

IX. LINEAR REGRESSION	2
EXCEL SUPPLIED REGRESSION TOOLS	2
WORKSHEET REGRESSION FUNCTIONS.....	2
SINGLE VALUE OUTPUT FUNCTIONS.....	2
ARRAY OUTPUT FUNCTIONS	3
WORKSHEET DATA ANALYSIS TOOL PAC REGRESSION ROUTINE	3
MULTIPLE UNIVARIATE REGRESSION	5
APPLICATIONS:.....	5
THE MULTICOLINEARITY ISSUE	5
SINGULARITY	6
THE CORE WORKSHEET ROUTINE – LINEST	6
THE LINEST FUNCTION IN EXCEL 2000 AND EARLIER VERSIONS.....	6
NORMAL DATA	6
SINGULAR DATA	7
REGRESSION THROUGH THE ORIGIN.....	7
THE LINEST FUNCTION IN EXCEL 2003 AND 2007	7
NORMAL DATA	7
SINGULAR DATA	9
POLYNOMIAL REGRESSION	9
TRANSFORMATIONS TO A LINEAR SYSTEM.....	9
REPORTED PROBLEMS AND FAULTS:.....	10
GENERAL PROBLEMS:.....	10
EXCEL 2000.....	10
EXCEL 2003 SPECIFIC:	12
EXCEL 2007 SPECIFIC:	12
TESTS ON REGRESSION FUNCTIONS AND ROUTINES	12
ASSUMPTIONS.....	12
TESTS CONDUCTED:.....	12
TEST RESULTS:.....	13
EXCEL 2000.....	15
REGRESSION OUTPUT VALUES	15
TEST FOR INACCURATE MULTIPLE RESULTS:	15
CONCLUSIONS:	15

EXCEL 2003 AND EXCEL 2007	16
REGRESSION THROUGH THE ORIGIN.....	16
DATA ANALYSIS REGRESSION ROUTINE OUTPUT FAULTS.....	16
REGRESSION OUTPUT VALUES	16
REPORTED FAULTS AND ERRORS	16
MISSING DATA	17
THE MULTICOLINEARITY ISSUE	17
NIST DATA SETS	17
SINGULARITY	19
DATA ANALYSIS TOOL PAC REGRESSION ROUTINE	22

IX. LINEAR REGRESSION

EXCEL SUPPLIED REGRESSION TOOLS

WORKSHEET REGRESSION FUNCTIONS

SINGLE VALUE OUTPUT FUNCTIONS

FORECAST – Given an X range of data and a Y range of data, returns a predicted Y value for a given X value.

INTERCEPT – Returns the intercept of a least mean squares straight line through an imputed range of X and Y values.

SLOPE – Returns the slope of a least mean squares straight line through an imputed range of X and Y values

STEYX - Returns the standard error of the residuals from a least mean squares straight line through an imputed range of X and Y values

INDEX(LINEST(range of y values, range of x values),1) – Returns the slope or b value of a linear regression on the entered x and y values.

INDEX(LINEST(range of y values, range of x values),2) – Returns the intercept or a value of a linear regression on the entered x and y values.

INDEX(LOGEST(range of y values, range of x values),1) – Returns the base value (b, equation 6 above or m value as listed in Help) of an exponential regression on the entered x and y values.

INDEX(LOGEST(range of y values, range of x values),2) – Returns the multiplier (a in equation 6 above or b value as listed in Help) of an exponential regression on the entered x and y values.

SUM({a,b}*(x,1)) – Returns the results of the equation $= a * x + b$. {a,b} can also be entered as (LINEST(range of y values, range of x values))

ARRAY OUTPUT FUNCTIONS

GENERAL – The limit on the number of variables for these following functions is 16, which has not changed since Excel 4. Multiple X variables (up to 16 x variables) can be fitted, and by transformations, polynomial and some non-linear models can be fitted. For Excel 2007, the 16 independent variable limit still is in effect.

LINEST – Fits a least mean squares straight line to a vector of Y values and a range of X values. Returns a block of coefficient values, an intercept value, and regression statistics in a certain order. Does not include labels for each cell. You have to use help to assist in determining what each cell with numbers represents.

TREND – Fits a linear least mean squares line to a given range of X and Y values, and returns a range of predicted new Y values for an input range of new X values based on the fitted line. Accepts multiple X ranges.

GROWTH – Fits an exponential curve to a given range of X and Y values, and returns a range of predicted new Y values for an input range of new X values based on the fitted curve. Accepts multiple X ranges. Fits linear regression to the logs of the X values and transforms back LOGEST provides the fitted constant values. GROWTH is essentially a “FORECAST” of a LOGEST fit. As McCullough and Heiser (2008) and Hesse (1983 and 2006) point out, this is a totally WRONG computation since it is not a true fit or projection of the non-linear exponential function.

LOGEST – Fits an exponential function to a vector of Y values and a range of X values and returns a vector of exponential values, a coefficient, and regression statistics in a certain order. It is the application of LINEST on the logs of Y and X values, and conversion to an exponential system. The Help write up is essential to figure out how to properly use LOGEST. As McCullough and Heiser (2008) and Hesse (1983 and 2006) point out, this is a totally WRONG computation since it is not a true projection of the non-linear exponential function.

