Forward Stepwise Regression

FORWARD STEPWISE REGRESSION is a stepwise regression approach that starts from the null model and adds a variable that improves the model the most, one at a time, until the stopping criterion is met. The criterion for predictor entry into the model is based on the F-statistic and corresponding p-value (p-value must be less than the Alpha-to-Enter). It is also known as **Forward Selection** regression.

How To

- ✓ Run: **STATISTICS->REGRESSION -> FORWARD STEPWISE REGRESSION...**
- ✓ Select the **DEPENDENT VARIABLE (RESPONSE)** and **INDEPENDENT VARIABLES (PREDICTORS).**
- ✓ **ENTER IF ALPHA <** option defines the *Alpha-to-Enter* value. At each step it is used to select candidate variables for entry, with partial F p-value less or equal to the alpha-to-enter. The default value is 0.05.
- Select the **SHOW CORRELATIONS** option to include the correlation coefficients matrix to the report.
- Select the **SHOW DESCRIPTIVE STATISTICS** option to display the mean, variance and standard deviation of each term.
- Select the **SHOW RESULTS FOR EACH STEP** option to show the regression model and summary statistics for each step.

Model

The criterion to enter a variable at each step is the partial F p-value (that is over the alpha-to-enter). When none of the unselected variables meet the entry criterion, the forward selection command terminates the process. Forward stepwise algorithm is greedy version of the best subsets regression, so it may not result with the best model (model with lowest SSE). It is sensitive to choice of the alpha-to-entry value.

Results

The report shows regression statistics for the final regression model. If the **SHOW RESULT FOR EACH STEP** option is selected, the regression model, fit statistics and partial correlations are displayed for all variables entered at a selection step. A correlation coefficients matrix and descriptive statistics for predictors are included to the report if the corresponding options are selected.

R^2 (COEFFICIENT OF DETERMINATION, R-SQUARED) - is the square of the sample correlation coefficient between the **PREDICTORS** (independent variables) and **RESPONSE** (dependent variable).

ADJUSTED R2 (ADJUSTED R-SQUARED) - is a modification of R^2 that adjusts for the number of explanatory terms in a model. While R^2 increases when extra explanatory variables are added to the model, the adjusted R^2 increases only if the added term is a relevant one. It could be useful for comparing the models with different numbers of predictors. Adjusted R^2 is computed using the formula:
$Adjusted\ R^2 = 1 - \frac{SS_{error}}{SS_{total}}\frac{df_{total}}{df_{error}} = 1 - (1 - R^2)\frac{N-1}{N-k-1}$, where *k* is the number of predictors.

S – the estimated standard deviation of the error in the model.

MS (MEAN SQUARE) - the estimate of the variation accounted for by this term,

$$MS = SS/DF.$$

F - the F-test value for the model.

P-LEVEL - the significance level of the F-test. Values less than α (0.05) show that the model estimated by the regression procedure is significant.

VIF – variance inflation factor, measures the inflation in the variances of the parameter estimates due to collinearities among the predictors. It is used to detect multicollinearity problems. The larger the value is, the stronger the linear relationship between the predictor and remaining predictors. VIF equal to 1 indicates the absence of linear relationship with other predictors (there is no multicollinearity). VIF value between 1 and 5 indicates moderate multicollinearity, and values greater than 5 suggest that a high degree of multicollinearity is present. It is a subject of debate whether there is a formal value for determining presence of multicollinearity: in some situations even values greater than 10 can be safely ignored – when high values caused by complicated models with dummy variables or variables that are powers of other variables. But in weaker models even values above 2 or 3 may be a cause for concern: for example, for ecological studies Zuur, et al. (2010) recommended a threshold of VIF=3.

TOL - the tolerance value for the parameter estimates, it is defined as *TOL = 1 / VIF*.

Partial Correlations are correlations between each predictor and the outcome variable excluding the effect of other variables.

References

Hocking, R. R. (1976) "The Analysis and Selection of Variables in Linear Regression," *Biometrics, 32*

Nargundkar R. (2008) Marketing Research: Text and Cases. Third edition. Tata McGraw-Hill Publishing Company Ltd.

Neter, J., Wasserman, W. and Kutner, M. H. (1996). Applied Linear Statistical Models, Irwin, Chicago.

Zuur, A. F., Ieno, E. N. and Elphick, C. S. (2010), A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1: 3–14.