

XX. GRAPHICS, CHARTS AND VISUAL DISPLAYS

THE SITUATION

- 1. THE DEFAULT EXCEL 2007 GRAPHIC CHARTS ARE ALL BAD.**
- 2. EXCEL 2007 CHARTS ARE LOADED WITH CHARTJUNK.**
- 3. EXCEL 2007 LACKS THE CAPABILITY TO CREATE SOME STANDARD STATISTICAL CHARTS.**
- 4. CHARTS AND GRAPHICS BUILT UNDER EXCEL 2000 OR 2003 MAY NOT WORK IN EXCEL 2007.**
- 5. BUILDING GOOD STATISTICAL CHARTS IN EXCEL 2007 IS MORE COMPLEX AND DIFFICULT THAN IN PREVIOUS VERSIONS.**

INTRODUCTION

This section has been difficult to write. For Excel 2007, the basic problem is to convert all the 2007 chart-junk into decent displays of data and results (see Su 2008 and Few 2004). In many cases, it just can't be done. It also takes up so much time to recall all the arcane and illogical menus, commands and steps to get decent statistical charts.

The other problem is about the use of Excel 2000/2003 charts in Excel 2007. In many cases, where the data sources are different, there will be an incompatibility. The 2007 menus and defaults may prevent their use. What frequently happens is a reversion to a chart form not recognized by Excel 2007, and an inability to fix the problem. In all cases it is best to start over from scratch and build new charts in Excel 2007 to fit the Excel 2007 menus.

Cryer (2001) described all the distracting elements and chart-junk problems with Excel 2000. For Excel 2007, the problem has just gotten worse.

Stephen Few, a consultant who specializes in data visualization for business and who has written books (Few 2004) on the subject recently wrote of Excel 2007 graphics:

“I learned who at Microsoft was responsible for the development of the new charting engine and quickly sent him an e-mail. In it, I introduced myself and offered to send him my suggestions for shaping Excel's new charting functionality. With his polite assent, I submitted a list of recommendations.

In response, I received a very courteous e-mail. Here's an excerpt: ‘Thanks. This was quite interesting and useful. I think you'll be pretty happy with some of the changes we are making for Office 12. I did read your book for inspiration (as well as Tufte, Cleveland, Wilkinson, Zelazny, etc). ‘ “ (Few, 2006)

Su (2008) is more optimistic and said, “However to be fair about Excel, it is still possible to use this program to make effective charts, Users just have to clean up ‘chartjunk’ in Excel on their own.”

Good charts can be obtained by putting in the effort to learn the obscure menu terms and list options, and making the necessary changes to the default charts (using the default as a starting point). Not everything can be fixed. Some chart details may not be changeable. **There are still some common graphics used in statistical publications that cannot be built in Excel.**

The process of creating a chart can be saved and in many cases can be used over again with new data¹.

SPECIFIC CHART AND GRAPHICS PROBLEMS

Table 20-1 is a summary of issues and problems reported in the literature. These problems are directed primarily at the default charts.

Table 20-1: General Excel Graphics Problems

Application or Function	Problem	Source
Chart Graphics	Default charts are Inadequate.	Cryer 2001
Chart Graphics	No boxplots, dotplots, stem-plots, labeled X-Y, scatterplot arrays, data brushing	Hunt 1996
Data Analysis Histogram	X Axis labels wrongly positioned	RSS 1996
Data Analysis Histogram	Cannot set in unequal class intervals	RSS 1996
Data Analysis Histogram	Not clear what tick marks apply to X axis values	Goldwater 2001
Data Analysis Histogram	There are gaps between the histogram bars	Cryer 2001

¹ Excel is different from the other statistical programs, in that the menu captions selected and the keystrokes used can be captured in a macro. These macros can be rerun to form similar charts with other data sets, saving the time to build each chart from scratch. A chart built on a worksheet with the data on the same worksheet, represents a storable and reusable object. Charts can be repeated with different data sets, by copy and paste operations. With macros, VBA functions and VBA subroutines, a long series of similar charts can be quickly built. This gives a lot of flexibility to chart building, and many similar charts can be built quickly.

GRAPHICS:

OVERVIEW, THE ISSUES:

Criticism has essentially come from the academic view, where publication and instruction are the dominate modes. The objective then of graphics is to support these modes. For publications, there is a lot of high quality graphics software that will give publication quality graphics. Excel 2007 lacks this ability.

With respect to the instruction mode, there is the “presentation” (lecture) mode and the student exercise mode (homework), There appears to be two views here, that of instruction and that of paper publication.

Instruction: For a beginning student, it takes a lot of time to get them to be able to prepare clear, standard charts from the complicated Excel Chart set-up commands. What shows up in student exercises often comes from Excel’s default graphics formats or the students choice to be outrageous. They also tend to get carried away (Dawson 2003) with all the chart options and tend to turn in “weird” charts.

Presentation: Excel charts will not meet some publication or paper presentation requirements. There is a lot of presentation graphics that Excel just can’t do.

Neil Cox (Statistician, AgResearch Ruakura, 2000) states, “While some statistics packages have much more powerful exploratory graphing capability, Excel can often do all that is needed quite easily”. There are also a vocal group of teachers on the Internet lists who feel strongly that these high quality graphics packages disguise the data and distract the student, and take the position that the simpler packages such as Excel should be used instead. (Note that for Excel 2007, a lot of this simplicity has been lost.)

However Excel is not capable of doing some of these simpler descriptive statistics forms (i.e. Tukey’s box and whisker plots).

Criticism has mainly focused on the default charts, which if not changed can give some dreadful looking charts. (The jargon is “defart” for the display of the default chart.)

EXCEL 2000 AND 2003 CHARTS

Actually, some fairly good charts can be prepared in Excel 2000 and 2003. Some examples follows.

The Excel file, NewChart.xls² on this URL, gives some convenient tools to create better Excel charts for general use. The tables are included on each worksheet to directly generate the numbers for the included chart. These only represent a starting point for a desired chart, and should not be used without changes to fit your requirements.

NewCharts.xls gives the sequences to build good histograms and frequency polygons of the type shown in Moore and McCabe (2002) and in Larson (2003). Histograms illustrated in other statistics books differ in appearance, and Excel cannot duplicate all of

² The set of charts is still under construction, and some charts are not yet included. The current set (Excel 2003) does not work properly in Excel 2007.

them. Also the histograms shown in Larson (2003) cannot be exactly duplicated, since the left hand broken axis mark cannot be put in.

NewChart.xls also has the corrected exponential smoothing calculations and chart that are described in McCullough and Heiser (2008).

New Chart.xls is incomplete in the sense that many other charts used in statistical analysis are not covered. In a sense, it is incomplete.

Peltier (Peltier 2006a, 2006b and 2006c) has some excellent basic charts for use under Excel 2003

Peltier (2006c) also shows how a very presentable box and whisker plot can be generated.

Peltier (2006a) shows how the individual points in a given series can be separately formatted (shape, form, outline, size, color) to distinguish them from other points in the series. For a given point, there are over 153,000 combinations of size, shape, form, outline and color to arrive at an effective way to show point differences.

EXCEL 2007 CHARTS

Charts prepared under Excel 2000 and 2003 may appear entirely different or not even work. All the old commands have been changed, and the tree structure of commands has also changed, both in wording and in locations.

Excel 2007 expanded the graphics capability, but this was for business applications, where the mode is presentation before a “business audience”. This is not an “academic” or “student” audience. It can vary from a one-on using a computer display, to a larger group with projectors using Power Point. These business charts (forms, shapes and the colors and primarily bar charts) may be inappropriate for scientific, technical or statistical applications, but they are what business has asked of Microsoft to provide.