WORKSHEET DATA ANALYSIS TOOL PAC REGRESSION ROUTINE

The Data Analysis Tool Pac provides a single linear regression routine that gives an output in the form of a report on a worksheet.

REGRESSION - Fits a least-mean-squares straight line to a vector of Y values and a range of X values. Multivariate (up to 16 x variables) can be fitted, and by transformations, polynomial and some non-linear models can be fitted. Returns a block of cells with the results and labels for all the cells.

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

The output is in the specified output block, and looks this;

Table 9-1: Typical Data Analysis Regression Output (The Longley Data Set)

SUMMARY OUTPUT

<i>Regression Statistics</i>					
Multiple R	0.997736942				
R Square	0.995479005				
Adjusted R Square	0.992465008				
Standard Error	304.8540736				
Observations	16				

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

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>
Intercept	-3482258.635	890420.3836	-3.910802918	0.003560404	-5496529.479
GNP Deflator	15.06187227	84.91492577	0.177376028	0.863140833	-177.0290349
Gross National Product	-0.035819179	0.033491008	-1.069516317	0.312681061	-0.111581102
Unemployment	-2.020229804	0.488399682	-4.136427356	0.002535092	-3.12506664
Military Employment	-1.033226867	0.214274163	-4.82198531	0.000944367	-1.517948699
Population	-0.051104106	0.2260732	-0.226051145	0.826211796	-0.562517213
Year-Time	1829.151465	455.4784991	4.015889813	0.003036803	798.7875174

Table 9-1 (Continued)

Upper 95%	Lower 95.0%	Upper 95.0%
-1467987.79	-5496529.479	-1467987.79
207.1527794	-177.0290349	207.1527794
0.039942744	-0.111581102	0.039942744
-0.915392968	-3.12506664	-0.915392968
-0.548505035	-1.517948699	-0.548505035
0.460309002	-0.562517213	0.460309002
2859.515412	798.7875174	2859.515412

The Data Analysis Regression provides all the information that normally is needed to evaluate the regression results.

MULTIPLE UNIVARIATE REGRESSION

APPLICATIONS:

This includes polynomial fits of single variables, where the power terms are generated as separate variables. Given the scope of Excel, it is best to take a simple approach to multiple regression. Excel lacks the tools to properly evaluate these more complex fits. Also to be recognized is that the more complex fits may fit the data very well within the range of the data, but give totally wrong results when predictions are made using variable values beyond the range of the generating data. This is especially true of polynomial fits.

In Excel 2000, for polynomial fits it is best to center the X and Y data, then derive the power terms as additional variables from the centered X values. This was how a good solution to the Filip data set was obtained. Note U, Polynomial Regression gives Filip analysis outputs and provides some additional information. It also provides directions for converting the centered coefficients back to direct coefficients. In Excel 2003 this does not need to be done. However this is some very limited data, which suggests that by doing the above centering in Excel 2003, the resulting coefficient values have about a 0.5 to 1.0 LRE value increase.

In Excel 2000 for accuracy, it is best to center the data after any transformations, since the $X'X$ matrix is otherwise dominated by the squares of the absolute values, and will result in inaccurate results. Note R, Linear Regression again may provide some assistance in returning back to the original coordinate system.

One insurmountable problem with Excel is that multiple regression is restricted to a maximum of 16 variables. Under Excel 2000 there may be reasons due to matrix inversion accuracies. For Excel 2003 this restriction should be eliminated.

THE MULTICOLINEARITY ISSUE

This is where two or more variables are highly correlated. When this occurs, the regression coefficient values are likely to be in great error. This is generally a fault of the data itself, not of the algorithm or regression computation itself.

This is frequently stated in the references as a problem, and the reason why something did not work. Section 8 discusses this problem. Tests on the Excel regression functions and routines given multicollinear data are presented below.

SINGULARITY

This is where two or more variables are so closely related or have a linear relationship with no error. The regression algorithm completely fails or is not able to come up with valid coefficient values, which also is a failure. Tests on the Excel regression functions and routines when the data has singularity are presented below

THE CORE WORKSHEET ROUTINE – LINEST

LINEST is the core regression solution algorithm for all the Excel regression tools except the graphics trendline tool. The graphics trendline module-routine is a separate module that does not use the Excel LINEST function supplied with the given version.

All of the other above functions and routines has as its kernel, the LINEST function, which returns an array of values. For those functions returning a single value, the function selects the appropriate single value from an internal LINEST output array. Table 9-2 shows what the LINEST outputs are.

Table 9-2 The LINEST Output Array

Row	Col 1	Col 2	Col 3	Col m
1	Coefficient n	Coefficient n-1	Coefficient n-2	Intercept
2	Std. Error Coefficient n	Std. Error Coefficient n-1	Std. Error Coefficient n	Std. Error of Intercept
3	Coefficient of Determination, r^2	Std. Error of a calculated Y value	#N/A	#N/A
4	F value	Degrees of Freedom for the F test	#N/A	#N/A
5	Sum Squares of Regression	Sum of Squares of Residuals	#N/A	#N/A

THE LINEST FUNCTION IN EXCEL 2000 AND EARLIER VERSIONS

NORMAL DATA

Details on the LINEST algorithm are not known. Microsoft in KBA 828533 and 829249 has stated that the “normal equations” were used. KBA gives the source of LINEST as Hemmerle 1967. Based the behavior of the function and information in some of the older KBAs, it appears to be by the normal equations, expressed as matrices, $b = (X'X)^{-1} (X'Y)$ where b is the vector of regression coefficients, X is the independent variable value

matrix and Y is the dependent variable value vector. From the normal equations, the standard errors of the coefficients can be easily calculated.

