"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518, USA), Vol.7, Issue 2, 68-78 | Science Signpost Publishing

Improving Stepwise Logistic Regression Using a SAS Macro

Jian Sun[1,2]

1. School of Public Health, University of Alberta, Edmonton, Canada

2. Department of Medicine, University of Calgary, Calgary, Canada

Abstract

Stepwise covariate selection is a popular method for multivariable regression model building. Based on the different significance levels pre-specified by statisticians, different covariates are included in the model. Further analyses with these models might introduce biases. This paper proposes a novel method to select covariates for stepwise logistic regression without pre-setting a significance level. Multiple models containing different numbers of covariates were outputted for final model selection. A user-oriented SAS macro was developed. Users of the macro may determine the final models, based on estimated characteristic changes of the overall models, the variances of the covariate effects on the response variable and their special needs. With this method, model selections are much easier than with purposeful or the best subsets method. This method improved stepwise covariate selection processes. Broad applications are expected.

Keywords: logistic regression, model building, multivariate statistics, SAS Programming, Statistical computation.

Introduction

Logistic regression is the most frequently used statistical model for the analysis of data with a discrete outcome in various research areas. If the data consist of a few covariates, there may be only one best model, which is easy to build. However, if covariates increase, multiple “best” models may exist.

Author’s address: Jian Sun, PhD, Health Technology and Policy Unit, School of Public Health, University of Alberta, 3-021 RTF Building, 8308-114 Street, Edmonton, Alberta T6G 2V2, Canada.
Email: jsun9@ualberta or sunjian@hotmail.com

Building a best model with data consisted of numerous covariates is difficult.

What is the best model? A best regression model should be a model with the correct covariates and the most precise estimates for them. As Hosmer et al. described, “The traditional approach to statistical model building involves seeking the most parsimonious model that still accurately reflects the true outcome experience of the data” [1]. A model with fewer covariates is numerically stable and easier to use. In contrast, the more covariates included in a model the greater the standard errors of the coefficients and the wider the confidence intervals of the corresponding odds ratios.

Stepwise selection is a popular and effective statistically driven method to reduce covariates of multivariable model including logistic regression. In each stage of the selection procedure, a covariate is added to (or subtracted from) the set of covariates based on a pre-specified significance level by the statistician. With different significance levels, different models are determined. Further analyses with these models might introduce biases.

This paper proposes a novel method to select covariates for backward stepwise logistic regression without pre-setting a significance level. A user-oriented SAS macro was developed. To validate the proposed covariate selection method and the macro, a randomly generated hypothetical dataset was used.

Material and methods

The proposed method is an improvement of the traditional backward elimination method. All covariates were included in the model first. SAS LOGISTIC procedure was used to remove covariates one at a time. The process was repeated until only one covariate was left in the model. Multiple models contained different numbers of covariates were outputted for final model selection. The detail of the method is described by means of the macro.

Table 1 is a SAS macro for data generation. The first DATA step (lines 2-8) generated nine uniform distributed probabilities (*p*s) ranged from zero to one. The second DATA step (lines 11-23) generated nine random samples as covariates *x*1, *x*2, …, *x*8 and response variable *y* with 100 observations from Bernoulli distribution with the variant *p*s derived from the first DATA step. Seeds for both RAND functions were set to "2".

Table 1. SAS macro for data generation

```
1  %macro case(seed=2,obs=100,cov=8);
2   data uniform(keep=p);
3    call streaminit(&seed);
4    do i=1 to %eval(&cov+1);
5     p=rand("Uniform");
6     output;
7    end;
8   run;
9   proc transpose data=uniform out=p prefix=p;
10  run;
11  data Bernoulli (keep=x1-x&cov y);
12   set p;
13   array p{%eval(&cov+1)} p1-p%eval(&cov+1);
14   array x{%eval(&cov+1)} x1-x%eval(&cov+1);
15   call streaminit(&seed);
16   do i=1 to &obs;
17    do j=1 to %eval(&cov+1);
18     x(j)=rand("Bernoulli",p(j));
19    end;
20    output;
21   end;
22   rename x%eval(&cov+1)=y;
23  run;
24 %mend;
25 %case();
```

To determine whether there were any complete or quasi-complete separations [2, 3], missing values, or all ‘0” or all “1” variables in the dataset, a general LOGISTIC procedure for the model containing all variables in the dataset was performed. After evaluating the data, the Macro Variate (Table 2) with file name Bernoulli was invoked.

