

Stepwise Regression as a Multiple Comparison Procedure

Rodney Sparapani, Duke Clinical Research Institute

ABSTRACT

Stepwise Regression is a valuable tool for exploratory model building. This paper will examine how Stepwise Regression fits into the Multiple Comparisons paradigm. SAS will be used to demonstrate how these models will be handled in practice.

Keywords: Multiple Comparisons, Forward Selection, Backward Elimination, Stepwise Regression, Step-down Procedure, Step-up Procedure

INTRODUCTION

"When many significance tests are performed, each at a level of, say 5 percent, the overall probability of rejecting at least one true null hypothesis is much larger than 5 percent" (SAS Institute, Inc. 1990). This dilemma is known as the problem of Multiple Comparisons. Stepwise Regression is one of the many techniques that suffer from this problem. The following discussion intends to address this problem in the framework of Multiple Linear Regression, but the conclusions are general enough that the methodology can be extended to Generalized Linear Models.

METHODS

Definitions

Type I Error: Reject H_0 , when H_0 is True

Type II Error: Fail to Reject H_0 , when H_0 is False

Family-wise Error Rate (FWE): Probability of committing one or more Type I Errors

Set Subtraction: $(C=A-B)$ set C consists of every object in A that is not in B

Bonferroni's Method

$$\alpha_{FWE} = 1 - (1 - \alpha)^k$$

$$\alpha = 1 - (1 - \alpha_{FWE})^{1/k}$$

$$f(x) = 1 - (1 - x)^{1/k}$$

$$f(x) = x/k + (1 - 1/k)x^2/2k + \dots > x/k$$

Suppose there are k independent null hypotheses

$(H_{01}, H_{02}, \dots, H_{0k})$

Their corresponding p-values $(p_{01}, p_{02}, \dots, p_{0k})$ can be ordered as $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$

Union-Intersection (UI) principle (Roy and Bose 1953)

H_0 : Intersection Hypothesis

Rejection Rule: Reject H_0 if $p_{(1)} < \alpha_{FWE}/k$

Intersection-Union (IU) principle (Lehmann 1952)

H_0 : Union Hypothesis

Rejection Rule: Reject H_0 if $p_{(k)} < \alpha_{FWE}$

Notation

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \iff \{1, 2, \dots, k\}$$

$$|\{1, 2, \dots, k\}| = k$$

$$(k)_m = k! / (k - m)!$$

Extra Sum of Squares

model m: m parameters, n-m degrees of freedom

model m+1: m+1 parameters, n-m-1 degrees of freedom

$$F_{(m+1)} = (SSE_m - SSE_{(m+1)}) / SSE_{(m+1)} / (n - m - 1) \sim F_{1, (n - m - 1)}$$

RESULTS

Suppose there are k potential regressors under investigation. Without loss of generality, assume that the intercept is in the model.

Forward Selection

Let $(k-i)$ represent the steps of the process. At each step, a decision must be made whether to stop or to continue on to the next step.

Table 1 Critical Values for Forward Selection

Step (k-i)	Model Parameters (k-i)	Critical Value (α_{FWE}/i)
0	$ \{\} = 0$	α_{FWE}/k
1	$ \{(1)\} = 1$	$\alpha_{FWE}/(k-1)$
•	•	•
•	•	•
•	•	•
k-1	$ \{(1), (2), \dots, (k-1)\} = k-1$	α_{FWE}
k	$ \{1, 2, \dots, k\} = k$	

Step-down Procedure

In multiple comparisons terminology, this is a Step-down procedure (Holm 1979). At each step, you Reject the null hypothesis if the corresponding p-value is less than the critical value. The procedure stops at the first step where you Fail to Reject the null.

Intersection Hypothesis

Here, the null hypothesis is the Intersection hypothesis by the UI principle. For each variable under investigation that is not already in the model, an Extra Sum of Squares statistic can be computed corresponding to an individual

hypothesis. A Intersection hypothesis can be constructed from the individual hypotheses. The Intersection hypothesis is Rejected if at least one of the individual hypotheses is Rejected. This approach is consistent with Forward Selection. Forward Selection chooses the variable corresponding to the smallest p-value. Therefore, at each step this variable is selected, provided it is less than the critical value. If not, the process should be stopped.

Union Hypothesis

The critical value at each step is determined by Bonferroni's Method (see Table 1). Therefore, at each step the FWE is controlled at the level α_{FWE} . But what about the whole process, i.e. more than one step? A Union hypothesis can be constructed from each step's Intersection hypothesis by the IU principle. When each Intersection hypothesis is Rejected, we Reject the Union hypothesis. This is the case for every step, except the last one. Since you Fail to Reject the final Intersection hypothesis, you could not have committed a Type I Error. Therefore, we have a procedure that controls the FWE at some arbitrary level, α_{FWE} .

Backward Elimination

Let i represent the step of the process. At each step, a decision must be made whether to stop or to continue to the next step.

