The Multiple Inference Problems in Stepwise Regression Analysis

Parys, Dariusz
University of Lodz, Chair of Statistical Methods
Rewolucji 1905 r. 41 Str.
Lodz, Poland
E-mail: katstat@krysia.uni.lodz.pl

Domañski, Czes³aw
University of Lodz, Chair of Statistical Methods
Rewolucji 1905 r. 41 Str.
Lodz, Poland
E-mail: katstat@krysia.uni.lodz.pl

1. The problem

The problem of multiple inference arises often in statistical analysis. The statistical problem treated here are mainly the ones of multiple comparisons and multiple linear regression. Since the inference from this analysis is based on the model, it is important to select a proper model. If the choice of model is restricted to linear ones, the problem reduces to that of telling which regressor variables should be entered into the model.

A response variable Y is measured along with a number of variables $X_1, X_2, ..., X_L$ that perhaps have some influence on Y.

The aim of analysis is to tell which X-variables can be shown to have such on influence.

The final product of the analysis could be regarded as a classification of the X-variables into K groups. All but one, the K-th, of these groups are showing significant impact on Y.

The basic assumption or hypothesis is that all X-variables belong to the latter group.

A number of null hypotheses $H_{01}, H_{02}, ..., H_{0k}$ is to be tested against the alternatives $H^*_{A1}, H^*_{A2}, ..., H^*_{Ak}$.

The multiple level of significance is the bound of the probability of constructing a group where none of the variables are really correlated with Y.

The main problem is to tell which of the possible regressors that should be included in the final model.

2. The multiple stepwise procedure

In step one the overall null-hypotesis is

$$H_0 : Corr\ (X_i, Y) = 0,\ \ i = 1, ..., L$$

As the output of the first step should be the regressor with the greatest impact on Y, the alternative hypothesis is formulated as:

$$H_A : \max\ |Corr\ (X_i, Y)| > 0,\ \ i = 1, 2, ..., L$$

with means that even if two or more X-variables are significantly correated with Y, only the one with the greatest correlation should be picted out.

If H_0 in step one is not rejected the problem is solved. If on the other hand, H_0 in k-th step becomes

$$H_0 : Corr\ (X_i, Y - \alpha - \beta_1 X^*_1 - \beta_2 X^*_2 - \beta_{k-1} X_{k-1}) = 0$$

where X^*_j are the X – variable judged as the most significant in step j.

If the procedure is performed at least some steps, the structure of dependencies give us the prediction model

$$Y = \alpha + \beta_1 X^*_1 + \beta_2 X^*_2 + \ldots + \beta_K X^*_K + \varepsilon$$

where K is the number of groups of entered variables.

This procedure is likely to keep the multiple level of significance.

REFERENCES

Holm, S. (1977). Sequentially Rejective Multiple Test Procedures, Statistical Research Report 1977.

Miller, R. G. Ir (1980). Simultaneous Statistical Inference, 2nd ed: Springer Verlag, New York.

Thompson, M. L. (1978). Selection of Variables in Multiple Regression: Part I, A review and evaluation. Part II Chosen procedures, computations and examples, Inst. Stat. Rev. 46, 1-19, 129-146.

RESUME

Le procédé multiple présenté dans cet article est fondé sur la méthode de test, fait pas a pas. Cette méthode de test maintient le niveau d'effectivité au niveau fixé auparavant. Elle permet de découvrir ces X variables indépendents laquelles sont les plus fortes corrélés avee un Y variable dépendent, et aussi de créer d'eux, un nouveau modele de régression.