. Compute R^2 .

The value of R^2 can be calculated by subtracting the ratio of the sum of the squares of the residuals and the total variation in the y -data. So, in cell D2, we place the label " R^2 " and in cell E2 we enter the formula "`=1 - E28/D28`". This will probably result in

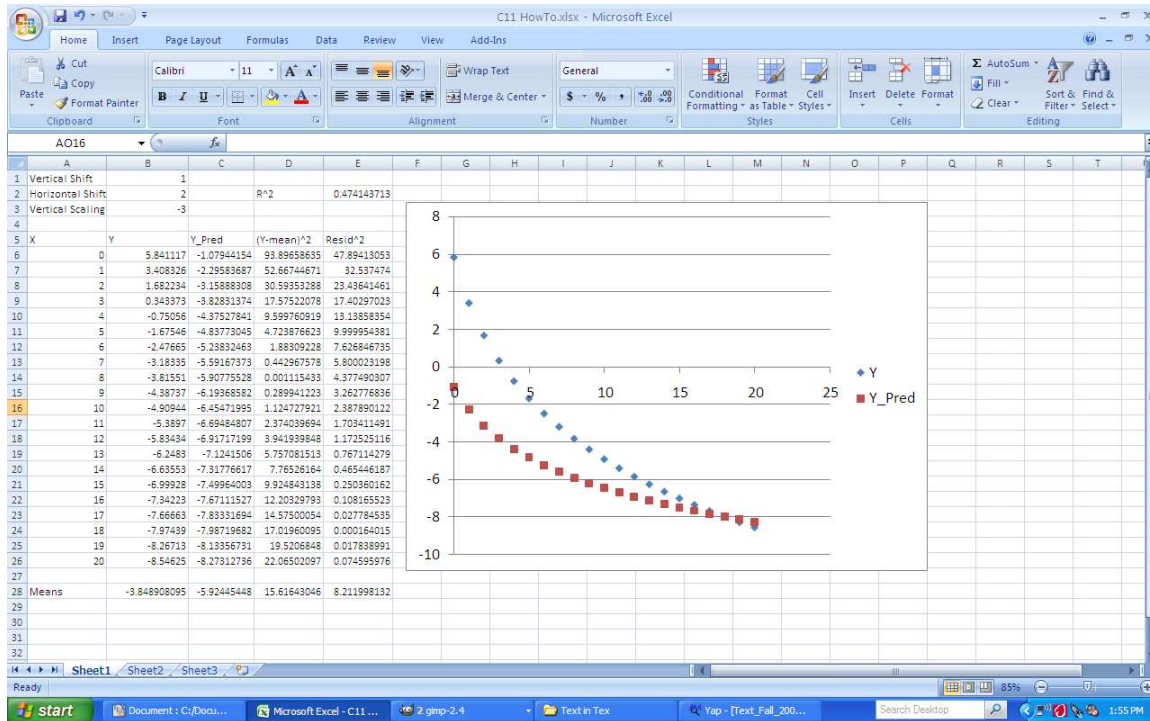


Figure 21.1: Spreadsheet for using parameters to attempt finding a best fit to a set of data.

an R^2 value that is not between 0 and 1. This is because we have not actually created a "best fit" line until we do some adjustment on the parameters.

When you're done, the spreadsheet should look like figure 21.1 (you can add a graph in as well, by selecting the first three columns of data and making a scatterplot of them).

Chapter 22

Using Goal Seek to find the best values for the shifts and scales

Once you have the spreadsheet configured to compute a set of y values based on your parameters for the shifts and scales, and have added the calculation for R^2 into the sheet to see how well your shifted and scaled curve fits the data, you can try to find the best values of the shifts and scales in one of two ways: trial and error or using goal seek. Trial and error is tedious, and can be difficult, if the data is tightly packed. Often, changing a value doesn't produce any noticeable change in the graph, making this process even harder. We recommend that you start by trial and error to get roughly the right values of the parameters, then refine the values with goal seek. (Goal seek was introduced in the computer how-to for chapter 7B.)

Suppose your spreadsheet is set up in the following way: You have values for three parameters (horizontal shift, vertical shift, and scale) in cells B1, B2 and B3. You have a calculated value of R^2 from the data and your best fit in cell E2. Now, there's no guarantee that the data can be "perfectly fit" so we can't try to goal seek for $R^2 = 1$. We also don't know ahead of time what the "best R^2 " is, so we have to work our way up to it. Select the cell with the R^2 value in it. Then activate goal seek with "Tools/ Goal Seek". In the blank for "To value" enter a number that is a little better than the current value of R^2 . For example, if the current value is 0.90, try entering 0.91. Now, the "By changing" cell should be one of the three parameters. Do this systematically: goal seek by changing the first parameter (B1) then repeat this, changing B2, then repeat changing B3, and start again at B1. Eventually, you'll get the R^2 up close to the "best possible" value. Each time, try increasing the "To value" amount a little, say 0.01 at a time. At some point, you'll know you're close to the best when the routine spits out "Goal seek may not have found a solution." When this happens, goal seek again with the same target value, but using the other parameters.

Chapter 23

Computing S_e and R^2 for nonlinear models

This is appropriate only for models in which the y variable (the response variable or dependent variable) is transformed. For example, if the model equation is of the form

$$\begin{aligned}\ln(y) &= A + Bx \\ e^{\ln(y)} &= e^{A+Bx} \\ y &= e^A e^{Bx}\end{aligned}$$

you will need to complete the process below in order to determine the actual S_e and R^2 of the model, since the summary statistics from the regression output are not based on the actual y variable at all. Thus, you only need to go through these steps when producing exponential models (like the example above), power function models, or multiplicative models.

This process will be broken down into several steps, listed in order below. A sample worksheet with all the formulas listed is shown as well, be sure you complete the steps in the proper order. The numbers on the sheet show the order in which the columns and formulas should be constructed.

Recall that for exponential and multiplicative (power) regression models, the original response variable Y has been transformed by the log function. This means that the standard error of estimate for the model cannot be used to interpret how accurately the model fits the original Y data. The following activity will guide you through the process of finding a standard error of estimate that conforms with the original Y data.

Bring up the data found in C12 Power.xls. We will fit an exponential model to this data and then compute the that is consistent with the original cost data. After we have S_e , we can compute the R^2 .

1. The exponential model can be constructed with the following steps
 - StatPro
 - Data Utilities
 - Transform Variable(s)

- Select "cost" to transform
- Select Log transformation

	A	B	C	D
1	Data on cost versus production level			
2				
3	Month	Cost	Units	Log(Cost)
4	1	45623	601	10.72817
5	2	46507	738	10.74736

2. Perform regression using Log Cost as the response variable. Be sure to check off the box to have the routine create columns for "fitted values" and "residuals". If you do not do this, start step (2) over. We must have the fitted values in order to complete this process.

The Standard Error of Estimate of the model with response variable Log(Cost) cannot be directly used in interpreting how accurately the model fits the original Cost data. StatPro will insert a blank column (E) between the calculations. In order to remain consistent with the screen images and formulas below, you should right click on column E (in the original data) and select "Delete" from the context menu that appears.

Your data sheet now looks like this:

	A	B	C	D	E	F
1	Data on cost versus production level					
2						
3	Month	Cost	Units	Log(Cost)	Fitted Values	Residuals
4	1	45623	601	10.72817	10.6373	0.0908

3. Next, we transform the Fitted Values produced by the model, which approximate the actual Log(Cost) values, back into the original cost units by applying the Antilog function (exp) to the fitted values with Exp(Fitted Values). Thus, in cell G4, type the formula " $=EXP(E4)$ ". Then fill this formula down to copy it to all the data.
4. Create the columns indicated below, in order, from left to right. In this case, remember that (a) the "Actual" values are the cost data. Column H "Actual minus EXP(Fitted)" is the residual data that is squared to produce the data in column I.

