Wizard: Data Analysis for the Non-Statistician

By Evan Miller

Version 1.4

Table of Contents

I. Introduction 3

II. Getting started 5

III. Summary statistics and single-column tests 8

IV. Scatterplots, category break-downs, and two-column tests 12

V. Building a multivariate model 14

VI. Making predictions 18

I. Introduction

Wizard is the first statistics program designed to make multivariate data analysis *easy* and *fun*. The statistical reporting is accessible: p-values are accompanied by pictures, and if you can move a mouse, you can perform a t-test. With interactive graphics, contextual menus, and ubiquitous drag-and-drop, it looks and feels like a desktop productivity program. If you have occasional analysis needs that aren't met by spreadsheet programs, or if you feel a combination of fear and loathing whenever you fire up a beefy statistical package such as R or SPSS, you'll find a friendly companion in Wizard.

I should first explain what Wizard is *not.* Unlike existing statistical software, Wizard is not a graphical shell bolted on top of an ancient command-line program. In fact, there is not a command line to be found in the entire program, nor is there any kind of "console" that prints out reams and reams of numbers in a fixed-pitch font. All of the program interaction, and all of the statistical reporting, occurs through native Mac menus and widgets. Because all of Wizard's statistical routines have been written from scratch, Wizard does not require any other program to be installed first.

Professional statisticians will be disappointed to learn that Wizard is quite limited in its range of functions. Wizard generates only a few kinds of plots (scatterplots, pie charts, histograms) and performs only a handful of multivariate regressions (ordinary least squares, probit and logit, and a couple of models for count data). But then Wizard is not designed for the professional statistician. It is designed for the researcher or analyst who needs to perform a run-of-the-mill analysis on prepared data, and then make sense out of the results.

Of course, when performing an ordinary analysis, it can be easy to get lost in the details -- at least with other programs. Wizard, however, is designed to keep your data and results as organized as your iTunes library. You can have as many data tables as you like in a single document. You can join tables horizontally and stack them vertically through simple interfaces, and you can always ask to see the underlying linking logic. You can rearrange the variables in a table by dragging and dropping. You can make recode variables, and edit the recoding logic later on. If you accidentally delete a variable, you can press Undo to get it

back from Variable Valhalla. Wizard keeps your models organized, too. It treats each regression as an object inside the document, so you can name each regression, and then rearrange them, duplicate them, edit them, and delete them as you wish.

In some ways, Wizard is more powerful than the heavy-duty programs. Once you've built a model and computed its coefficients, Wizard helps you perform back-of-the-envelope calculations and simulate "what if?" scenarios. What if the dosage is doubled? What if the customer is new? What if the person is 60 years old and married? Wizard gently prods you into asking interesting questions about your data. It also helps you organize the answers; each "what-if" scenario is an object in the document that can be named, edited, and rearranged.

Finally, Wizard is fast and responsive. Most graphics appear instantly, and simple regressions are computed in the blink of an eye, even with hundreds of thousands of rows in your data set. Wizard's secret is that it takes advantage of all the cores on your computer. (So if you are working with a particularly large data set or estimating complex models, upgrading your hardware can noticeably improve performance.)

Some scientists will probably sniff at Wizard's limited functionality and its use of "eye candy". But it was with a very serious purpose that I decided to make Wizard fun, easy-to-use, and inexpensive. Today, data analysis feels like a chore, like going to confession in a large cathedral. I think it should feel more like a treat, like an interesting conversation among friends. My hope is that Wizard will make you want to start analyzing data of out curiosity rather than obligation. The result, I think, will be better data, better understanding, and better science.

II. Getting started

Like word processors or spreadsheet programs, Wizard is a document-oriented application; you can open, save, and close documents from the File menu, and you can have as many independent documents open at once as you like. Wizard documents contain both data and analyses, and have the extension `.wizard`.

Although you can enter data into Wizard manually, the program is primarily designed to work on prepared data. Wizard can import the following data formats:

* Comma-separated variables (CSV) with extension `.csv`
* Tab-delimited files with extension `.tab`, `.txt`, or `.tsv`
* Excel spreadsheets with extension `.xls` or `.xlsx`
* Numbers spreadsheets with extension `.numbers`
* R workspaces with extension `.rdata`
* SQLite database files with extension `.sqlite`
* Access database files with extension `.mdb` or `.accdb`

Wizard can also import data by connecting directly to MySQL and PostgreSQL databases. Wizard Pro can additionally import these formats:

* Stata binary files with extension `.dta`
* Stata dictionary files with extension `.dct` (and an accompanying data file with any extension)
* SPSS binary files with extension `.sav`
* SPSS portable files with extension `.por`
* SPSS command files with extension `.sps` (and an accompanying data file with any extension)
* SAS binary files with extension `.sas7bdat`

If you are using Wizard Pro, it's best to import data from either the Stata or SPSS binary formats. Binary imports are faster than plain-text imports, and Wizard can glean additional information (such as value labels) from these formats that is not present in plain-text files.

