PharmaSUG 2019 - Paper PO-077

Backward Selection-a way to final model

Changxia (Connie) Shao, Boehringer Ingelheim (China) Investment Co., Ltd

ABSTRACT

In the absence of prior knowledge about population relations, investigators frequently employ a strategy that uses the data to help them decide whether to adjust for a variable. In statics, statisticians and programmers usually choose stepwise regression so that the choice of predictive variables could be carried out by an automatic procedure. Especially when you have all predictors but could not make a decision that which is the more useful predictor in a short time. More often than not expert knowledge of the therapeutic area will point to these variables, but can a data driven approach help to inform and supplement these decisions? Then backward selection or backward elimination could help a lot.

This paper provides the background of backward selection. And how to implement it in clinical trial. At the same time, this paper will demonstrate the algorithm of the backward selection in SAS statistical procedures by an example.

INTRODUCTION

Backward selection was introduced in the early 1960s (Marill & Green, 1963). It is one of the main approaches of stepwise regression. In statistics, backward selection is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. It involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.

The aim of selection is to reduce the set of predictor variables to those that are necessary and account for nearly as much of the variance as is accounted for by the total set. In essence, selection helps to determine the level of importance of each predictor variable. It also assists in assessing the effects once the other predictor variables are statistically eliminated.

AN OVERVIEW OF BACKWARD SELECTION

Backward selection is the simplest of all variable selection procedures and can be easily implemented without special software. In situations where there is a complex hierarchy, backward elimination can be run manually while taking account of what variables are eligible for removal.

1. Start with all candidate variables in the model.
2. Remove the variable with the largest p-value, that is, the variable that is the least statistically significant.
3. The new (p - 1)-variable model is t, and the variable with the largest p-value is removed.
4. Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value de fined by some significance threshold.

In order to be able to perform backward selection, we need to be in a situation where we have more observations than variables because we can do least squares regression when n is greater than p. If p is greater than n, we cannot fit a least squares model. It's not even defined. Implementing this technique appropriately using statistical software such as SAS(Table 1), R can be a quick and efficient way of identifying correlations and trends of interest within data.

Table 1: Summary of the model selection techniques available in SAS version 9.3/9.4 by SAS procedure

SAS procedure	Brief summary
PROC LOGISTIC	Fits linear logistic regression models for discrete response data by the method of maximum likelihood. It can also perform conditional logistic regression.
PROC REG	A general purpose procedure for fitting regression models.
PROC PHREG	Performs regression analysis of survival (or 'time to event') data based on the Cox proportional hazards model.
PROC GLMSELECT	Performs effect selection in the framework of general linear models. The procedure bridges the gap between PROC GLM, which doesn't include effect selection functionality, and PROC REG, which doesn't include a CLASS statement.

DATA STRUCTURE

A screenshot of some binary data appropriate for this procedure is shown in the table below.

PTNO	TEST1	TEST2	TARGET	TEST3	TEST4	TEST5	GROUP
1001	0	0	0	3	4	4	1
1002	0	0	0	2	1	4	0
1003	0	0	1	5	5	4	0
1004	0	0	0	4	3	3	0
1005	0	0	0	2	5	4	1
1006	0	0	1	3	4	2	1
1007	1	0	1	3	2	6	0
1008	0	0	0	2	5	2	1
1009	0	0	0	1	4	4	1
1010	0	1	0	5	4	3	0
1011	0	0	1	4	5	4	0
1012	0	0	1	4	4	3	1
1013	0	0	0	2	3	6	0
1014	0	0	0	3	2	3	1
1015	0	0	1	1	3	3	0
1016	0	0	1	1	3	3	1
1017	0	0	0	4	5	4	1
1018	0	0	0	3	3	4	0
1019	0	1	1	2	4	6	1
1020	0	0	0	3	5	4	0
1021	0	0	0	3	3	4	1

MODEL SELECTION PROCEDURE OPTIONS

The following invocation of PROC LOGISTIC illustrates the use of backward selection to identify the prognostic factors or "TARGET" variable. Assumed that a significance level of 0.1 is required for a variable to stay in the model (SLSTAY=0.1). A detailed account of the variable selection process is requested by specifying the DETAILS option. The Hosmer and Lemeshow goodness-of-fit test for the final selected model is requested by specifying the LACKFIT option. Please be noted that once an effect is removed from the model, it remains excluded in the following steps.

