

XIV. THE ANALYSIS TOOL PAK DATA ANALYSIS ROUTINES

GENERAL INFORMATION

This is an add-in package which provides the following routines

Table 19-1: DATA ANALYSIS TOOL ROUTINES

| | |
|----|---|
| 1 | ANOVA: Single Factor |
| 2 | ANOVA: Two-Factor With Replication |
| 3 | ANOVA: Two-Factor Without Replication |
| 4 | Correlation |
| 5 | Covariance |
| 6 | Descriptive Statistics |
| 7 | Descriptive Statistics, Confidence Level |
| 8 | Exponential Smoothing |
| 9 | F-test: Two Sample for Variances |
| 10 | Fourier Analysis |
| 11 | Histogram |
| 12 | Moving Average |
| 13 | Random Number Generator |
| 14 | Rank and Percentage |
| 15 | Regression |
| 16 | Sampling |
| 17 | t-Test: Paired Two Sample for Means |
| 18 | t-Test: Two-sample Assuming Equal Variance |
| 19 | t-Test: Two-sample Assuming Unequal Variances |
| 20 | z-Test: Two Sample for Means |

This is a program (as the Tool-Pak add-in) that was originally a developer package for Excel 4.0 from Grey Matter International Inc, Cambridge, MA. KBA 213939 is a list of the books that were used as the sources of the equations that were programmed. The programming language was a vba like macro language for Excel 4.0. However KBA 829215 shows that it is possible to change this macro language to vba and to incorporate Excel functions.

The only known changes were the migration of a set of functions to the basic native VBA directory (KBA 912719) and changes to the ANOVA routines described in KBA

829215. All the ATP functions, including those that operate with complex numbers are now accessible as direct VBA callable functions. There are small differences in returned values between the original and revised version. KBA 912719 describes these differences.

ROUTINES

ANOVA: SINGLE FACTOR

Returns a range of values. This routine was tested and evaluated in section 6. There is no chart output, only a table of values.

ANOVA: TWO-FACTOR WITH REPLICATION

Returns a range of values. This routine was tested and evaluated in section 6. There is no chart output, only a table of values.

ANOVA: TWO-FACTOR WITHOUT REPLICATION

Returns a range of values. This routine was tested and evaluated in section 6. There is no chart output, only a table of values.

CORRELATION

Returns a range of values. This routine was tested and evaluated in section 7. There is no chart output, only a table of values.

COVARIANCE

Returns a range of values. This routine was tested and evaluated in section 7. There is no chart output, only a table of values.

DESCRIPTIVE STATISTICS

Returns a column array of values. This routine was tested and evaluated in section 5. There is no chart output, only a table of values.

EXPONENTIAL SMOOTHING

This routine and the results of tests on it are covered in section 12. Forecast.

F-TEST: TWO SAMPLE FOR VARIANCES

A test of significance. Returns a table of values. This routine was tested and evaluated in section 17. There is no chart output, only a table of values.

FOURIER ANALYSIS

This is not a statistical test or statistical analysis. No tests were run on this routine. It returns a column of complex numbers, which then require other complex functions to return magnitude and angle (phase).

The XNUMBERS free routine is an array function that will give the Fourier analysis magnitude in db and the angle (phase) in degrees as two separate columns.

HISTOGRAM

THE ROUTINE

This Data Analysis routine takes column data from two columns on a worksheet and generates a preset histogram.

Excel does have limitations. It takes a lot of buttons to obtain acceptable histograms from the Data Analysis → Histogram → ‘buttons’ (Dretzke 2003, pages 35 to 40, 12 buttons to complete a histogram). For a beginning student this is a formidable task. To fix the tic mark problem, it takes another 3 buttons.

For the Cultural Literacy data (Dretzke and Heilman 1998, page 31) in column A and the specified bin sizes in column B

Column A: Cultural Literacy, 24, 35, 22, 29, 17, 29, 26, 16, 26, 20, 28, 19, 25, 19, 22, 24, 26, 24, 23, 24, 28, 22, 21, 22, 23, 31, 28, 25, 19, 25, 24, 23, 21, 21, 26, 23, 20, 17, 22, 15, 21, 23, 30, 24, 30. 18, 22

