Spreadsheet software for linear regression analysis

Robert Nau
Fuqua School of Business, Duke University

Copies of these slides together with individual Excel files that demonstrate each program are available at **http://regressit.com/data.html** (which also includes other data analysis examples) or at the following direct links:

http://regressit.com/Comparison_of_add-ins_for_regression.pdf (these slides)

http://regressit.com/Beer_sales_with_RegressIt_analysis.xlsx

http://regressit.com/Beer_sales_with_StatTools_analysis.xlsx

http://regressit.com/Beer_sales_with_Analyse-it_analysis.xlsx

http://regressit.com/Beer_sales_with_XLSTAT_analysis.xlsm

http://regressit.com/Examples_of_regression_forecasts_from_4_add-ins.xlsm

- Issues that may arise with statistical analysis software in general:
  - The syntax, analysis options, and output have not been optimized from the viewpoint of today's user. They often have idiosyncratic origins and have been unchanged for many years out of mere inertia or to cater to an old client base.
  - In many programs, output consists of a linear log file that mingles code with results. This is good for editing and audit trail purposes, but the code can get in the way when you are trying to focus on results and compare models. Also, the code may not be intelligible to nonspecialists who try to read the files themselves.
  - Some programs produce chart output in separate windows, but these may lack permanence within or between sessions, or portability across computers, and/or side-by-side comparability among models.
  - Default tables and charts are often not good quality, not well formatted, nor adequately labeled to identify the variables, units, and models. It may be hard to correctly interpret what they show if they are copied elsewhere without modification. Numbers in tables often have far more decimal places than are relevant.
  - Of course, output can always be edited down and sharpened up for presentation (and you can do anything with R) , but often that is saved for the end stage of the analysis and is not done uniformly well, especially by students. Even a low barrier can be a nuisance and a distraction that gets in the way of careful work.
  - *High-bandwidth default output is important for more than presentation purposes: it helps the analyst to do a more efficient, thoughtful, and error-free analysis.*

- Excel add-ins for statistical analysis
  - Advantages
    - Excel users can work within their familiar environment and take advantage of all of its modeling and presentation features, besides those of the stat software.
    - Excel files are a common language for sharing data and models among individuals with different computer configurations and different levels of technical expertise.
    - Worksheets can be formatted to highlight the results and minimize the distractions
    - Users can easily navigate the results in 2 and 3 dimensions by paging among worksheets and scrolling around within them.
    - Live formulas can be intermingled with values: nice for teaching the math.
  - Disadvantages
    - Feature sets are typically much smaller than those of programming languages.
    - Audit trails can be problematic due to lack of visible code. It's easy to end up with a disorganized jumble of similar-looking, similarly-named worksheets.
      - What was done when and by whom and in what order? Which model produced the chart that is currently in view? What other models were tested beside the final one, and how close were their results? Can someone else easily replicate the analysis or perform their own variations on it?
    - Selection and formatting of output for printing can be a headache.

- My wish list for an Excel add-in
  - Lots of output on one sheet: complete results of descriptive data analysis or regression [or other analytic procedure], including all relevant tables and charts that provide different views of the same data sample or same model, and *as much of it as possible in the same field of view*.
  - Support for long, descriptive variable names.
  - Support for nonlinear transformations and time transformations of regression data. Transformations should be easy and be named in a systematic way.
  - Good worksheet design:
    - No more white space than necessary
    - Charts big enough but not too big
    - Most important tables and charts at the top
  - Presentation-quality tables with intelligent scaling of column widths, titles and numbers fully visible, enough digits but not too many.
  - Presentation-quality charts in native Excel format with intelligent scaling and thorough labeling of axes and effective use of color and point/line formats.
    (Tufte should approve.)