A word of caution is in order when importing from CSV or tab-delimited files. Wizard will attempt to infer whether a column consists of numeric data or textual data by looking at the first row of data. If it guesses incorrectly (for example, if numeric-looking data in a text column causes Wizard to think the column is actually numeric), it is up to you to go back and "trick" Wizard into making the correct inference by modifying the input file, for example, by rearranging the rows.

Enough about importing; go ahead and import your first file by clicking the "Open New Document" button in the Welcome screen, and then clicking "Import Data" in the toolbar. For the rest of this tutorial I will be using a "classic" data set, one that should be familiar to most Stata users: the Stata automobile data set. It can be downloaded from http://wizard.evanmiller.org/static/auto.csv and imported into Wizard without trouble.

When you import a file, you'll first be presented with the raw data. This table looks something like a spreadsheet, but there are important differences. It cannot be edited, for one. In fact, this table has relatively few features; it is solely designed for sorting and inspecting data.

Try clicking the column headings to sort the records by columns of interest. If you're using the automobile data set, sort by "price" to see the most and least expensive cars in the table. You can always return to the original ordering by clicking the first column heading (which corresponds to the row number).

Because data sets are often large and cumbersome, this table can be filtered by column or by row. First try filtering the columns by using the search box in the bottom left. For example, typing the letter "m" into the search box will show you only the columns that contain the letter "m". This feature is not especially useful in tables with only a handful of columns, but for wider data sets, it is a godsend.

Next try filtering by row. Click the "Data Filters" button on the top right to reveal the filters pane. You can filter the raw data by any field that is treated as a category (more on categories later). For example, in the automobile data set, selecting "foreign" from the pop-up button will let you view domestic cars, foreign cars, or both. Up to three filters can be applied at once. You can make the filters disappear by clicking the "Data

Filters” button again, or by dragging the divider up toward the top of the window.

That is about all there is to do in the Raw Data view. If the data looks like it was correctly imported, we can start having fun!

III. Summary statistics and single-column tests

To start analyzing the imported data set, click the “Summary” button in the toolbar.

The Summary view is a bit overwhelming at first sight, but it will all make sense in just a few minutes. To keep things simple, close the filters pane if you haven’t already by dragging the divider up. Now let’s take a look around.

On the left is a list of all the columns in the data table. You can view and edit each column’s short name (for example, “make”) and description (if present), as well as view summary statistics. By default, no summary statistics are visible; to see them, just control-click the table header. Available statistics include the count (number of non-missing values), mean (average), standard deviation, variance, standard error of the mean, minimum, maximum, and variance percentiles.

The Summary view includes a couple of familiar controls. You can again filter the columns using the search box in the bottom left. As before, you can use the filters pane to analyze only a subset of the data. Try it.

But the fun begins when we select a column to analyze. Go ahead and click one.

The top-right view will immediately give you a visual summary of the column you clicked. The summary style depends on the type of column selected. There are four basic kinds of data columns in Wizard:

* Category columns. A category column can contain textual data or numeric data, but usually has a limited number of discrete values. Summarized with clickable pie charts.
* Textual, non-category columns. Summarized as “(M observations, N values)”
* Numeric, non-category columns. Summarized with a histogram and interactive t-test of the column mean.
* Date and time columns. Summarized with a histogram and confidence intervals around the estimated arrival rate in each time period.

Text columns are rather boring if they are not treated as categories.

Next try selecting a numeric column (you can tell whether a column is numeric by whether it has a mean displayed in the left-hand column table). Now things start to get interesting!

The top half of the column-summary view shows a histogram of all the values in this column. You can adjust the number of bins by using the slider in the bottom-right corner of the column-summary view. This histogram should give you a quick overview of the distribution of values in the data set.

But the real action is in the bottom half of the column-summary view. This view helps you answer the following question:

"Assuming my data is a random sample from a population, what do we know about the population mean?"

Statistics is an uncertain science, so you need to choose a confidence level to say anything useful. Use the confidence slider to pick a confidence level and see the corresponding amount of statistical uncertainty that surrounds the estimated population mean. Note that the confidence level simply corresponds to the percent of the curve's total area that is dark blue, that is, between the two dashed lines, which represent the bounds of the confidence interval. (To perform a formal, one-sample t-test, simply see whether the hypothesized mean falls between these two confidence bounds.)