SINGULAR DATA

When there is singularity in the data set, LINEST will completely fail, which is good, in the sense that singularity is very clearly indicated.

REGRESSION THROUGH THE ORIGIN

There is no intercept parameter. Excel allows for this type of linear regression, and calculates the correct coefficients and standard error. However in Excel 2000 it incorrectly computes R squared, the correlations, degrees of freedom and F test values. The internal coding is in error. See Note S.

THE LINEST FUNCTION IN EXCEL 2003 AND 2007

NORMAL DATA

KBA 828533 describes the new LINEST algorithm in general terms as a QR decomposition. It is done on the centered X matrix, which improves the accuracy of the resulting computations. "QR Decomposition performs a sequence of orthogonal linear transformations on the X matrix." An example is given in a spreadsheet on how the new algorithm handles singularity.

In general, the term QR refers to a structural form, where $QX = R$ where Q, R and X are matrices and R is upper triangular (Stewart, 1995) In general the following are QR methods.

QR decompositions, using Householder transformations

QR Gram-Schmidt Algorithm, Classical or Modified

In KBA 828533, the specific method used is described, and illustrated by operations on an Excel spreadsheet. A name for it is not given. The method described however and illustrated is a Householder triangularization method with column interchanges for stability. The matrix identified as P in the KBA is the Householder transformation matrix. Stewart (1995) uses the term H for it. The method in the KBA is not the same as that in Stewart (1995), p263-265, and the math is somewhat different. The generation of the transformation vector (V in the KBA and u in Stewart (algorithm 1.1) is done differently.)

V represents a column vector after centralizing the values.

$$P = I - (2 / V^T * V) * (V * V^T) \quad (\text{Stewart: } H = I - u * u^T)$$

I is the identity matrix (1's on the diagonal, 0's elsewhere)

$2 / V^T * V$ is a constant

P is the Householder transformation matrix

$$P \dots P(X) = R$$

$P \dots P$ represents successive pre-multiplications on X

X is the n x p matrix of x values

R is the desired upper triangular matrix

The coefficients come from $R \times b = Y$, and the coefficient values (b) are obtained by back-substitution. This is shown in the lower part of the Excel spreadsheet in the KBA. In the actual algorithm, it is a little more complicated since the column interchanges have to be taken into account, and there is built in logic to identify coefficients that are from singular columns. The singularity logic deletes singular variables and resets degrees of freedom to fit the new rank of the X matrix.

It is not clear why Microsoft chose this method. The modified Gram-Schmidt is a simpler algorithm, and Stewart (1995) says it has superior numerical properties compared to the Householder method. I programmed a modified Gram-Schmidt (in VBA) using Stewart's algorithm 1.11. I found full agreement to 15 digits with Microsoft's KBA 828533 illustration spreadsheet R matrix values. However the data set is elementary.

There is a lot of commercial statistical software on the market. It appears from messages on sci.stat.math and in related discussions that the SVD decomposition is mainly used. One claim is that it is more accurate than the QR decomposition. However its biggest advantage is that it gives eigenvalues and eigenvectors which are very important these days. Not only to work data having colinearity problems but it provides direct values for Principle Components Analysis (PCA, PCR, and others).

KBA 828533 does not identify the algorithm to calculate coefficient standard errors. Stewart does not discuss this problem or give any clues on how to generate coefficient standard errors from the R matrix. One method is to generate a $X^{-1} * R$ matrix and use diagonal values to develop coefficient error values.

A theoretical error analysis of values of b in Stewart (1995) is not done directly on the b values. The error is given in terms of a new column vector as $x + e$ that represents a column vector that under exact arithmetic {IR} gives the same b value that {IF} arithmetic gives for the x vector.

$$\| e \|_2 / \| x \|_2 \leq \phi * \epsilon \quad (\text{Stewart, 1995, p.269})$$

Where

The left hand side represents the ratio of two vector norms

ϕ is some undefined slowly growing value (i.e. a constant)

ϵ is the value of the rounding error in {IF} (2E-16)

The theoretical analysis gives no help here. Under Excel 2003 (and Excel 2007), there is then no way to estimate the accuracy of the results from LINEST without having some standard test results. Stewart (1995), making some general observations says that generally the QR methods tend to give more accurate results but the difference from normal equations (the edge) is small. However the QR methods are much more stable. There are conditions where the normal equations are not even positive definite. (Stewart, 1995, p 318)

If any cell in the data input range is either empty or contains a non-numeric variant, LINEST will return an error message.

If there are zero's (or blanks) in the X or Y ranges, LINEST may return wrong values or fail to compute.

SINGULAR DATA

If the data set is singular, LINEST will calculate a set of coefficient values, arbitrarily setting one (or more) of the coefficient values (row 1) and the corresponding standard error value (row 2) to zero. There is no statement or other method that shows when singularity occurs. The outputs when singularity occurs may be misleading and have errors. This is discussed toward the end of this section, below.

POLYNOMIAL REGRESSION

Fitting a polynomial to data is usually done by creating new variables for each of the power terms and then by use of linear multivariate regression on all the variables.