The SAS code of logistic regression model is shown as follows:

```
proc logistic data=INDATA  descending;
 class GROUP TEST1 TEST2 TEST3 TEST4 TEST5;
 model TARGET=GROUP TEST1 TEST2 TEST3 TEST4 TEST5
                   / selection=backward
                     slstay=0.1
                     details
                     lackfit;
run;
```

Details of the model selection steps are shown in Outputs below:

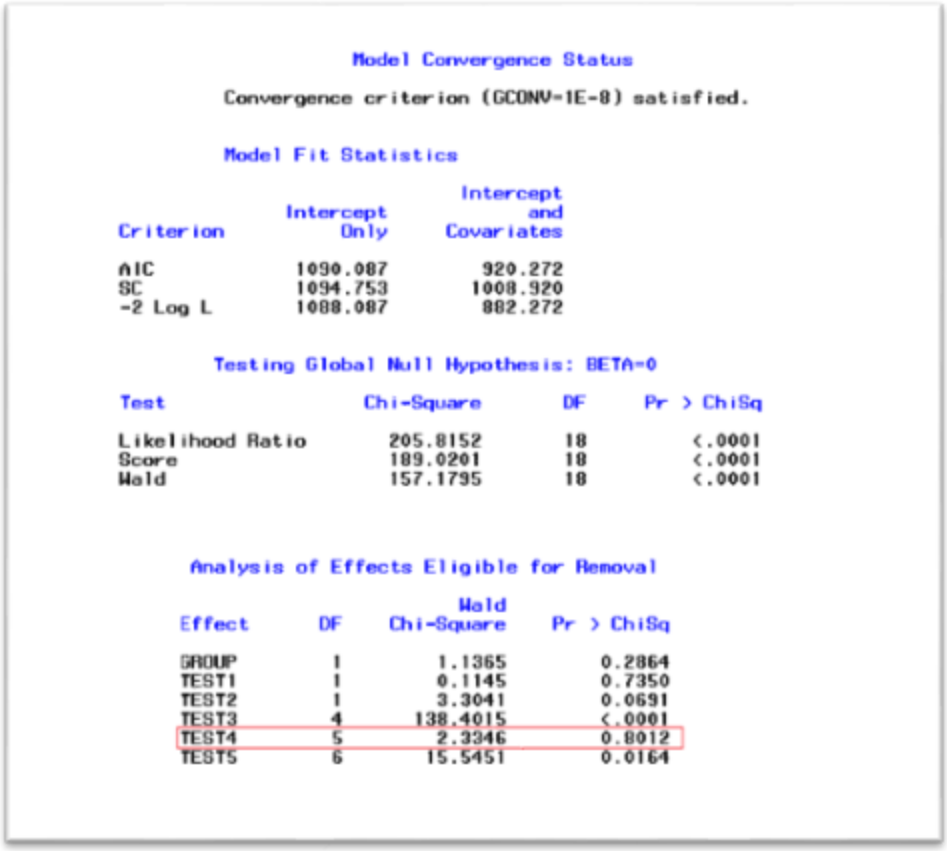


Model Information

Data Set	WORK.INDATA	
Response Variable	TARGET	TARGET
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Table 2: Startup model

Step 0. The following effects were entered: INTERCEPT GROUP TEST1 TEST2 TEST3 TEST4 TEST5.



Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1090.087	920.272
SC	1094.753	1008.920
-2 Log L	1088.087	882.272

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	205.8152	18	<.0001
Score	189.0201	18	<.0001
Wald	157.1795	18	<.0001

Analysis of Effects Eligible for Removal

Effect	DF	Wald Chi-Square	Pr > ChiSq
GROUP	1	1.1365	0.2864
TEST1	1	0.1145	0.7350
TEST2	1	3.3041	0.0691
TEST3	4	138.4015	<.0001
TEST4	5	2.3346	0.8012
TEST5	6	15.5451	0.0164

Table 3: Step0 of Backward Selection

In step 0(Table 3) "TEST4" is identified as the largest p-value among all effects.