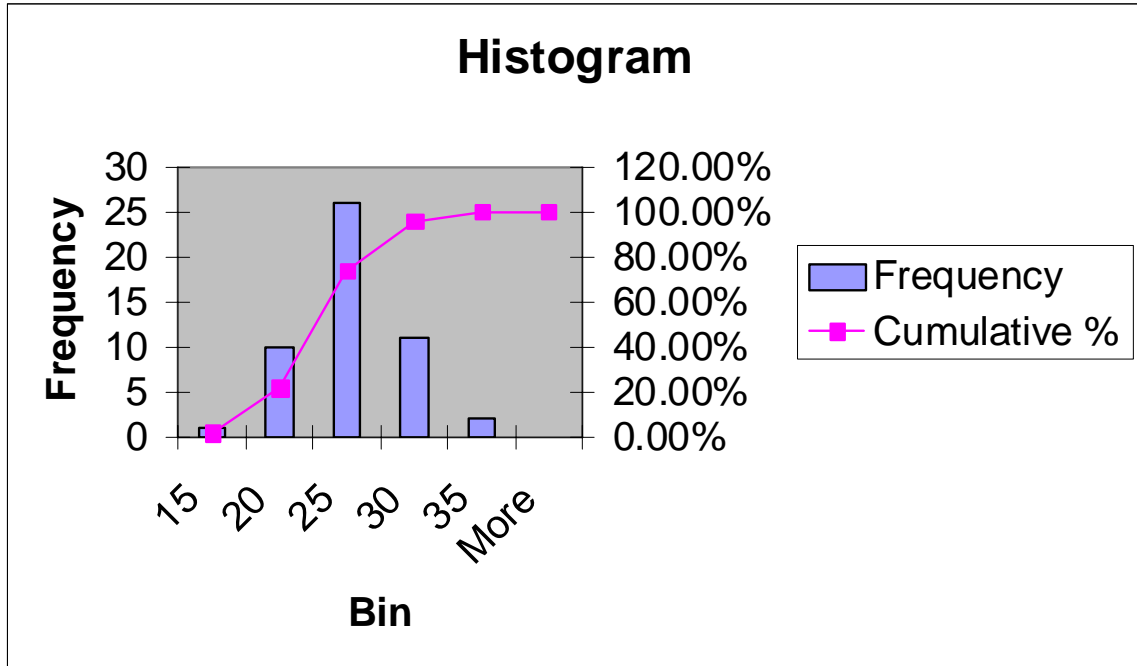
Column B: bins, 15, 20, 25, 30, 35

The routine generates a data table

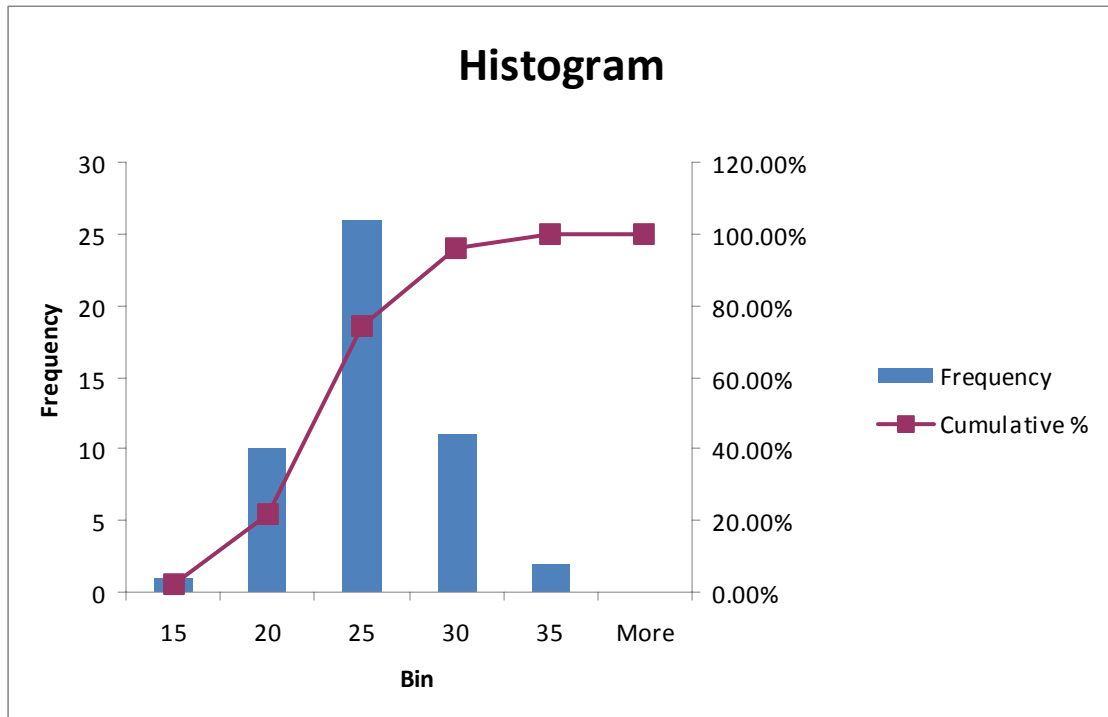
| <i>Bin</i> | <i>Frequency</i> | <i>Cumulative</i> <i>%</i> |
|------------|------------------|-------------------------------|
| 15 | 1 | 2.00% |
| 20 | 10 | 22.00% |
| 25 | 26 | 74.00% |
| 30 | 11 | 96.00% |
| 35 | 2 | 100.00% |
| More | 0 | 100.00% |

