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-part (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 and Best Subsets Regression.** All subsets regression (and closely related "best subsets regression" in SPSS) 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). One difficulty is deciding the optimal criteria to use in choosing the “best” model. Researchers may use a variety of criteria for picking the best possible model, including the highest R-square, the lowest MSE, Akaike's Information Criterion (AIC or AICC), or the lowest Mallow's 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). AICC is based on the likelihood function and also takes the number of predictors into account. The all subsets procedure is available in SAS with PROC REG and the best subsets procedure is available in SPSS under the LINEAR procedure. In R, you can use the leaps package or regsubsets() base function with the method=c("exhaustive")
**Bagging and Boosting.** Bagging and boosting are model selection approaches derived from machine learning decision-tree processes and commonly used in a few areas such as neural networks (see Sutton, 2005 for a review). Bagging or bootstrap aggregating (Breiman, 1996) is a resampling method based on subsets of cases in the data set. Multiple samples are combined to reduce error in prediction, but each sample is combined with equal weight to decide on inclusion of variables in the model (i.e., "voting"). Boosting (Freund & Schapire, 1996) also combines results from many analyses, but cases are weighted based on how well they are predicted and analyses are recomputed in an iterative process. Bagging and boosting are relatively new, but some evidence suggests that they can improve prediction, with some papers suggesting boosting outperforms bagging (Quinlan, 2006). These are relatively new processes that do not seem to be widely used yet in the social sciences, and work is still being done to decide which are the best performing algorithms used are still being improved.

**Comments**

Hierarchical regression can be useful for estimating variance accounted for sets of variables. But, 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.” There are no substantial differences in the two approaches, however. Keep in mind that hierarchical regression is simply running several different regression models at the same time, each one with added predictors. The one difficulty I have with hierarchical regression merely involves 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. Your textbook (Cohen et al., 2003, Chapter 5, Section 5.3) makes a case for the order of entry with hierarchical regression based on causal ordering of variables. In general, I think hierarchical regression provides little advantage in understanding causal ordering and that causality should be explored in other ways (more detail when I discuss longitudinal analysis and mediation).

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 or boosting approaches may be preferred over the other exploratory approaches.