- My wish list, continued
  - Ease of use
    - Efficient and self-explanatory menu design.
    - Minimal typing or range-selection to be performed by clumsy fingers.
    - Ability to explore variations on any previous descriptive analysis or model (not merely the last one) without having to re-enter all of its specs from the beginning.
  - Unique titling of tables and charts so that they can be traced to their source if copied elsewhere or used in presentations.
  - Ability to print results to standard-width paper with minimal effort.
  - A clear audit trail for all analysis in the workbook: who did what, and when?
    Related: *it should not be easy for students to pass off others' work as their own.*
  - It should not be necessary to edit the default tables and charts in order to bring them up to acceptable standards for analysis, communication, internal documentation, and authentication. This is true even if the "client" with whom you are communicating is just your own future self.

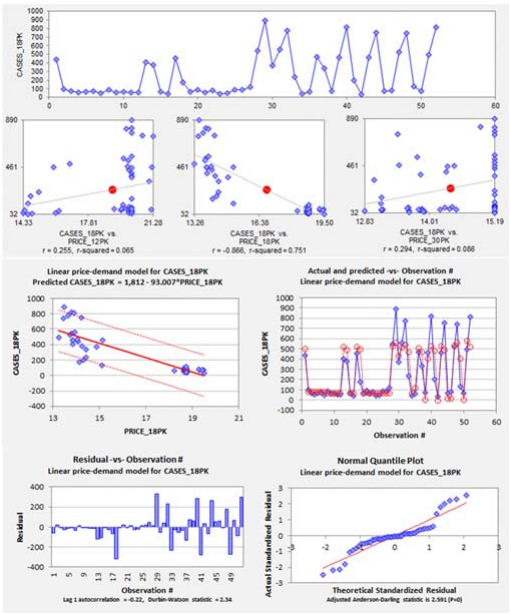
Most widely used Excel add-ins for statistics

- XLSTAT
  - Over 30,000 (50,000?) users.
  - A full-service stat program with a very complete menu of features, including ARIMA models for time series.
- Analyse-it
  - Over 25,000 users.
  - Includes descriptive statistics, linear and logistic regression, several kinds of ANOVA, and a wide range of statistical tests
  - No time series models or graphs.
- StatTools
  - Part of the Palisade "Decision Tools" suite, which has over 100K users.
  - Probably the add-in most used by MBA's (due to bundling).
  - Menu is similar to Analyse-it + exponential smoothing models for time series.
  - Includes a nice data transformation tool that the others lack.

They can all be tried for free for 30 days—visit their web sites for details.

- The RegressIt project
  - Excel add-in developed at the Fuqua School of Business over the last 6 years by myself and John Butler (McCombs School of Business, University of Texas) Why? Because commercial products didn't fulfill enough of my wish list.
  - Offered for free as a public service: visit **http://regressit.com**
  - Designed to support good modeling practices in a way that is good for both teaching and applications.
  - Used in teaching an advanced MBA elective course on statistical forecasting, as well as a core statistics course in a one-year master's program.
  - Menu is limited (descriptive data analysis and regression) but it is intended to be a "concept car" for design features that would be nice to have in general-purpose software. *Makes a good companion to other statistical add-ins.*
  - Written in Visual Basic, it currently runs only on PC's. *A Mac version is under development and is expected to be released within another couple of months.*

RegressIt



Distinctive features of RegressIt

- The user is prompted to enter an analyst/project name every time the program is launched. It is stamped in *bitmap* form (together with run time) on every worksheet for authentication.
- Variables are named column ranges. Names can be assigned in the usual way via Excel's formula menu or via a button on the RegressIt toolbar. Names can be up to 30 characters long with full visibility in all tables, charts, and dialog boxes.
- Charts and tables are in grouped rows so those not informative may be individually hidden.
- The data analysis and regression procedures produce a *lot* of output on a single worksheet, and all of it is presentation quality and fully labeled. Each analysis or model can be given a customized name which becomes the worksheet name and is included in all titles.
- A *data analysis* worksheet can include a summary stats table, autocorrelation table, correlation matrix, scatterplot matrix, and time series plots. Individual scatterplots can include regression lines and parameters of equations.
- A *regression analysis* worksheet includes many charts and tests, and many of the numbers are produced with live formulas for pedagogical value and to allow interactive manipulation of confidence levels. All table and chart titles include not only the unique model name but also the dependent variable name, sample size, and # variables.
- Forecasts can be automatically generated for missing values of the dependent variable and they are shown in two plots as well as a table, with std. errors & conf. limits for means & forecasts.
- A versatile data transformation tool is included, which has many time series options. Transformed variables are automatically assigned self-descriptive names.
- A "model summary" worksheet maintains an audit trail consisting of summary stats and coefficients for all models in the workbook, arranged side-by-side in a table suitable for presentation. *No other program offers audit trail features like these.*