The inherent problem is that the size of these power term x values are very much larger than the linear terms and significant errors in the resulting parameter values can occur. This is the reason why several regression programs failed on fitting the NIST Filip data. (Sequence 27, Tables 9-7 and 9-9 below). There is really no way around this problem, since the fault is fundamental to the limitations of the IEEE-754 floating point double.

If all the X data values are evenly spaced (apart), then an alternate method of solving for coefficient values can be used. Microsoft does not provide this known alternate method in Excel.

Also see note U on Polynomial Regression.

There are two solutions here:

- 1) Reposition the Y and the X variables that have linear and power terms about the X and Y means (or rounded to a convenient number). The power terms will be positive and negative. Then do a multivariate linear regression on these centered variables, or,
- 2) Go to xnumbers or Mathematica or other exact methods, where computations can be done with numbers having substantially more than the 15 decimal digits of the IEEE-754 standard. However calculating Y values from X values will require doing it in xnumbers or Mathematica, in order to retain accuracy when large numbers are subtracted.

If 1) is used, then any calculation must be based on new x value that represents the difference from the set central X value, and the resulting Y value requires that the central Y value be added to the equation calculated Y value.

In the case of NIST test data sequence 27, the new 2003 LINEST retains sufficient accuracy so that method 1 does not have to be applied. The old 2000 LINEST does not have the accuracy, and therefore method 1 has to be used.

TRANSFORMATIONS TO A LINEAR SYSTEM

This is the area where the equation is transformed to a linear system, and LINEST is used to obtain parameter values. The LINEST parameter values of course have to be transformed back to the original system, and this may be a difficult task.

The primary area here is in the use of LOGEST (and GROWTH) to model non-linear exponential systems of equations. LOGEST was developed specially to be able to convert exponential equations to a linear system that could be solved using LINEST. However the errors in the reported constants may be substantial..

Fitting to a “least-mean-squares” is done on the transformed values, not of the original values, and hence the fit is truly not “a-least-mean-squares-fit” with reference to the original data. This in many cases gives a false feeling of fit adequacy and also can give wrong parameter values. This is a true fault in Excel, and is discussed below under Reported Problems and Faults.

REPORTED PROBLEMS AND FAULTS:

GENERAL PROBLEMS:

The following is a summary table of problems with linear regression for Excel 2000 and earlier versions. No faults have yet surfaced on the Excel 2003 or 2007 versions.

EXCEL 2000

Most of the faults found in the literature were specific to Excel 2000. It was common knowledge that linear regression in Excel 2000 was faulty. The comments and fixes in table 9-4 apply only to the Excel 97 and 2000 versions. The presumption here is that the user has only Excel 2000, and has not updated to 2003 or 2007.

Table 9-4: Excel Faults in Linear Regression (All on Excel 2000)

Application or Function	Problem	Source	Fix or Comments
Linear Regression	Accuracy	Cryer 2000	*Centering X and Y data will help. See KBA 277585 and Note R.
Linear Regression	Accuracy, 15 digit X and/or Y values.	KBA 277585	*Centering X and Y data will help. See KBA 277585 and Note R.
Linear Regression through the origin (zero intercept models)	R squared, correlations, dfs and F test values are in error, due to incorrect formulas.	Cryer 2000 McCullough, Hunt 1996, RSS 1996, KBA 214230,	*Ignore these reported values or manually recalculate the correct values. See Note S. Do not specify a zero constant in Excel 2000. Fixed in Excel 2003
Linear Regression	Negative sums of squares.	Cryer 2000	*Center the X and Y data about their means and regress on these values. See Note R on centering.

Application or Function	Problem	Source	Fix or Comments
Linear Regression	Does not handle multi-colinearity correctly.	Crier 2000	Center the X and Y data about their means and regress on these values. Crier (2000) does not say what is incorrect about it. See Note R on centering.
Linear Regression	Zero values in the X and Y input data.	KBA 215559	No fix for Excel 2000.
Linear Regression	Dependent X values	Simonoff, KBA 309326	No fix for Excel 2000. Corrected in Excel 2003
Linear Regression	Computes Standardized Residuals Incorrectly	Cryer 2000	Cryer (2000) does not say which “standardized” he is referring to. See Note H.
Linear Regression	Displays normal probability plots that are completely wrong	Cryer 2000	Change the generated chart to a correct chart; see Note V on how to do that.
Linear Regression	Makes variable selection very difficult. (Variables have to be in contiguous columns)	Cryer 2000	The Excel method has some validity and is necessary to work with Excel’s range concept. By making copies of the data set on additional worksheets, the simple process of deleting columns (and rows) can be done to form contiguous data sets, specific to one analysis.
Moving Average Trend line Routine	Moving Average is out of Phase	RSS 1996	Help formula is out of phase. Actual values are in phase.
Regression Analysis, Singular X Matrix	Zero standard-coefficient-error value & negative adjusted-R	Simon 2000	*Excel 2003 fixed the singularity problem, In Excel 2000, the problem indicates singularity and gives no solution. See note T.

Application or Function	Problem	Source	Fix or Comments
Polynomial curve fitting, using multiple LINEST	Numerical instability	CISE 27/99	*See Polynomial Regression below:
LOGEST and GROWTH functions.	Does not calculate a true LMS fit.	CISE 27/99	Use Non-Linear Regression. See Section 10 about this fault.
Multiple Regression	Cannot use more than 16 independent variables.	(Several Sources)	This restriction by Microsoft should be removed.
LOGEST, GROWTH and Trendline equations	Incorrect model parameter values	Hesse 2006	Use Non-Linear Regression. See Section 10 about this fault.