If Excel is used, there is no way around the problem of learning all of the Excel graphics commands, and what the terms and words mean, and what they actually do. There is no logic to what the words in the menus and command appear to “say or do” and the actual effects. The “Microsoft mind” is different from a “statisticians mind”.

The defaults have to be changed. For example, lines have to be made distinct (they are fuzzy by default). Colors have to be changed to bold or to black for publication and fill options have to be carefully evaluated .

SPECIFIC ISSUES AND CONSIDERATIONS ABOUT CHARTS AND GRAPHICS

The other side of the debate about Excel charts, has to do with personal opinions on/about the display itself. Some of these issues and opinions³ are:

³ These basically come from Cryer (2001), Su (2008) and Few (1996). The tic mark/number dislocation problem can't be fixed in Excel 2007.

1. To show grid lines or not. Has to do with clutter, sparseness and objective of the display. Do they add to the “argument”? What distracts and what is “covering up”? Grid line weight, color, texture and spacing are all important considerations.
2. Legend box or direct labeling of “curves”. Excel does have the ability to put in overlay text (message boxes) to identify specific features of the data. The legend box is the simpler and common tool. The legend takes up a lot of page or slide space, compressing the actual chart, and leaving a distinct loss of visual balance.
3. Range of axis, limit to data range or to include the zero point? Does zero have a meaning? Are the axis’s overstretched?
4. Is the background important or just chartjunk?
5. Is 3-D important or is it just chartjunk? Does it just obscure parts of the data?
6. The color schemes may be too light or dark for publications, or they may tend to obscure the relationships between data sets.
7. How much is actually needed to “label” the axis’s, with respect to the chart title? It is common in ASA publications to not put in a chart title, but to refer to the chart as a figure-number. However in Su (2008) each example has a chart title.
8. How should a histogram be shown? Should unequal intervals be allowed? Numbered intervals along the x-axis have to have equal intervals for Excel 2007.
9. Should specific points in a series have separate different characteristics (size, color, outline, shape, etc.) from the other points? In this case there is something about these points that is different from to others.
10. Should error bars be added-in to show (for example) confidence intervals? The stock chart may be adapted to show intervals about a point.

SUPPORT FOR CREATING GOOD CHARTS

MICROSOFT SUPPORT

KBA’s 213930, 214033 and 155130 describe how to construct other descriptive statistics graphics.

ADD-INS

Some of the textbook add-ins include routines to build descriptive statistics graphics.

PHStat2 (1999) an add-in, has a boxplot button and a stemplot button, but they are crude. The PHStat2 stemplot is shown as a group of cells in a standard worksheet, not as a true chart.

PREVIOUS VERSION PROBLEMS THAT HAVE BEEN FIXED IN EXCEL 2007

The Histogram gap problem that Cryer (2001) complains about has been fixed in Excel 2007. Histogram columns can be separated by a gap or not separated, and with a border or with no borders.

PROBLEMS THAT HAVE NOT BEEN FIXED

THE X-Y PLOT CONFUSION

This problem keeps coming up, over and over and over again. There are essentially two basic plot schemes shown in the menus, “Line” and “xy(scatter)”. Users constantly get these mixed up, and can’t understand why subsequent “TRENDLINES” are wrong.

THE “LINE”:

In this plot, the actual Y values given in the cells are used, but the X values are taken as categorical, and are assigned the values 1,2,3,4,etc, INDEPENDENT of the values in the cells assigned as the X values. The spacing between X values is identical (uniform). However the appearance of the chart looks like it should, but it is deceiving, since it does not reflect the actual X values. The actual X values are taken as “categorical” values. Tic marks represent spacings between successive Y values. It is essentially a “Bar” chart form, changed to draw a straight line from point to point.

The confusion occurs, because the actual X and Y values appear on the chart.

THE “XY (SCATTER)” :

In this plot, the actual Y values given in the cells are used, and the actual X values are used. Tic marks represent uniformly spaced X intervals, and have inherent assigned X axis values.

HISTOGRAM TIC MARK PROBLEM

The histogram tic mark problem is an Excel fault that is repeatedly found in the literature, and one that Microsoft never changed. It remains a problem in Excel 2003 and Excel 2007. This is because the basic histogram chart is The “line” chart, described above. This is essentially unfixable in Excel, since Chart only has the two chart forms “Line” and “xy(scatter)”, and obviously the “xy(scatter)” form can’t be used for a histogram.

The Excel histograms are generated from a column type display. For a column display, a value (such as the number of items) is placed centrally below the column. This is automatic and can’t be changed.

Tic marks can be generated and placed either above or below the x axis. For all statistical applications, it should be below the axis.

There are now two choices of where to place the tic mark, either at the borders of the column or at the middle of the columns. On the basic 2007 “Axis Options” menu (from the “Format Axis” menu), selecting “Position Axis, On Ticmarks”, the tic mark moves to the center of the column, above the number. Selecting “Position, Between Tic Mark”, the tic marks move to the column boundaries, but the number remains in the middle. None of the other selections on the “Axis Options” menu can change this.

For statistical applications, the histogram column must be identified by both the lower and upper boundaries, and clearly Excel, can’t do it. The number representing each column is a category name.

If the tic mark is set to the upper boundary, the number remains at the midpoint. The number cannot be repositioned to be below the set tic mark. Microsoft “fudged” on this problem by tipping the number to lie at an angle, so that it appeared to the value at the right hand tic mark. This was done for the 2000 and 2003 versions. They reverted to the

above disaster instead of fixing the problem, that of being able to move the x axis numbers. To just below the tic marks.

You cannot add spaces to the number display to move the numbers.

To tilt and expand the numbers, so that the connection to the upper tic mark is clearer, select the “Alignment” menu and change “Text Layout”. Set the “Custom Angle” to about 45 (or other angle) so that the tic connection is apparent. Also go to the “Number” menu and expand the number of digits so that there is an apparent connection of the right digit to the upper tic mark. The numbers will tend to be fuzzy and indistinct,

This is really not a good solution, but that is our only option in Excel 2007.

COMMON STATISTICAL GRAPHS AND CHARTS THAT HAVE NOT BEEN ADDED

TUKEY’S BOX-WHISKER PLOTS

This is one of the cases where common statistical graphics are used and where Microsoft provides absolutely no capabilities.

Boxplot graphics are found in mainstream statistical publications, such as The American Statistician, and should be considered basic for displays of data. However they represent the old pre-computer technology approaches that Tukey invented for quick pencil and paper views of data. They are still used however in mainstream publications() in an augmented form to show particular characteristics of data sets that are being investigated.

The elementary form in NEWCHART.xls does not do the job very well. It uses a method of generating “points” and constructing lines between the points. The points are not relative, and consequently the example is limited to the five boxplots per chart.

Microsoft Knowledge Base Article 153130 describes how a stock type chart can be used to make a box and whisker plot. This not a true Tukey box-and-whisker plot.

TUKEY’S DOT-PLOTS AND STEM-PLOTS:

There is no direct support for these graphics.

Dot-plots and Stem plots can be created in an Excel chart. Examples are in NEWCHART.xls. The problem is that specific identifications, notes or observations cannot be annotated by overlays and message boxes.

A true stemplot in accordance with Tukey’s method of expression of the extreme ends has to be done as a drawing. Consequently some of the finer structure in the data set may not be easily shown with these graphics.

PLACING A CHART INTO A WORD FILE

In Excel build the chart. Use values that do not change. If the chart values are from equations or functions, copy the rows/columns with the values and do an Edit → Paste Special → Values on a clean worksheet. This sets up the chart source data as being invariant and not excessively link dependent.

Build the chart on this worksheet.

Select the finished chart and do an Edit → Copy. The file and object links between the Excel chart, the data source and the word document are formed.