	E	F	G	H	I
1					
2					
3	Fitted Values	Residuals	Exp(Fit)	Resid	Resid ²
4	10.6373	0.0908	41661.43	3961.57	15694061.87
5	10.7463	0.0010	46458.42	48.58	2360.07
6	10.7049	-0.0280	44575.85	-1232.85	119923.07

5. We are ready to apply the formula for the standard deviation of the sum of the squares of the residuals (SSR):

$$\sqrt{\frac{\text{Sum of Squares of Residuals}}{n - p}} = \sqrt{\frac{\text{SSR}}{n - p}}$$

The bottom right corner of your data sheet should look something like this:

	H	I
37	2387.82	5701660364
38	-36.70	1346.99
39	2485.24	6176425.81
40		
41	Sum of Resid^2	296201375.38
42	Parameters (p)	2.00
43	Data Points	36.00
44	S_ E	2951.58

The number of parameters refers to the number of coefficients that your regression model was used to predict. In this case, there are two parameters. In general, the number of parameters for this step is the number of independent variables plus 1. This number is the p in the formula for standard error. The number of data points is the actual number of observations contained in the data. This is the n in the formula for standard error above.

Thus, about 2/3 of the actual cost data will fall within \$2951.58 of the cost predicted by the exponential model.

To compute the R^2 value, we need to first compute the mean of the original y data, in this case, the cost data. It's also useful to have the standard deviation of the y data in order to compare the and determine how accurate the model really is. Once we have the actual mean of the y -data, we need to create another column to the right of our previous calculations. This column should compute the square of the actual values minus the mean of the y data. Thus, if you have the mean of the y -data in cell B41, the formula for the first entry in column J (J4) will be " $= (B4 - \$B\$41)^2$ ". Fill this formula down the column to get all the data.

Next, compute the total variation, the sum of all the results in column K. We have placed this information in cell J46. To get R^2 , we then use the formula for R^2 :

$$R^2 = 1 - \frac{\text{SSR}}{\text{Total Variation}}$$

Thus, in cell J47, we enter " $= 1 - I41/J46$ " to get the actual R^2 for this model.

NOTE: The following figures show a final result of these calculations, all on one spreadsheet, with all the formulas written out and explained. However, the formulas are different than this example. The reason is because in the example above, we have deleted

the blank column E that StatPro inserts during the regression. In the image below, it has not been deleted.

CELL	Formula	CELL	Formula
B41	=average(B4:B39)	J44	=sqrt(J41/(J43-J42))
B42	=stdev(B4:B39)	K46	=sum(K4:K39)
J41	=sum(J4:J39)	K47	=1-J41/K46

COLUMN	Information
D	Created by "Data Utilities/Transform Variables" - This is the Log of the COST Variable.
F and G	Created by the Multiple Regression routine ONLY IF you remember to click the last check box in the diagnostic options during the regression setup. The "Fitted Values" are fitted LOG(Cost) values. The residuals are the difference between Log(Cost) and Fitted - not what we need.
H	These are the model's actual predictions for the Cost. To get these, we need to "unlog" the Fitted Values Column. We do this with the exponential function, which is EXP in Excel. Thus, in cell H4 we enter =EXP(F4) and copy this down the H column.
I	This column shows the real residuals - the difference between Cost and Exp(Fit). In I4 type =B4-H4 and copy this to the whole column.
J	This column computes the square of each of the residuals. Simply enter the formula =I4^2 and copy this down the column.
K	This column computes the square of the difference between each data point's Y-value (COST) and the mean of all the actual Y-Values. This is the variation of the data point. To compute this, in cell K4 enter =(B4-\$B\$41)^2 and fill down the column.

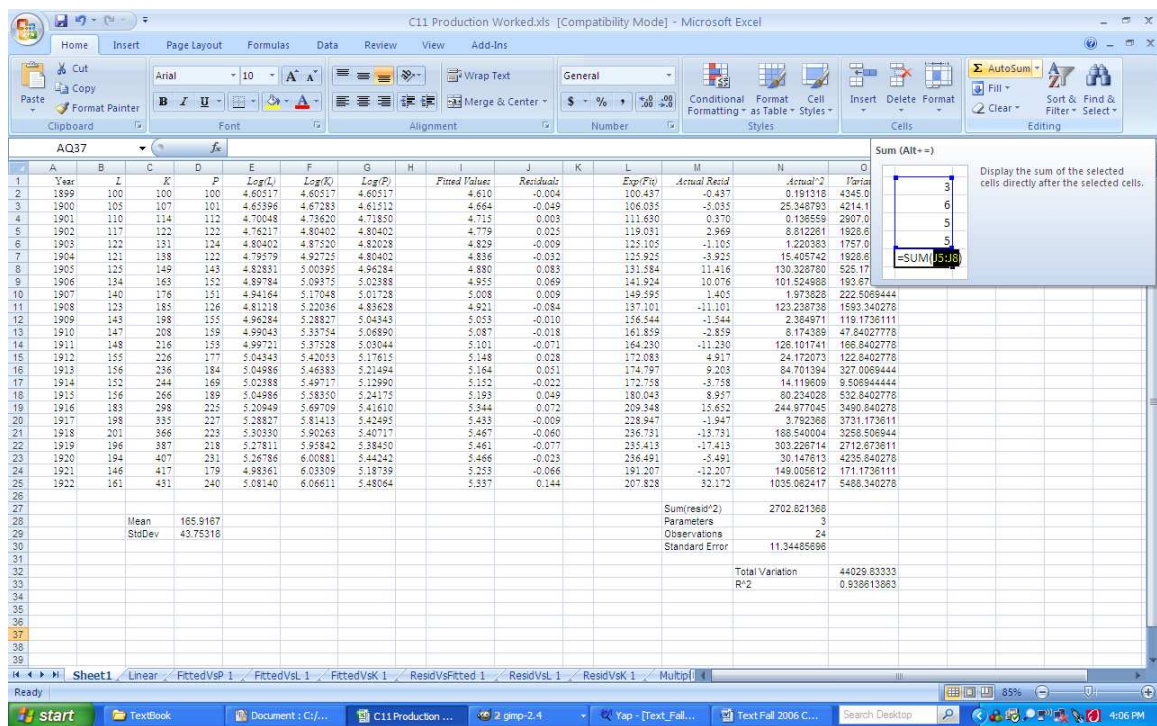


Figure 23.1: Screen showing the set up for calculating R^2 and S_e for nonlinear models.

Chapter 24

Creating a column-oriented, one-variable data table

Excel makes it easy to create tables of data from formulas using a data table. Essentially, you create a column of values you want to substitute into the formula, and enter a sample computation of the formula. Then you tell Excel which cell is the input cell and it substitutes all your values into the formula at once, producing a table of outputs corresponding to each input.

	A	B	C	D
1	A	2	-1	
2	B	2	2	
3	C	6	-3	
4				
5	X	4		
6				
7	X	46	-11	
8	1			
9	2			
10	3			
11	4			
12	5			
13	6			
14	7			
15	8			
16	9			
17	10			

Figure 24.1: Setting up a column-oriented, one variable data table.