Most of the problems, faults and errors have been fixed for Excel 2003. The fixes above in the last column with an *, represent faults fixed in Excel 2003. For users who still have Excel 2000 there are some fixes and workarounds that can correct most of the faults and errors. They are given briefly in the last column of table 9-3. Those faults and errors not fixed by Excel 2003 are discussed under the Excel 2003 section.

EXCEL 2003 SPECIFIC:

None found or reported.

EXCEL 2007 SPECIFIC:

None found or reported.

TESTS ON REGRESSION FUNCTIONS AND ROUTINES

ASSUMPTIONS

The assumption here is that the functions FORECAST, SLOPE, INTERCEPT, STEYX and LINEST have the LINEST core algorithms as given in the function manuals and in Help. That is, the equation given is exactly how the function obtains its values. The assumption also is that GROWTH and LOGEST transform the data going into these functions and transform the results back to the source coordinates.

TESTS CONDUCTED:

Table 9-8 gives the LRE results from Excel reported by McCullough (1997, 1998 and 2000) on the NIST data sets. Also included are LRE values for some commercial

software packages reported by McCullough (1997, 1998 and 2000) and Creighton and Ding (2002). Also included are my results using the same data sets in Excel, using the fixes described below.

Table 9-8: Tests on Regression Equations

Sequence	Source	Dataset	Category	Difficulty	Size	Number Significant Figures	Number of X Variables	Number of Coefficients
23	NIST	Norris	Linear	1	36	4	1	2
24	NIST	Pontius	Quadratic	1	40	6	1	3
25	NIST	NoInt1	Linear, zero intercept	2	11	3	1	1
26	NIST	NoInt2	Linear, zero intercept	2	1	1	1	1
27	NIST	Filip	Polynomial	3	82	10	1	11
28	NIST	Longley	Multiple	3	16	6	6	7
29	NIST	Wampler1	Polynomial	3	21	7	1	6
30	NIST	Wampler2	Polynomial	3	21	7	1	6
31	NIST	Wampler3	Polynomial	3	21	7	1	6
32	NIST	Wampler4	Polynomial	3	21	7	1	6
33	NIST	Wampler5	Polynomial	3	21	8	1	6
34	STERN	1	Linear		10	13	1	2
35	STERN	Simonoff	Linear (with singular X matrix)		54	4	4	5
36	STERN	Eakin	Linear		6	9	1	2
37	STERN	Simon	Linear		10	3	1	1

TEST RESULTS:

Rather than use the lowest LRE value for the coefficients and the lowest LRE value for the standard errors as the test value, Stewart's suggestion for a combined assessment of all coefficient values is a better method. However this was never used in the literature on statistical software tests.

Table 9-9: Test Results on Stata and JMP

Sequence	Stata		JMP 4.0.5 Fit Y By X		JMP 4.0.5 Fit Model	
	Lowest LRE for Coefficients	Lowest LRE for Standard Errors	Lowest LRE for Coefficients	Lowest LRE for Standard Errors	Lowest LRE for Coefficients	Lowest LRE for Standard Errors
23	12.8	13.5	12.2	11.7	13.3	10.4
24	11.5	13.0	11.2	8.4	11.8	9.0
25	14.7	15.0	14.7	13.5	14.7	13.5
26	15.0	15.0	15.0	14.6	15.0	14.6
27	No solution	No solution	No solution	No solution	No solution	No solution
28	12.1	12.9	No solution	No solution	8.3	10.0
29	6.9	15.0	8.0	15.0	7.0	15.0
30	9.7	15.0	10.6	15.0	9.5	15.0
31	6.5	10.8	8.0	10.8	7.0	10.2
32	6.5	10.8	8.0	10.9	7.0	9.9
33	6.4	10.8	8.0	10.9	7.0	9.9

Table 9-10: Test Results on Excel 2000 and 2003

Sequence	Excel 2000 (U)		Excel 2000 (C)		Excel 2003 (U)		
	Lowest LRE For Coefficients	Lowest LRE For Standard Errors	Standard Error	Final F Statistic	Lowest LRE For Coefficients	Lowest LRE For Coefficients	Lowest LRE For Standard Errors
23	12.1	13.8	13.8	13.5	12.8	12.0	14.1
24	11.2	14.3	14.7	13.9	12.5	12.0	12.7
25	14.7	15.0	15.5	15.0	14.7	14.7	14.7
26	15.0	15.0	15.5	15.0	15.3	15.3	14.8
27	0.0	0.0	14.5	14.3	10.3	7.8	7.5
28	7.4	8.6	14.9	14.6	12.4	13.4	14.7
29	6.6	7.2	9.1	(2)	8.6	9.9	10.4
30	9.7	11.8	13.2	(2)	11.1	13.4	15.0
31	6.6	11.2	14.0	13.7	8.6	10.1	11.4
32	6.6	11.2	14.8	15.2	8.5	8.1	11.8
33	6.6	11.2	14.9	13.7	8.4	6.1	12.0
34					6.2		
35							
36	-8.4		8.1	7.8	8.6		
37	8.0						

Notes: (2) indicates that the p value was beyond the range of the F distribution function.

