Multivariable Systems

Lawrence Hubert

July 31, 2011

Whenever results are presented within a multivariate context, it is important to remember that there is a system present among the variables, and this has a number of implications for how we proceed.

Automated analysis methods that cull through collections of independent variables to locate the "best" regression equations (e.g., by forward selection, backward elimination, or the hybrid of stepwise regression), are among the most misused statistical methods available in all the common software packages.

They offer a false promise of blind theory-building without user intervention, but the incongruities present in their use are just too great for this to be a reasonable strategy of data analysis:

a) one does not necessarily end up with the "best" prediction equations for a given number of variables;

b) different implementations of the process don't necessarily end up with the same equations;

c) given that a system of interrelated variables is present, the variables not selected cannot be said to be unimportant;

d) the order in which variables enter or leave in the process of building the equation does not necessarily reflect their importance;

e) all of the attendant significance testing and confidence interval construction methods become completely inappropriate.

Several methods, such as the use of Mallow's C_p statistic for "all possible subsets (of the independent variables) regression," have some possible mitigating effects on the heuristic nature of the blind methods of stepwise regression.

They offer a process of screening all possible equations to find the better ones, with compensation for the differing numbers of parameters that need to be fit.

Although these search strategies offer a justifiable mechanism for finding the "best" according to ability to predict a dependent measure, they are somewhat at cross-purposes for how multiple regression is typically used in the behavioral sciences.

What is important is in the structure among the variables as reflected by the regression, and not so much in squeezing the very last bit of variance-accounted-for out of our methods.

More pointedly, if we find a "best" equation with fewer than the maximum number of available independent variables present, and we cannot say that those not chosen are less important than those that are, then what is the point?

The implicit conclusion of the last argument extends more generally to the newer methods of statistical analysis that seem to continually demand our attention, e.g., in hierarchical linear modeling, nonlinear methods of classification, procedures that involve optimal scaling, and so on.

When the emphasis is solely on getting better "fit" or increased prediction capability, and thereby, modeling "better," the methods may not be of much use in "telling the story" any more convincingly – and that should be the ultimate purpose of any analysis procedure we choose.

Also, as Roberts and Pashler (2000) note rather counterintuitively, "goodness-of-fit" does not necessarily imply "goodness-of-model."

Even without the difficulties presented by a multivariate system when searching through the set of independent variables, there are several admonitions to keep in mind when dealing with a single equation.

The most important may be to remember that regression coefficients cannot be interpreted in isolation for their importance using their size, even when based on standardized variables (i.e., those that have been Z-scored).

Just because one coefficient is bigger than another, does not imply it is therefore more important.

For example, consider the task of comparing the relative usefulness of the Scholastic Aptitude Test (SAT) scores and High School Grade Point Averages (HSGPA) in predicting freshmen college grades.

Both independent variables are highly correlated; so when grades are predicted with SAT scores, a correlation of about 0.7 is found.

Correlating the residuals from this prediction with HSGPA, gives a small value.

It would be a mistake, however, to conclude from this that SAT is a better predictor of college success than HSGPA. If the order of analysis is reversed, we would find that HSGPA correlates about 0.7 with freshmen grades and the residuals from this analysis have only a small correlation with SAT score.

If we must choose between these two variables, or try to evaluate a claim that one variable is more important than another, it must be from some other basis.

For example, SAT scores are like the product of an experiment; they can be manipulated and improved.

Flawed test items can be discovered and elided. But HSGPA is like the result of an observational study; they are just found, lying on the ground.

We are never sure exactly what they mean.

If one teacher someplace harbors a secret bias, and thus gives students of a particular ilk grades that do not represent their true accomplishments, how are we to know?

There are some formal methods that can at times help reduce our ignorance.

We will mention these next, but first remember that no formal procedure guarantees success in the face of an unthinking analysis.

The notion of importance may be explored by comparing models with and without certain variables present, and comparing the changes in variance-accounted-for that ensue.

Similarly, the various significance tests for the regression coefficients are not really interpretable independently, e.g., a small number of common factors may underlie all the independent variables, and thus, generate significance for all the regression coefficients.

In its starkest form, we have the one, two, and three asterisks scattered around in a correlation matrix, suggesting an ability to evaluate each correlation by itself without consideration of the multivariable system that the correlation matrix reflects in its totality.

Finally, for a single equation, the size of the squared multiple correlation (R^2) gets inflated by the process of optimization, and needs to be adjusted, particularly when sample sizes are small.

One beginning option is to use the commonly generated Wherry "adjusted R^2," which makes the expected value of R^2 zero when the true squared multiple correlation is itself zero.

Note that the name of "Wherry's shrinkage formula" is a misnomer because it is not a measure based on any process of cross-validation.

A cross-validation strategy is now routine in software packages, such as SYSTAT, using the "hold out one-at-a-time" type of mechanism.

Given the current ease of implementation, such cross-validation processes should be routinely performed.