In Word, put in sufficient blank lines (CR's) to allow space for the chart. Select a line in the middle of this group and do a paste.

Resize the chart to the desired size. Put any titles above and footnotes below the chart. Delete excess lines. In Word you can resize the chart or delete it only. If these links between the source data, chart and word document are broken, the chart may disappear or become impossible to edit. The links may be broken as a result of word document changes, revisions, or changes to the Excel worksheets, or changes to the sources of the data cells that the chart depends on.

If you are going to use Excel, it pays to be well versed in how to make graphs and to make changes.

CHARTS FROM THE TOOL PAC DATA ANALYSIS ADD-IN

These are sets of default based charts that can be easily output from the routines in the Data Analysis tool-pak. These charts were changed for the 2007 version, and are different in appearance from the previous versions. Only the routines HISTOGRAM, EXPONENTIAL SMOOTHING, MOVING AVERAGE AND REGRESSION output charts (when selected by checking the appropriate boxes).

HISTOGRAMS

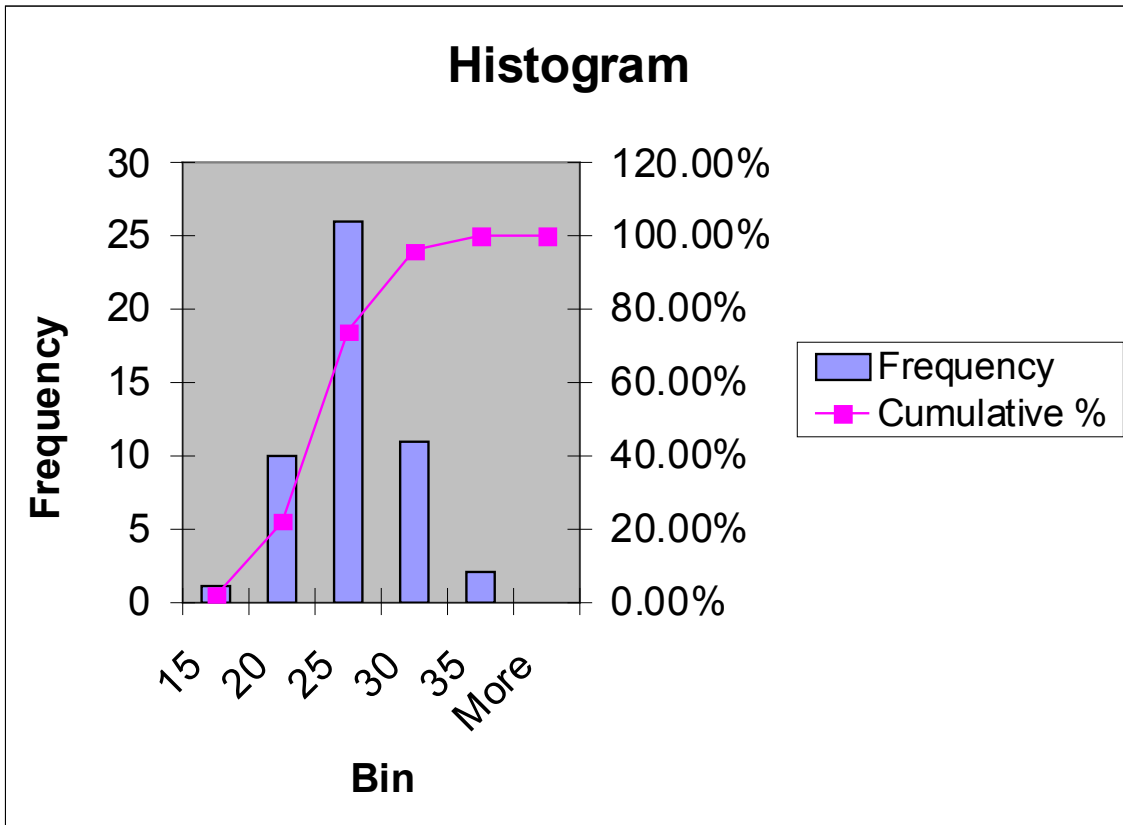
Histograms are treated in Excel as if they represent categories. Any numbers used as "names" for each vertical bar, are not treated as true X values. They are used as a "name" for the bar. The true X values (used for example as X values for a trend line on the histogram) are simply a count from left to right. This constantly surprises users, who expect that the given numbers are used in forming a trend line.

The cumulative values are based on the numerical Y values input.

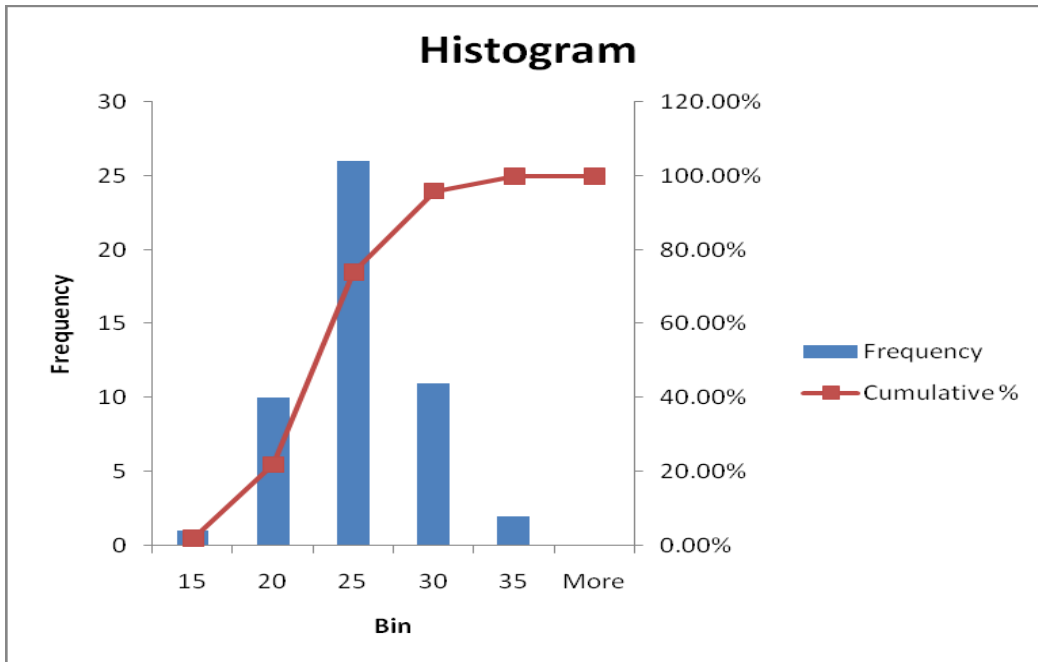
EXCEL 2000 AND 2003 DATA ANALYSIS HISTOGRAM

Using the example in Chapter 4 of Dretzke and Heilman (1998), the Data Analysis Histogram routine gives the following output.

EXCEL 2003 DATA ANALYSIS HISTOGRAM



EXCEL 2007 DATA ANALYSIS HISTOGRAM



The chart is different for Excel 2007. It does appear to be clearer and not as “clunky” as the 2003 Histogram.

The tic marks clearly define the interval boundaries. There is still the tic mark confusion, and the uncertainty of just what the actual ranges are for each bar. This is of course the use of numbers as categories, where a “range” for each bar is meaningless.

