Computational Statistics and Data Analysis 54 (2010) 1190–1196

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Polynomial Trendline function flaws in Microsoft Excel

Bruce R. Hargreaves [a,*], Thomas P. McWilliams [b]

[a] Department of Earth and Environmental Sciences, Lehigh University, 31 Williams Drive, Bethlehem, PA, 18015, USA
[b] LeBow College of Business, Drexel University, USA

ARTICLE INFO

*Article history:*
Received 25 October 2009
Accepted 26 October 2009
Available online 1 November 2009

ABSTRACT

Numerous statistical and graphical problems have been reported for different versions of Microsoft Excel, including the newest version (Excel 2007). We report newly discovered problems with Excel 2007 when generating polynomial trend line equations, having a user-specified (forced) intercept, from graphed data. We also remind users of Excel's Trendline function of problems with Excel 2003 that have not been corrected in Excel 2007. Excel will "fit" nonsense trend lines to data presented on column and line charts, and can report an inadequate number of significant digits for polynomial trend lines. We provide suggestions for avoiding these continuing problems, but are unable to identify an Excel 2007 workaround solution for the forced-intercept polynomial trend line errors.



## 1. Introduction

Statistical routines in Microsoft Excel have a long and well-documented history of problems. A series of articles by McCullough and Wilson (1999, 2002, 2005) exposes deficiencies in Microsoft Excel 97, 2000, and 2003. The series continues with McCullough and Heiser's 2008 report on Excel 2007. That Excel 2007 critique and related articles illustrate a wide variety of errors involving topic areas that include statistical distributions (Yalta and Talha, 2008), random number generation (McCullough, 2008a), estimation, and probability plots, plus the appropriateness of Excel for teaching statistics (Nash, 2008) and statistical graphics (Su, 2008).

In each new article, McCullough and co-authors discuss which previously reported problems have or have not been fixed and identify new deficiencies. When introducing new versions of Excel, Microsoft does not always correct computational problems, including those that are easily corrected and have been publicized to the point of becoming common knowledge in the statistical community. Additionally, new problems often crop up in new versions of the software. David Heiser maintains an active website http://www.daheiser.info/excel/frontpage.html dealing with Excel "Faults, problems, workarounds, and fixes". Some of the information contained in this article has appeared on that site.

Some Excel graphics, such as scatterplots and column charts, have a feature, enabled by right-clicking on a point in a data series, which allows the user to fit one of six types of trend lines to the data: Linear, Logarithmic, Polynomial, Power, Exponential, and Moving Average. McCullough and Heiser briefly mention problems with incorrect results for the Linear Trendline when working with ill-conditioned data (Pottel, 2003) and for the Exponential and Power Trendlines even when the data are not ill-conditioned (Hesse, 2006). They pose the question "Does any user of Excel wish to bet that the remaining three trendlines are correctly implemented?"

In this article we provide a partial answer to that question by examining the performance of the polynomial Trendline function in Excel 2003 and 2007. A new problem in Excel 2007 is that the forced-intercept version of this function, where the user specifies rather than fits a value for the intercept, invariably yields incorrect results for polynomials of degree three

* Corresponding author. Tel.: +1 610 758 3683; fax: +1 610 758 3677.
*E-mail addresses:* brh0@lehigh.edu (B.R. Hargreaves), tpm23@drexel.edu (T.P. McWilliams).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.10.020

