RegressIt

## MORE FEATURES AND MODELING ADVICE[1]

Here are some additional details of RegressIt's features as well as some advice on how to use them. Also see the **RegressIt slide show** and the **User Manual** if you haven't already. If you have encountered any problems, see the **Tech Support** page for more information.

**1. Be sure you are using the latest version of RegressIt.** Go to the **Download page** for your program and check the current version dates there. To check your own version date, **click the Info button on the right end of the RegressIt ribbon**. Updating is easy: just replace your existing program file with the one on the corresponding download page. We periodically make enhancements, and they are always backward-compatible.

**2. Use descriptive variable names.** Excel supports the use of long variable names (range names). In RegressIt they have full visibility in all tables and charts, unlike what you get with other Excel addins, and you should take advantage of this when naming your variables. **The upper limit on variable names is 30 characters.** The definitions and units of your variables should be clear from their names, and they should be appropriate for use in table and chart titles. (The name of the dependent variable is shown in the title of each table and chart.) However, **don't start with very long names if you are planning to use transformations of variables**. When you create transformed variables, the default names for the new variables are the original ones with the transformation name tacked on the back.

**3. Use descriptive model names.** It is good to assign descriptive model names to regression analyses at run time. These will become worksheet names and will also appear in table and chart titles and in the various audit trail displays. Also be aware of RegressIt's default naming convention. RegressIt's own default model names are a numbered sequence, "Model 1", "Model 2" etc. You can type any other name you want (up to 30 characters), and **if your model name ends in a number preceded a space or period, RegressIt's default name for the next model launched from that sheet will be the same except with that number increased to the next higher unused number.** For example, if the most recent name in a model sequence is " Beer Sales 3.4", the default name that RegressIt will use for the next one is "Beer Sales 3.5". **Choose your desired model name at run time: do not change the names of existing worksheets nor edit the variable names or model names that appear on them.** Making those sorts of changes ex post will mess up the audit trail: it may not be possible to trace copied charts to the models that created them and the model and variable names on the Model Summaries worksheet may not agree with worksheets.

**4. You don't have to save the worksheet for every regression model:** RegressIt makes it easy to explore variations on regression models, and it is not uncommon to create a large number of model worksheets in the same workbook. It's OK to delete the ones that are inferior: their summary statistics and coefficient estimates are stored on the Model Summaries worksheet, so they will remain part of the audit trail. **Also, you can re-launch a deleted model from the Model Summaries sheet** (linear regression version only). Place the cursor on the seemingly-blank cell just above the model name. This cell should contain a text string that includes the model name and statistics and the list of variables. If you hit the Linear Regression button while positioned on this cell, it will jump you to the worksheet of that model if it already exists, and otherwise it will pre-select the variables in this list for a new model.

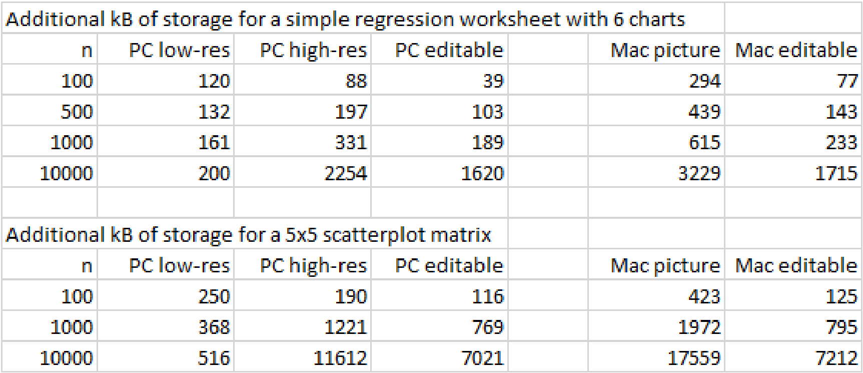
[1] November 23, 2020, Robert Nau. Visit https://regressit.com for complete documentation and the free software

**5. Visualize your data.** Don't just look at test statistics and their P-values. Take advantage of RegressIt's high-quality chart output to see what your variables look like, how their patterns line up, how well the model accounts for those patterns, what adjustments to the data or the model may be needed, and how best to illustrate your findings for others.