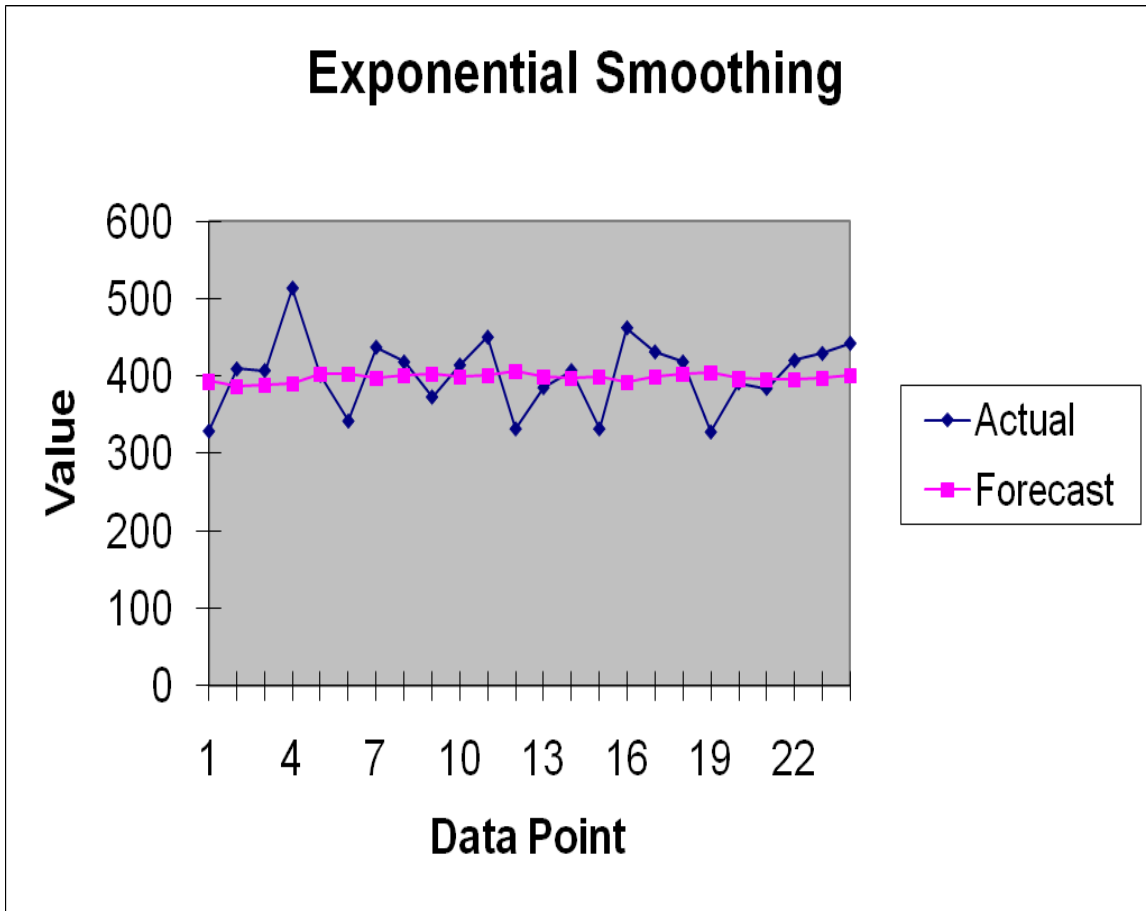
The cumulative scale does not need to go above 100%.

Appearance is improved.

The 2007 histogram chart can be moved around on the worksheet, as a complete entity. The fracturing, characteristic of the other Excel 2007 data analysis charts does not occur here.

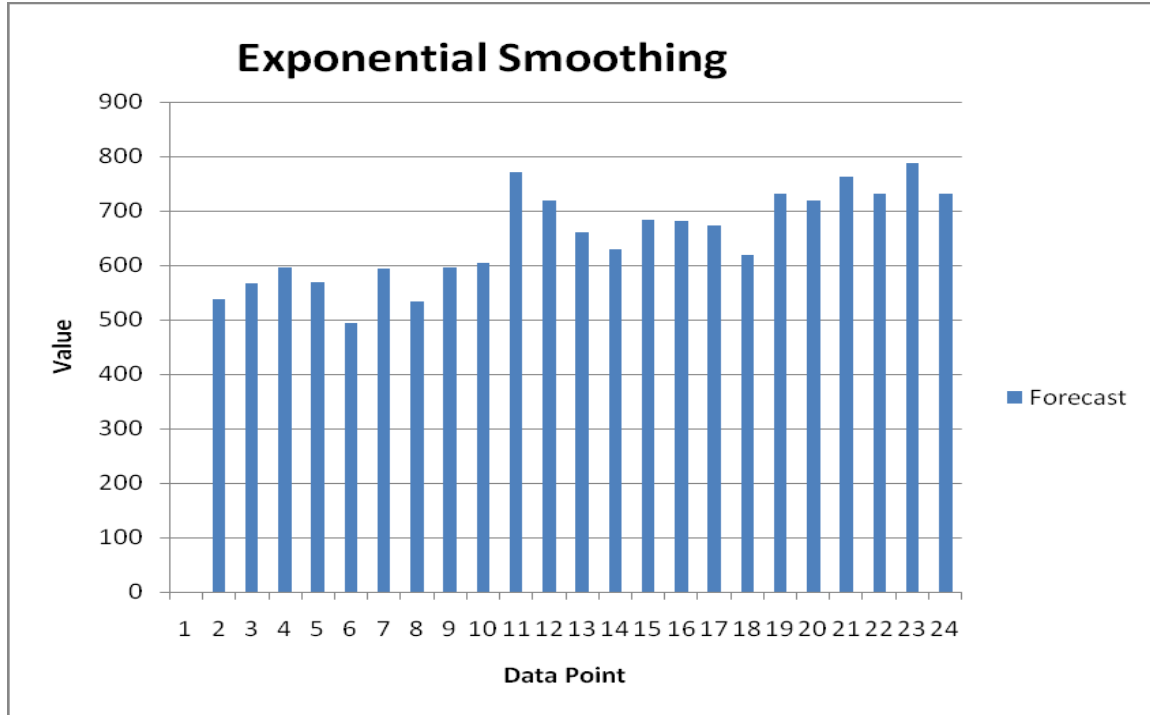
EXPONENTIAL SMOOTHING

EXCEL 2003 EXPONENTIAL SMOOTHING CHART



Both the data and the smoothed forecast are plotted. The 2003 default chart is poor and requires editing effort to change it into a good chart.

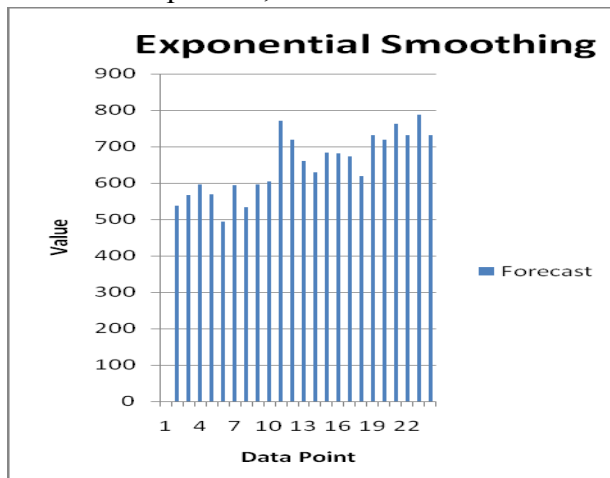
EXCEL 2007 EXPONENTIAL SMOOTHING CHART



In 2007, only the smoothed forecast is plotted, not as an XY scatter plot, but as a column plot. The actual data could be added as different colored columns, but the presentation would not be clear. Excel 2007 Chart can not superimpose two different chart types into one chart. This is a disadvantage.

You can see that changing to columns, creates a lot of visual clutter. The data set here is different, but it still is a forecast. The source data is not shown as it is in the 2003 plot. The x axis numbers are jammed together. An x-y plot would be clearer. The inherent defects of Excel 2007 charts show up here, and they may not be easily corrected, such as the closing right side axis.