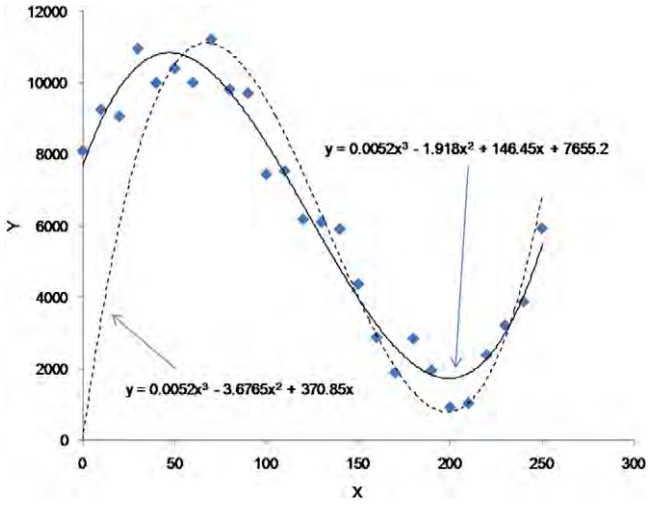


**Fig. 1.** 3rd order polynomial trend lines fitted by Excel 2007 with an erroneous forced-intercept trend line equation. Trend lines and equations for a set of data (solid symbols) show the correct equation (solid trend line, $y = 0.0052x^3 - 1.918x^2 + 146.45x + 7655.2$) when the $y$-intercept is calculated but an erroneous equation (dashed trend line, $y = 0.0052x^3 - 3.6765x^2 + 370.85x$) when the $y$-intercept is forced (in this case, through zero). The valid equation for the zero-intercept trend line is $y = 0.0092x^3 - 3.6765x^2 + 370.85x$.

to six (the maximum allowed in Excel 2007). We also illustrate a number of other problems with this function that are specific to Excel 2007. We then discuss the use of the Trendline function, polynomial or otherwise, for 'fitting' mathematical functions to data presented on column charts where the independent variable is categorical. This is truly a bizarre "feature" of Excel 2003 and 2007.

## 2. Deficiencies of the Excel 2007 polynomial Trendline function incorrect polynomial trend line equation and $R^2$ value with forced intercept

To create such a trend line, the user first creates a scatterplot of the data. The user then right-clicks on any data point and selects 'Add Trendline…'. To create a 'with-intercept' trend line, specify the order of the polynomial and check the 'Display Equation…' box. For a 'forced-intercept' trend line, follow the same procedure but also check the 'Set Intercept =' box and specify a value for the intercept. Fig. 1 illustrates examples of with-intercept and forced-intercept 3rd order polynomial trend line fits, with the forced-intercept value set equal to zero.

The equation reported in Fig. 1 for the with-intercept fit (solid line) is the correct least-squares regression fit, as verified by the statistical software package Minitab. However, the equation specified for the forced-intercept fit (dashed line) is incorrect and is in fact a mix of the correct least-squares with-intercept and forced-intercept fits. The correct regression equations, with and without the intercept term are:

$$\text{With-intercept:} \quad y = 0.0052x^3 - 1.918x^2 + 146.45x + 7655.2$$
$$\text{Forced-intercept (equal to zero):} \quad y = 0.0092x^3 - 3.6765x^2 + 370.85x.$$

The incorrect forced-intercept equation reported by Excel is a mix of these two equations, given by:

$$y = 0.0052x^3 - 3.6765x^2 + 370.85x.$$

The cubic term in this equation comes from the correct with-intercept fit, while the lower order terms come from the correct forced-intercept fit. Note that while the reported equation is incorrect, the plot of the fitted forced-intercept regression line appears to be correct. We verified this by adding, to our plot, a series of new values (not shown) calculated using the correct forced-intercept equation, and this series nicely tracked the trend line created by Excel.

The Excel 2007 Trendline function includes the option of reporting the value of $R^2$. For both the with-intercept and forced-intercept fits the formula $R^2 = 1 - SSE/SSTO$ is used, where $SSE$ is the sum of squared errors and $SSTO$ is the sum of square deviations of the response variable values around the mean. Unfortunately, in the forced-intercept application this calculation can lead to a negative result, so $R^2$ cannot be interpreted in the usual way, namely as the percentage of variation in the response variable that can be explained by the regression model.

Kutner et al. (2005) present an alternative method of calculation in the forced-intercept case, where $SSTO$ is replaced by $SSTOU$, the uncorrected sum of squared response values. This measure has the advantage of being bounded between zero and 100 percent, but has other problems with regards to its value and interpretation. For some data sets, the $R^2$ value calculated in this fashion will exceed that of the $R^2$ value from the with-intercept fit, which would suggest that more of the variation in the response variable can be explained by constraining a parameter in the model! Kutner et al. (pp. 165) state