A New Statistics and Forecasting Toolset for your Spreadsheet

Buy Now

Free Trial Download

Have you ever needed forecasting, regression, quality control charts, or other statistical analyses beyond the basics that are provided with Excel? Have you ever doubted the accuracy of some of Excel's statistical results? StatTools addresses both of these issues, providing a new, powerful statistics toolset to Excel.

StatTools covers the most commonly used statistical procedures, and offers unprecedented capabilities for adding new, custom analyses. StatTools replaces Excel's built-in statistics functions with its own calculations. The accuracy of Excel's built-in statistics calculations has often been questioned, so StatTools doesn't use them. All StatTools functions are true Excel functions, and behave exactly as native Excel functions do. Over 30 wide-ranging statistical procedures plus 9 built-in data utilities include forecasts, time series, descriptive statistics, normality tests, group comparisons, correlation, regression analysis, quality control, nonparametric tests, and more.

StatTools features live, "hot-linked" statistics calculations. Change a value in your dataset and your statistics report automatically updates. There is no need to manually re-run your analyses.

Learn how to get started quickly in StatTools
Watch videos of StatTools features

StatTools has also been fully translated into Spanish, German, French, Portuguese, Russian, Japanese, and Chinese.

> *"I've worked with Minitab before, and now abandoned it altogether since StatTools is so much better!"*
>
> *Alex Lebedev*
> *Lebedev Consulting*
> *Pretoria, South Africa*

Distinctive features of StatTools

- Variables are named Excel ranges, as in RegressIt.
- Includes a nice variable transformation tool, similar in concept to the one in RegressIt. Transformed variables are new descriptively named ranges.
- Data values can be updated interactively.
- Summary stats, correlations, scatterplot matrix, and series plots all require separate worksheets. Scatterplot matrix consists of an array of fully-labeled Excel charts, as in RegressIt, but they are bigger than they need to be.
- Axes are often not well scaled in charts. (Excel's default system is used, which tries too hard to anchor one end at zero.) Data points are small crosses or circles (always the same size) in dark blue font and do not stand out well for small data sets.
- Regression output does not include residual distribution or time plots or tests of model assumptions. Simple regression output does not show the fitted line on an X-Y plot.
- Forecasts are generated from regression models by setting up additional data worksheets with values of independent variables for forecasting. Forecasts are written to these same worksheets. This means that a separate forecast worksheet needs to be produced in addition to the model worksheet.
- Forecast output consists only of a table of forecasts and lower and upper confidence limits for forecasts. No standard errors and no plots.

Analyse-it

Analyse-it
Standard Edition

Transform Excel into a world-class statistical software package. Discover more about your data with powerful statistical analysis and visualization.

- Descriptive statistics, histograms, box-whisker plots, normal plots, CDF plots and more.
- Estimate parameters and test hypotheses for location, dispersion, and proportions
- Discover relationships with correlation, scatter plot matrices, principal component analysis (PCA) and biplots.
- Fit models and make predictions with simple linear, multiple linear, and logistic regression.
- From US$ 269 per-user, or US$ 99 annual