EXCEL 2000

REGRESSION OUTPUT VALUES

Excel appears to perform fairly well except on the NoInt1, NoInt2 and Filip data sets. It also shows that the performance of Excel can be improved. Although McCullough (2000) did not report it, the Excel regression through the origin (routine) gives faulty values of the correlation coefficients and faulty values in the variance table. The zero's inserted in rows 25 and 26 of table 9-9 reflect this error.

The improvement in Excel accuracies shown in the last column was due to a preliminary centering of all data, and the centered data used in the routine. Note R, Linear Regression provides further information. For regression through the origin, the fix given in Note S, Regression Through the Origin was also applied.

It should be noted that the Best Linear Unbiased Estimator (BLUE) on the variables in deviation form (centered) represents the best estimate of the coefficients, because the inversion is on the $p' \times p$ matrix rather than the $(p+1)' \times (p+1)$ matrix. The $(p+1)' \times (p+1)$ matrix is used to get a value of the intercept. The intercept value in this case can be obtained by a linear solution on all the means.

TEST FOR INACCURATE MULTIPLE RESULTS:

Excel does not provide a message when there is an obvious problem or error in the regression. McCullough (2001) criticizes Excel for not automatically providing information when a solution is worthless. This is a valid criticism. For general users, they would not know that there were rank/singularity/multi-colinearity issues with their data, which are giving him bad regression output. A typical user would not be aware of this.

When using Excel, the user is responsible for checking the output. The obvious check is the regression statistics column (R squared...). If there are any negative values, then the regression has some error in the results.

As a check on the choice of ranges used as inputs, always put in a label as the first column of the range, and check the label box on the regression input menu. This label comes out in the results and should be a check on the selection of the right columns for the input. (Simon 2000)

If you suspect errors in the output values, or if the standard deviation of the coefficient values appears to be large, then a test for singularity (within the capability of Excel) such as that in Note R may be helpful to detect solution problems.

CONCLUSIONS:

If you look at the results, Excel comes out rather well, even without centering. An LRE of 6.6 for linear regression fits on the data typically used in problem solving situations is good. If the data is centered, Excel has more accurate solutions than Stata or JMP 4.0.5. McCullough states "The Stata results for the higher difficulty problems are the best that can be achieved using conventional algorithms and double precision calculation on a PC" (McCullough 2000). I have shown that pre-centering in some cases can better this standard.

EXCEL 2003 AND EXCEL 2007

The Excel 2003 functions and routines were not changed for Excel 2007. Therefore the following analysis applies to both versions.

REGRESSION THROUGH THE ORIGIN

The errors and faults with regression through the origin in previous versions of Excel have been fixed. The regression-through-the-origin data set in Simon (2000) results in an Excel 2003 output that agrees with the published Minitab output values to the very short number of digits given in the Minitab outputs (generally displayed to only 3 digits). The difference in intercept values are basically due to the fact that Minitab uses FORTRAN which is not IEEE-754 limited, while Excel uses Visual Basic which is IEEE-754 limited.

DATA ANALYSIS REGRESSION ROUTINE OUTPUT FAULTS

The faults and problems described above under DATA ANALYSIS REGRESSION ROUTINE OUTPUT FAULTS were not fixed for Excel 2003. Excel 2003 and Excel 2007 continues to have these faults on the routine output sheets.

REGRESSION OUTPUT VALUES

If you compare table 9-9 with the other tables, you can see that the Excel 2003 algorithm clearly has improved stability. It will give results when the Excel 2000 algorithm fails. Also one gains in accuracy when the Excel 2003 algorithm is used. The gain in accuracy primarily comes from the initial centering, not the QR method. The centered Excel 2000 algorithm gives in some cases better accuracy than the Excel 2003 algorithm, and in other cases the reverse is true.

If you compare the reported STATA and JMP results (reported above), it appears that the Excel 2003 algorithm will generally give more accurate results, (8+ and 3- when compared to JMP), but the differences are small. As Stewart says, it is hard to say that any one of the different methods will generally give more accurate results. The real advantage is that one can get a good set of coefficient values for the Phillip data set.

REPORTED FAULTS AND ERRORS

The functions FORECAST, STEYX, SLOPE AND INTERCEPT depend on values generated from LINEST. Microsoft corrected a coding error in these four functions that gave incorrect values. The correction is in the SR-1 upgrade to Excel 2003. Consequently it is important that all Excel 2003 uses get SR-1 installed.

Microsoft reports a problem with Excel 2003 LINEST in KB887964 (7 Nov 2005). It has to do with an internal rounding of output values to approximately 9 significant digits. An error occurs when the data is such that the full 15 digit computed slopes, intercepts and other computed values must be viewed as having more than 9 significant digits. A hotfix is available (See KB903240). The fix appears to be some changes to a registry file as described in KB887964. Microsoft stated that the next service release will automatically fix this problem.

MISSING DATA

If any cell in the selected data range has an empty cell, or a non-numeric cell, an error message appears. If the Data Analysis routine “Regression” is used, the error message is “Regression – LINEST() function returns error. Please check input ranges again.”