THE CHARTS

EXCEL 2000 AND 2003



EXCEL 2007



Note that the 2007 chart has a moving and editing problem

THE TIC MARK PROBLEM

In all these charts, the value at the tic marks on the X axis is **NEVER** identified. The default charts always put the number in-between tic marks, so that the reader does not know if the number corresponds to the right or left tic mark

The Data Analysis histogram default 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 2000, 2003 and 2007.

MOVING AVERAGE

THE ROUTINE

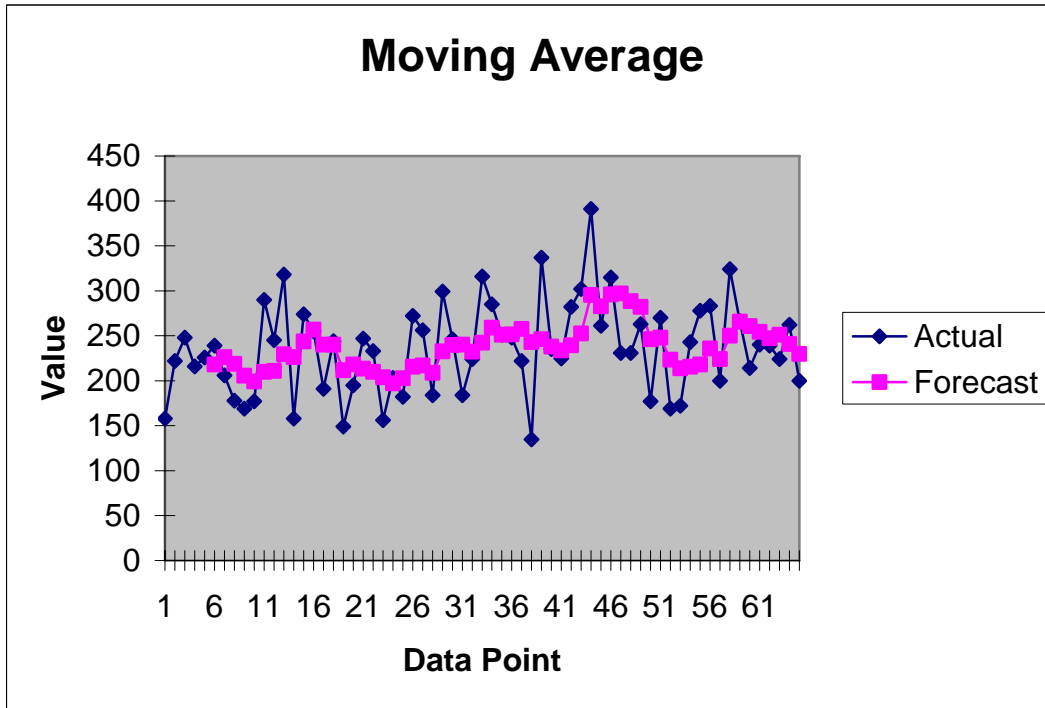
Moving average is considered in this situation as a statistical analysis of data. Both means and error values are calculated. The equations actually used for values are in the output cells.

The moving average values are correct, based on the data in table 2-2 of Montgomery and Johnson (1976). Montgomery and Johnson (1976) do not give equations for values of the error.