LOGISTIC procedure (lines 23-25 in Table 2) estimated the coefficient and the *p*-value of the Wald statistic [1, 4] for each covariate. The least significant effect (i.e. a covariate with the largest *p*-value) was removed. For the remaining covariates, the LOGISTIC estimated their effects, and the covariate with the largest *p*-value was removed again. The process was repeated until only one covariate was left in the model.

DATA step Compare&*k* (lines 61-68) merged the full model with each reduced model, and calculated the *delta-beta-hat-percent* [1]

$$\Delta\hat{\beta}_{ik}\% = 100*(\hat{\beta}_{ik}-\hat{\beta}_{i1})/\hat{\beta}_{i1},\ i = 1, 2, \ldots, 8,\ k = 2, 3, \ldots, 8, \quad (1)$$

Table 2. SAS macro for model building

```
%macro variate(mydata=Bernoulli);
 data _null_;
  set &mydata;
  column=n(of _all_)-1;
  call symput('num',put(column,2.));
 run;
 %macro x;
  %do i=1 %to &num;
   %global x&i;
   %let x&i=x&i;
  %end;
 %mend x;
 %macro names;
  %do i=1 %to &num;
   &&x&i
  %end;
 %mend names;
 %macro logistic;
  %x;
  %do j=1 %to &num;
   ods select none;
   ods output ParameterEstimates=_pe FitStatistics=_fit;
   proc logistic data=&mydata;
    model y(event='1')=%names;
   run;
   ods output close;
   proc sql;
    create table output1 as
    select variable, estimate, ProbChiSq
    from _pe
    where variable^='Intercept'
    order by ProbChiSq desc;
   quit;
   data output2;
    set output1;
    max=0;
    if _n_=1 then max=1;
   run;
   proc sql;
    create table output&j._1 as
    select variable, estimate as estimate2,
     ProbChiSq as ProbChiSq2, max
    from output2
    union all
    select Criterion as variable,
     InterceptAndCovariates as estimate2,
     . as ProbChiSq2, 0 as max
    from _fit
    where Criterion='-2 Log L'
    order by variable;
   quit;
   data output&j._2;
    set output&j._1;
    rename ProbChiSq2=ProbChiSq1
     estimate2=estimate1;
    if variable='-2 Log L' then variable='_2LogL';
    if max=1 then do;
     call symput(variable,' ');
     delete;
    end;
    if variable='_2LogL' then variable='-2 Log L';
   run;
  %end;
  %do k=2 %to &num;
   data compare&k(drop=estimate1 ProbChiSq1);
    merge output1_2 output&k._1(in=a drop=max);
    by variable;
    if a;
    delta&k=100*(estimate2- estimate1)/estimate1;
    rename estimate2=estimate&k
     ProbChiSq2=ProbChiSq&k;
    if variable='-2 Log L' then delta&k=estimate2;
   run;
  %end;
  data output1_3;
   set output1_1(drop=max);
   rename estimate2=estimate1
    ProbChiSq2=ProbChiSq1;
  run;
  %macro names2;
   %do m=2 %to &num;
    compare&m
   %end;
  %mend names2;
  data delta;
   merge output1_3 %names2;
   by variable;
   if variable='-2 Log L' then variable='D';
   temp=input(substr(variable,2,3),3.);
  run;
 %mend logistic;
 %logistic;
 proc sort data=delta out=delta1(drop=temp);
  by temp;
 run;
 ods select all;
 proc print data=delta1 noobs;
  title 'Table 3. Delta-beta-hat-percent between full
 and reduced model';
  var variable delta:;
  where variable^='D';
 run;
 proc print data=delta1 noobs;
  title 'Table 4. P-values of the Wald test';
  var variable ProbChiSq:;
  where variable^='D';
 run;
 proc print data=delta1 noobs;
  title 'Table 5. Deviance and coefficient estimate';
  var variable estimate:;
 run;
%mend variate;
%variate();
```