THE MULTICOLINEARITY ISSUE

NIST DATA SETS

The Longley StRD data set is the only true multiple set in the StRD suite. This set is frequently cited for its co-linearity (see Hadi and Ling 1998), and used to describe the multi-collinearity problems. Table 9-11 gives values of the correlation coefficients among the seven variables

Table 9-11: Longley Data Set Variable Correlations

	Response	GNP Deflator	Gross National Product	Unemployment	Military Employment	Population	Year-Time
Response	1.000000						
GNP Deflator	0.970899	1.000000					
Gross National Product	0.983552	0.991589	1.000000				
Unemployment	0.502498	0.620633	0.604261	1.000000			
Military Employment	0.457307	0.464744	0.446437	-0.177421	1.000000		
Population	0.960391	0.979163	0.991090	0.686552	0.364416	1.000000	
Year-Time	0.971329	0.991149	0.995273	0.668257	0.417245	0.993953	1.000000

The multi-collinearity of this data set did not cause any problems to LINEST (2003). LRE values of the resulting calculated coefficients and standard errors are given in table 9-12.

Table 9-12: LRE Values of Excel 2003 Regression on the Longley Data Set

Variable	Measures	Coefficient Values	Coefficient Standard Errors
R Square	15.48		
Standard Error	15.25		
F Ratio	15.07		
Intercept		14.73	15.11
GNP Deflator		13.36	15.48
Gross National Product		14.24	15.68
Unemployment		14.36	15.25
Military Employment		15.37	15.11
Population		13.68	14.66
Year-Time		15.60	15.90

One of the reasons for the good performance here is that the X matrix has widely spaced eigenvalues. Hadi and Ling (1998) give the eigenvalues for the Longley X matrix as 4.60338, 1.17534, 0.203425, 0.0149283, 0.00255207 and 3.76708E-04. If there were some values close together, then the fit would have been poorer.

The FILIP data set has strong multi-colinearity from the fact that it is a fit of a tenth order polynomial to a X-Y data set. In this case the X values are expanded by multiplications to a 10-column matrix, and LINEST is used in a multinomial sense. Table 9-13 gives the correlations between the 10 new variables.

Table 9-13: Correlations Between Filip Data Set Expanded X Values

	X	X2	X3	X4	X5	X6	X7	X8	X9	X10
X	1.000000									
X2	-0.991250	1.000000								
X3	0.968545	-0.992822	1.000000							
X4	-0.937670	0.974860	-0.994449	1.000000						
X5	0.903573	-0.951027	0.980855	-0.995859	1.000000					
X6	-0.869587	0.924958	-0.962985	0.985794	-0.996954	1.000000				
X7	0.837553	-0.898917	0.943395	-0.972530	0.989526	-0.997752	1.000000			
X8	-0.808266	0.874110	-0.923613	0.957830	-0.979630	0.992214	-0.998317	1.000000		
X9	0.781903	-0.851066	0.904464	-0.942745	0.968485	-0.984718	0.994110	-0.998713	1.000000	
X10	-0.758328	0.829925	-0.886333	0.927862	-0.956832	0.976117	-0.988312	0.995444	-0.998993	1.000000

The Excel 2003 Regression gives good values as shown in table 9-14

Table 9-14 LRE Values of Excel 2003 Regression on the Filip Data Set

Variable	Measures	Coefficient Values	Coefficient Standard Errors
R Square	10.60		
Standard Error	8.42		
F Ratio	8.11		
X		7.90	7.55
X2		7.89	7.55
X3		7.88	7.54
X4		7.87	7.53
X5		7.86	7.53
X6		7.85	7.52
X7		7.83	7.52
X8		7.82	7.51
X8		7.80	7.51
X10		7.79	7.51

The eigenvalues for the X matrix are not known. Microsoft did not include an eigenvalue computation routine in Excel.

Actually, the Excel routines and functions are fairly robust to strong multi-colinearity. The fact that a good solution to the Filip data set was obtained (10 highly correlated X variables), and several successful multiple trials with X variables having a correlation of 0.9999999 suggests that this is not that big of a problem when using Excel 2003.

As discussed in section 8, the statistical problem is not accuracy, but that the high correlations among the data gives statistically incorrect values for the coefficients. The NIST tests only test computational accuracies.

SINGULARITY

None of the NIST data sets test for singularity. Therefore other data sets have to be used. One set is that submitted by Simon and is included in the Simonoff (2000) reference.

Simonoff (2000) presented a data set by Simon of 54 rows, one Y value column, and four x value columns, representing some dental data. Column 2 and 3 are essentially a binary representation of the treatments, and column 4 (an accident) is the number 4 minus the binary combination of 2 and 3 as a decimal.

Table 9-15: Simon's Dental Data

Y	1	2	3	4
5.88	1	1	1	1
2.56	6	1	1	1
11.11	1	1	1	1
0.79	6	1	1	1
0.00	6	1	1	1
0.00	0	1	1	1
15.60	8	1	1	1
3.70	4	1	1	1
8.49	3	1	1	1
51.20	6	1	1	1
14.20	7	1	1	1
7.14	5	1	1	1
4.20	7	1	1	1
6.15	4	1	1	1
10.46	6	1	1	1
0.00	8	1	1	1
10.42	2	1	1	1
17.36	5	1	1	1
13.41	8	1	1	1
41.67	0	1	1	1
2.78	0	1	1	1
2.98	8	1	1	1
9.62	7	1	1	1
0.00	0	1	1	1
4.65	5	1	0	2
3.13	3	1	0	2
24.58	6	1	0	2