Below the summary view, you'll see what appears to be a scatterplot. This is actually known as a Q-Q plot, and it is useful for determining whether data in a numeric column seem to follow a normal (Gaussian) distribution. If the dots follow a straight line, the data are more or less normally distributed; otherwise, the data are not normally distributed.

We'll finish the basic tour of data types by selecting a category column. In the automobile data set, "foreign" is the only category column. However, we can treat any column as categories by control-clicking it and choosing "Treat as Categories". If you're using the automobile data set, go ahead and do this for the "rep78" column, which can only take one of 5 values.

Whereas numeric columns are visualized with a histogram and t-test, category columns are visualized with two pie charts. The pie chart on the left is a summary of the actual data. You can choose to display missing values in the left pie chart with the "Show Missing" checkbox. Missing values will then show up as a white slice of the pie, starting at 12 o'clock. (If you're using the automobile data set, try toggling this checkbox with the "rep78" column selected.)

However, the pie chart on the right is not affected by missing values, because it has to do with population inference. It attempts to answer the question:

"Assuming my data is a random sample from a population, what do we know about the proportion of values in the population?"

This pie chart has white slices labeled "Uncertain" that reflect the statistical uncertainty at the given confidence level. (In statistician's terms, each colored slice reflects the lower bound of a simultaneous multinomial confidence interval, and the "Uncertain" slices represent whatever is left.) Move around the confidence slider to see the effect on the statistical uncertainty.

The view in the bottom center will display a bar chart of all the category values in this column. You can choose whether to sort these bars from largest to smallest using the "Sort by amount" checkbox. You can also choose to view raw count data, relative proportions, or the confidence intervals around the relative proportions by toggling the radio button in the bottom-right.

There is one more feature worth introducing here. The bottom of the document window has my second-favorite feature in the entire program, known as "The Bottom Line." The Bottom Line provides a plain-English interpretation of whatever statistical analysis is currently being performed, and will occasionally provide usage hints. For example, when a category variable is selected, the Bottom Line will tell you whether the values appear to exist in equal proportions in the underlying population. If you have selected the "foreign" column in the automobile data set, it will say "Unequal proportions (2 values)".

You've now seen the most important Wizard features available for looking at individual columns. There are many more features in the Column menu (which can also be accessed by control-clicking a data column); from there you can perform various recodes and computations on existing columns. The curious reader is encouraged to explore these options. But for now, we'll press ahead with multivariate analysis.

In the next section we'll look at interactions and correlations between two columns, which can start to reveal patterns in the data. From there it will be a quick jump to building formal multivariate models (which will help us confirm or refute the patterns we think we see), and finally, we'll use our models to make predictions.

IV. Scatterplots, category break-downs, and two-column tests

The table in the bottom-right corner of the document window is called the covariate table. It lets us examine the interaction of the selected column with another column in the data set.

Wizard can create four kinds of two-column summaries, which will appear in the bottom-center view:

* If a category column is selected on the left, Wizard will display a set of “flattened” population pie charts. These are broken down by category (or in the case of numeric columns, quantile) of the column selected on the right
* If a date or time column is selected on the left, Wizard will display a set of histograms broken down by category or quantile of the column selected on the right
* If a non-category numeric column is selected on the left, Wizard will display a set of histograms, t-tests, dot plots, or box plots, again broken down by category or quantile of the column selected on the right
* If two non-category numeric columns are selected, Wizard will display a scatterplot of the two values when the “Dots” button is selected

We’ll look at a quick example of each of these charts. To start, select a category column from the main column table on the left, and any category or numeric column from the covariate table. Recall that category columns can be identified by the gray expansion triangle next to their names. If you’re using the automobile data set, select “foreign” in the left-hand (summary) table and “weight” in the right-hand (covariate) table.

You’ll now see a flattened version of the “Estimated Population” pie chart for each category or quantile. “The Bottom Line” will tell you whether the distribution of one column appears to be affected by the value of the other column. (In statistician’s terms, it is performing a chi-square test of the independence of both columns when the covariate is a category column; otherwise, it is performing an analysis-of-variance of the numeric covariate across values of the category column.)

You can focus on a particular category value by clicking it in the pie chart, or by expanding the gray triangle and selecting the value of interest. When you do, that value will be highlighted in the bottom-center view. If you'd like to see correlation coefficients between the selected category and each covariate, control-click the header of the covariate table and select "Correlation". Correlation coefficients that appear in black are statistically significant; coefficient in gray are not statistically significant. Note that the desired significance level is inferred from the confidence slider. If you'd like to see individual p-values, just control-click the column heading of the covariate table and choose "P-value".