Learn more | Download 30-day trial

Analyse-it
Method Validation Edition

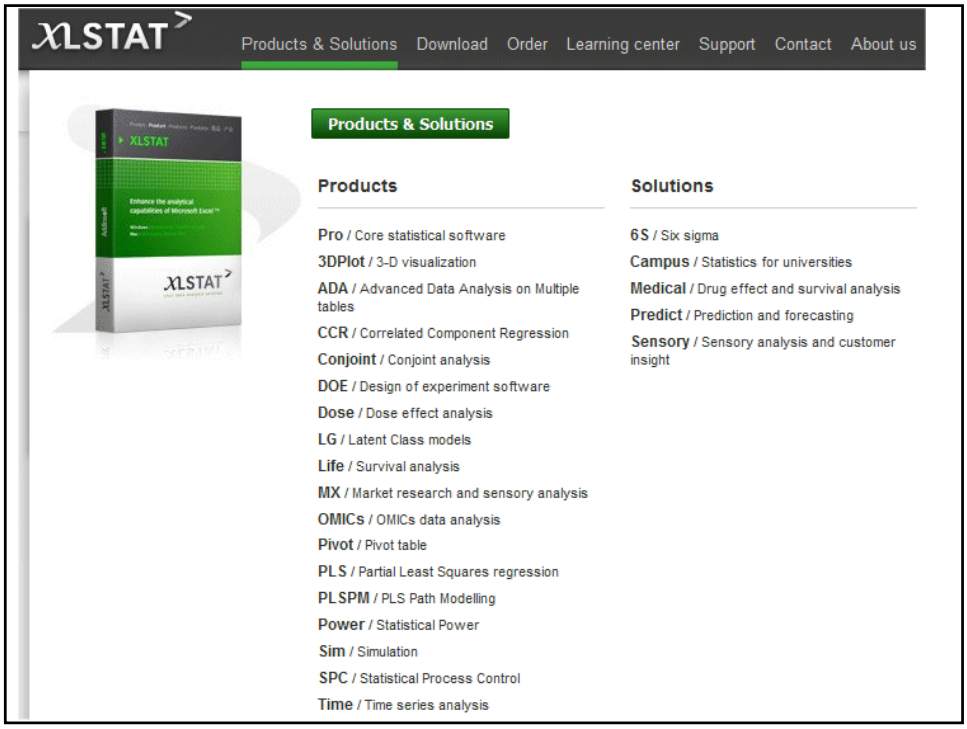
Meet laboratory regulatory compliance demands. Statistics software for method validation, to establish & verify analytical and diagnostic method performance.

- Compare methods with Bland-Altman, Linear regression, Deming regression, Weighted Deming regression and Passing-Bablok regression.
- Establish and verify within-device / within-laboratory precision and repeatability.
- Establish the linear measuring interval (reportable range), or calibration verification with linearity.
- Establish medical decisions levels with ROC curves and reference intervals (reference range).
- From US$ 699 per-user

Learn more | Download 30-day trial

Distinctive features of Analyse-it

- Variables have names that are automatically picked up from the first row of the data worksheet. (They do not become Excel range names, though.)
- Once an analysis is created, it remains "live". When its worksheet is revisited, the specifications can be edited, resulting in changes to the results, and the sheet name may change too. Nice for playing around, but it leaves no audit trail! Also, crazy things may happen if you make copies of worksheets.
- Uses some intelligence in determining how many decimal places to display.
- "Correlation & covariance" report includes summary stats, scatterplot matrix, and color-coded correlation matrix. The color-coding of the correlation matrix is effective, and the scatterplot matrix looks nice, but it doesn't have titles or axis scales on individual plots. There are no procedures for time series plots.
- The multiple regression output worksheet is dominated by two overly-large graphs: actual vs. predicted (with optional y=x and y=constant lines) and residuals vs. predicted. Optional output includes smaller residual distribution plots. A line fit (scatter) plot can be shown at the top for a simple regression model.
- Forecasts can be generated from regression models only by entering additional values for independent variables *by hand.*
- Forecast output consists of a table of forecasts, standard errors for means and forecasts, and confidence intervals for means (not forecasts). There are no forecast plots.