where $\hat{\beta}_{i1}$ and $\hat{\beta}_{ik}$ are the estimated coefficient of the covariate *xi* in the full model and the *k*th reduced model, respectively. In this paper, a full model means a model containing all covariates in the dataset rather than that defined by McCullagh and Nelder [5]. It is not a saturated model, in which there are as many estimated parameters as observations [6]. The reduced model is the model obtained by setting certain parameters in the full model equal to zero [7].

DATA step Delta (lines 79-84) merged the deviances $D = -2LogL$ [1], *the delta-beta-hat percent*, *p*-values and coefficient estimates of each covariate in the full and the reduced models. *L* is the likelihood of a fixed model. PRINT procedures (lines 91-104) printed them out for final model selection. Each column except the "Variable" column in any of the three output tables represented a model. They were named model 1, model 2, …, and model 8. They are in descending order of the number of covariates from left to right.

Based on the outputs of the macro, the process of final model selection, which had three steps, was performed:

(1) Observe *p*-values of each covariate in each model. Choose the smallest model that contains at least one covariate with $p < 0.05$. If the statistician wants to include some important covariates, choose the smallest model containing those covariates.

(2) Check *delta-beta-hat-percent,* $\Delta\hat{\beta}_{ik}\%$ *for the* covariates in the model with $p < 0.05$. If any $|\Delta\hat{\beta}_{ik}\%| >$ 20% or another criterion pre-set based on special requirement, shift one column left in the output tables and repeat this step, until arriving at a column that does not include any $|\Delta\hat{\beta}_{ik}\%|$ which is larger than the pre-set criterion for the covariates with $p < 0.05$.

(3) Compare deviance $D = -2logL$ for each model with $\varkappa^2$ in the Chi-Squared distribution table for $p =$ 0.05. If any $D > \varkappa^2$ with the same degree of freedom (*DF*), shift one column left and then compare *D* with $\varkappa^2$ again with the reduced *DF*. The comparisons will repeat until we find a model with $D < \varkappa^2$ and then return to Step 2. The iteration between steps 2 and 3 continues until a model that meets both $\Delta\hat{\beta}_{ik}\%$ and *D* requirements is found.

An alternative method for step 3 is to conduct the likelihood ratio test [7]:

$$LR_k = -2LogL_k - (-2LogL_{k-1}),\ k = 2, 3, \ldots, 8, \qquad (2)$$

where L_k and L_{k-1} are the likelihoods of the models *k* and *k*-1, respectively. Because deviance $D = -2logL$, which provided by LOGISTIC directly, and $-2logL$ cannot be used as a SAS variable name, *D* was used

for output instead of -2*logL*.

Although the process described above appears complicated, the steps can be completed within minutes.

To validate the methodology and the macro, the best subsets regression [1] was performed with the same dataset using a SAS code created by King [8]. The method and macro were also used for a project to detect factors associated with recommendations for rare disease drug in Canada [9].

All analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC).

Results

Table 3 shows the *delta-beta-hat-percent* $\Delta\hat{\beta}_{ik}\%$ between the full model and each reduced model. *delta*2, *delta*3, …, *delta*8 are the variable names of $\Delta\hat{\beta}_{ik}\%$. Tables 4 and 5 show *p*-values (*ProbChiSq*1, *ProbChiSq*2, …, and *ProbChiSq*8) of the Wald tests and the coefficient estimates (*estimate*1, *estimate*2, …, and *estimate*8) of all covariates in each model, respectively. Deviance (*D*), a criterion rather than a variable is included in Table 5 for convenience.