**6. Considerations in choosing chart formats:** There are 3 chart formats: editable and low-resolution and high-resolution pictures. Low-res format is only available on PC's.

- Low-res charts require very little storage space in general (a few hundred kB each). They are bitmaps with little increase in size for larger data sets.
- High-res and editable charts contain the plotted data so that they remain perfectly sharp when scaled up, and hence they require an amount of space that depends on the sample size.
- On a PC, high-res charts actually take up *less* space than low-res ones for sample sizes up to around 150, so there is no benefit in using low-res format for small data sets. On a Mac, the size of high-res pictures goes up more rapidly with sample size but is not great in absolute terms for sample sizes in the hundreds.
- Editable charts require about the same amount of space on PCs and Macs.
- Editable charts on a PC are around 60% of the size of high-res charts, and editable charts on a Mac are around 40% of the size of pictures. *You should be sure to save your work frequently when running models with large data sets and producing a lot of chart output on a Mac.*
- Heights and widths and titles and point and line sizes scale up proportionally when low-res or high-res pictures are enlarged, while they do *not* scale up proportionally when an editable chart is enlarged, so you should use one of the picture formats if you want chart element proportions to remain the same under enlargement.
- You will generally want to use a picture format when copying charts to Word or Powerpoint in order to maintain their integrity.

The following table illustrates how the increase in file size (in kB) due to the addition of analysis worksheets depends on the sample size and type of chart output:



| Additional kB of storage for a simple regression worksheet with 6 charts | | | | | | |
|---|---|---|---|---|---|---|
| n | PC low-res | PC high-res | PC editable | | Mac picture | Mac editable |
| 100 | 120 | 88 | 39 | | 294 | 77 |
| 500 | 132 | 197 | 103 | | 439 | 143 |
| 1000 | 161 | 331 | 189 | | 615 | 233 |
| 10000 | 200 | 2254 | 1620 | | 3229 | 1715 |
| | | | | | | |
| Additional kB of storage for a 5x5 scatterplot matrix | | | | | | |
| n | PC low-res | PC high-res | PC editable | | Mac picture | Mac editable |
| 100 | 250 | 190 | 116 | | 423 | 125 |
| 1000 | 368 | 1221 | 769 | | 1972 | 795 |
| 10000 | 516 | 11612 | 7021 | | 17559 | 7212 |

***Bottom line: if your sample sizes are small (less than a few hundred), you should use high-res format on a PC or picture format on a Mac for routine work. For large samples on a PC, use low-res format. For large samples on a Mac, editable charts use less than half as much space as pictures and they are produced more quickly because there is not an extra step for converting editable to picture format. For logistic models you should stick with editable format (the default) because most logistic regression***

***charts are interactive*. *When copying and pasting editable charts to Word and Powerpoint documents, use paste-special-picture or bitmap format to preserve chart appearance and avoid broken links.***

**7. One more thing to note about charts: if the data range contains more than 32K rows, only the first 32K points will be plotted.** This is an Excel constraint. If you want to produce charts for datasets larger than that, and if the ordering of rows is not important (e.g., not time series data) you may want to sort the rows randomly on the data sheet so that the 32K points that get plotted are reasonably representative of the whole sample. Charts for data sets this large are only feasible in low-res format on PC's.

**8. Running the data analysis procedure immediately after a regression yields matching descriptive statistics:** If you launch the data analysis procedure from a regression model worksheet, the default variable selections are those of the regression model, with the dependent variable designated as the variable to list first in the table and chart arrays. This allows a descriptive statistics report to be instantly generated for the same variables and sample used in a given regression model.

**9. Confidence levels and cutoff values can be adjusted interactively after fitting a model:** In RegressItPC and RegressItMac**,** if the "Formulas" option has been selected at run time, the confidence level that is stored in cell I10 on a regression model sheet can be interactively adjusted after the model is fitted, and all confidence limits in tables will be updated accordingly. If the "Editable" graph format option has also been chosen, confidence bands on charts will also be updated. In RegressItLogistic, formulas are always used by default because most of the outputs are intended to be interactive. In particular, the cutoff value for binary classification is always interactive. On a RegressItLogistic output worksheet, spinners next to the charts and tables are used to adjust the values. Graphs are updated if the "Interactive" option has been chosen. It is especially interesting to watch what happens on the ROC curve when the cutoff value is varied: a red square marks the current position on the curve and it will slide up and down. This is very useful for helping to determine a good value for the cutoff level rather than just accepting the default value of 0.5. The interactive ROC curve is a unique feature of RegressItLogistic.