For the next trick, select a numeric column from the main column table and a category column from the covariate table. You'll now see a set of dots for each category representing the distribution of the numeric column within that category. Click the "Bars" too view a histogram for each category, and click "t-test" to see the familiar blue curves with dashed lines, which represent confidence intervals on the numeric column for each value of the category column. If there is a lot of uncertainty in an estimated mean, its curve will be wide; if the estimate is precise, its curve will be narrow. As before, you can move the confidence slider and see the effect on the confidence intervals.

The Bottom Line tells us the result of performing an analysis of variance (ANOVA) across all of the categories.

You can reduce a many-sample test to a two-sample test by selecting an individual category value from the covariate table. The two samples will be the selected value, and "Other".

Next, select a numeric column from the left table, and a numeric column from the covariate table, and click the "Dots" button below the covariate view. You'll now see a simple scatterplot of the two columns. The covariate table will display relevant correlation coefficients, and the Bottom Line will tell you whether the selected correlation is statistically significant.

We've now covered the entire summary view -- almost. You might be wondering why there are checkboxes in the covariate table, and what the big "Build Model" button is for. In the next section, we'll find out.

V. Building a multivariate model

A model in Wizard has two parts: something to be explained (variously known as a dependent variable, explained variable, or outcome variable), and a set of things that do the explaining (known as independent variables, predictor variables, or explanatory variables). In Wizard, we'll use the terminology *outcome variable* and *explanatory variables.*

Building a model in Wizard is easy: in the Summary view, first click the Lock button in the bottom-right corner. Next select the desired outcome variable on the left, check the box of all the covariates that will serve as explanatory variables on the right, and then click "Build Model".

If you're using the automobile data, try a very simple model: select "price" on the left, check the box next to "mpg", and press "Build Model". This will attempt to model each car's price as a function of its gas mileage.

You'll see a new screen, known as the Modeler. Let's take a quick tour.

The Modeler is divided in half, with information about the outcome variable on the left, and information about the explanatory variables on the right. The outcome variable table (top left) displays information about the number of observations used in the regression. Additional diagnostic numbers, such as the R-squared, can be added to the table by control-clicking the column header of the outcome variable table.

Since "price" is a non-category numeric column, Wizard has modeled it with a linear regression. We can perform a weighted linear regression by checking the "Weight by" box and choosing a column to use as analytic weights. (Category columns, on the other hand, are modeled with a logistic regression.)

Below the outcome variable table is a scatterplot of model residuals (that is to say, actual values of the outcome minus the values predicted by the model). You can plot model residuals against any column in the data set with the pop-up button at the bottom. This kind of residual analysis can provide clues about whether we've left out anything important from our model. Just for fun, try adding "headroom" to the model with the "Add to

Model" button; this will model "price" as a function of both "mpg" and "headroom".

Next, take a look at the explanatory variable table over on the right. This is a table of estimated model coefficients with point estimates, standard errors, and a significance indicator for each explanatory variable in the model. The significance indicator will show one green bar if the coefficient is significantly different from zero at the 10% level; two green bars if significant at the 5% level; and three green bars at the 1% level. (Other programs, and academic papers, usually display the same information with one, two, or three asterisks next to the coefficient.) Additional information about each coefficient is available by control-clicking the column header of the explanatory variable table.

The checkbox next to each coefficient can be used to individually exclude variables from the model without deleting them altogether. Try it. To remove a variable from the model entirely, control-click it and choose "Remove from Model" (or select it in the explanatory variable table, and then choose Model > Explanatory Variable > Remove from Model). Removing variables from the model, of course, will not drop it from the underlying data table.

If your analysis requires robust or clustered standard errors (common in the social sciences, but a not always a bright idea in the physical sciences), use the checkboxes below the explanatory variable table.

Finally, the bottom-right panel displays a visualization of the coefficient selected in the explanatory variable table. The visualization requires a bit of explanation.

The drawn curve is the actual t distribution used in testing hypotheses about the selected estimate. Technically, it should appear to be centered at zero (the location of the null hypothesis), but centering it at the point estimate makes the interpretation more intuitive. The point estimate is indicated by a thick black dashed line, and the locations of standard errors are marked by thin gray dashed lines. Point estimates with a lot of uncertainty (that is, large standard errors) will have wide curves; precise point estimates will have narrow curves.