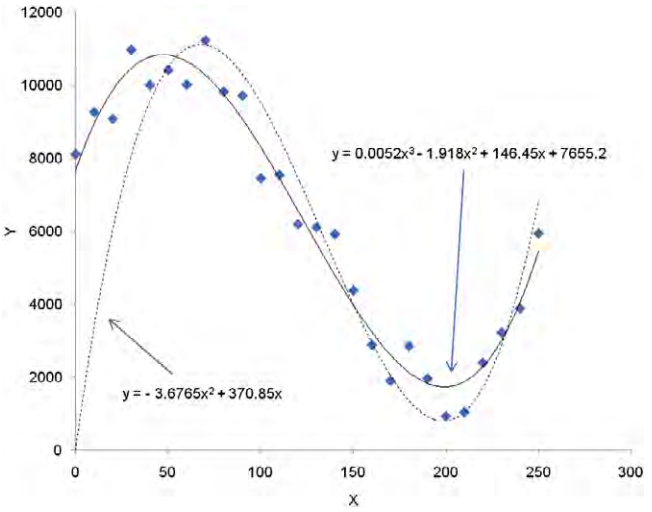


**Fig. 2.** 3rd order polynomial trend line with missing cubic term. When the graph in Fig. 1 is copied and pasted, or when the spreadsheet or a word document containing the graph is closed and then re-opened, the cubic term disappears from the forced-intercept equation, in this case displaying $y = -3.6765x^2 + 370.85x$. See text for variations on this error for the forced-intercept equation when the order of the polynomial or one of the data values is changed.

that this measure "lacks any meaningful interpretation". Note that SPSS and SAS both report this value, with SPSS including a warning that the result is not comparable to an $R^2$ value calculated from a with-intercept regression fit.

As mentioned above, in Excel 2007 Microsoft's approach for the forced-intercept trend line is to calculate $R^2$ via $1 - SSE/SSTO$. This formula was also used in Excel 2003. In January 2006 Microsoft's developers decided that this was incorrect, due to the possibility of obtaining negative results, and published an article (http://support.microsoft.com/kb/829249) recommending that when performing forced-intercept regression the user should use the LINEST function or the Analysis Toolpak (ATP), both of which use the *SSTOU*-based formula, in order to obtain the 'correct' value of $R^2$. However, the use of the *SSTO*-based formula within the Trendline function continues in Excel 2007. Apparently Excel's developers have decided that this error is unimportant, is not worth fixing, and that it is acceptable for different regression-based routines in Excel to yield different results! In any case, we agree with Kutner et al. and recommend against the use of either $R^2$ measure for forced-intercept regression.

### 2.1. Disappearing terms and constants that vary

We also experienced problems with an incomplete display of the trend line equation when working with forced-intercept trend lines in Excel. In our worksheet, we copied the scatterplot and pasted it into another region on the same page. The result is given in Fig. 2. In this case the stated equation for the forced-intercept fit does not even include a cubic term, even though the sketch of the fitted relationship is clearly cubic!

However, when we right-clicked on the trend line, selected 'Format Trendline..', and then toggled the 'Set Intercept = 0.0' option off and on, the cubic term magically reappeared! We also experienced loss of the cubic term on the forced-intercept trend line equation when we closed and then re-opened our Excel worksheet, when we copied a scatterplot from Excel and pasted it into a MS Word document, or when we closed and re-opened the Word document containing the scatterplot. To create Fig. 1 with the cubic term included we generated the forced-intercept trend line equation and $R^2$ value in Excel, copied the results into a new text box on the scatterplot so that it was no longer linked to the trend line calculation, and then opted within 'Format Trendline' to not display the equation or $R^2$. This was necessary to generate a stable version of the displayed results.

### 2.2. Changing the degree of the polynomial

The behavior of the highest order term for the forced-intercept equation was also bizarre when the order of the polynomial fit was changed. In experimenting with changes in the order of the polynomial, we started with a cubic fit, with the usual incorrect 3rd order coefficient. We then used 'Format Trendline..' to change to a fourth order fit. The initial reported equation was missing the coefficient of $x^4$, with all other coefficients displayed and correct. Toggling the 'Set Intercept = 0.0' option off and on led to a displayed fourth order fit, but once again the highest order coefficient was incorrect and came from the correct with-intercept fit. Going back down to a third order polynomial, a cubic equation was displayed but the coefficient of $x^3$ reported was incorrect and was actually the correct $x^3$-coefficient from the forced-intercept 4th order fit.