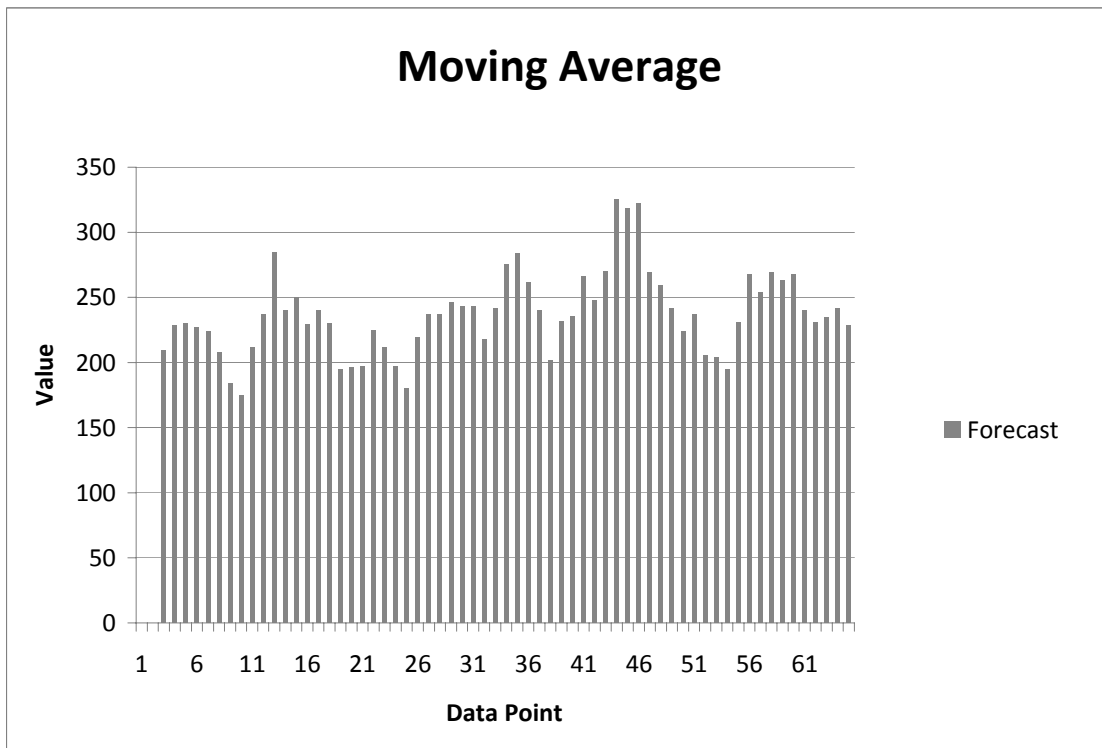
There is an error in the equation used to calculate the error values. The error should be a standard deviation and the sum of the squares (SUMXMY2) should be divided by N-1, not N.

THE OUTPUT CHARTS

EXCEL 2000 AND 2003



EXCEL 2007



Note that the 2007 version charts only the forecast. A new chart would have to be built from scratch here, to show both data and the moving average.

RANDOM NUMBER GENERATOR

This set of separate functions are all defective, and should not be used. See KBA 829208. The underlying RNG is defective. This is covered fully in section 16.

RANK AND PERCENTAGE

This routine has been tested and is covered in Note N.

REGRESSION

There are numerical and conceptual faults in the Data Analysis regression routine output that occur in Excel 2000, 2003 and 2007

The Data Analysis routine here uses the native Excel functions to build an expanded table of regression information.

TOOLS → DATA ANALYSIS → REGRESSION → OK

An Input Data Box appears with the following Inputs

Input Y Range

Input X Range

Labels

Constant is zero

Confidence Level

%

Output Options

Output Range

New Worksheet Ply

New Workbook

Residuals

Residuals

Residual Plots

Standardized Residuals Line Fit Plots

Normal Probability

Normal Probability Plots

Note: symbolizes an input, either as a check or data and symbolizes a selection box

RESIDUALS

When Residuals is checked, the regression output sheet gives three columns of data below the regression output information block. The first column is observation and is just a sequence number referring the sequence of input data values. The second column is the predicted Y value, and the third column is the difference between the actual data Y value and the calculated Y value, based on the regression coefficients listed. The respective Y values are not shown. This is a fault. The residuals list is entirely dependent of the order

of the original Y values, which is lost, because there is no link backwards to the source data. If the data set is sorted, the link is lost.

RESIDUAL PLOTS

When residual plots is checked, the output regression gives a column of values titled “Standardized Residuals”, which come from the division of the residual by the standard deviation of the residuals.

Residuals (as Y values) are plotted against values of each of the X variables. These charts would only convey information when there was a gross misfit (or error) with respect to one or more values of the indicated variable. The gross misfit (large residual) would show up on each of the plots as a noticeable vertical separation between the observed Y value and the predicted Y value. If the large residuals tended to be found at either the low or high ends of the range of one x variable (for a specific chart), then some useful information can be inferred. If there was no pattern, one would not be able from these charts to determine which X variable was a major contributor, or if all of them were significant contributors. One would not be able to infer model improvements from adding products of variables terms from these plots.

The conclusion is that for multiple data, these charts do not convey any useful information if the model is correct. For single X variables, the information is useful.

This routine uses the LINEST function and other testable Excel functions to output a tables of values. LINEST is covered in section 9.