Suppose you want to make a table of data so that you can compare two functions. If your functions are both quadratic, we might set up the spreadsheet as shown below. In cells B1:C3, we have entered two sets of A , B , and C values for a quadratic function of the form

$$y = Ax^2 + Bx + C$$

In cell B5 we have entered a sample x value. This x value will be used to generate sample calculations for the data table in cells B7 and C7. These cells contain the calculation of the functions, using the formulas

$$B7 = B1*B5*B5+B2*B5+B3$$

$$C7 = C1*B5*B5+C2*B5+C3$$

Once you have the sample calculations, you then create a list of the values you want to substitute into the formulas for x (the sample x in cell B5 will be replaced by the x values in your list). In the screen shot in figure 24.1, these values are listed in cells A8:A17.

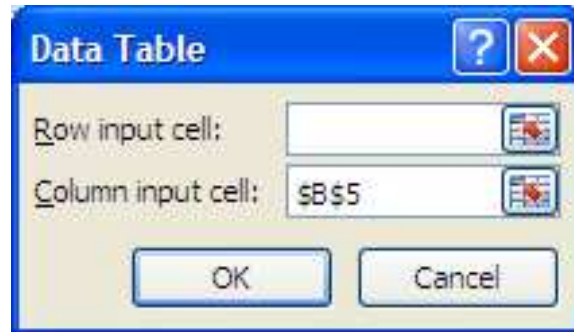


Figure 24.2: Dialog box for data tables.

To complete the data table, highlight the region of cells containing your list of values and the sample calculations, in this case A7:C17. Activate the data ribbon, click on "What if analysis" and select Data Table on the menu and you will see a dialog box like the one shown in figure 24.2.

Since your table is set up as a column-input table, in the "Column input cell" box enter a reference to the cell with the sample value of x used in your formulas. In this case, the sample value is in B5, so enter that. When you click "OK" the table will automatically compute the result of substituting values into the formulas.

FINAL NOTE: If you select any cell of the data table, you will not see a calculation or a value. Instead, you will see something like `=TABLE(,B5)`. This is because the table is live, so changing the cells in B1:C3 or the input cells in A8:A13 or the formulas themselves in B7 or C7 will instantly propagate through the table. To freeze the values, first select the table, then copy it, and instead of pasting it, use "Edit/ Paste Special". Selecting "Values" from the list will copy just the resulting values and not the formulas.

Chapter 25

Using mixed cell references to compute a table of function values

To compute a table of values for a function of two variables, say $f(x, y)$, it is very efficient to use mixed cell references. These are cell references with either the column or the row fixed (by placing a \$ in front of it) but not both of them (that would be an absolute cell reference).

For example, to make a graph of $z = 4 + 2x - 3y + x^2 + 2y^2 - 5xy$, we first set up the spreadsheet with values of x across the first row (skipping the first cell) and values of y down the first column (skipping the first cell). So if we wanted to graph this function for values of x from -5 to 5 and values of y from -10 to 20, we might set the spreadsheet up as shown (see "C13 HowTo.xls" for the data and the sample graph; the labels for the "X Values" and "Y Values" were done using the "merge and center" feature). Note that the x values run left to right, and the y values run top to bottom. Also note that we have left the first cell blank; this is important in helping Excel "guess correctly" when it formats your graph (See "Making a 3D surface plot" in the next chapter).

D4		=4+2*D\$2-3*\$B4+D\$2^2+2*\$B4^2-5*D\$2*\$B4												
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1			X Values											
2			-5	-4	-3	-2	-1	0	1	2	3	4	5	
3	Y Values	-10	-1	42	87	134	183	234	287	342	399	458	519	
4		-8	-29	4	39	76	115	156	199	244	291	340	391	
5		-6	-41	-18	7	34	63	94	127	162	199	238	279	
6		-4	-37	-24	-9	8	27	48	71	96	123	152	183	
7		-2	-17	-14	-9	-2	7	18	31	46	63	82	103	
8		0	19	12	7	4	3	4	7	12	19	28	39	
9		2	71	54	39	26	15	6	-1	-6	-9	-10	-9	
10		4	139	112	87	64	43	24	7	-8	-21	-32	-41	
11		6	223	186	151	118	87	58	31	6	-17	-38	-57	
12		8	323	276	231	188	147	108	71	36	3	-28	-57	
13		10	439	382	327	274	223	174	127	82	39	-2	-41	
14		12	571	504	439	376	315	256	199	144	91	40	-9	
15		14	719	642	567	494	423	354	287	222	159	98	39	
16		16	883	796	711	628	547	468	391	316	243	172	103	
17		18	1063	966	871	778	687	598	511	426	343	262	183	
18		20	1259	1152	1047	944	843	744	647	552	459	368	279	

Figure 25.1: Table of values for a function of two variables.

The next step is to enter the formula, paying careful attention to the types of cell references we need. Every time we refer to the x value in our formula, we need to fix the row (like C\$2 in the formula highlighted above), since the x values are always in row 2, but will be in different columns depending on the specific x value we want to use. The y values are always in column B, so when we refer to a cell with a y value in it, we need to freeze the column (like \$B3 in the formula shown above). The full formula (in cell C3) is shown below to make it easier to read. Of course, we could have set up cells with parameters for the function instead of typing the 4, 2, -3, etc. This would have allowed us to easily see how changing these numbers changes the shape of the graph, but it would have made the formula harder to read for this example. In general, you should always use parameters, rather than "hard coding" the numbers into the formula.

$$B2 = 4+2*C\$2-3*\$B3+C\$2\hat{2}+2*\$B3\hat{2}-5*C\$2*\$B3$$

Once these references are correct, we copy the formula to all the other cells in the table, as shown. The result shows us the value of z for a given value of x (the column) and y (the row). Thus, when $x = -2$ and $y = 10$, we find that $z = 274$. This procedure will work with any type of function of two variables, linear or nonlinear.

Chapter 26

Adjusting security settings for macros in Excel

If you open a file with macros (for example, the sliders in exploration 13B) you will see one of three things happen. Either:

1. You will be asked if you want the macros to be active (medium security),
2. The macros will be automatically active and you will see nothing (low security), or
3. The macros will automatically be disabled (high security).

Macros are simply collections of instructions (a small program, basically) that have been connected to make them easier to run together, rather than having to repeat all of the commands each time you want to reproduce that set of actions. Macros are a common way to distribute computer viruses, so many recommend that you think carefully about enabling them all the time; we prefer to use medium security, so that we are asked before macros are enabled in a particular file.

To adjust your security settings, open the Excel Options menu by clicking on the Office button in the upper left corner of the Excel window and clicking on "Excel options" at the bottom. In the dialog box, click on the Trust Center (see figure 26.1) and click the "Trust center settings" button in the lower right corner. Select the level of security you wish to have from the screen in figure 26.2, then click "OK" and "OK" again to apply these settings.