Note: AIC: Akaike's Information Criteria, AIC estimates a measure of the difference between a given model and the "true" model. The model with the smallest AIC among all competing models is deemed the best model.

Step 1. Effect TEST4 is removed.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1090.087	912.628
SC	1094.753	977.947
-2 Log L	1088.087	884.628

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
2.3459	5	0.7995

Analysis of Effects Eligible for Removal

Effect	DF	Wald Chi-Square	Pr > ChiSq
GROUP	1	1.1007	0.2941
TEST1	1	0.0917	0.7620
TEST2	1	3.2118	0.0731
TEST3	4	140.8456	<.0001
TEST5	6	15.2850	0.0182

Table 4: Step1 of Backward Selection

In step1 (Table 4) “TEST1” is identified as the largest p-value among all other effects. The AIC is 912.628 smaller than all effects entered (920.272).

Step 2. Effect TEST1 is removed.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1090.087	910.720
SC	1094.753	971.373
-2 Log L	1088.087	884.720

Analysis of Effects Eligible for Removal

Effect	DF	Wald Chi-Square	Pr > ChiSq
GROUP	1	1.1677	0.2799
TEST2	1	3.2362	0.0720
TEST3	4	140.9387	<.0001
TEST5	6	15.3176	0.0179

Table 5 Step2 of Backward Selection

In step2 (Table 5) “GROUP” is identified as the largest p-value among all other effects. The AIC is 910.720 smaller than step1 (912.628).

Step 3. Effect GROUP is removed.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	1090.087	909.889
SC	1094.753	965.877
-2 Log L	1088.087	885.889

Analysis of Effects Eligible for Removal

Effect	DF	Wald Chi-Square	Pr > ChiSq
TEST2	1	3.2182	0.0728
TEST3	4	141.4565	<.0001
TEST5	6	14.7160	0.0226

Table 6: Step3 of Backward Selection

In Step 3(Table 6) No (additional) effects met the 0.1 significance level for removal from the model.

Step 4. Summary of Backward Selection.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	TEST4	5	5	2.3346	0.8012	TEST4
2	TEST1	1	4	0.0917	0.7620	TEST1
3	GROUP	1	3	1.1677	0.2799	GROUP

Table 7: Summary of backward selection

Partition for the Hosmer and Lemeshow Test

Group	Total	TARGET = 1 Observed	TARGET = 1 Expected	TARGET = 0 Observed	TARGET = 0 Expected
1	79	12	10.52	67	68.48
2	68	16	13.63	52	54.37
3	88	19	20.25	69	67.75
4	97	32	35.94	65	61.06
5	95	42	45.13	53	49.87
6	84	54	55.14	30	28.86
7	84	60	57.51	24	26.49
8	74	56	51.76	18	22.24
9	116	96	97.14	20	18.86

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
3.6080	7	0.8237

Table 8: Hosmer and Lemeshow Test results

Results of the Hosmer and Lemeshow test are shown in Table 8. There is no evidence of a lack of fit in the selected model (p=0.8237).

CONCLUSION

In order to select prognostic factors for this binary data. Initially, a full model containing all six effects is fit to the data. In next 5 steps, PROC LOGISTIC removes *TEST4, TEST* and *GROUP* step by step. This leaves *TEST2, TEST3* and *TEST5* as the only variables in the final model.

Finally you will see that backward selection is sufficiently simple that 'manual' implementation is not overly time consuming and can allow a more pragmatic approach to the removal of explanatory variables from the full model.

REFERENCES

Halinski, R. S. & Feldt, L. S. 1970. "The selection of variables in multiple regression analysis". Journal of Educational Measurement, 7 (3). 151-157.

Peter Williams, Veramed, Biohub (Alderley Park), UK 2017 "The Model Statistician, Insights into model selection techniques in Clinical trials" *Proceeding of PhUSE 2017*

WIKIPEDIA "Stepwise regression". Available at https://en.wikipedia.org/wiki/Stepwise_regression

Selection Process for Multiple Regression. Available at https://www.statisticssolutions.com/selection-process-for-multiple-regression/

NCSS Stepwise Regression. Available at https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Stepwise_Regression.pdf

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS® For Dummies®*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Changxia (Connie) Shao
Boehringer Ingelheim
Connie.986@hotmail.com
www.boehringer-ingelheim.com