**10. Regression line formulas are behind the line fit plot:** On the output sheet for a simple regression model, the table of values used for plotting the regression line and its confidence bands is located behind the plot itself, nicely formatted. If you grab the line fit plot and drag it to the right, you can see the table. It shows the calculations of predictions as well as standard errors and confidence limits for both means and predictions for 5 equally spaced values of the independent variable. If the "Formulas" was chosen at run time, the cells in the table contain live formulas, and you can plug in different values for the independent variable in the first column if you wish to see the corresponding results for them. If the "Editable" graph format option was also chosen, the line fit plot will be updated when the confidence level in cell I10 is changed. It is easy to rescale the axes on the plot to show new values that are outside the original ranges of the variables if necessary: just right-click on an axis scale, choose "Format Axis" from the pop-up menu, and set the minimum and maximum and crossing point to "Auto."

**11. Check the "Time series statistics" box when working with time series data:** Both the data analysis and regression procedure menus include a check-box for time series statistics. You should generally check this box when working with time series data. It has several functions. First, it causes a table of autocorrelations to be added to a data analysis output worksheet and a table of residual autocorrelations to be added to a regression model output worksheet. (See items 14 and 15 below for more about these.) Second, it activates the time transformation options (such as lag, difference, percent-change, etc.) in the variable transformation procedure. **If you have not checked the time series statistics box first, the menu of options that you see when you click the variable-transformation button will not include time**

**transformations.** And third, for a regression model, checking this box will cause connecting lines to be drawn between points on the actual-and-predicted-vs-observation-number chart.

**12. How to copy and paste tables and charts:** Many tables and charts on regression model worksheets contain numbers computed with live formulas. This allows the interactive adjustment of confidence limits, as noted above, and it has instructional value in showing the equations by which various statistics are computed, including R-squared and adjusted R-squared, t-stats, F-stats, P-values, standardized regression coefficients, standardized residuals, and standard errors of residual autocorrelations. It also has important implications for copying and pasting results. If you copy a table from the output of a regression model to another part of the Excel worksheet, you should paste the results as *values* rather than formulas, so that links do not get misdirected. **When copying and pasting a table of regression statistics to another place in the Excel file, use the Paste-Special/Values-and-Number-Formats command, and immediately afterward, while the range of pasted cells is still selected, also execute the Paste-Special/Formats command to get identical text formatting such as boldface column headings.** If you are pasting a table or chart to Word or Powerpoint, you can use any of the paste options or the paste-link option depending on how much or little further editing updating you wish to be able to do. You can even embed the Excel file in a Powerpoint file by pasting any portion of it as an Excel object, although this is not recommended for large data sets. **When copying and pasting tables and editable charts into Word or Powerpoint files, it is generally best to use one of the picture formats in order to preserve the integrity of the results**. If you wish to edit a chart for presentation, we recommend that you make a second copy of it in the Excel file and edit it there as desired. Then copy and paste the edited version to your document in picture form. You can either keep the copy on the original model worksheet (say, off to the right) or put it on a new worksheet. If you do the latter, you should move the new worksheet to the right of the Model Summaries worksheet, as described in item 13 below.

**13. The best way to move charts to new sheets:** It is often desirable to enlarge an editable chart to fill an entire sheet, using the move-or-copy sheet command. The charts are formatted so that they look good by default when enlarged in this way, and they are subject to further editing if needed. It helps to be systematic when doing this, in order to preserve a clean audit trail in the workbook. The following approach is recommended.

- **Make a second copy of the chart on the original worksheet** using the copy-and-paste command.
- **Move the copied chart to a new worksheet** by right-clicking on it and choosing "Move Chart" from the pop-up menu and selecting the "New Sheet" option.
- **Move the new worksheet to a position to the right of the Model Summaries worksheet** by clicking on its tab at the bottom of the window and dragging it to the right.

