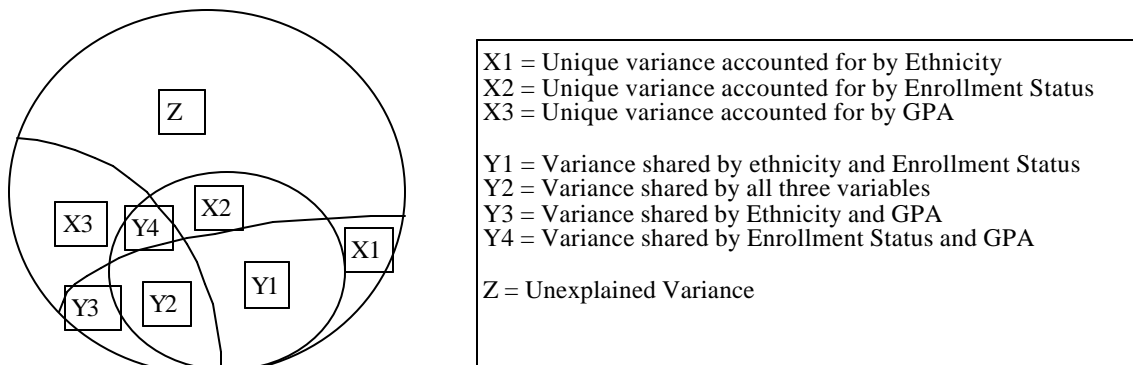


Using Multiple Regression Models to Evaluate the Unique Contributions of Variables to Effects

Since the onset of computer-based statistical analyses, the use of multiple regression to evaluate the relative influence of multiple variables on a single dependent variable (e.g. retention) has become increasingly common. While correlations indicate the simple linear relationship between two variables, multiple regression allows one to determine the unique contribution of a specific variable in explaining (predicting) the variability of a dependent variable. Explaining the variability in a dependent variable is the primary purpose of most research. Regression theoretically requires a continuous dependent variable, but will work with even a low-level ordinal scale (e.g. three scale points from high to low). Conover & Iman (1981) explain the use of “rank” data for these purposes. Independent/predictor variables can be categorical (e.g. sex).

As an example of the use of multiple regression, I will use three common variables to predict the retention of a college student (where retention may be defined as number of years in college up to 5). Three relevant predictor variables are prior institution grade point average (GPA), enrollment status (part-time/full-time) and race/ethnicity. Say that the simple relationship of GPA with retention is .25, of enrollment status with retention is .27, and of race/ethnicity with retention is .29. By simply looking at these numbers, one might decide that ethnicity had the greatest relationship with retention. However, as the Venn diagram below shows:



Although each variable accounts for approximately the same portion of the total area of the dependent variable's (retention) variability, for variables X1 (Ethnicity) and X2 (Enrollment Status), much of the total area they explain is shared/accounted for either between the two or is also accounted for by X3 (GPA). The unexplained variance (Z) is not predictable by the regression equation's variables. The unique variance accounted for by each variable makes it obvious that GPA (X3) is the single most important predictor variable, that Enrollment Status (X2) explains considerably less, but even this is considerably greater than that explained by Ethnicity (X1), which, when the other variables are added to the model, reduces in influence from the most important, to the third best predictor variable.

Using Backward Elimination Rather than Stepwise Regression

I generally recommend the use of a backward elimination regression model rather than the typical forward stepwise model. In a simple stepwise model, the variable that has the greatest simple correlation with the dependent variable (in the example above, Ethnicity) becomes the first variable entered into the regression model. The total variance accounted for by this variable (in the example, X1, Y1, Y2 and Y3) is then removed from consideration, and the remaining variables partial correlations (variance of X1 excluded) with the dependent variable are then computed. The variable with the greatest remaining partial correlation is entered next, its variance is excluded from the model, and this continues either until all variables are entered or until none attain an adequate level of significance to be included. Although this is sometimes a legitimate approach, it fails to consider two important effects:

- Suppressor effects - Suppression occurs when a variable itself has a very small simple relationship (correlation) with the dependent variable, however, including this variable increases substantially the overall model's explained variance.
- A variable's unique variance is the only variance that may be evaluated in a simple fashion, because shared variance is exceedingly difficult to evaluate (how much effect did each of the three variables have on the individual student).

The backward elimination method uses the amount of unique variance a variable adds to the complete model (all remaining variables) as the criterion for exclusion from the model. In this method, the full model (with all variables included) is computed first. Then all variables are removed from the model, and the variable that causes the least reduction in accounted variance by its removal from the model is the first to be eliminated. This process continues until all remaining variables contribute a significant amount of unique variance to the final model.

Applying Regression to Categorical Dependent Variables

In the model discussed, the dependent variable, student retention is usually treated as a dichotomous variable or low-level ordinal variable, that is, it takes one of two values: retained (good) or non-retained (bad). If treated as a dichotomous variable (without high/low order), prior to the development of Logistic regression methodologies, researchers would have been constrained to use either a form of Discriminant analysis or Hotelling's T to evaluate the relative relationships of independent variables. However, since the development of Logistic regression models, it has become feasible to use this computational method and thereby either avoid or, at least, reduce some of the problems that associate with categorical dependent variables under multivariate assumptions..

References

Conover, W.J. and Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35, pp. 124-129