Hierarchical versus Stepwise Regression

Stepwise multiple regression, also called statistical regression, is a way of computing regression in stages. In stage one, the independent best correlated with the dependent is included in the equation. In the second stage, the next remaining independent with the highest partial correlation with the dependent, controlling for the first independent, is entered. This process is repeated, at each stage partialling for previously-entered independents, until the addition of a remaining independent does not increase R^2 by a significant amount (or until all variables are entered, of course). Alternatively, the process can work backward, starting with all variables and eliminating independents one at a time until the elimination of one makes a significant difference in R-squared. In SPSS, select Analyze, Regression, Linear; set the Method: box to Stepwise.

Stepwise regression and theory. Stepwise regression **is used in the exploratory phase of research or for purposes of pure prediction, not theory testing**. In the theory testing stage the researcher should base selection of the variables and their order on theory, not on a computer algorithm. Menard (1995: 54) writes, "there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data and produces results that tend to be idiosyncratic and difficult to replicate in any sample other than the sample in which they were originally obtained." Likewise, the nominal .05 significance level used at each step in stepwise regression is subject to inflation, such that the real significance level by the last step may be much worse, even below .50, dramatically increasing the chances of Type I errors. See Draper, N.R., Guttman, I. & Lapczak, L. (1979). For this reason, Fox (1991: 18) strongly recommends any stepwise model be subjected to cross-validation.

Other problems of stepwise regression. By brute force fitting of regression models to the current data, stepwise methods can overfit the data, making generalization across data sets unreliable. A corollary is that stepwise methods can yield R^2 estimates which are substantially too high, significance tests which are too lenient (allow Type 1 error), and confidence intervals that are too narrow.

Hierarchical multiple regression (not to be confused with hierarchical linear models) is **similar to stepwise regression, but the researcher, not the computer, determines the order of entry of the variables**. F-tests are used to compute the significance of each added variable (or set of variables) to the explanation reflected in R-square. This hierarchical procedure is an alternative to comparing betas for purposes of assessing the importance of the independents. In more complex forms of hierarchical regression, the model may involve a series of intermediate variables which are dependents with respect to some other independents, but are themselves independents with respect to the ultimate dependent. Hierarchical multiple regression may then involve a series of regressions for each intermediate as well as for the ultimate dependent.

SPSS data analysis. For hierarchical multiple regression, in SPSS first specify the dependent variable; then enter the first independent variable or set of variables in the independent variables box; click on "Next" to clear the IV box and enter a second variable or set of variables; etc. One also clicks on the Statistics button and selects "R-squared change." Note that the error term will change for each block or step in the hierarchical analysis. If this is not desired, it can be avoided by selecting Statistics, General Linear Model, GLM-General Factorial, then specifying Type I sums of squares. This will yield GLM results analogous to hierarchical regression but with the same error term across blocks.

Assumptions. Multiple regression shares all the assumptions of correlation: linearity of relationships, the same level of relationship throughout the range of the independent variable ("homoscedasticity"), interval or near-interval data, absence of outliers, and data whose range is not truncated. In addition, it is important that the model being tested is correctly specified. The exclusion of important causal variables or the inclusion of extraneous variables can change markedly the beta weights and hence the interpretation of the importance of the independent variables.