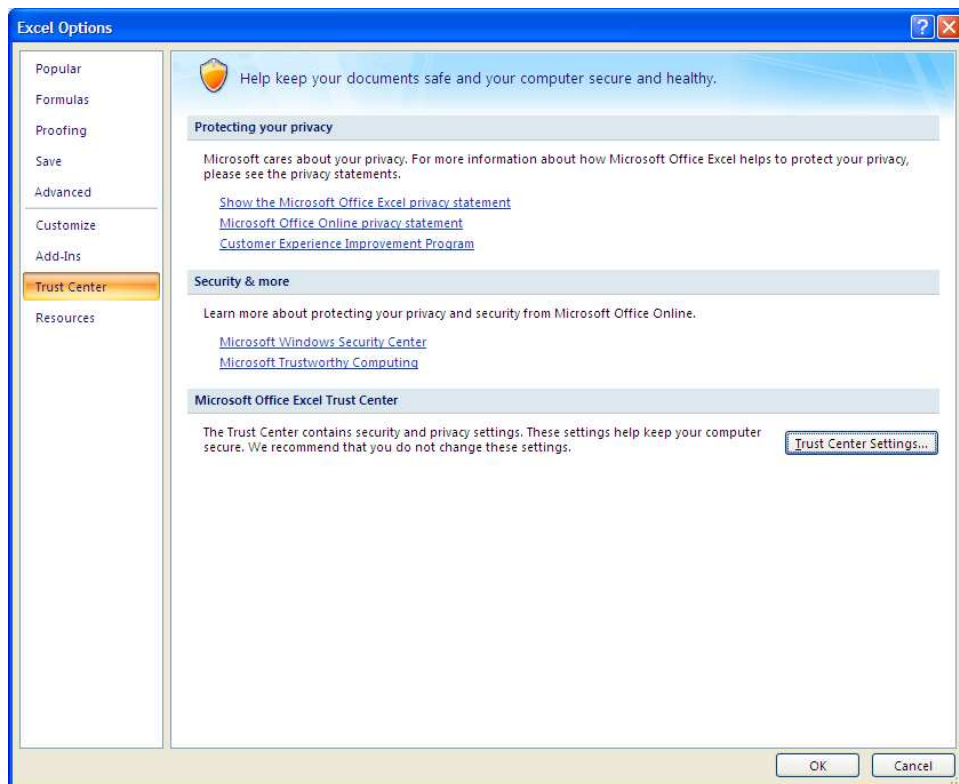


Figure 26.1: The Excel options menu, showing the trust center active.

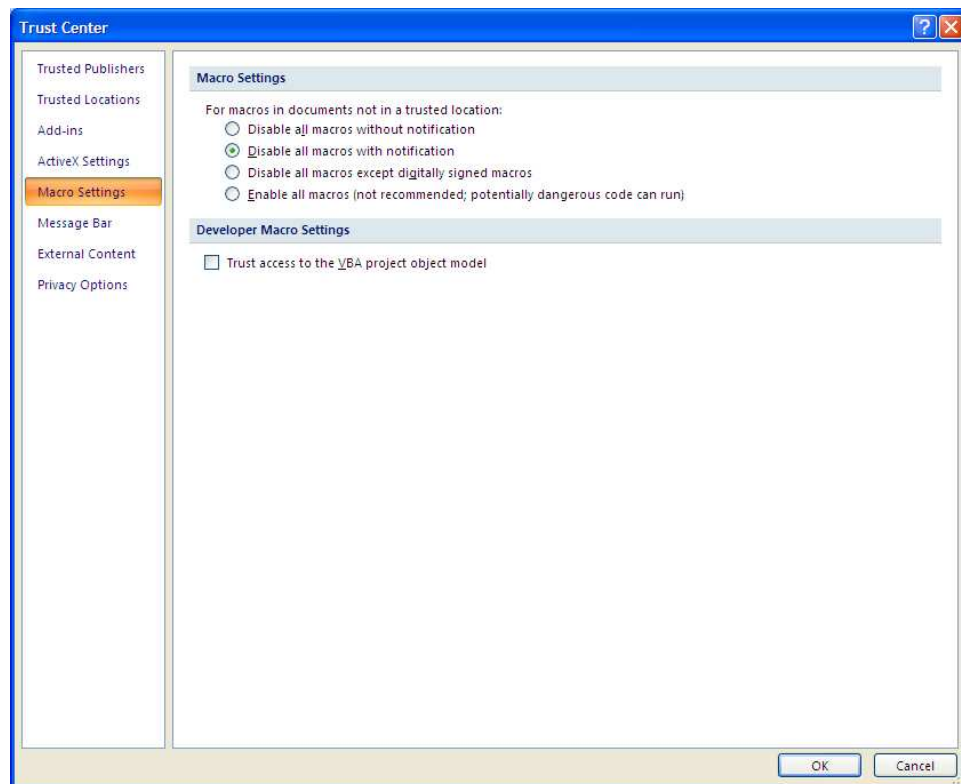


Figure 26.2: Adjusting the macro settings in the trust center.

Chapter 27

Typesetting and formatting equations in MS Word

Unknown to most people using Microsoft Word, there is a way to type set complicated mathematical formulas. To activate the Equation Editor in Word, first select the point where you want to insert the equation. Then activate the "Insert" ribbon, and choose "Equation". You can then choose to insert a pre-set equation from the drop-down list that appears, or you can insert a new equation from scratch by selecting "insert new equation" at the bottom of the list.

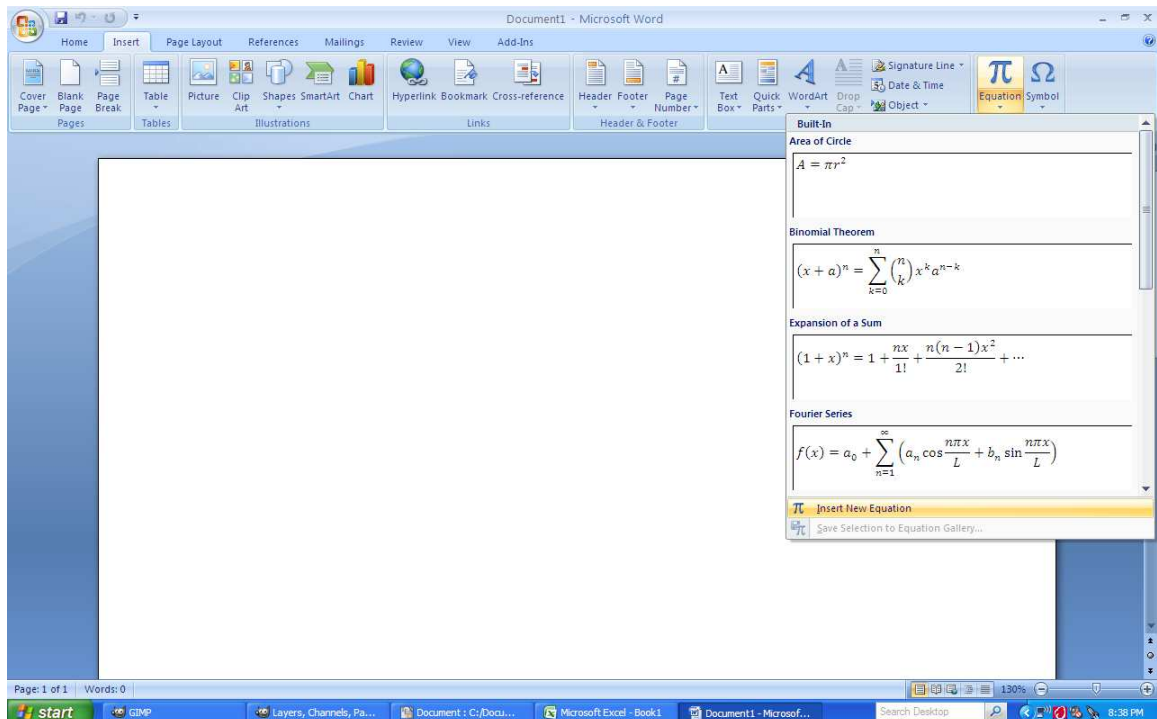


Figure 27.1: Inserting a new equation from the insert ribbon.