To show this and the faults with the charts an actual regression using the NIST Longley is shown:

Table 17-1: The NIST Longley Data Set

| Response | GNP Deflator | Gross National Product | Unemployment | Military Employment | Population | Year-Time |
|----------|--------------|------------------------|--------------|---------------------|------------|-----------|
| 60323.00 | 83.00 | 234289.0 | 2356.0 | 1590.0 | 107608.0 | 1947.0 |
| 61122.00 | 88.50 | 259426.0 | 2325.0 | 1456.0 | 108632.0 | 1948.0 |
| 60171.00 | 88.20 | 258054.0 | 3682.0 | 1616.0 | 109773.0 | 1949.0 |
| 61187.00 | 89.50 | 284599.0 | 3351.0 | 1650.0 | 110929.0 | 1950.0 |
| 63221.00 | 96.20 | 328975.0 | 2099.0 | 3099.0 | 112075.0 | 1951.0 |
| 63639.00 | 98.10 | 346999.0 | 1932.0 | 3594.0 | 113270.0 | 1952.0 |
| 64989.00 | 99.00 | 365385.0 | 1870.0 | 3547.0 | 115094.0 | 1953.0 |
| 63761.00 | 100.00 | 363112.0 | 3578.0 | 3350.0 | 116219.0 | 1954.0 |
| 66019.00 | 101.20 | 397469.0 | 2904.0 | 3048.0 | 117388.0 | 1955.0 |
| 67857.00 | 104.60 | 419180.0 | 2822.0 | 2857.0 | 118734.0 | 1956.0 |
| 68169.00 | 108.40 | 442769.0 | 2936.0 | 2798.0 | 120445.0 | 1957.0 |
| 66513.00 | 110.80 | 444546.0 | 4681.0 | 2637.0 | 121950.0 | 1958.0 |
| 68655.00 | 112.60 | 482704.0 | 3813.0 | 2552.0 | 123366.0 | 1959.0 |
| 69564.00 | 114.20 | 502601.0 | 3931.0 | 2514.0 | 125368.0 | 1960.0 |
| 69331.00 | 115.70 | 518173.0 | 4806.0 | 2572.0 | 127852.0 | 1961.0 |
| 70551.00 | 116.90 | 554894.0 | 4007.0 | 2827.0 | 130081.0 | 1962.0 |

Table 17-2: Outputs

| Regression Statistics | |
|-----------------------|-------------|
| Multiple R | 0.997736942 |
| R Square | 0.995479005 |
| Adjusted R Square | 0.992465008 |
| Standard Error | 304.8540736 |
| Observations | 16 |