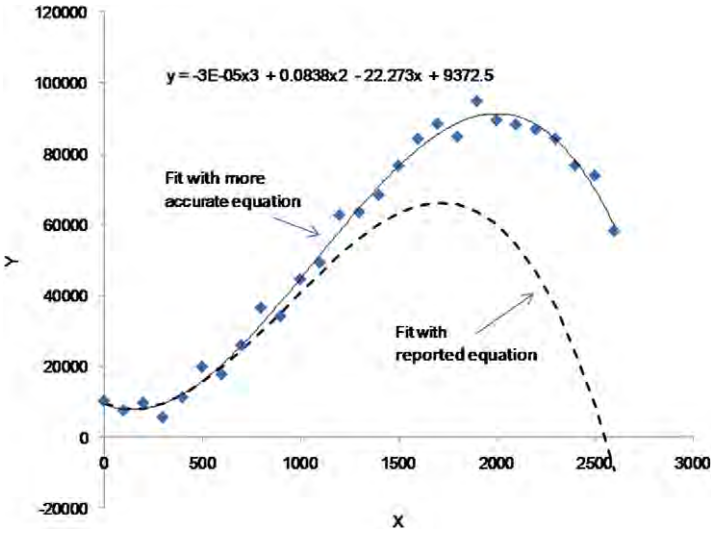


**Fig. 3.** 3rd order polynomial trend line equation reported with one significant digit. The reported trend line equation (dashed curve, $y = -3E-05x^3 + 0.0838x^2 - 22.273x + 9372.5$) uses only 1 significant digit in scientific notation. An equation with adequate significant digits ($y = -2.6088E-05x^3 + 0.083813x^2 - 22.273x + 9372.5$) fits the data set (solid symbols) much better.

Finally, toggling the 'Set Intercept = 0.0' option off and on once again yielded the original incorrect cubic fit, where the reported coefficient of $x^3$ was equal to the correct coefficient for the with-intercept fit!

### 2.3. Trend line updating

Another interesting "feature" relates to updating of the trend line equation. We changed the response $Y$ at $X = 200$ from $Y = 936$ to $Y = 3000$. In this case Excel correctly updated the with-intercept regression equation. For the forced-intercept equation, Excel once again reported correct values for the lower order terms. However, the cubic term coefficient retained the value of the *original* with-intercept cubic term coefficient of 0.0052 rather than updating to the new value. Once again, toggling the 'Set Intercept = 0.0' option off and on 'corrected' this problem in that the invariably incorrect leading coefficient switched to being equal to the leading coefficient of the correct updated with-intercept fit.

### 2.4. Lack of significant digits (a problem in Excel 2003 and 2007)

To fit, for example, a cubic polynomial using Excel's ATP, it is necessary to create columns of $X^2$ and $X^3$ values and then use the regression tool to perform a multiple regression. The use of the trend line option on the scatterplot requires much less effort, so it is likely that this approach would be preferred by users if the only result of interest to the user is the least-squares regression equation. This is disturbing, as the use of the equation found via 'Add Trendline..' for forecasting could lead to highly inaccurate results due to the lack of significant digits reported in cases where coefficients have small values.

Fig. 3 shows a cubic trend line fit using the 'Add Trendline..' command. The reported least-squares regression equation is

$$y = -3E-05x^3 + 0.0838x^2 - 22.273x + 9372.5.$$

The coefficient of the cubic term, which is of course the dominant term in the relation, is only reported to one significant digit. The actual value could be as low as $-3.5 \times 10^{-5}$ or as high as $-2.5 \times 10^{-5}$. A more accurate representation of the regression relation, using five significant digits for all coefficients, is given by

$$y = -2.6088E-05x^3 + 0.083813x^2 - 22.273x + 9372.5.$$

The 'Add Trendline' command resulted in the function shown by the solid line shown on the graph, which was clearly calculated using a more accurate version of the regression relation. We added the function shown by the dotted line, which depicts the fitted values obtained when the reported equation is used. Clearly, the use of the reported equation to generate a forecast could lead to disastrous results! For example, if a forecast were desired at $X = 2700$, the use of the more accurate equation leads to a forecast value of $Y = 46{,}720.2$ while the use of the reported equation leads to a forecast value of $Y = -30{,}352.6$. We do not understand why Excel's programmers felt that the use of only one significant digit was appropriate in this application.

Fortunately, this flaw is easy to correct. This is done by right-clicking on the fitted equation and selecting 'Format Trendline Label..'. The user then has the option to display the equation using either standard decimal (choose Number format in Excel) or scientific notation (choose Scientific format) with as many displayed decimal places as desired.