The area under the curve on the far side of zero, and its mirror image in the other direction, correspond to the actual p-value of the statistical significance test. This area is colored dark blue. If you don't see any dark blue, select an insignificant coefficient (such as "headroom" in the automobile example). Highly significant coefficients have very low p-values and therefore almost no dark blue visible. If you'd like, you can perform a one-sided test instead of a two-sided test using the "Null hypothesis" radio button at the bottom. Play with it to see the effect on the p-values.

This coefficient view makes it easy to see how an individual coefficient's estimated value and standard error are affected by changes in the model. To see what I mean with the automobile data set, try selecting "headroom" in the explanatory variable table and then checking and unchecking the box for "mpg".

Now that you've gotten a feel for the interface, we'll talk a little bit about the model and see what we can do to improve it. Note that the coefficient on "mpg" is *negative*; that is, cars with better mileage are actually *cheaper* than cars with worse mileage. Seems counterintuitive, right? Other things being equal, wouldn't people pay for cars with better mileage?

The trouble is, other things are *not* equal until we've added them to the model. Go ahead load up the model with other explanatory variables using the pop-up button and "Add to Model" button on the bottom left. (Note that when you add a category column to the model, a coefficient is estimated for every category value except the first; these coefficients are relative to the first value, which is deliberately omitted from the model.)

Once you add these additional variables, a new picture will emerge: a car's price can be explained by its weight and whether it is foreign, but the other variables (including the mileage) don't seem to matter very much. This solves a little puzzle from before: the price coefficient on mileage was negative because cars with good mileage tend to be small, but small cars tend to be cheap! After controlling for the car's weight, the price-mileage effect is not statistically significant.

With most statistical programs, estimating a model is usually the end of the story. With Wizard, it is the beginning. In the next section, we'll use

our model to make concrete predictions with an interactive interface known as the Predictor.

VI. Making predictions

Once you've made a model that you are satisfied with, you can use it to predict values for the chosen outcome variable given specific values of the explanatory variables. You can make your first prediction by clicking "Predict" in the toolbar.

Here you'll see the Predictor, which displays an interactive equation. In a linear model, each line of the equation consists of a coefficient multiplied by an input value, and the lines are simply added up to produce a prediction at the bottom. (In non-linear models, the linear sum is transformed in some way in order to produce a final prediction.) You can confirm that the coefficients are the exactly the same ones that appeared in the Modeler.

You can use conservative or generous estimates of particular coefficients with the sliders on the left. Each tick mark on the slider corresponds to a standard error on the coefficient. In this way, you can see how sensitive a prediction is to the statistical uncertainty on the input coefficients.

But the real fun is in the controls on the right side. These let you manipulate input values themselves and make predictions about cars that are not in the data set. How much would a heavy, foreign, head-room-intensive car cost? What about a short, light, domestic car with terrible mileage? You can ask and answer these all of these questions with the provided controls and by clicking directly on the graph in the bottom-right corner. Try it!

With these tools in hand, it's easy to generate a number of predictions quickly. But what to do with all of them? So far I have omitted any discussion of Wizard's features for organizing an analysis. I should introduce them before you get too carried away with the Predictor.

Click the "Predictions" button in the toolbar to bring up the Prediction organizer. This is a table of hypothetical input values along with the predicted value of the outcome variable. You can delete or duplicate a specific prediction with the buttons on the right-hand side. Selecting a prediction in the Prediction organizer loads it into the Predictor, where it can be edited and explored.

Predictions aren't the only things that can be easily organized in a Wizard document. You probably noticed the navigational pane on the left side of the document window; if you've been following along so far, you should see "auto" and "price model" in outline form. This pane is where we can select, name, and store multiple models and multiple tables.

Try following your work backwards by clicking on "auto" to make sure the table is still there. You can give your model a more creative name just by double-clicking it. You can also duplicate or delete any object in the navigational pane by control-clicking it and choosing "Duplicate" or "Delete"; note that all of the object's children are copied (or deleted) as well. Try duplicating and editing some of these objects, or rearranging them by dragging and dropping; you should quickly appreciate how Wizard can be used to organize complex analyses with many data tables, many variations on each model, and many concrete predictions.

***

I believe that some congratulations are in order. You are now an official Wizard user! You know how to import, organize, summarize, analyze, and interpret data. In just an hour or two, you've learned how to do all the things needed to start performing real research with Wizard, or just gain some interesting insight from a data set. Of course, there is more to Wizard than what we've seen in this guide, and you are encouraged to explore the menus (especially the Help menu) to discover all of Wizard's features. But you've seen the most useful parts, and you are sufficiently equipped to go out now and start making your own discoveries.