The screen will appear almost normal, except for two things - you will now see an area

marked "type equation here" and the equation ribbon is activated (see figure 27.2). You can start typing the equation with the keyboard and then use the buttons on the ribbon to access more symbols and tools. Some of the more common symbols are directly accessible. The remaining symbols are grouped into categories such as fraction, script, radical, integral, and bracket. For example, to enter a formula to show x^2 , we first set up a placeholder for the symbols using the script category. From the script menu, select the button with the large empty box in the center and a small empty box in the upper right corner - the one that looks like a superscript. You can now enter the "x" into the large box (the base) and the exponent 2 into the small box.

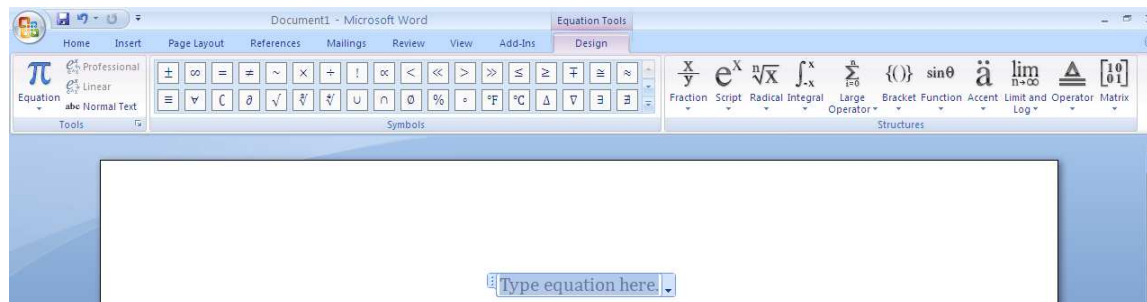


Figure 27.2: The equation ribbon.

To make a more complicated formula, say $x^2 + \frac{1}{x}$, start as before, then, after you type the superscript of 2, hit the right arrow button on the keyboard - this will move the cursor out of the exponent so that you can continue typing in the main part of the formula. Type a "+" then add a fraction with the tools in the second group on the bottom row. The first fraction (with two empty boxes, one above and one below the line) is best. Click on each box and type in the symbols you need (a "1" on top and an "x" on bottom). When you are done editing the formula, click the mouse outside the boxed in region of the equation. Once inserted, you can copy, paste, and modify the equation. (To modify it, just double-click on it.)

Chapter 28

The PMT Function

Excel can easily help you compute monthly payments on loans using the PMT function. This function requires at least three arguments, and has two optional arguments that can be used in different situations. The format of the function is shown below.

`=PMT(Rate, NPer, PV, [FV], [Type])`

Rate is the interest rate per period that the interest is compounded. If you are working with most loans, this will be the APR divided by 12, since the interest is compounded monthly. NPer represents the number of periods over which the loan is being computed. PV is the present value of the loan; in most cases, this is the principal, the amount that you have borrowed. The two optional arguments are FV and Type. FV is the future value of the loan; use this if you do not want the amount to be 0 at the end of the loan period. If you leave this argument out, Excel assumes that you want it to be 0. Type is either 0 or 1; 0 means that the payments are due at the end of each loan period; 1 means that payments are due at the beginning of each loan period (default is 0).

Thus, if you have a \$10,000 loan, at 6% APR for 2 years, we can compute the monthly payments with the formula below (assuming payments are due at the end of each period). Notice that the number of periods is multiplied by 12 to make sure that we are computing everything in months (2 years * 12 months/year = 24 months for the number of payments).

`=PMT(6%/12, 2*12, 10000)` (should be \$443.21)

Notice that in two years you will pay a total of $\$443.21 \times 2 \times 12 = \$10,636.95$ to the loan, which is the original loan of \$10,000 plus about \$636 in interest. This function can also help you plan savings or retirement accounts using the FV argument. Suppose you want to know how much to save each month, starting with nothing, if you get 5% APR and want to have \$250,000 in 10 years. The formula below computes the monthly payments you should make to the account (payments at beginning of each period).

`=PMT(5%/12, 10*12, 0, 250000, 1)` (should be \$1,603.29)

Notice that if we multiply the PMT amount by the number of periods ($10 \times 12 = 120$) we get \$192,394.90, indicating that over the ten years, we will pay about \$192 thousand, but

the interest will give us a balance of \$250 thousand. Thus, we earn about \$58 thousand over the savings period.

FINAL NOTE: PMT returns the answer in accounting format. This means that it is almost always a negative number displayed in red and in parentheses. This makes it easier to use in computations for loan amortization schedules, since you normally want to subtract the loan payment each period.

Chapter 29

Using SOLVER

29.1 Introduction to using SOLVER to minimize and maximize a function.

Excel has a very powerful equation solving tool built into it. This routine has limitations, and it certainly won't work for solving equations that don't have numerical values for the parameters, but it is a powerful tool for solving specific problems.

To use the solver, you need to have two things set up on your spreadsheet:

1. A cell that calculates something (the target cell)
2. Other cells (virtually any number of them) whose values are used in the calculation of the target cell (the parameter cells)

Using the solver is easy, once it's set up. Select the target cell. Then activate the solver routine with "Tools/ Solver". In the dialog box, make sure the "Set Target Cell" refers to the correct target cell. Then, click on the option for what solution you want: either maximum, minimum, or exact value - like goal seek. Finally, click in the space next to "By changing cells" and then highlight the parameter cells on the worksheet (use the control key to select multiple, non-adjacent parameter cells). Finally, hit the SOLVE button and let Excel compute.

Since the process is numerical, there are some errors that may occur. First, Excel may not find a solution. This can happen for a variety of reasons, but most often it's related to having the starting values of the parameter cells too far from the solution, so try changing the values of the parameter cells and starting over. You might also get problems if your target cell involves calculations with logarithms, since the process may need to try a variety of values for each parameter and this may lead to computing the log of a negative number, which is impossible.

29.2 Setting up constraints in Excel

In order to use Excel's routines for solving constrained optimization problems, you must first set up the spreadsheet so that (a) all the information is present and easily understood by

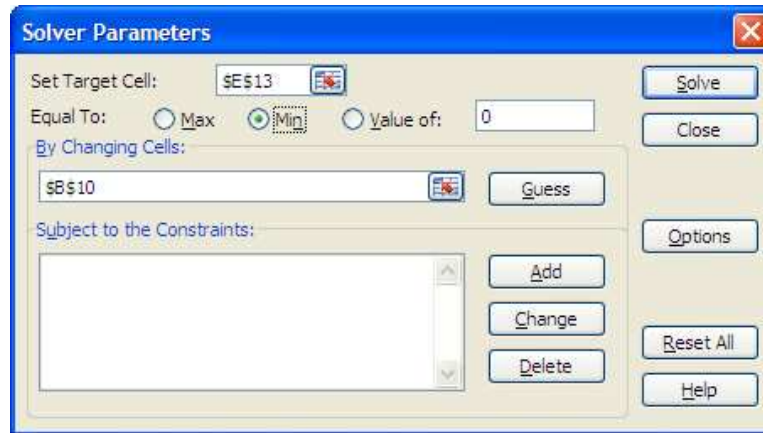


Figure 29.1: Setting the parameters for the solver routine.

a person reading it and (b) it contains all the formulas needed to connect the optimization variables and constraints to the objective function.

Let's return to examples ?? (page ??) and ?? (page ??), where we have translated an optimization problem into mathematical language as a preparation for solution. If you recall, in these problems we were trying to determine the optimum mix of three products (chairs, tables and juice carts) to manufacture in order to maximize our profits.