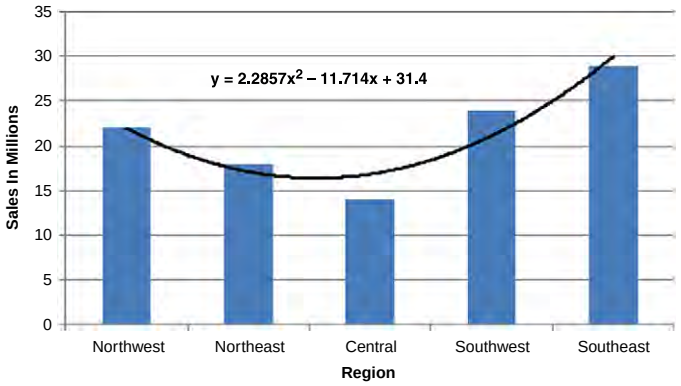


**Fig. 4A.** Column chart with Nonsensical Quadratic trend line Fit. A quadratic trend line is fitted with Excel 2007 (or Excel 2003) to a column chart that depicts sales broken down by region. Excel fits the quadratic equation by arbitrarily assigning the value $X = 1$ to the first category, $X = 2$ to the second, and so on.

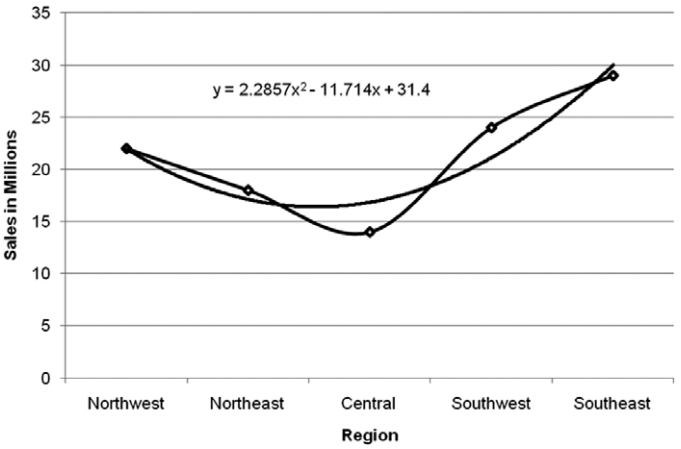


**Fig. 4B.** Line chart with Nonsensical Quadratic trend line Fit. A quadratic trend line is fitted with Excel 2007 (or Excel 2003) to a line chart that depicts sales broken down by region. As in Fig. 4A, Excel fits the quadratic equation by arbitrarily assigning the value $X = 1$ to the first category, $X = 2$ to the second, and so on.

## 3. Nonsense trend lines for graphs having categorical $X$-variables

Excel 2003 and 2007 treat the horizontal axis ($X$) variable in column charts and line charts as being categorical, requiring text-formatted input values, regardless of whether the actual values of that variable are categories or numbers. Nonetheless, the 'Add Trendline..' command is enabled for these charts. The trend line calculation is done by ignoring the values of the $X$-variables and replacing them with the ordinal values 1, 2, 3, … This can lead to results that are clearly meaningless. What is much more disconcerting is that it can also lead to results that may appear to be meaningful when it fact they are nonsense.

Figs. 4A and 4B illustrate nonsensical trend line results, obtained by fitting a quadratic trend line to a column chart and a line chart that depicts sales broken down by region. The fitted quadratic equation was calculated by arbitrarily assigning the value $X = 1$ to the Northwest category, $X = 2$ to the Northeast, and so on. The result is clearly meaningless.

Fig. 5A illustrates a more disturbing situation. Suppose that we have data on the number of employees at six companies, expressed in units of 1000, and also on the annual cost of employee benefits for these companies, expressed in units of one million. We decide that a column chart would be an appropriate tool for displaying this data. This chart was created in two steps. First, a column containing the cost values was selected and used to create the column chart. Then we right-clicked on a bar in the column chart, chose the 'Select Data…' option and specified that the cells containing the number of employees data be used as the horizontal axis labels.