If the chart is compressed, it looks like this:



The open, borderless left side gives the impression that the chart is unfinished.

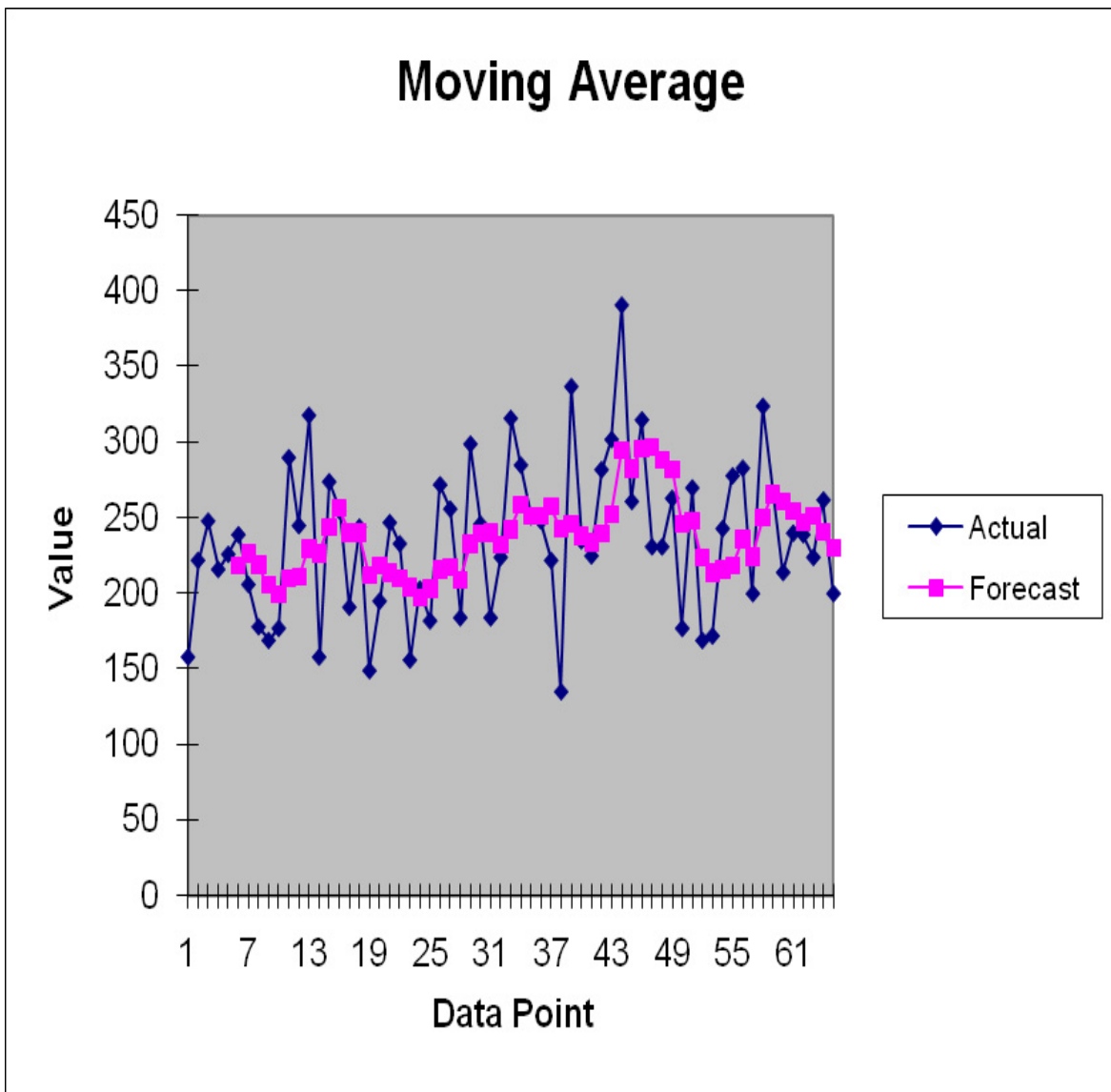
Here the actual bar corresponding to the number is not just above the number, but “near” the number, One has to go to the data to identify the periods for the two highest peaks.

The other problem is that with the 2007 charts, we can’t see the actual data, so it is hard to see the extent of the actual data that drove the two peaks.

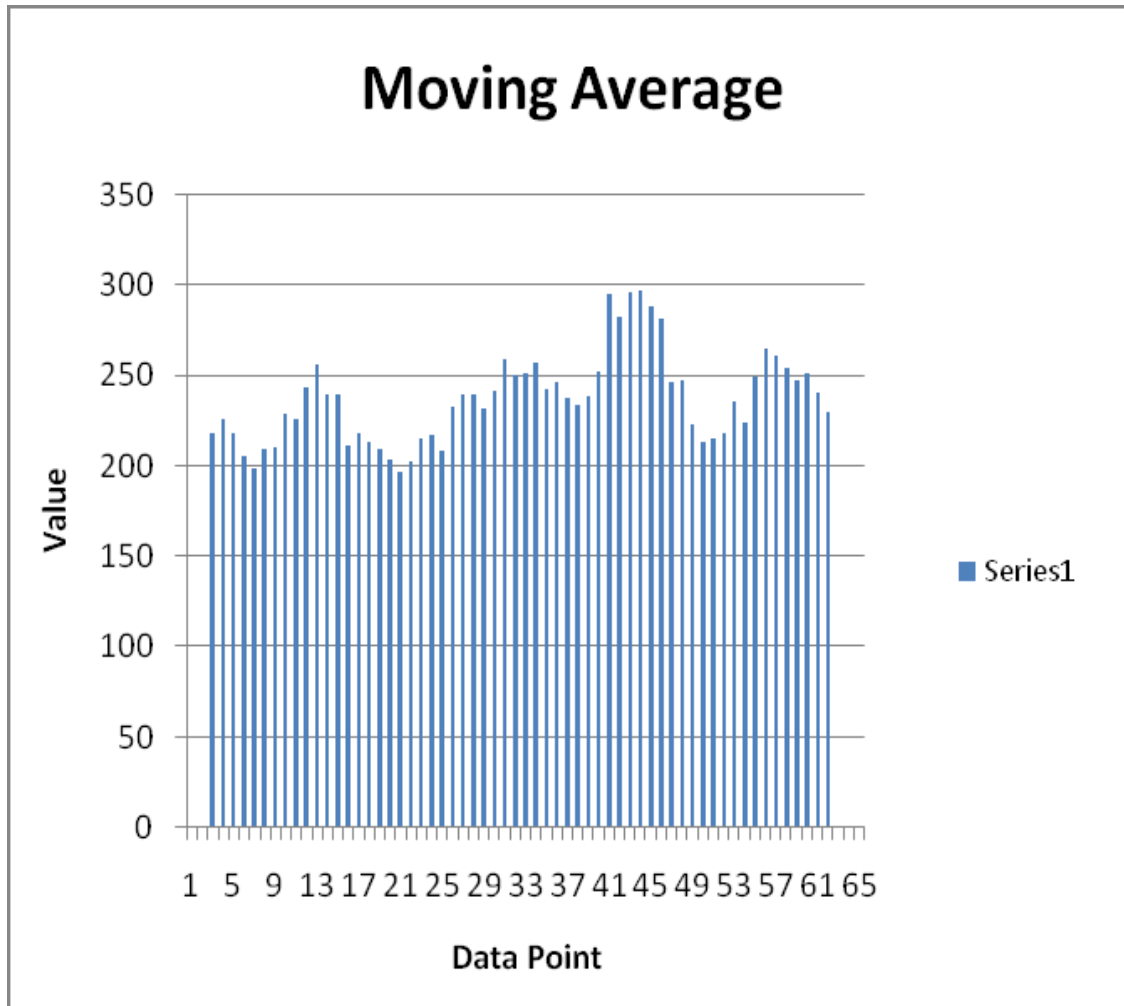
The compressed view shown in the lower figure that stretches the vertical and compresses the time scale is what likely would be shown in a presentation.