The screenshot below shows this problem set up (see C16 Furniture.xls). Notice that the first block of data, in cells B3:E8 is the given information about producing each of the three products (chairs, tables and juice carts). The only part of this that is not given is the second row (B4:E4) where we start with a guess regarding the number of each product to make in order to maximize our profit. In this case, we have simply guessed at 50 of each.

One of the reasons for setting up the spreadsheet as shown is that calculating the various costs and total times for assembly and finishing can be done efficiently with the SUMPRODUCT formula in Excel.

The next block of information contains the labor statistics (cells B10:F12). A few words about this are needed. First, the symbols in E11:E12 are not used by Excel in solving this problem; these are given strictly to remind the user about the set up of this problem and some of the constraints present. Second, the "total labor used" in cells D10:D12 are calculations, based on the number of each item produced and the number of hours each step takes for that item. In cell D11, the following formula has been entered:

$$=\text{SUMPRODUCT}(\$C\$4:\$E\$4, C5:E5)$$

This calculation is an Excel shorthand formula for " $=C4*C5 + D4*D5 + E4*E5$ " which goes through the list of products and computes the assembly time needed for each product (number of items*hours per item) and then adds these times together. Using the sumproduct formula makes it easy to generalize this calculation to any number of products you might be considering, without having to type an unnecessarily long formula in. Using the absolute cell references for the product amounts (C4:E4) means that you can then directly copy the

	A	B	C	D	E	F
1						
2						
3		Product Information	Chairs	Tables	Juice Carts	
4		# Produced	50	50	50	
5		Assembly Time (hours)	1	1	2	
6		Finishing Time (hours)	1	4	2	
7		Materials Cost	\$5	\$15	\$10	
8		Selling Price	\$18	\$54	\$36	
9						
10		Labor Information	Cost (\$/hr)	Labor Used		Available hours
11		Assembly	\$4	200	<=	250
12		Finishing	\$7	350	<=	350
13						
14		Profit Information				
15		Total Revenue	\$ 5,400			
16		Total Material Cost	\$ 1,500			
17		Assembly Cost	\$ 800			
18		Finishing Cost	\$ 2,450			
19		Net Profit	\$ 650			
20						
21						
22						

Figure 29.2: Setting up the constraints and functions for optimization with solver.

formula in D11 to cell D12 to compute the total hours of finishing labor used for this mix of products.

The third block of information (B14:C14) contains the calculations for the total revenue from selling the items, the total labors costs (broken down by assembly and finishing), the total materials costs and the net profit from sales. Each formula is a sumproduct, multiplying the row of information about quantities if each item produced ($\$C\$4:\$E\4) by the appropriate information (for revenue and materials costs) or it is a simple product (e.g., assembly cost in cell C17 is simply the product of total assembly hours used and cost per assembly hour = $D11 \times C11$). The net profit is then the revenue (C15) minus the total costs involved. Thus, profit is given by

$$=C15-(C16+C17+C18)$$

It is this quantity, profit, that we seek to maximize (objective function) by changing the amounts of each product we make (input variables) subject to various constraints.

29.3 Installing the SOLVER add-in

The solver add-in is not always installed when you load Excel. To make it available, click on the "Tools" menu and select "Add-ins". Regardless of what other add-ins you have installed, you should see "Solver add-in" in the list of available add-ins. Check the box next to it, and

then hit "OK" and from now on, solver should load in Excel and be available for solving optimization problems.

29.4 Adding constraints in Solver

In the solver dialog box, clicking on the "Add" button under "constraints" brings up a dialog box like the one in figure 29.3.

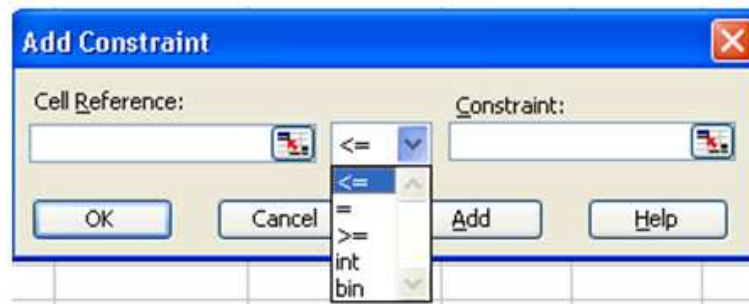


Figure 29.3: Adding a constraint in the solver routine.

Notice that there are three pieces of information needed for each constraint: the "Cell reference" for the constraint (which must somehow be connected to the input variables so that it changes as different input variable values are tested), the "constraint" (the value that you want to meet) and how the cell is constrained with respect to the value you provide. You have five possible options for this:

- <= Forces the value in the cell reference to be less than or equal to the constraint
- = Forces the value in the cell reference to be equal to the constraint
- >= Forces the value in the cell reference to be greater than or equal to the constraint
- Int Forces the cell value to be an integer
- Bin Forces the cell value to be a binary (either 0 or 1)

Clicking "Add" will enter your constraint and allow you to enter another constraint. Clicking "OK" enters the current constraint and returns you to the solver dialog box.

There is also a clever way to enter several constraints at once, if each is essentially the same type of constraint and the spreadsheet is properly organized. Consider the constraints shown in figure 29.4. (These are taken from a modified form of the solver scenario in C16 Furniture.xls.) Rather than entering one constraint for each of the product quantities to be integer, we have entered that "\$C\$4:\$E\$4 = integer" which forces each of the cells in the cell reference (C4, D4, and E4 in this case) to be an integer; this is much more efficient than entering each one separately, especially when you could have hundreds of products in the scenario.

Likewise, instead of entering the labor constraints separately, one for assembly hours and one for finishing hours, we have entered a single constraint "\$D\$11:\$D\$12 <= \$F\$11:\$F\$12". Solver goes through the list of cells on the left of the constraint and pairs them up with the corresponding cell on the right. Very efficient!

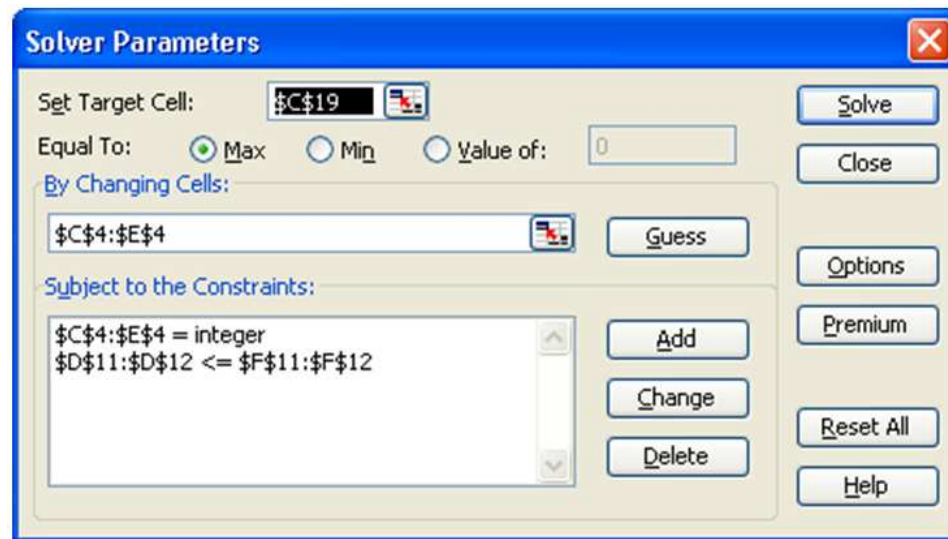


Figure 29.4: Setting up the solver parameters and constraints efficiently.

