Best Subsets Regression

The **BEST SUBSETS REGRESSION** command involves examining all the models for all possible combinations of predictor variables and determines the best set of predictors for each subset size. It can be used as an alternative to the stepwise regression procedures. The best subsets regression is also known as *all possible subsets regression*.

How To

- ✓ Run: **STATISTICS->REGRESSION -> BEST SUBSETS REGRESSION...**
- ✓ Select **DEPENDENT** (**RESPONSE**) variable and **INDEPENDENT** variables (**PREDICTORS**)**.**
- ✓ Select the **Show correlations** option to display the partial correlation matrix at each step.
- ✓ Select the **Show descriptive statistics** option to show the descriptive statistics (mean, variance and standard deviation) for each term.
- ✓ **Casewise** deletion method is used for missing values removal.

Results

Best subset regression command selects the subset of predictors at each step that fits best, based on the criterion of having the largest R^2. The report includes a set of best fitted models with standardized regression statistics and ANOVA summary for each subset size (from 1 to the number of predictors). Correlation coefficients matrix and descriptive statistics for predictors are displayed if the corresponding options are selected.

R^2 INCREMENT – is the increment of R^2 in comparison with the previous subset.

R^2 (COEFFICIENT OF DETERMINATION, R-SQUARED) - is the square of the sample correlation coefficient between the **PREDICTORS** (independent variables) and **RESPONSE** (dependent variable). In general, R^2 is a percentage of response variable variation that is explained by its relationship with one or more predictor variables. The definition of the R^2 is $R^2 \equiv 1 - \frac{SS_{error}}{SS_{total}}$.

ADJUSTED R^2 (ADJUSTED R-SQUARED) - is a modification of R^2 that adjusts for the number of explanatory terms in a model. 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. Identifying the model with the smallest mean square error (MSE) is equivalent to finding the model with the smallest S.

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

$$MS = SS/DF$$

F - the F-test value.

P-VALUE – p-value for a F-test. A value less than α level (0.05) shows that the model estimated by the regression procedure is significant.

References

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

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

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

[OMK] Olejnik, S. Mills, R. and Keselman, H. “Using Wherry’s Adjusted R2 and Mallows' Cp for Model Selection from All Possible Regressions”, The Journal of Experimental Education, 2000, 68(4), 365-380.