Distinctive features of XLSTAT

- Runs on Macs as well as PC's, a big advantage over the competition.
- A full-service stat package with a broad and deep menu: another advantage.
- Big *dis*advantage: *variables do not have names.* They are merely identified by cell coordinates. Input data for procedures is selected by pointing and clicking on columns or cell ranges. Names to show in the output are picked up from the first row if the range includes it, but range names are not remembered from one procedure to the next. (A named-variable option is promised soon.)
- Another disadvantage: *variable names and column titles are often not fully visible in tables* because a uniform narrow column width is used, irrespective of name or title length. The user has to manually adjust column widths to see all the details and to format the worksheet for any sort of presentation.
- "Correlation test" procedure creates a worksheet very similar to the data analysis worksheet in RegressIt, minus the series plots.
- A newly added feature produces series plots on a separate worksheet.
- Regression output worksheet is dominated by a very large column chart of standardized coefficients . (Not the best use of prime space, especially for a simple regression.)
- Out-of-sample forecasts for regression models are generated by selecting the ranges that contain additional values of the independent variables. These selections must *not* include names.
- Forecast table includes standard errors and confidence limits for both means and forecasts, but they are not plotted.

Forecasting from a regression model: big differences in procedure options and outputs

- Additional data for forecasting may need to be entered or selected in an additional step.
- Forecast data may or may not be read from the original data worksheet. It may require manual entry or loading from a separate worksheet.
- Forecast output may or may not include standard errors and confidence limits for both means and forecasts
- Chart output may or may not be provided

This data set will be used to compare the forecasts from regression models that can be produced by four Excel add-ins: RegressIt, Analyse-it, StatTools, and XLSTAT.

Row	Y	X1	X2	X3
1	27	4	8	6
2	12	1	5	1
3	28	7	4	2
4	13	0	4	5
5	24	3	5	8
6	27	7	2	4
7	25	5	8	4
8	29	6	6	9
9	37	8	5	7
10	18	2	8	1
11		1	6	8
12		2	5	1
13		3	9	4
14		4	4	3
15		5	8	6

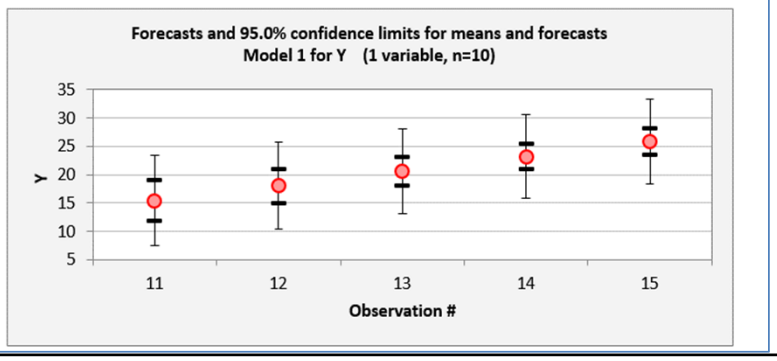
The objective is to generate forecasts for 5 additional values of Y (rows 11-15 of the data set) using two different models: a simple regression of Y on X1 and a multiple regression of Y on X1, X2, and X3.

The simple regression model results are shown on the following slides. See the accompanying Excel file for the multiple regressions.

- In **RegressIt**, forecasts can be generated automatically for rows on the data worksheet where the dependent variables are missing and the independent variables are all present. Both a table and a plot of the forecasts are produced.
- The confidence level used in the forecast table and charts can be interactively adjusted on the worksheet after fitting the model.

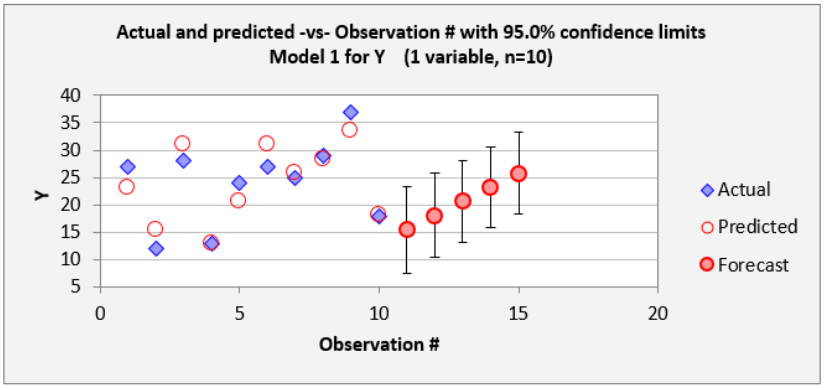
Forecasts: Model 1 for Y (1 variable, n=10)