29.5 Changing constraints in solver

You can easily change an existing constraint in solver. Select the constraint in the solver dialog box, click "Change" and you can set all three of the options available when adding a new constraint.

29.6 Deleting constraints

Deleting a constraint is easy. Simply click on the constraint so that it is highlighted in the solver dialog box, and then click the "Delete" button.

29.7 Options in solver

Most of the time, you will not want to change many of the options in the solver options dialog box (see figure 29.5.) The first block of options all deals with the routine's limitations. The "Max time" is simply the maximum amount of time you want solver to search for a solution. The iterations relates to the number of times it loops through its procedures - there is almost no need to change these two options. Precision, tolerance and convergence all relate to the fact that, as numerical calculations, there is some rounding being done. Since all answers are approximate, these numbers let you specify how close solver must get to the specified values before it considers its work to be complete.

You will frequently encounter the need to set the "Assume linear model" and "Assume non-negative" options. Most of the rest of the options relate to the specific solution techniques used by solver. Details of these are fairly technical, and you rarely need to change them, but if you encounter a stubborn problem that defies solution, you may want to try

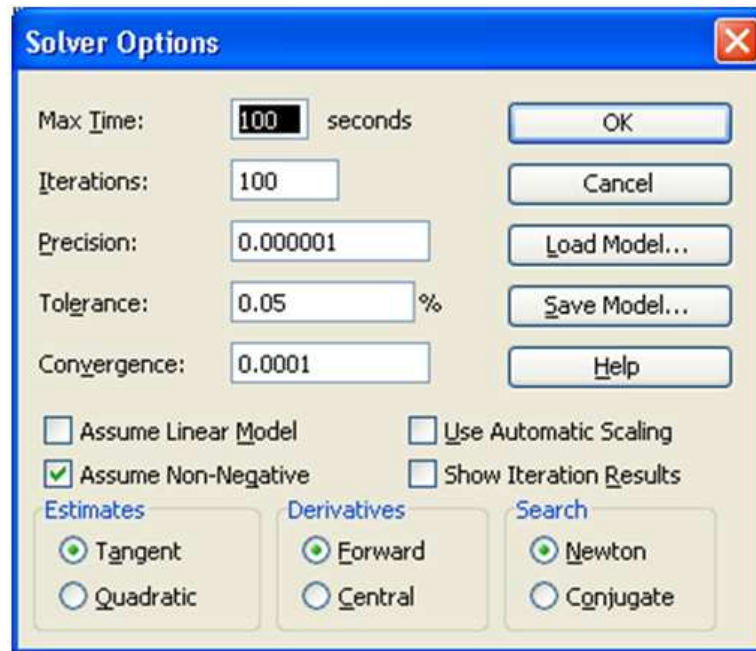


Figure 29.5: Clicking on the "Options" button in solver brings up this dialog box.

solving it with a variety of options on these, to see if a solution can be found using alternative methods.

29.8 Errors in Solver

In example ?? (page ??), suppose we had union requirements that led to a minimum number of labor hours at our company. If we change the constraints from "Assembly hours ≤ 250 " and "Finishing hours ≤ 350 " to be minimum hours, swapping all the " $\leq$ " for " $\geq$ " what happens?

Solving this model (with the labor hours as minimums, rather than maximums) produces the following screen. Excel tells you that "the set cells do not converge" which means that you need to produce an infinite amount of each item in order to maximize profit. This is not a mistake or error in Excel; it is a problem with the scenario we set up. After all, if we have no maximum number of labor hours, then we can make as many of each item as we want; making more always results in more profit, so we should make as much as possible, which is infinite in this case. We can return this to a problem with a feasible solution by adding a constraint. For example, we may have a maximum amount of money available for materials.

Another possible error that could occur is "Solver could not find a feasible solution". This doesn't mean the computer is broken; it simply means that the constraints you have specified over-determine the problem and cannot all be met simultaneously. You will have to carefully consider what constraints can be relaxed, and re-run the solver.

There are other possible errors that Solver may encounter. For more information, consult the help features in Excel under "Troubleshoot Solver".

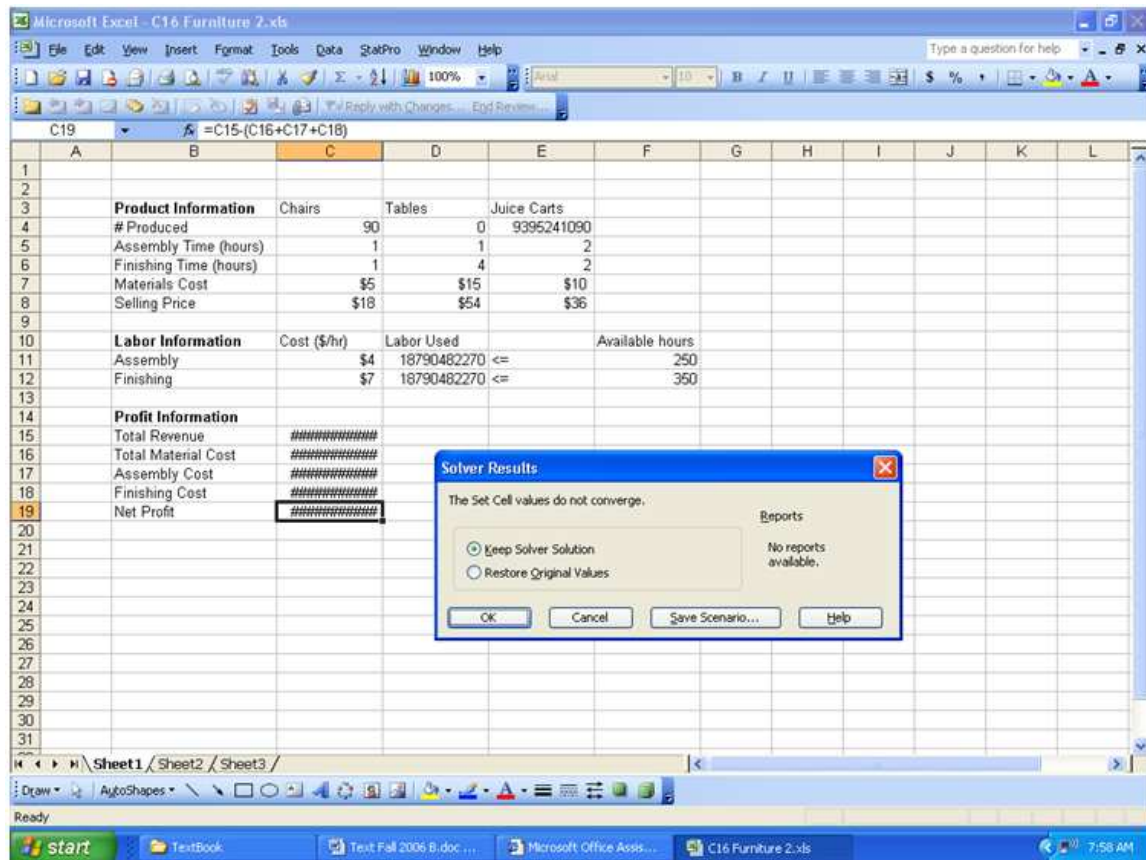


Figure 29.6: A typical solver error when there is no solution to the optimization problem.