Y	1	2	3	4
0.00	1	1	0	2
5.56	4	1	0	2
9.26	3	1	0	2
0.00	0	1	0	2
0.00	0	1	0	2
3.13	1	1	0	2
0.00	0	1	0	2
7.56	5	0	1	3
9.93	6	0	1	3
0.00	8	0	1	3
16.67	6	0	1	3
16.89	7	0	1	3
13.71	6	0	1	3
6.35	5	0	1	3
2.50	3	0	1	3
2.47	7	0	1	3
21.74	3	0	1	3
23.60	8	0	0	4
11.11	8	0	0	4
0.00	7	0	0	4
3.57	8	0	0	4
2.90	5	0	0	4
2.94	3	0	0	4
2.42	8	0	0	4
18.75	4	0	0	4
0.00	5	0	0	4
2.27	3	0	0	4

The LINEST 2000 results by Simonoff (2000) are given in table 9-16. Also given are the results reported by Simonoff (2000) for Minitab

Table 9-16: Results of Regression on Simons's Data

	Excel 2000	Minitab
Columns Used	1,2,3,4	1,2,3,4
Intercept	0.384972384	4.194
Variable 1	0.386246607	0.3862
Variable 2	2.135547339	0.231
Variable 3	4.659552583	3.707
Variable 4	0.952380952	-----

The LINEST 2003 output for table 9-15 is as follows:

Table 9-17: LINEST Output For Simon's Dental Data (1,2,3,4)

0	-0.11539272	3.591778913	0.386246607	4.656067061
0	1.579488423	3.741180996	0.565167745	4.956396239
0.047673146	10.13673187	#N/A	#N/A	#N/A
0.834327453	50	#N/A	#N/A	#N/A
257.1897798	5137.666652	#N/A	#N/A	#N/A

The Excel 2003 routine automatically set the coefficient for variable 4 to zero, and arrived at a reasonable solution. However the values did not agree with the published Minitab output, since the 1,2,3,4 set of Excel coefficients did not agree with the Minitab set of 1,2,3,4 coefficients

The R output for the 1,2,3,4 set was (McCullough 2008)

Residuals:

Min	1Q	Median	3Q	Max
-11.222	-5.821	-2.546	3.171	40.750

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.1945	3.9749	1.055	0.296
X1	0.3862	0.5652	0.683	0.497
X2	0.2308	3.1590	0.073	0.942
X3	3.7072	2.9922	1.239	0.221
X4	NA	NA	NA	NA

Residual standard error: 10.14 on 50 degrees of freedom
Multiple R-Squared: 0.04767, Adjusted R-squared: -0.009466
F-statistic: 0.8343 on 3 and 50 DF, p-value: 0.4814

The R output agrees with the Minitab output, but does not agree with the Excel output. Therefore we have to conclude, Excel still does not deal correctly with singularity

If variable 4 is excluded, then we get

Table 9-18 LINEST Output For Simon's Dental Data (1,2,3)

3.70717163	0.230785434	0.386246607	4.194496194
2.99221431	3.158976845	0.565167745	3.97490643
0.047673146	10.13673187	#N/A	#N/A
0.834327453	50	#N/A	#N/A
257.1897798	5137.666652	#N/A	#N/A

Table 9-18 agrees with the Minitab and R outputs. However the intercepts and coefficient values are interchanged. This just points out the importance of deleting the right singular variable from the data set and re-running LINEST. This also clearly indicates the importance of the investigator to determine which of the singular variables to delete. Letting LINEST choose, is not a good solution.

You cannot assume that Excel has picked the right singular variable to be deleted.

DATA ANALYSIS TOOL PAC REGRESSION ROUTINE

The regression routine as described above just uses the LINEST output, calculating the other measures and p values from the LINEST output. For the table 9-17 LINEST output, the regression routine output is shown as table 9-19 .

Table 9-19: Excel 2003 Data Analysis Regression Output

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.218341811				
R Square	0.047673146				
Adjusted R Square	-0.029466465				
Standard Error	10.13673187				
Observations	54				
ANOVA					
	df	SS	MS	F	Significance F
Regression	4	257.1897798	64.29744495	0.834327453	0.509972412
Residual	50	5137.666652	102.753333		
Total	54	5394.856431			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	4.656067061	4.956396239	0.939405737	0.352038686	-5.29914757
X1	0.386246607	0.565167745	0.683419411	0.497496766	-0.74892619
X2	0	0	65535	#NUM!	0
X3	3.591778913	3.741180996	0.960065529	0.341642088	-3.92260412
X4	-0.115392717	1.579488423	-0.07305702	0.942052222	-3.28788852

The gray cells have wrong values. The X2 row should not even show, since there are no values associated with variable X2.

When LINEST returns a standard error of zero, a “t Stat” value cannot be calculated. The 65535 value is from a fault in the algorithm. A P-value cannot be calculated when there is zero df.

The Data Analysis package was never changed to handle the two zero LINEST returns, and ends up having a #NUM return when a p value calculation is tried with a zero standard deviation. Note that the mean square value from LINEST is correct. This is why the DA mean square value is correct on the DA display, but the other values are wrong.

The Significant F p value is also wrong, because it is based on the wrong df value.

The Data Analysis routine was never changed to handle zero standard error values from LINEST.. This is a fault that has carried over to Excel 2007.