MOVING AVERAGE

EXCEL 2003 MOVING AVERAGE CHART



EXCEL 2007 MOVING AVERAGE CHART



Both charts suffer from the lack of an evident grid structure that would allow easy visual identification of peaks and valleys to the specific X axis values. One has to go back to examining the data to make these generalizations, defeating the concept of visual identification of relationships.

The 2007 chart looks cluttered because of the way the data is presented. If you are only concerned about peak values, why is there so much emphasis in the chart of the space below the peaks? It just is a case of introducing chartjunk to be impressive.

A viewer's question as to why some lines are darker than others can never be answered, because it occurs as a result of some unknown Microsoft logic, or a default due to a default "dot" spacing.

REGRESSION

THE BASIC CHARTS THAT SHOULD BE GENERATED:

Based on Kutner, Nachtshein, Neter and Li (2005), the type of charts that are of interest are:

1. Scatter plots between any two variables
2. Residual plots between residuals and input variable values
3. Normal probability plot of residuals

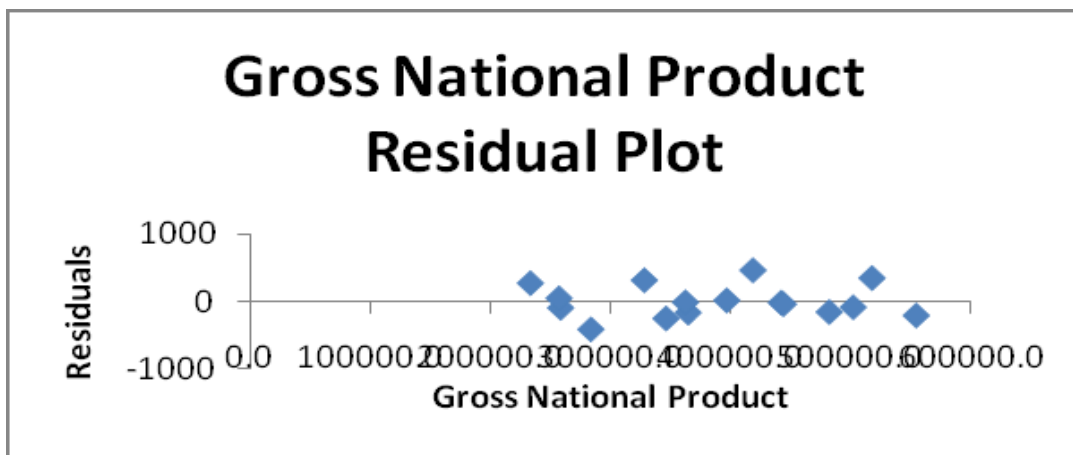
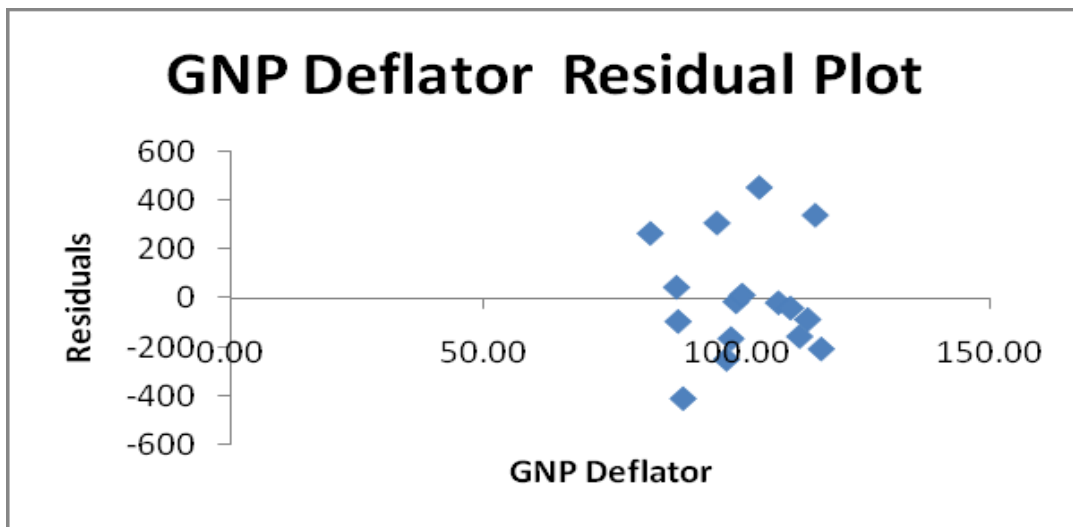
THE ACTUAL OUTPUT CHARTS GENERATED

The ATP regression routine allows one to generate a long dump of charts.

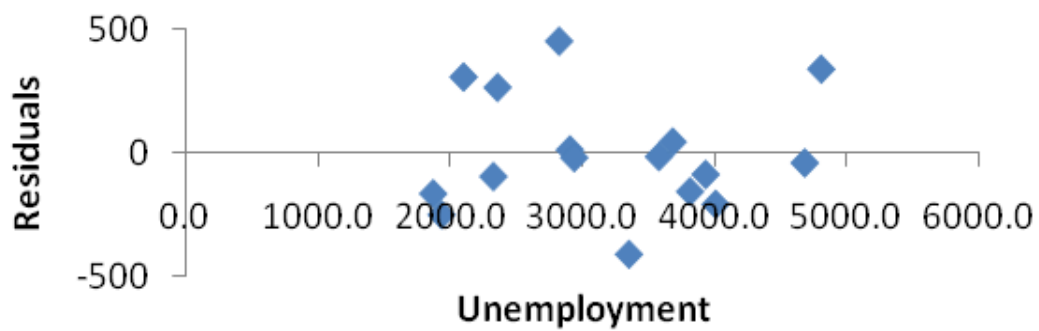
Type 1 charts are not output. Type 2 charts are output as the residual charts. The type 3 normal probability plot chart that is output is wrong. The line fit plots that can be selected are not important and should not be selected. Type 1 charts have to be generated separately by the user

For the Longley set (see section 9 for the regression and data set), thirteen charts can be generated.

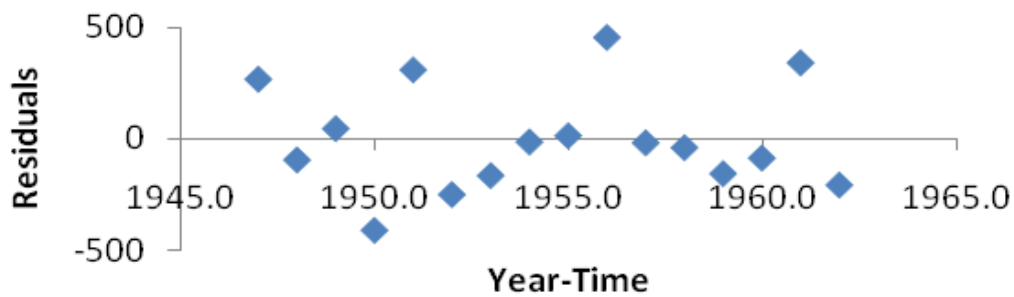
The following charts are Excel 2007 ATP Regression outputs, repositioned to see each of the 13 charts. These charts can only be seen in the print layout view.



