Best Subset Methods

In Regression Modeling

Prof. Tom Fomby
Department of Economics
Southern Methodist University
Dallas, TX 75275
5/27/05

Reference: SAS/STAT User's Guide, Volume 2, GLM – VARCOMP, Version 6, Fourth Edition, 1990 (SAS Institute: Cary, NC), pp. 1397 - 1400.

1. Methods of model selection are implemented in PROC REG in SAS by using the SELECTION = option in the MODEL statement.
2. Forward Selection (FORWARD): The forward-selection technique begins with no variables in the model. For each of the independent variables, FORWARD calculates F statistics that reflect the variable's contribution to the model if it is included. The p-values for these F statistics are compared to the SLENTRY = value that is specified in the MODEL statement (or to 0.50 if the SLENTRY = option is omitted). If no F statistic has a p-value less than the SLENTRY = value, FORWARD stops. Otherwise, FORWARD adds the variable that has the smallest p-value less than SLENTRY (equivalently the largest F statistic). FORWARD then calculates F statistics again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant F statistic. **Once a variable is in the model, it stays in the model**.
3. Backward Elimination (BACKWARD): The backward-elimination technique begins by calculating F statistics for the full model that includes all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce F statistics significant at the SLSTAY = level specified in the MODEL statement (or at the 0.10 level if the SLSTAY = option is omitted). At each step, the variable showing the smallest contribution to the model (i.e. having the smallest F statistic or equivalently the largest p-value) is deleted. **Once a variable is dropped from the model, it never enters again**.
4. Stepwise (STEPWISE): The stepwise method is a modification of the forward-selection technique and differs in that **variables already in the model do not necessarily stay there**. As in the forward-selection method, variables are added one by one to the model, and the F statistic for the variable to be added must be significant at the SLENTRY = level. After a variable is added, however, the stepwise method looks at all the variables already included in the model and deletes any variable that does not produce an F statistic significant at the SLSTAY = level. Only after this check is made and the necessary deletions

accomplished can another variable be added to the model. The stepwise process ends when none of the variables outside the model has an F statistic significant at the SLENTRY = level and every variable in the model is significant at the SLSTAY = level, or when the variable to be added to the model is the one just deleted from it. **When using this technique it is possible for a variable to enter the model, later leave it, and then possibly re-enter the model to stay**.

5. With respect to the Forward Selection technique, the tuning parameter is the SLENTRY probability level (default SLENTRY = 0.50). The larger this probability level, the more independent variables that will eventually enter into the final model selected. In contrast, the smaller this probability level, the more parsimonious the final selected model tends to be.
6. With respect to the Backward Elimination technique, the tuning parameter is the SLSTAY probability level (default SLSTAY = 0.10). The smaller this probability level, the more parsimonious the final selected model will be. The larger this probability level the less parsimonious the final selected model.
7. The Stepwise Selection technique has two tuning parameters, SLENTRY and SLSTAY. The smaller the SLENTRY and SLSTAY p-values (SLENTRY > SLSTAY), the more parsimonious the final selected model. The larger the SLENTRY and SLSTAY p-values, the less parsimonious the final selected model.
8. The Adjusted R^2 criterion (ADJRSQ) selection method looks at **all possible subset models** and determines the model that maximizes the Adjusted R^2 criterion. The Adjusted R^2 criterion is defined as

$$\overline{R}^2 = 1 - \frac{n(1-R^2)}{n-p}$$

where n = the number of observations, p = number of independent variables, and R^2 = the coefficient of determination = $\frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y})^2}$, that is, the regression sum of squares divided by the total sum of squares.

9. Mallows' C_p (CP) Selection Criterion: This criterion is defined as

$$C_p = \frac{SSE_p}{s^2} - (n - 2 * p)$$

where s^2 is the MSE for the full model, SSE_p = is the error sum of squares for a model with p parameters including the intercept and n denotes the number of observations. For each model size p we can plot the C_p's against p. Mallows recommends choosing the model whose C_p is closest to p upon the first approach of p (i.e. the first time a model's C_p is "close" to p).

10. Notice that there are no tuning parameters for the Adjusted R-square or CP selection criteria. **As we will later see, most Data Mining techniques involve**

the use of one or more tuning parameters in selecting a “best” model from a class of models.