Exploratory analysis: what to do first

Robert W Platt

One of the more important but often overlooked parts of statistical analysis is the very first step—an exploratory and descriptive analysis. Typically, researchers take a quick look at the data and then dive into more complex regression models or *t* tests. In this column, I discuss preliminary analysis in general and look at some techniques less well known than others, but which provide interesting and useful results.

The first step in understanding your data is to establish the kinds of variables you have. Are they continuous (ranging over several values, like weight or height) or categorical (taking only a few values)? Are the continuous variables bounded (like age, which can't be less than zero) or unbounded? Are there any outliers or strange values?

This last question can be looked at in a simple way. Calculate the mean and standard deviation of a variable, and examine values that are more than three, or if you want to be very careful, two, standard deviations from the mean. If there are outliers, they need to be investigated, and either eliminated (if they are errors) or treated carefully (if they are valid data points). Next, for the continuous variables, look at histograms of your data, and for the categorical variables, look at frequency tables. These will tell you roughly what the distributions of the variables are and this influences the statistics you can use.

The next thing to consider is bivariate analyses of the data. First, what do we do with continuous variables? A common mistake is to examine correlations first. But these are usually an inefficient way of inspecting the data, because significant correlations depend on a linear relationship between the variables and if the true relationship is curved, the correlation may not indicate the association. Another

approach is to graph a scatterplot of the two variables and check for a relationship.

For categorical variables, it is easiest to inspect bivariate (for example 2×2) cross tabulations to identify patterns and potentially interesting relationships. These relationships provide the baseline for further analyses.

Finally, a multivariate exploratory analysis may be needed to detect possible confounding (the mixing of effects of an outcome, an exposure and a third variable that is associated with the primary predictor and also affects the outcome) or effect modification (when the effect of an exposure on the outcome differs for different levels of a third variable). The easiest way to do this is with a bivariate analysis stratified by the third variable. If the latter is categorical just look at the relationship between the other two variables restricted to the levels of the third, and if it is continuous, create a new categorical variable. If there is important confounding or effect modification (the definition of "important" here is arbitrary and depends on the needs of the analysis) these must be accounted for in the formal models when computing estimates of the primary predictor.

After these preliminary analyses, the patterns and relationships in the results should be reasonably clear and the analyses that need to be done should be obvious. If this is the case, then the rest is simple—for continuous variables, *t* tests, ANOVA, or linear regression can be used to confirm the exploratory work. Similarly, for categorical data, χ^2 or non-parametric tests can be used.

If patterns in the results are not clear, two things are possible: either there aren't any interesting relationships, or there are but they are complex and you need to consult a statistician!

**McGill
University/Montreal
Children's Hospital
Research Institute,
2300 Tupper, Montreal,
PQ H3H 1P3, Canada**

Correspondence to:
Dr Platt
(robertp@epid.lan.mcgill.ca).