This procedure will serve two purposes. First, it will keep all the enlarged charts together, which makes it easier to find them and page back and forth among them. Second, and more importantly, it will prevent the worksheet counter from getting confused about where to insert the next data analysis or regression sheet. **If a new chart sheet is left in the middle of the existing sequence of data analysis and regression worksheets, any additional data analysis or regression worksheets will get out of order.** It is also advisable to rename a chart worksheet immediately after it is created. The default name it is assigned will be "Chart XXX", where XXX is a large number that is determined by the total number of individual charts on all the analysis worksheets that have been created so far. You should assign it a more descriptive name such as "Model 2 line fit".

**14. Calculation of out-of-sample forecasts:** When (and only when) the "Forecasts for missing or additional values" box is checked at the time a regression model is fitted, a forecast table and chart will be created on the output worksheet, and forecasts will be calculated by default for every row in which the dependent variable is missing and the independent variables are all present. If the forecast table and chart are maximized (i.e., unhidden), the forecasts will also be plotted on the actual-and-predicted-vs-observation-number chart. The forecast table and chart can also be minimized by clicking the "-" box next to them, in which case the forecasts will also disappear from the actual-and-predicted chart if it is in editable format.

**15. How to create an autocorrelation chart:** When the time series statistics box is checked for a descriptive analysis or regression model, a table of autocorrelations or residual autocorrelations is produced. To create a plot of the autocorrelations, proceed as follows. Select the cell range for the table, including the name(s) in column A and the lag values in the top row. Click the Insert/Column-chart button on the Excel toolbar to create a column chart from the selected data. This takes about 30 seconds.

**16. Residual autocorrelations in standard-errors-from-zero format:** When the time series statistics box is checked for a regression model, the output includes a table of residual autocorrelations. The values are also converted to standard-errors-from-zero units to easily judge their significance. These numbers are stored in the second row below the autocorrelations themselves, although by default their font color is set to white in order to hide them so as to prevent visual clutter. If you want to see them, just click the font or color button on the RegressIt ribbon. The standard error of the lag-k autocorrelation is 1/SQRT(n-k), where n is the sample size, and an autocorrelation is significantly different from zero at the 0.05 level if its absolute magnitude is greater than 2 standard errors. This value is calculated with a formula in the cell that includes the standard error, so the standard error itself is also available in the output.

**17. How to calculate standard errors and confidence intervals for predicted values (all of them, not just out-of-sample forecasts).** When a regression model is run, it is possible to have its predicted values and residuals stored in a table at the bottom of the model sheet, and you can use the Filter tool to sort them on absolute standardized values or influence or leverage to highlight the most significant errors. It is also possible to have their values written back to the data worksheet, where they will be stored in new columns inserted right next to the dependent variable. However, there is no menu command for computing or storing the corresponding standard errors and confidence limits. (These are calculated for out-of-sample forecasts, if any, but not for those within the sample.) If you want to do it anyway, it is not hard, provided the data set is not too large. First, copy the entire data range on the data worksheet and use the Paste-Special/Values command to paste the copy immediately below the original range, so that you now have a duplicate set of values for each variable, i.e., twice as many rows and the same number of columns. Second, go to Formulas/Name Manager on the Excel menu and delete all the existing range names. Third, select the entire new data range (including the duplicate rows for all the columns) and hit the Create-Variable-Name button to re-assign the original names to the now-twice-as-long variables. Finally, *delete the duplicate values for the dependent variable*. Then run or re-run your models using the "Forecast missing or additional values" option. Forecasts and confidence limits for the missing values of the dependent variable in the extra rows will be computed automatically and shown in the forecast table. (Note: these calculations can be time-consuming for large data sets, so after duplicating the data, you will probably not want to turn the forecasting option on for every model.) **Alternatively, you can use the R interface to fit the same model in RStudio and choose the "Export predictions for training set" option.** This will produce a table of predicted values with confidence intervals and standard errors that is written to the clipboard and can be pasted back into the Excel file. It only takes a few seconds to do this and you do not need to write any code.