Obs#	Forecast	StErrFcst	Lower95%F	Upper95%F	StErrMean	Lower95%M	Upper95%M	X1_
11	15.471	3.440	7.539	23.404	1.562	11.868	19.074	1
12	18.056	3.326	10.386	25.725	1.292	15.077	21.035	2
13	20.640	3.250	13.145	28.136	1.083	18.143	23.137	3
14	23.225	3.216	15.808	30.641	0.976	20.975	25.474	4
15	25.809	3.225	18.373	33.245	1.003	23.495	28.123	5



The forecasts are also shown on the actual-and-predicted-vs-observation # chart if the forecast table is currently maximized (not hidden). If the forecast table is minimized, the forecasts are not shown on this chart either.

Actual and predicted -vs- Observation #



- In **Analyse-it**, forecasts can be generated after fitting a model only by entering additional values for the dependent variables by hand.
- The output consists of the table shown here, which includes forecasts, standard errors for means and forecasts, and confidence limits for means.

Predict Y for X

	Predicted Y	95% CI	Mean SE	Individual SE
X1=1	15.5	11.9 to 19.1	1.56	3.4
X1=2	18.1	15.1 to 21.0	1.29	3.3
X1=3	20.6	18.1 to 23.1	1.08	3.3
X1=4	23.2	21.0 to 25.5	0.98	3.2
X1=5	25.8	23.5 to 28.1	1.00	3.2

- In **StatTools**, a second data set is specified for purposes of forecasting, and the forecast output is written to a new worksheet.
- The output consists of the table shown here, which includes only forecasts and confidence limits for forecasts.

	A	B	C	D	E	F	G
1	Row	X1	X2	X3	Y	LowerLimi	UpperLimit95
2	11	1	6	8	15.47137	7.538697	23.40403
3	12	2	5	1	18.0558	10.38637	25.72523
4	13	3	9	4	20.64023	13.14492	28.13555
5	14	4	4	3	23.22467	15.80804	30.64129
6	15	5	8	6	25.8091	18.37273	33.24548

- In **XLSTAT,** the data ranges to be used for forecasting are selected by hand, separately from the selection of the variables in the model. (It is important to select columns in the same order when doing this.)

- The forecast output consists of the table shown here, which includes forecasts as well as standard errors and confidence limits for both means and forecasts. Column headings are not fully visible, though.

Predictions for the new observations:

Observatior	X1	Pred(Y)	v. on pred.	ound 95%	ound 95%	n pred. (Ol	nd 95% (Ol	nd 95% (Observation)
PredObs1	1.000	15.471	1.562	11.868	19.074	3.440	7.539	23.404
PredObs2	2.000	18.056	1.292	15.077	21.035	3.326	10.386	25.725
PredObs3	3.000	20.640	1.083	18.143	23.137	3.250	13.145	28.136
PredObs4	4.000	23.225	0.976	20.975	25.474	3.216	15.808	30.641
PredObs5	5.000	25.809	1.003	23.495	28.123	3.225	18.373	33.245

Summary

- Statistical add-ins still have a useful role to play in environments where Excel is used to any extent, even if R is taking over the larger world.

- Demand better: If you are a user of an Excel add-in, think about how its user interface and/or its output could be improved, and let the developers know what you think.

- RegressIt illustrates some features that you might want to have. It can be operated simultaneously with other add-ins, effectively adding more tools and better-looking output to their menus for regression and multivariate data analysis.
  Try it out if you are a PC user.

- Copies of these slides plus Excel files that demonstrate the programs in detail are available on the Data page at http://regressit.com.

- Also visit http://statforecasting.com for extensive teaching notes.