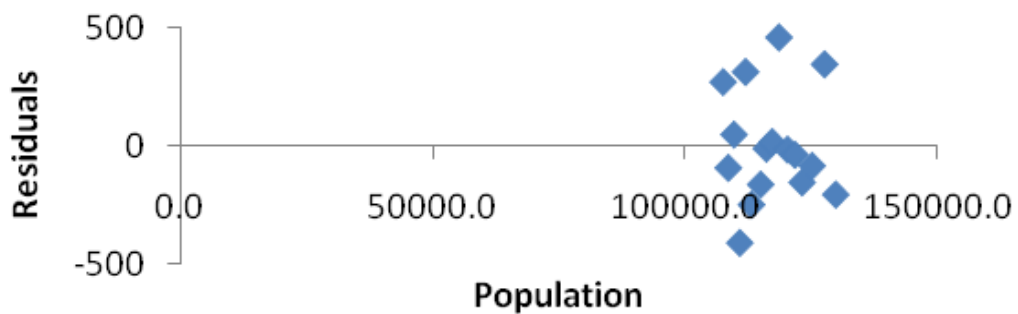
Unemployment Residual Plot



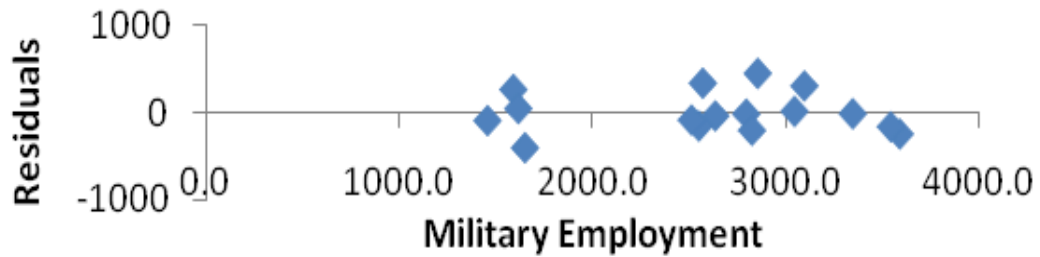
Year-Time Residual Plot



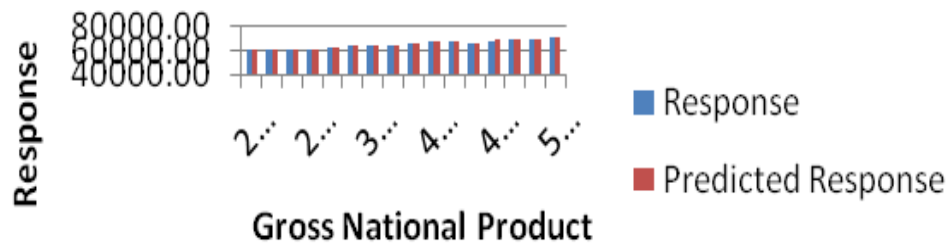
Population Residual Plot



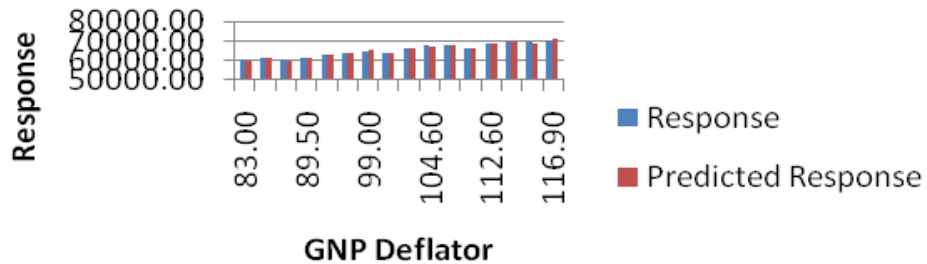
Military Employment Residual Plot



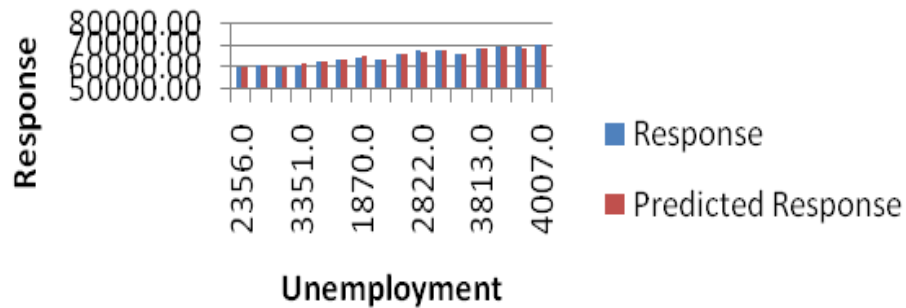
Gross National Product Line Fit Plot



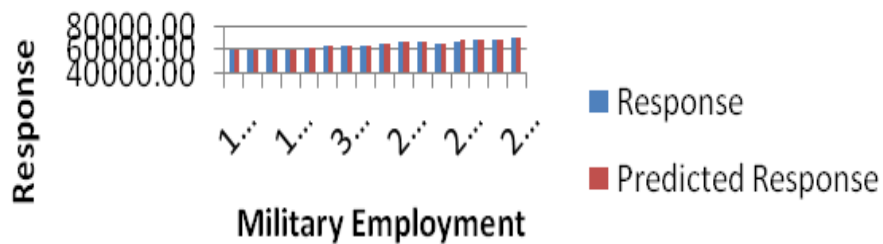
GNP Deflator Line Fit Plot



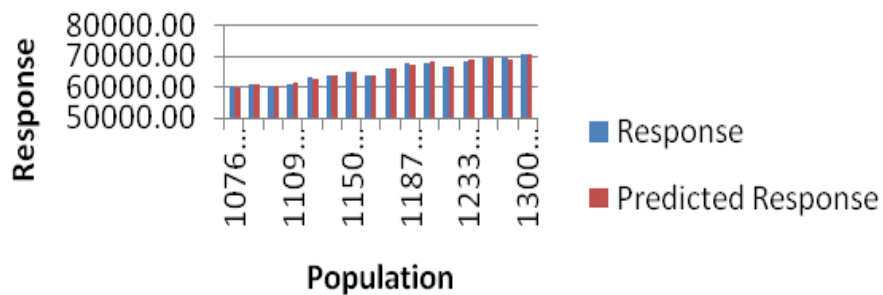
Unemployment Line Fit Plot



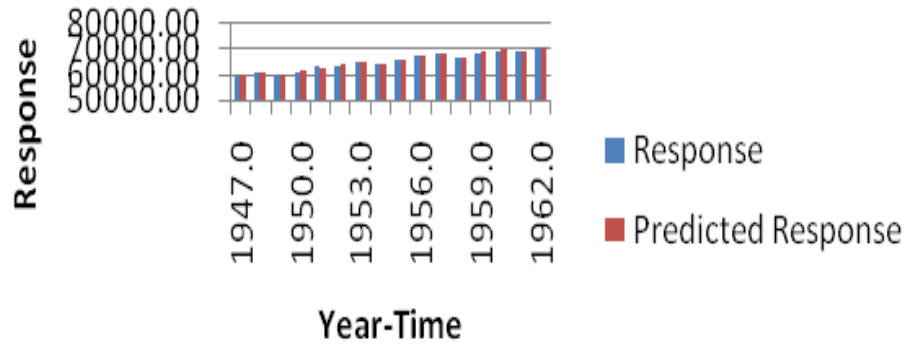
Military Employment Line Fit Plot



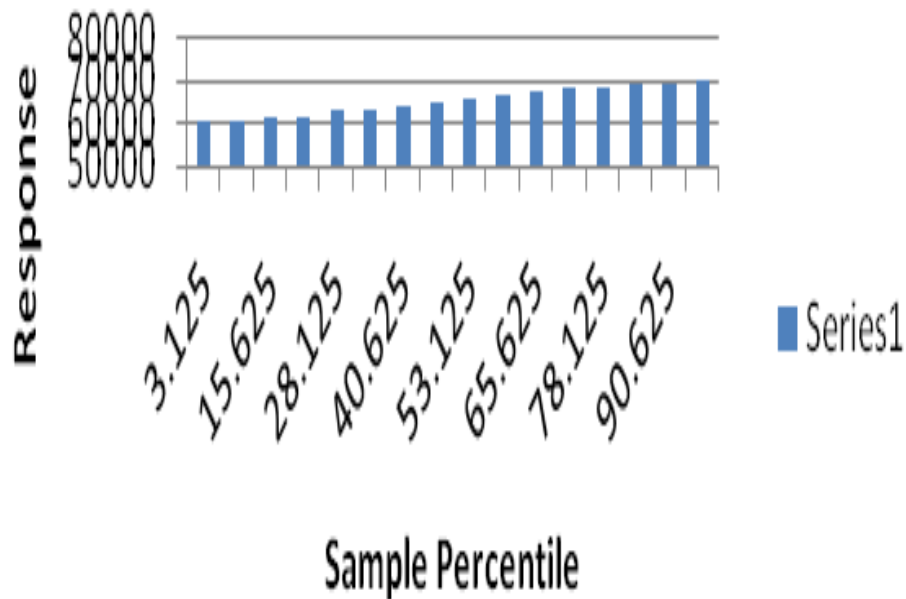
Population Line Fit Plot



Year-Time Line Fit Plot



Normal Probability Plot



BASIC PLOT DEFECTS

All of the plot charts are defective in that the central basic data cannot be enlarged to where the variations can be clearly seen. If the charts are expanded as given above, the expansion is to the titles and number values. The basic data section remains too small to really identify equation fit problems with specific variables.

RESIDUAL PLOTS

These plots are useful to see if there are some non-linear relationships between the listed X variable and the outcome Y variable. By resizing the title and the axis's names, resizing the chart and points, the chart could be shown. However as-is they are crude and require re-working to be included in any written analysis.

LINE FIT PLOTS

The line fit plots all minimize the data and exaggerate the titles, the X and Y coordinate names and the values. The charts, as output from the regression, fail to clarify the basic question, does this variable have an effect on the response? They can be useful when vertically expanded, to find the specific points that fail to fit the linear relationship with the specific variable. Normally these plots would not be used

NORMAL PROBABILITY PLOT

The above Excel Normal Probability Plot is **wrong** (see McCullough 2003)

The correct residual plot is as follows. It is somewhat different from the Minitab chart shown in Kutner, Nachtshein, Neter and Li (2005) and described on pages 110-112 of Kutner, Nachtshein, Neter and Li (2005).