Table 2 Critical Values for Backward Elimination

Step (i)	Model Parameters (k-i)	Critical Value ($\alpha_{FWE}/(i+1)$)
0	$ \{1, 2, \dots, k\} =k$	α_{FWE}
1	$ \{1, 2, \dots, k\}-\{(1)\} =k-1$	$\alpha_{FWE}/2$
•	•	•
•	•	•
•	•	•
k-1	$ \{1, 2, \dots, k\}-\{(1), (2), \dots, (k-1)\} =1$	α_{FWE}/k
k	$ \{\} =0$	

Step-up Procedure

This is a Step-up procedure (Hochberg, 1988). At each step you Fail to Reject the null hypothesis if the corresponding p-value is greater than or equal to the critical level. The procedure stops at the first step where you Reject the null.

Critical Values

For every variable that is currently in the model, an Extra Sum of Squares statistic is calculated. The null hypothesis is that the variable is zero. Backward Elimination selects the variable with the largest p-value to be omitted, provided that the statistic is not significant. The UI principle, which we relied upon in the Forward Selection context, doesn't fit Backward Elimination. However, the critical values seem to be pre-determined via symmetry with Forward Selection (see Tables 1 and 2). Therefore, we will just need to demonstrate that these critical values are correct. Consider an example with 4 variables (see Table 3).

Table 3 Example with 4 Variables

Step (i)						
0	{1,2,3,4}					
1	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}		
2	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
3	{1}	{2}	{3}	{4}		
4	{}					

Notice that the model {1, 2} has two routes of selection in a single step, i.e. from {1, 2, 3} or {1, 2, 4}. In other words, there are two ways of arriving at {1, 2}. Similarly, there are 3 ways of arriving at {1} in a single step, i.e. from {1, 2}, {1, 3}, or {1, 4}. This can be seen in the general case as well. At each step, you have i parameters that have been removed from the model. There are $(i)_i = i!/(i-1)! = i$ possible ways of selecting a variable for omission. You can think of this as a classic counting problem (see Table 4).

Table 4 Example of Ways to Reach Each Outcome

Step (i)	1	2	3	4
Model Parameters (k-i)	3	2	1	0
Ways to Reach Each Outcome (i)	1	2	3	4

Stepwise Regression

After laying the groundwork, Stepwise Regression is a simple extension. You start the process with Forward Selection, as discussed, and after each step, perform Backward Elimination on the current model.

DISCUSSION

Forward Selection

The first possible problem with this argument is the assumption of independence in the UI and IU principles. It is clear that in both cases, the individual hypotheses are not independent of each other. However, this is not a cause for concern, because the procedure arrived at is certainly more conservative than any procedure that does not take into account the problem of Multiple Comparisons.

Backward Elimination

Backward Elimination is easily identifiable as a Step-up procedure, yet there doesn't seem to be a simple motivation for it based on the UI and IU principles. Backward Elimination suggests the approach of counting the ways of reaching a given outcome. This seems to be a general result that can also be used to justify Forward Selection.

Stepwise Regression

Stepwise Regression has often been criticized for ignoring the problem of Multiple Comparisons. The approach, outlined here, controls the FWE. This is not to say that Stepwise Regression is an oracle and can be followed blindly. Other statistical approaches, such as R^2 Adjusted for the Degrees of Freedom, Akaike's Information Criterion and Mallows's C_p , can be employed in model building; however, these approaches require the exercise of judgment. This is not a problem where you have a small number of candidate variables under consideration, but quickly becomes cumbersome. Stepwise Regression's advantage is that it is an algorithm which is easily reproduced.

CONCLUSIONS

In the literature of Multiple Comparisons, many principles and procedures can be found that are useful in the analysis of Stepwise Regression. However, Stepwise Regression is not a perfect fit, and it is almost certain that a Multiple Comparisons procedure designed from scratch to perform this task would look very different. In that case, we would have to learn a new paradigm and patiently wait for new software. It is much better to try to understand Stepwise Regression in the context of Multiple Comparisons and use the software that is already tested and available. For instance, PROC REG can be used "as is" for Backward Elimination by setting the SLS option to α_{FWE}/k . It is not so easy to fool the software into performing Multiple Comparison procedures of Forward Selection or Stepwise Regression. These could be implemented with the Macro Facility or performed in an interactive session.

ACKNOWLEDGMENTS

Grace Lossman, Francis Kelley, Sally Muller and Jim MacDougall, without whose efforts, none of this would have been possible.

AUTHOR

Rodney Sparapani, Statistician III
(919)416-8043
spara002@mc.duke.edu

REFERENCES

- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, 75, 800-802.
- Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65-70.
- Lehmann, E.L. (1952). Testing multiparameter hypotheses. *Ann. Math. Stat.* 23, 541-552.
- Roy, S.N. and Bose, R.C. (1953). Simultaneous confidence interval estimation. *Ann. Math. Stat.* 24, 513-536.
- SAS Institute, Inc. (1990). SAS/STAT® User's Guide: Volume 2, Version 6, Fourth Edition, Cary, NC: SAS Institute Inc.