29.9 More tricks solver will play on you

One final note about solver. Since its routine is numerical in nature, essentially a sophisticated way of guessing an answer, checking whether it works, and then calculating how to adjust the guess, the results are highly dependent on your initial guess for the input variables. Changing them could drastically change your optimal solution, especially if your situation is highly sensitive, or if there are many possible solutions to the problem. For example, in file C16 Furniture.xls, if you change the initial values from 50, 50, and 50 for the production of chairs, tables, and juice carts, the solver routine will give very different results, summarized below.

Initial Guess (C, T, J)	Optimal Solution (C, T, J)
(50, 50, 50)	(62, 34, 76)
(20, 20, 20)	(52, 34, 81)
(20, 50, 45)	(42, 34, 86)
(0, 0, 0)	(44, 34, 85)

Notice, though, that regardless of the initial values of the three variables, the number of tables in the optimal solution seems to be 34. Thus, we are simply trading off juice carts and chairs in the various solutions.

Chapter 30

A Basic Integration Tool in Excel

Bring up the file C17 Integration Tool.xls. This file contains the numerical integration of the definite integral $\int_0^1 x^2 dx$; that is, it computes a reasonable numerical approximation to the exact area under the curve $f(x) = x^2$.

30.1 Cell References and Formulas in the Basic Tool

- A1 : is the lower limit 0, which is entered numerically in B1
 A2 : is the upper limit 1, which is entered numerically in B2
 A3 : N is the number of rectangles we will use to approximate the area under the curve from 0 to 1, which is entered numerically as 10000 in B3. Use this figure for all functions you wish to integrate numerically.
 A4 : h is the width Δx of a rectangle, which is entered numerically in B4 as $=(B2-B1)/B3$. You do not need to change this.
 A6 : The numerical value of the area under the curve from a to b , which is entered in B6 as $=SUM(D9:D10008)$
 A8 : n counts the number of rectangles from 1 to 10000 in a column. This column should remain unchanged when you compute other functions.
 B8 : indicates the left-hand endpoint of the base of the n th rectangle
 C8 : $f(x_n)$ evaluates the function at the left-hand endpoint x_n to give the height of the n th rectangle at x_n , which is entered in C9 as $=B9*B9$. This formula will change for different functions that you are integrating.
 D8 : $h \cdot f(x_n)$ multiplies the width h (i.e. Δx) and the height $f(x_n)$ of the n th rectangle to give the area of the n th rectangle, which is entered in D9 as $=B$4*C9$

30.2 How to Modify the Basic Integration Tool for Other Functions

Example: Find $\int_2^4 (x^2 + 1) dx$ by numerical integration by modifying the Basic Integration Tool.

1. Copy C17 Integration Tool.xls to a new worksheet and save under other file name.
2. Set B1 to 2
3. Set B2 to 4
4. Set C9 = B9*B9+1 This is the new function we wish to integrate.
5. Highlight C9 and place the cursor on the lower right corner (fill handle) until it becomes +. DOUBLE CLICK the plus. All calculations will be computed at this point.
6. The approximate area under the curve is 20.66547, which is found in B6. The exact area under the curve as found by the Fundamental Theorem of Calculus is $20\frac{2}{3} \approx 20.67$.

Appendix A

Excel Cursor Shapes

The pointer on the screen can take any of nine different shapes. The shape of the pointer is a clue to what actions will take place when you click the left mouse button.

Shape of pointer	Action when you left click
Normal arrow	Selects the current item
Fat plus sign	Selects the cell (either for entering data/formulas or for other purposes, like copying)
Skinny plus sign	Click and drag to copy the formula(s) or the pattern in the selected cell(s) to other cells on the worksheet
I-beam	Enter text
Short down arrow	Selects entire columns of data
Short right arrow	Selects entire rows of data
Double arrow (one line)	Click and drag to change cell widths (left-right arrows) or heights (up-down arrows)
Double arrow (two lines)	Allows you to split the worksheet into separate areas
Four-headed arrow	Click and drag to move toolbars around

Appendix B

Excel Errors

Under certain circumstances, even the best formulas can appear to have freaked out once you get them in your worksheet. You can tell right away that a formula's gone haywire because instead of the nice calculated value you expected to see in the cell, you get a strange, incomprehensible message in all uppercase letters beginning with the number sign (#) and ending with an exclamation point (!) or, in one case, a question mark (?). This weirdness is known, in the parlance of spreadsheets, as an error value. Its purpose is to let you know that some element - either in the formula itself or in a cell referred to by the formula - is preventing Excel from returning the anticipated calculated value.

Here is a list of some error values and their meanings:

#DIV/0! Appears when the formula calls for division by a cell that either contains the value 0 or, as is more often the case, is empty. Division by zero is a no-no according to our mathematical rules (you can divide a pizza into 2 slices, but you cannot divide a pizza into zero slices).

#NAME? Appears when the formula refers to a range name that doesn't exist in the worksheet. This error value appears when you type the wrong range name or fail to enclose in quotation marks some text used in the formula, causing Excel to think that the text refers to a range name.

#NULL! Appears most often when you insert a space (where you should have used a comma) to separate cell references used as arguments for functions.

#NUM! Appears when Excel encounters a problem with a number in the formula, such as the wrong type of argument in an Excel function or a calculation that produces a number too large or too small to be represented in the worksheet.

#REF! Appears when Excel encounters an invalid cell reference, such as when you delete a cell referred to in a formula or paste cells over the cells referred to in a formula.

#VALUE! Appears when you use the wrong type of argument or operator in a function, or when you call for a mathematical operation that refers to cells that contain text entries.

Appendix C

Other Features of Excel

Worksheet Tabs

Each workbook can contain multiple worksheets. For example, you may want a workbook containing data on your stock portfolio. The workbook could contain a separate sheet for each of the stocks in your portfolio showing past data and current information. One worksheet could contain a summary of all the current stock information, showing the current total value of your portfolio. To work with multiple sheets, use the worksheet tabs at the bottom of the data entry area. New workbooks contain three worksheets by default.

Simply click on the tab for the worksheet you wish to view. If there are too many worksheets to display all of them, use the arrow keys to move forward in the list (right arrows) or backward in the list (left arrows). The arrows with vertical bars take you to the first worksheet in the list (left arrow with bar) or the last worksheet in the list (right arrow with bar).

To add another worksheet either use the "Insert" menu, or right-click on the worksheet tabs and select "Add worksheet". You can also click on the worksheet tab to the right of the last worksheet in the workbook. To change the order of the worksheets, click and drag one of the tabs to a new place in the list; you will see a small sheet icon and a down arrow showing you where the sheet will be placed. It is also helpful to rename the worksheets with more meaningful names than "Sheet1" and "Sheet2". To do this, either

1. Right-click on the tab for the worksheet you want to rename. Select "Rename" from the pop-up menu, then type the new name over the highlighted text.
2. Double-click on the name of the sheet you wish to modify, then type the new name over the highlighted text.

Special Keys

There are several special key strokes that you will want to use on a regular basis. To display the actual formulas in each cell, rather than the results of the formulas, use CTRL + '. The ' (tilde) key is on the upper left of most keyboards. Hitting this key combination again will switch back to displaying the results of the formulas.

To use the keyboard to move around the worksheet:

- The arrow keys move the active cell in the obvious directions.
- Holding down SHIFT while navigating with an arrow key will both move the cursor and select (highlight) the cells as you move.
- Holding down CTRL while navigating with an arrow key will jump to the last cell containing data in that direction. If there is a blank cell in between blocks of data, CTRL + arrow will stop at the last cell before the blank cells begin.
- Holding down CTRL + SHIFT and using the arrow keys combines both of the last two features. This is especially useful for selecting a large block of data that scrolls on for many screens.