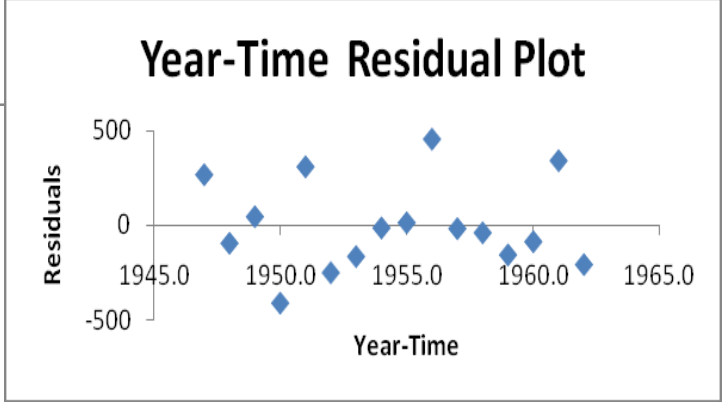
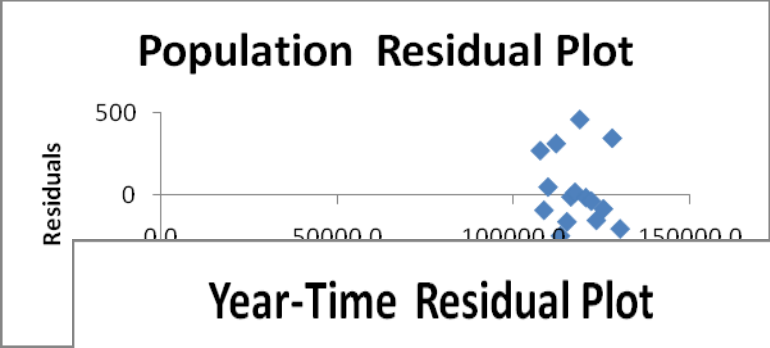
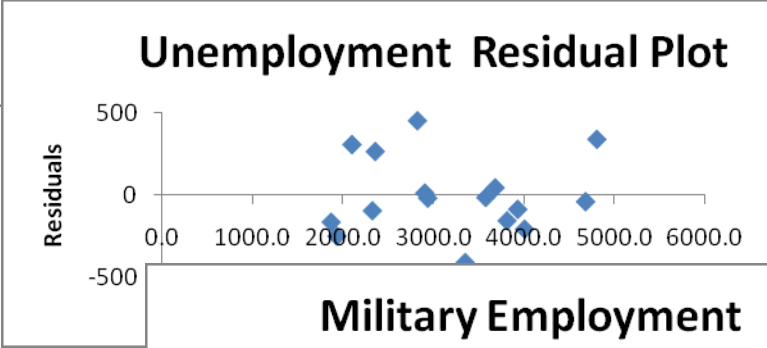
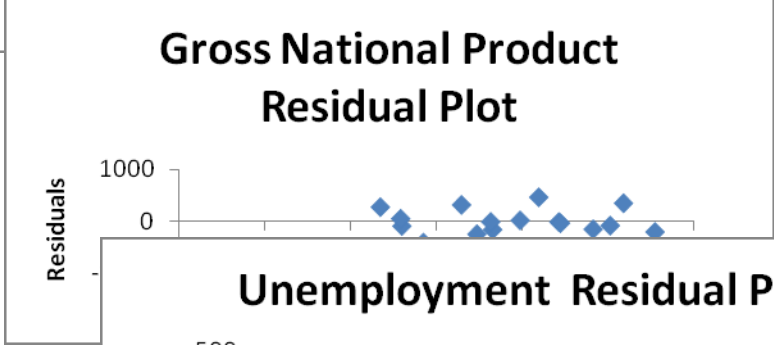
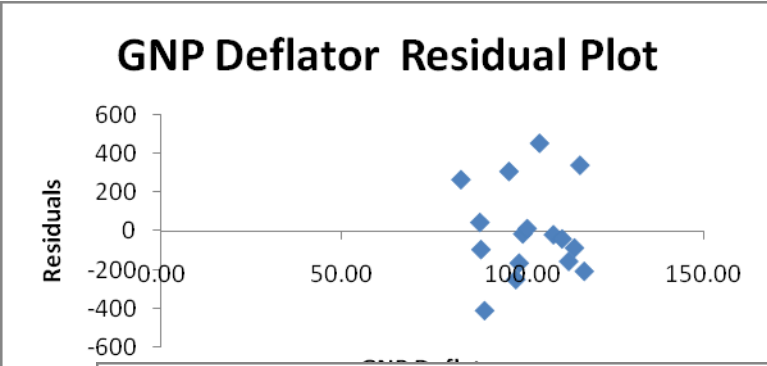
| ANOVA | | | | | |
|------------|----|-------------|-------------|-------------|----------------|
| | df | SS | MS | F | Significance F |
| Regression | 6 | 184172401.9 | 30695400.32 | 330.2853392 | 4.98403E-10 |
| Residual | 9 | 836424.0555 | 92936.00617 | | |
| Total | 15 | 185008826 | | | |

