

Model Building Procedures

Researcher Determines Model

Simultaneous

All predictor variables are entered at the same time. I typically use this approach. Simultaneous and hierarchical are probably the most common regression model testing procedures.

Hierarchical

Based on an *a priori* criteria, the researcher enters some number of variables into the model a step at a time. Any number of variables can be entered on each step, and any number of steps can be used. Each step is a separate regression model. The resulting model is identical to a model in which all variables were entered simultaneously. The major advantage of this approach is that a change in R-square is computed, allowing for a test of whether a significant amount of additional variance is accounted for by the variable or variables entered on each step. If a single variable is entered on a step, the R-square is equal to the semi-partial (a.k.a. "part") correlation coefficient, and the test of the R-square change is equivalent to the test of the regression coefficient for the new variable.

Data Determine Model

Forward Selection

Predictor variables are added to the model a step at a time. The first step evaluates all of the variables, and the variable with the largest correlation with the dependent variable is entered first. Then on each new step, the variable which will increase R-square the most will be entered on that step (other criteria for particular significance levels—termed "PIN" for the p-value needed to be entered—or F values can be used). This approach is rarely used anymore.

Backward Selection

Backward selection proceeds in the opposite manner to forward selection. All variables are entered and then the poorest predictor is eliminated. The process continues until all of the nonsignificant variables are removed. Usually by default variables that are not significant are removed on each step ("POUT" of .05), but any p-value or F-value can be used for the criteria. The model is reevaluated after each variable is removed. This approach is rarely used anymore.

Stepwise Selection

Stepwise, which uses a combination of forward and backward selection, is more commonly used than either forward or backward. Predictor variables are entered as they are in forward selection, but at each step the variables are evaluated to see if any can be removed. As with the others, the criteria can be changed to a particular PIN and POUT or FIN and FOUT values.

All Subsets Regression

All subsets regression picks the best combination of predictors by running regression analyses for all possible predictors (according to the list provided). That is, if five predictors are given, there will be one 5-predictor model, five 4-predictor models and so on ($2^5 = 32$ models). Researchers may use a variety of criteria for picking the best possible model, including the highest R-square, the lowest MSE, or the lowest Mallows' C_p . C_p is based on MSE but takes the number of predictors into account (models with more predictors always have higher R-square values regardless of how useful the variables really are). One difficulty is deciding the optimal criteria to use in choosing the "best" model. This model building option is currently available in SAS but not SPSS.

Comments

One reason I usually use simultaneous regression rather than hierarchical regression is that all coefficients are partial with respect to all other variables considered. Researchers who use hierarchical regression usually enter demographic variables on the first step as "control variables" or "covariates." One difficulty I see with this approach merely concerns the usual format of presentation. Because researchers often present results for only the new variables entered on a particular step, readers cannot tell what happens to variables entered on prior steps after new variables have been entered. For example, if age becomes nonsignificant after another variable is entered on the second step, the reader will conclude that age was an important predictor even though a variable entered later was responsible for the association of age with the dependent variable.

Forward and backward procedures are rarely used anymore, because stepwise selection is considered superior to either. Although stepwise selection is better than forward or backward alone, it still has problems. Simulation studies suggest that stepwise selection often leads to erroneous model choices (both Type 1 and Type 2 errors can occur; Freedman, 1983; Pope & Webster, 1972). I recommend that researchers use theory to decide the order that variables are entered into the model, rather than exploratory, data-driven approaches. If a researcher wishes to go completely exploratory, the all subsets approach is probably preferred because no models are missed.

Freedman, D.A. (1983) A note on screening regression equations. *The American Statistician*, 37, 152-155.

Pope, P.T., & Webster, J.T. (1972). The use of an F-statistic in stepwise regression procedures. *Technometrics*, 14, 327-340.