A true normal probability plot compares the distribution of the residuals with respect to a normal distribution. It is a means of assessing whether the residuals are normally distributed, and hence provides validity to statistical tests of the model.

There are different ways to show the relationship of the actual residuals to normally distributed residuals. They all are based on the ranking of the actual residuals. Here n is the number of residuals, and k is the rank (from 1 to n) of a given residual. The basis in both cases is a conversion of the rank of a residual to a position value (a percentile).

1. Excel: The equation is: $pctl = 100 * (k-0.5) / n$
2. Applied Linear Statistical Models: The equation is: $pos = (k - 0.375) / (n + 0.25)$

For method 1, $pctl$ represents a probability as a percentile (0 to 100) and for method 2 pos represents a probability value (0 to 1.0). The X axis values in the Excel output Normal Probability Plot (shown above) are $pctl$ values.

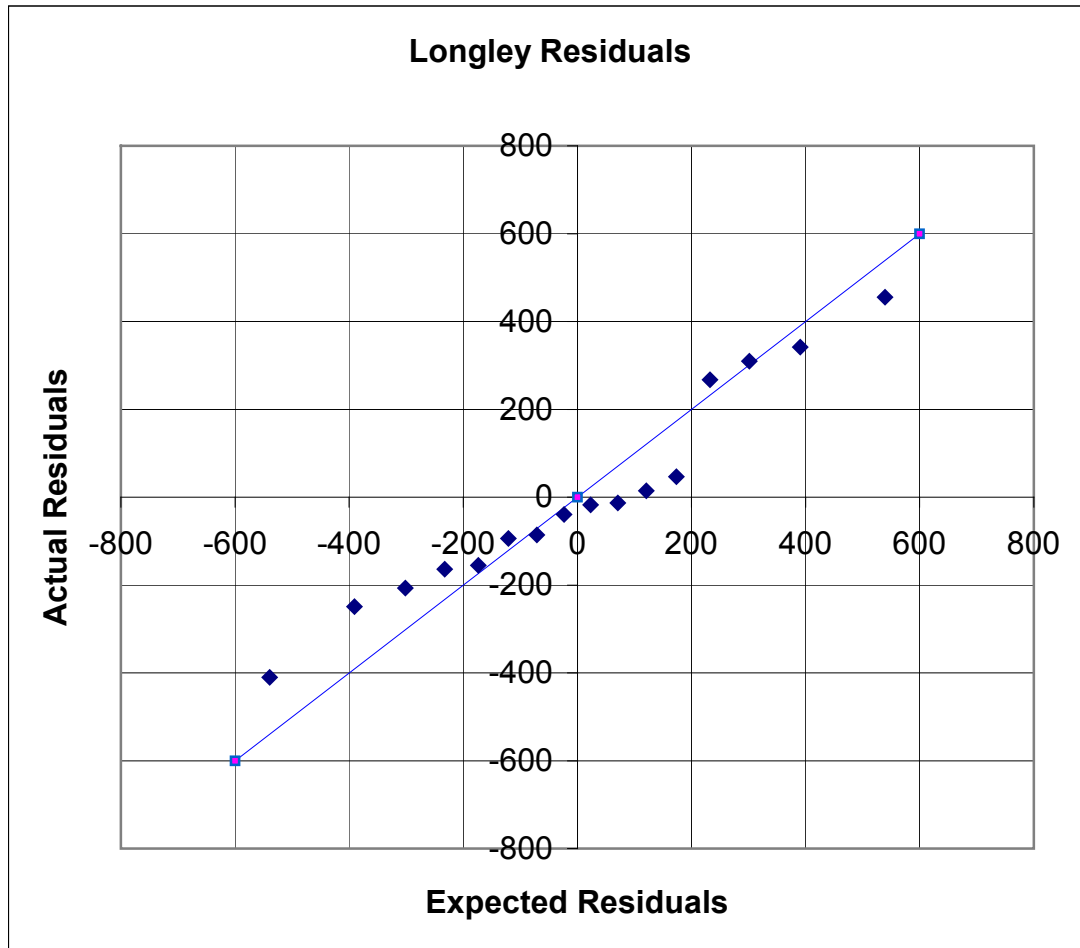
For method 2, the Excel NORMSDIST function can be used to convert pos to a normal distribution z value.

$$Z = \text{NORMSDIST}(pos)$$

Z here has both negative and positive values. Given Z then, a corresponding expected residual value (E) is calculated.

$$E = (\text{square root of the mean square regression error}) * Z$$

The chart for method 2 then is the actual residual (value) on the Y axis and the corresponding E value on the X axis. For the Longley data set the chart looks like this:



A chart showing actual percentiles versus expected percentiles could also be developed. However Kutner, Nachtshein, Neter and Li (2005) do not show this alternate form in their examples.

In all cases, an X-Y plot is the desired form. Going to a column chart (the Excel 2007 preference) is not a good visual representation of residuals.

If the points generally follow a straight line as shown above (on either expected-actual plot), then the residuals can be assumed to be basically normally distributed.