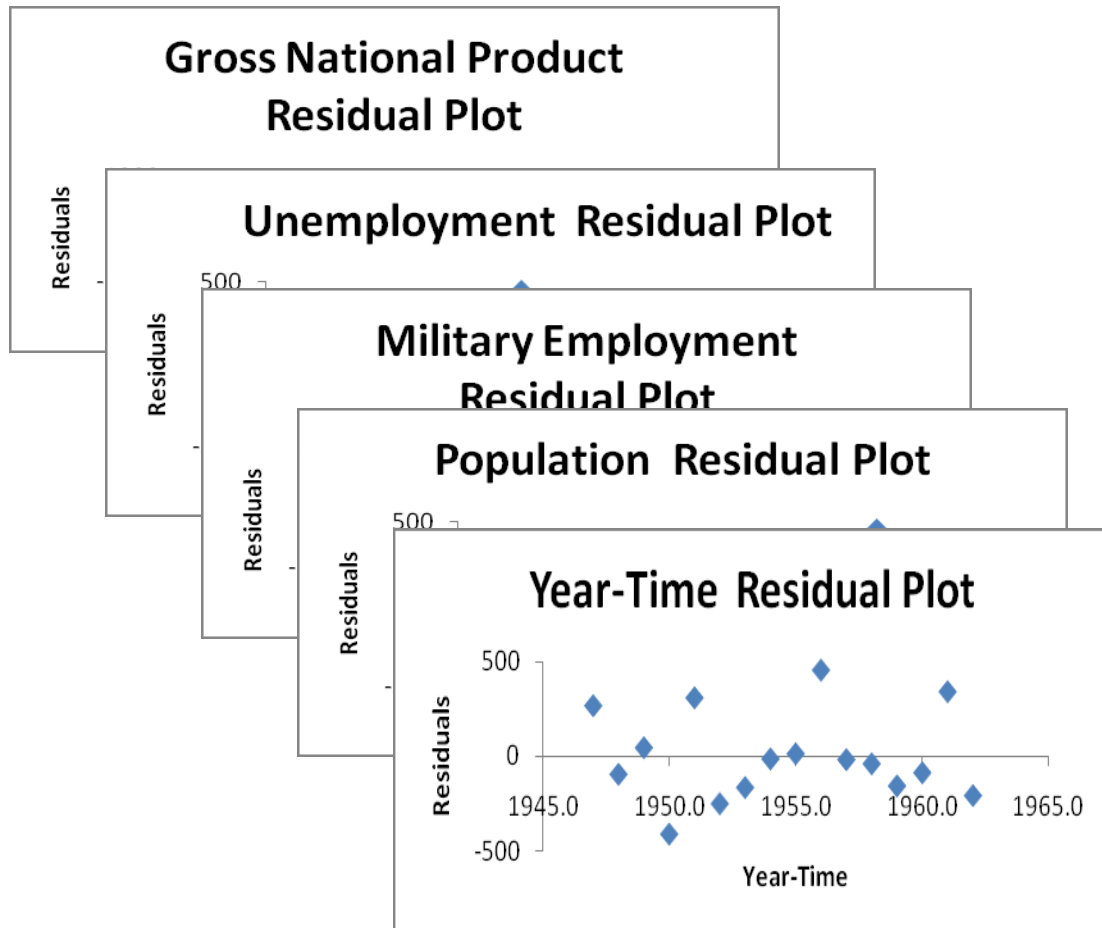
| | Coefficients | Standard Error | t Stat | P-value |
|------------------------|--------------|----------------|--------------|-------------|
| Intercept | -3482258.635 | 890420.3836 | -3.910802918 | 0.003560404 |
| GNP Deflator | 15.06187227 | 84.91492577 | 0.177376028 | 0.863140833 |
| Gross National Product | -0.035819179 | 0.033491008 | -1.069516317 | 0.312681061 |
| Unemployment | -2.020229804 | 0.488399682 | -4.136427356 | 0.002535092 |
| Military Employment | -1.033226867 | 0.214274163 | -4.82198531 | 0.000944367 |
| Population | -0.051104106 | 0.2260732 | -0.226051145 | 0.826211796 |
| Year-Time | 1829.151465 | 455.4784991 | 4.015889813 | 0.003036803 |

| | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
|------------------------|--------------|--------------|--------------|--------------|
| Intercept | -5496529.479 | -1467987.79 | -5496529.479 | -1467987.79 |
| GNP Deflator | -177.0290349 | 207.1527794 | -177.0290349 | 207.1527794 |
| Gross National Product | -0.111581102 | 0.039942744 | -0.111581102 | 0.039942744 |
| Unemployment | -3.12506664 | -0.915392968 | -3.12506664 | -0.915392968 |
| Military Employment | -1.517948699 | -0.548505035 | -1.517948699 | -0.548505035 |
| Population | -0.562517213 | 0.460309002 | -0.562517213 | 0.460309002 |
| Year-Time | 798.7875174 | 2859.515412 | 798.7875174 | 2859.515412 |

All of the above values are correct. They match the NIST reference values where they are given on the NIST data sets.

A series of overlaid charts are generated based on the regression equation calculated y values, that show the X-Y data points and the fitted line. The following are the charts from Excel 2007.





LINE FIT PLOTS

Calculated predicted Y values (\hat{Y}) are plotted versus each of the X variables. Does not show much. The comments under “Residual Plots” also apply here. Actually they represent a conceptual duplication of the residual plots.

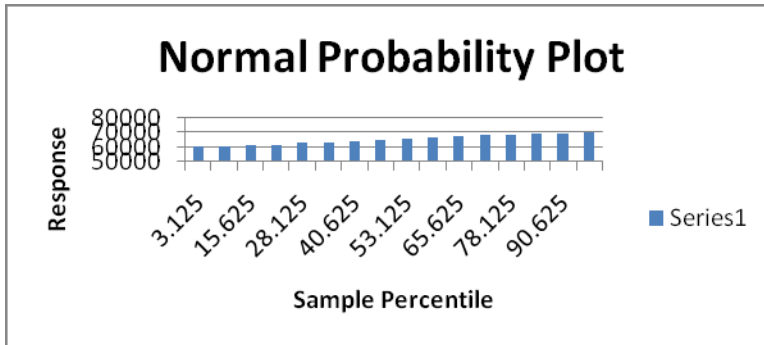
The inherent problem is that these charts can’t be moved, so that they can be seen. The only recourse is to copy, starting at the last one and paste it somewhere else on the worksheet. Then it can be resized for viewing.

