What's wrong with Excel's Analysis Toolpak for regression

The regression tool in the Analysis Toolpak (now called the Data Analysis add-in) was introduced in 1995. It was poorly designed even for that time, and remarkably it has not changed at all since then, despite tremendous increases in the power and sophistication of personal computers and the richness of data sources available for analysis. Yet is is still widely used because it is what comes for free with Excel, and this has shaped the teaching of regression analysis to several generations of students. Some of its major flaws, which were the motivation for the development of RegressIt are discussed below.

Here is a screen shot of a sample of the Analysis Toolpak's regression output from *Data Analysis with Excel 5.0* by Mike Middleton (c. 1995):


And here is a screen shot of the same model fitted with Excel 2016, along with its dialog box:


Not one character is different, nor has the dialog box changed, including use of the mysterious word "ply".
Not one character is different, nor has the dialog box changed, including use of the mysterious word "ply". An Excel file with a comparison of output from the Analysis Toolpak and RegressIt for this model can be found here.

An incomprehensible design flaw by modern standards is that the coordinates of data ranges for variables must be selected by hand for each model, rather than by choosing from a list of variables that are already defined as named ranges. Range names were a feature of Excel as far back as 1995 (the present Name Manager was later introduced in 2007), and they are ideal for use as variable names in statistical analysis performed within Excel, as is done in RegressIt and most other add-ins.

Even worse, the independent variables for a model are required to be in consecutive columns to form a single X range, so that data usually must be copied and pasted and/or columns must be inserted or deleted whenever variables are added to or removed from a multiple regression model, and the coordinates of the X range then need to be re-selected. If the file contains a single data sheet and separate model sheets, the current arrangement of variables on the data sheet will usually not agree with all the models, so they are not easily reproducible. The alternative is to have a different copy of the dataset on each model sheet.

And worse again, it can only handle up to 16 independent variables, which means it is a toy, not a tool for serious work. (This arbitrary limit and the single-X range requirement are legacies of program's original coding for the 1995 Excel environment.) Regression models often have many more than 16 independent variables, especially when dummy variables are used to model repeating patterns or special conditions (e.g., promotions, holidays, features, interventions) or combinations of categorical factors. For example, a model that uses dummy variables to estimate an hour-of-day effect would need 23 variables for that purpose alone. Regression analysis often proceeds in a backward stepwise fashion in which the first models fitted have some variables whose significance in the context of other variables is not known a priori, and one of the goals is to root out the less significant ones. Even if the final model turns out to have no more than 16 variables, many more could have been used in early stages of the analysis. (RegressIt can fit linear regression models in Excel with over 200 independent variables on a PC or 125 variables on a Mac, and its R interface can be used to fit large models much faster.)

Another very serious problem is the lack of any provision for forecasting from additional values of the independent variables. To do this, you must hand-enter a column of formulas next to the forecast data, which requires great care in constructing cell references, and this will give you only point forecasts, not confidence intervals. Also, if the formulas are to be copyable, the ranges of data for the variables in the forecast table must be horizontal so as to line up properly with the coefficient table, which does not agree with the vertical orientation of the original data ranges. Therefore the additional values of independent variables to be used for forecasting cannot merely be appended to the columns of the original data: they must be turned sideways. Alternatively, the coefficient table must be copied and transposed. This issue creates even more problems when variables are added to or removed from a model, because the original set of formulas for computing forecasts will not work. (RegressIt can automatically create tables and charts of forecasts and confidence intervals for any rows of data where the independent variables are all present and the dependent variable is missing.)

Another issue is the selection and presentation of charts. The available charts are small and primitive in quality (way below Excel's capabilities), and they are stacked on top of each other by default so you have to peek under the ones on top to see the ones farther down or drag them all over the screen to have them in view at once. The selection of charts that are offered is very strange: it consists of plots of residuals vs.
independent variables and actual-and-predicted-values versus independent variables ("line fit plots"), which are not routinely looked at in models with more than one independent variable, at least not first and foremost. In fact, the user could be overwhelmed. If the model has 10 variables, there would be 10 charts of each type in the stack. Probably only a few would convey useful information, but you wouldn't know until you searched through all of them. There's also a normal probability plot (good idea, except it is poorly scaled with no diagonal reference line). However, there is no plot of residuals vs. predicted values, which is the best plot to study when there is more than one independent variable for purposes of revealing unexplained nonlinear patterns and error variances that are not the same for all predictions. There is also no plot of actual and predicted values versus row number and no plot of residuals versus row number, which are essential to look at when working with time series data. (The latter 3 plots are all available in RegressIt, and all of the plots are appropriately sized and presentation-quality with model information in their titles. You can display them all at once or hide them all by clicking a single button, was well as display them selectively. Some of the tables and charts can be made interactive to show the effects of making changes to parameters such as confidence levels or cutoff values, another feature of Excel that is nice to exploit.)

Yet another annoyance is the lack of appropriate formatting of the displayed numbers. "General" format is used, which shows as many decimal places as will fit. Most of the displayed numbers in the output shown above have 6 or more digits after the decimal point, when only 3 are enough for R-squared, t-stats, and P-values. For coefficient estimates, the number of decimal places ought to be determined relative to their precision, as determined by their standard errors. In the example shown above, the coefficient of Assessed and its standard error are shown with 8 digits after the decimal point, when at most 3 are meaningful. The extra digits are just visual noise. Unlike software that produces plain text output (such as R and Stata) Excel can format numbers for display with any desired number of digits after the decimal point, but this capability has not been used. (In RegressIt, displayed digits after the decimal point are included in blocks of 3 according to their significance, and in this example both Assessed and its standard error would be shown with 3 decimal places.)

The clumsiness of this design wastes the user's time and attention, restricts the scale of problems that can be modeled, invites errors in selecting ranges and entering formulas, and is an impediment to thoughtful analysis and effective teaching. It may suffice in a course for beginners where only a few toy examples are considered and models are given rather than constructed through analysis, but it is very unfortunate if better software is not made available for more advanced or applied work. RegressIt has been designed to address all the issues discussed above for work performed in Excel. Nowadays more advanced work often gravitates toward the use of R, and RegressIt also provides a user-friendly menu interface for regression in R that is accessible even to non-programmers.