After creating the chart, it is natural to want to learn about the relationship between the two variables, so a quadratic trend line fit is obtained via 'Add Trendline …'. The results are shown on the chart and it appears that we have a reasonably good fit so the reported least-squares regression equation should be useful for forecasting. However, this is not the case. Once again, the numbers shown on the horizontal axis are treated as labels and were *not* used in the trend line calculation. Instead, the values $X = 1, 2, 3, 4, 5, 6$ were used. If a forecast of annual benefits costs were desired for a company with 11,000 employees and we did not understand this, we would plug $X = 11$ into Excel's trend line equation and obtain a forecast of $ 866.73 M! The correct quadratic regression equation can be found by fitting a quadratic trend line to a *scatterplot*

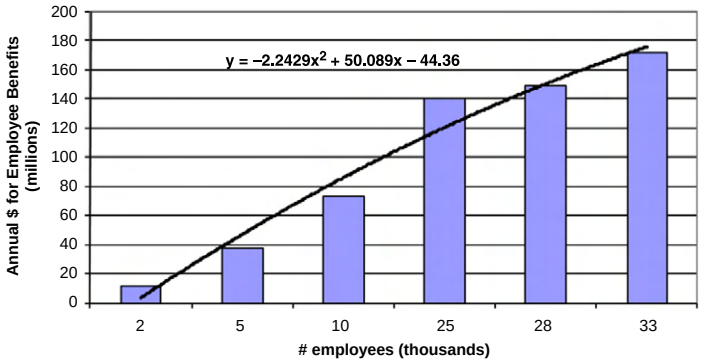


**Fig. 5A.** Incorrect trend line with column chart having numerical labels. A variation on Fig. 4A is when the horizontal axis categories are numeric. The numbers shown on the horizontal axis are treated as labels and are *not* used in the trend line calculation. Instead, the values X = 1, 2, 3, 4, 5, 6 are used.

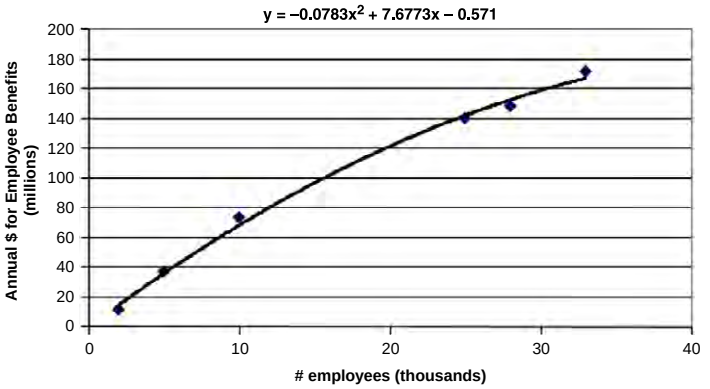


**Fig. 5B.** Correct trend line calculated from the scatterplot. When the data in Fig. 5A are graphed using the Excel scatterplot, the correct trend line and equation ($y = -0.0783x2 + 7.6773x - 0.571$) are calculated because the actual $X$ values are used. Note that an Excel line graph would appear similar except that the X-axis values would be spaced as in Fig. 5A, and the incorrect trend line equation would be displayed.

of the data rather than a column or line chart. This equation, shown in Fig. 5B, is given by $y = -0.0783x^2 + 7.6773x - 0.571$. At $X = 11$, the use of this correct equation would lead to a much more believable forecast value of $ 93.35 M.

The problem illustrated for a column chart in Fig. 5A also occurs when working with line charts. The only case for which trend lines might be meaningful for column and line charts is when the X-axis contains ordinal values. Considering the risks of having a user view a nonsensical result as meaningful, we recommend that Microsoft remove the 'Add Trendline…' option for charts that are not scatterplots.

## 4. Summary and conclusions

We have presented examples of and have discussed a variety of problems with Excel's Trendline function. When working with scatterplots and fitting forced-intercept polynomial trend lines of degree three to six, the primary problem is that the displayed equation is invariably incorrect. Other problems include missing terms in the displayed equation and erratic behavior when changing the order of the polynomial or when closing and re-opening the Excel file. For scatterplot Trendline function fits, regardless of whether the value of the intercept is fixed, the equation may be displayed with an insufficient number of significant digits. We also show that the use of the Trendline function when working with column or line charts yields results that are clearly meaningless or possibly results that appear meaningful but are in fact incorrect. Consistent with the results reported by McCullough et al. (2008), it is apparent that Microsoft's tradition of providing customers with software that is clearly flawed continues with Excel 2007.

### *4.1. Postscript*

During the second review cycle of this article, a referee was unable to replicate the problems we observed with the forced-intercept polynomial regression fits. We investigated and learned that the installation of The 2007 Microsoft Office Suite Service Pack 2 (SP2), which was originally published by Microsoft on April 24, 2009, corrects the problems associated with the highest order regression coefficient. The $R^2$ calculation inconsistency between the Trendline function and other Excel regression-based functions is still present.