NORMAL PROBABILITY PLOT

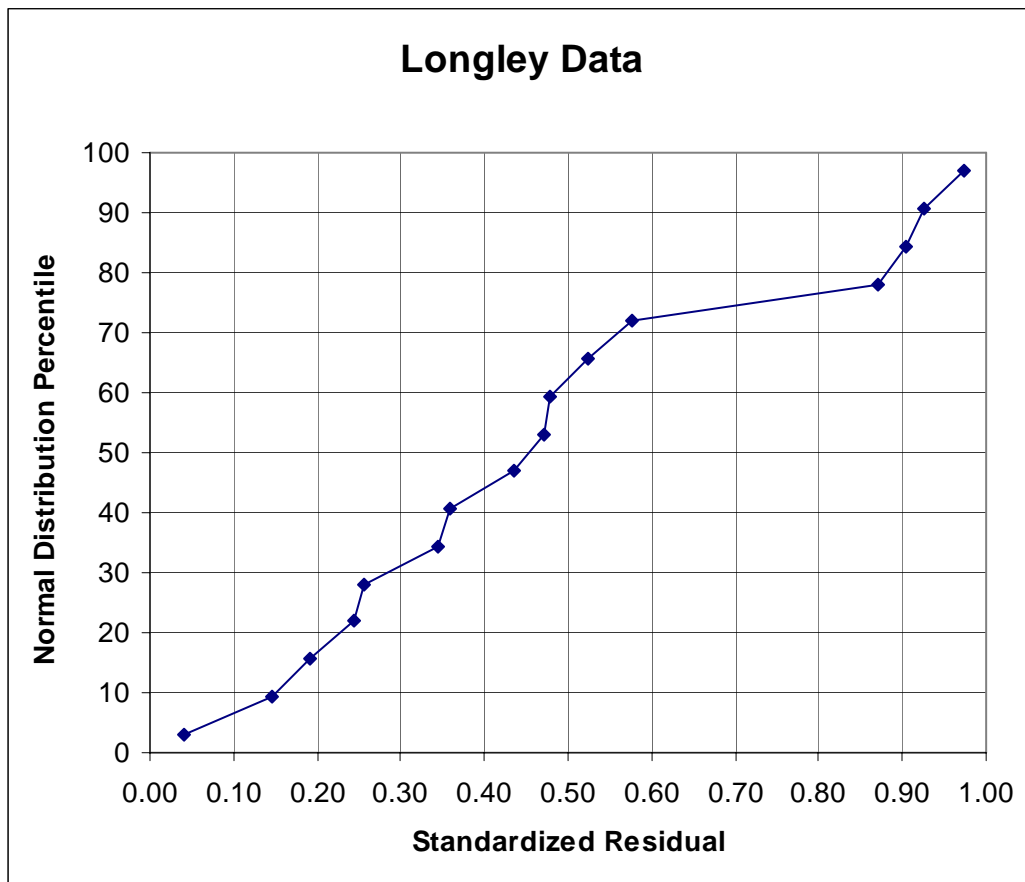
When this is checked, Regression outputs two columns of values under the heading “PROBABILITY PLOT” and two individual column headings of “Percentile” and “Response”. The resulting “Normal Probability Plot” is a chart of these two columns, with the X as the “Percentile” values and Y as the “Response” values.

“Response” values are a simple sort in ascending order of the actual Y data used as the input to the regression. The “Percentile” column is just the ranking of each sorted value, in relationship to the midpoint of an interval equal to $1/n$ in width. In the example, the

“Percentile” values are the midpoints of intervals 1/16 or 6.25% wide. The use of the word “Normal Probability Plot” here is misleading, since there is absolutely no relationship to a normal distribution plot.



The correct normal plot is a X-Y chart generated from two data columns, a Y column representing the sorted residuals, and an X column derived from the NORMSDIST function of the corresponding standardized residual. For the Longley data, this would be the correct normal probability plot.



CONCLUSIONS

The standard built-in regression charts have limited usefulness when multiple data is entered. Only for single X variable regressions, are the charts useful for determining outliers, and model misfits. The Normal Probability Plot conveys **no** useful statistical information.

SAMPLING

This routine was not tested, since the way it operates, does not allow for testing as a random number generator. Since it uses the defective Data Analysis random number generator, use of the function should be avoided.

T-TEST: PAIRED TWO SAMPLE FOR MEANS

This routine is covered and tested in section 17.

T-TEST: TWO-SAMPLE ASSUMING EQUAL VARIANCE

This routine is covered and tested in section 17.

T-TEST: TWO-SAMPLE ASSUMING UNEQUAL VARIANCES

This routine is covered and tested in section 17.

Z-TEST: TWO SAMPLE FOR MEANS

This routine is covered and tested in section 17.