**18. Calculation of R-squared and adjusted R-squared for a regression with no constant:** There are no universally accepted formulas for calculating R-squared and adjusted R-squared for a regression model that does not include a constant, as explained **here**, and some software packages do not even report them. RegressIt follows the same convention used by SPSS for no-constant models, namely: R-squared is defined as 1 minus {the sum of squared residuals divided by the sum of squared values of the dependent variable}, and adjusted R-squared is defined as 1 minus {the square of the standard error of the regression divided by the mean squared value of the dependent variable}. These formulas appear in the cells for R-squared and adjusted R-squared on the worksheet, and the regression statistics table shows the root-mean-squared value of the dependent variable rather than its standard deviation. **You should generally ignore R-squared and adjusted R-squared for no-constant models, and in particular do not try to compare them to the corresponding statistics for models that do include a constant.** Sometimes a naïve user will remove the constant from a model and think that this is an improvement because it gives a higher value of R-squared. No! Keep your eye on the bottom line, the standard error of the regression.

**19. Testing for normally distributed errors:** It is often (though not always) of interest to test whether the errors of a regression model are normally distributed. In particular, the formulas for calculating confidence intervals for predictions are based on the assumption of normally distributed errors, so it is important to test it in forecasting applications where confidence intervals are a desired output. The rationale for this assumption lies in the Central Limit Theorem, which states that a sum of independent identically distributed random variables converges to a normal distribution as the sample size grows large. In many regression applications the variables are constructed from sums or averages of large numbers of somewhat-independent somewhat-similar random quantities (e.g., monthly total sales is the sum of many independent random purchases by generic customers). In such cases, normally distributed errors might be reasonably expected in a good model. Also, violations of the normality assumption are sometimes indicative of failures of more important assumptions such as linearity. However, sometimes the unexplained variation in the data is just messy in ways that can't be fixed by cleaning or by transformations or by more complex model equations. So, a model should not necessarily be rejected or dispreferred to another model on the basis of the shape of its error distribution. *It's not the bottom line.* There are many different statistical tests for normally distributed errors. The one that is used in RegressIt is the **adjusted Anderson-Darling test**, which is considered by many to be the best for use in regression analysis over a wide range of sample sizes. The formulas for computing it, along with a discussion, can be found **here**, and an Excel worksheet that illustrates the calculations can be found **here.** A value greater than 0.752 indicates a deviation from normality that is significant at the 0.05 level and a value greater than 1.035 indicates that it is significant at the 0.01 level. However, as noted above, it is often too much to expect that a regression model's errors will be normally distributed enough to satisfy this test or any of the others, so it should not be used obsessively. If the Anderson-Darling statistic raises a red flag, the shape of the distribution of the errors (as revealed by the residual histogram and normal quantile plots) and the most extreme errors (which appear by default at the top of the residual table on the regression model worksheet) should be studied carefully to determine whether there is a systematic departure from normality or whether a few very large errors are to blame, and in the latter case, whether those very large errors had a lot of leverage with respect to estimation of the coefficients and whether they are likely to be repeated. You should also reflect on your model assumptions: does a linear/additive equation make sense for your variables? The Anderson-Darling stat takes a long time to compute in Excel for large sample sizes, and the alternative **Jarque-Bera test** (which is based only on skewness and kurtosis) is used in RegressIt for sample sizes greater than 2000, and it is always included in the cell comment for the normality test in the Model Summaries worksheet.

**20. Creating dummy variables:** The create-dummy-variable tool in the variable transformation procedure can be used to create dummy (0-1) variables for all distinct values of a given input variable. The input variable may have either text or integer values. The names that are assigned to the dummy variables are of the form X.Eq.zzz, where X is the input variable name and zzz is one of its unique values. **The cells in the data ranges for the created variables contain *formulas* to compute the 0's and 1's, not hard-coded values.** This is done in order to leave an audit trail and also allow for editing of the input values later. These formulas are of the form =IF(TEXT(inputvalue,"0") = zzz, 1, 0). For example, if you have a variable called Month that contains month-of-year data coded in text form (January, February, ...), the names of dummy variables created from it would be Month.Eq.January, Month.Eq.February, etc. If the Month variable is in column A, with data beginning in row 2, then the formula for the value of the Month.Eq.January variable in row 2 would be =IF(TEXT(A2,"0") = "January", 1, 0). The TEXT(.,"0") function rounds off numeric values. If the input values are numbers that are not integers, the rounded-off values will not match the actual data, and the values in the column will probably consist entirely of 0's. **If the data range for the dummy variables is very large, you may wish to convert the formulas to values by selecting the range, copying it to the clipboard, and then using the Paste-Special/Values command to paste it onto itself in value format.**