Microsoft is hardly forthcoming with the information that the forced-intercept Trendline function yielded incorrect results prior to installation of the service pack, and that this problem has been corrected. An overview of "improvements", which is apparently a Microsoft pseudonym for "corrected bugs", contained in the service pack was obtained at http://support.microsoft.com/kb/953195 and contains no reference to the Trendline function. In a more detailed list of improvements or corrections, which can be downloaded from that website, the trendline-related problems addressed by the service pack are described as "For charts with a trendline that has specific data, the first argument of the trendline is sometimes calculated incorrectly" and "For charts with polynomial trendlines, the first coefficient of the polynomial trendline equation is not being considered." These too-brief descriptions of the problems give no information to a user of the Trendline function regarding what results that may have been generated in the past can be relied upon, versus what results were incorrect and should be recalculated.

The Microsoft list includes 235 other Excel problems that have ostensibly been corrected by this service pack. Examples include:

The Moving Average Trendlines are not being drawn correctly for charts when they are based on data that contains #N/A or blank values

In Excel, elbow or curved connectors can be printed as a straight line.

In certain workbooks, formulas would not recalculate despite the workbook being in automatic calculation mode

Bar and Column charts that have some very small positive values relative to other values may appear as very small negative values.

This all serves to underscore the concerns raised by McCullough (2000) regarding commercial computational software in general, and by Yalta and Jenal (2008) regarding XLSTAT, a set of non-Microsoft commercial add-in programs for Excel, that commercial computational software should be viewed skeptically until proven to be accurate. McCullough (2008b) also comments regarding Microsoft's apparent lack of commitment to quality assurance in business software: "It is difficult not to think that if Microsoft tested business software the way it tested game software, then the statistical functions in Excel would be as accurate as those found in any other major software package". We can only speculate regarding how many undetected or detected but unfixed errors remain in Excel 2007, or how many new errors have been created by Microsoft Office Suite Service Pack 2.

## Acknowledgements

BRH is grateful for initial correspondence with David Heiser. Several anonymous reviewers and Bruce McCullough provided helpful suggestions.

## References

Hesse, Rick, 2006. Incorrect nonlinear trend curves in Excel. Foresight: International Journal of Applied Forecasting 3, 39–43.
Kutner, M., Nachtsheim, C., Neter, J., Li, W., 2005. Applied Linear Statistical Models, 5th ed. McGraw-Hill Irwin, New York.
McCullough, B.D., 2000. Is is safe to assume that software is accurate? International Journal of Forecasting 16, 349–357.
McCullough, B.D., 2008a. Microsoft Excel's 'not the Wichmann–Hill' random number generators. Computational Statistics and Data Analysis 52 (10), 4587–4593.
McCullough, B.D., 2008b. Editorial: Special section on Microsoft Excel 2007. Computational Statistics and Data Analysis 52 (1), 4568–4569.
McCullough, B.D., Heiser, David A., 2008. On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis 52 (10), 4570–4578.
McCullough, B.D., Wilson, Berry, 1999. On the accuracy of statistical procedures in Microsoft Excel 97. Computational Statistics and Data Analysis 31, 27–37.
McCullough, B.D., Wilson, Berry, 2002. On the accuracy of statistical procedures in Microsoft Excel 2000 and XP. Computational Statistics and Data Analysis 40 (4), 27–37.
McCullough, B.D., Wilson, Berry, 2005. On the accuracy of statistical procedures in Microsoft Excel 2003. Computational Statistics and Data Analysis 49 (4), 1244–1252.
Nash, John, 2008. Teaching statistics with excel 2007 and other spreadsheets. Computational Statistics and Data Analysis 52 (10), 4602–4606.
Pottel, Hans, 2003. Statistical Flaws in Excel, Mimeo, Innogenetics NV, Zwijnaarde Belgium.
Su, Yu-Sung, 2008. It's easy to produce chartjunk using Microsoft Excel 2007 but hard to make good graphs. Computational Statistics and Data Analysis 52 (10), 4594–4601.
Yalta, A., Talha, 2008. The accuracy of statistical distributions in Microsoft Excel 2007. Computational Statistics and Data Analysis 52 (10), 4579–4586.
Yalta, A.T., Jenal, O., 2008. On the importance of verifying forecasting results. International Journal of Forecasting 25 (1), 62–73.