Table 3. Delta-beta-hat-percent between full and reduced model

Variable	delta2	delta3	delta4	delta5	delta6	delta7	delta8
x1	4.67241	.	.	.	.	.	.
x2	0.60020	0.4229	7.21031	13.4218	18.0088	10.5127	.
x3	3.56317	13.9809	.	.	.	.	.
x4	-0.71575	-0.5464	0.86465	-2.1822	-3.2992	.	.
x5	0.41012	1.9325	-7.05382	.	.	.	.
x6	0.42983	-2.0565	-5.36167	-2.3109	.	.	.
x7	-0.07275	-0.6350	0.81951	1.2553	-3.6413	-10.9315	-18.3892
x8	.	.	.	.	.	.	.

Table 4. *P*-value of the Wald test

Variable	ProbChiSq1	ProbChiSq2	ProbChiSq3	ProbChiSq4	ProbChiSq5	ProbChiSq6	ProbChiSq7	ProbChiSq8
x1	0.90516	0.90072	.	.	.	.	.	.
x2	0.30450	0.30100	0.30185	0.25468	0.22172	0.20086	0.22428	.
x3	0.81040	0.80319	0.77825	.	.	.	.	.
x4	0.35729	0.35958	0.35939	0.35271	0.36482	0.36932	.	.
x5	0.65400	0.65237	0.64712	0.67263	.	.	.	.
x6	0.56257	0.56116	0.56824	0.58055	0.56805	.	.	.
x7	0.00943	0.00957	0.00944	0.00787	0.00731	0.00808	0.01114	0.016159
x8	0.91743	.	.	.	.	.	.	.

Table 5. Deviance and coefficient estimate

Variable	estimate1	estimate2	estimate3	estimate4	estimate5	estimate6	estimate7	estimate8
D	75.6505	75.6615	75.6767	75.7538	75.9349	76.2615	77.1366	78.6116
x1	0.1535	0.1606	.	.	.	.	.	.
x2	0.6631	0.6671	0.6659	0.7109	0.7521	0.7825	0.7328	.
x3	0.2364	0.2448	0.2694	.	.	.	.	.
x4	0.6900	0.6850	0.6862	0.6959	0.6749	0.6672	.	.
x5	0.2857	0.2868	0.2912	0.2655	.	.	.	.
x6	-0.3598	-0.3614	-0.3524	-0.3405	-0.3515	.	.	.
x7	1.7119	1.7107	1.7011	1.7260	1.7334	1.6496	1.5248	1.3971
x8	0.1203	.	.	.	.	.	.	.

Table 4 shows that in any of the eight models only $x7$ is significant ($p < 0.05$). Considering the definition of the best model, i.e. smaller is better if no other significant effect changes, the model with only one covariate $x7$ was chosen first and its *delta-beta-hat-percent* $\Delta\hat{\beta}_{78}\%$ (= -18.3892) in the column *delta*8 of Table 3 was checked. If $|\Delta\hat{\beta}_{ik}\%| <= 20\%$ was used as a criterion, this univariate model would be the final model. If the 18% coefficient change were not satisfied, this model would not be chosen. Instead, if 15% coefficient changes were accepted, $x2$ would be added to the model. More rigorously, if even a 10% variation for the coefficient of $x7$ were not permitted, $|\Delta\hat{\beta}_{ik}\%| <= 10\%$ would be set and the three covariate model with $x2$, $x4$ and $x7$ would be chosen. It would not matter that the value of *delta*6 for $x2$ increased to 18.0088% >10%, because $x2$ had no significant effect ($p = 0.20086$) on the outcome. However, although $x2$ had no direct significant effect on outcome, combining with $x4$, it adjusted the coefficient of $x7$ from 1.39711 to 1.64959. The odds ratio of $x7$ increased from 4.044 to 5.205 (not shown in the table).

If n (= 100 here) is the observation number and q (= 8 here) is the covariate number in the full model, the degree of freedom (*DF*) for the deviance test is $n - q = 92$. In Table 5, the *D* for each model is smaller than 80, while χ^2 with $DF = 92$ is 115.39 for $p = 0.05$. For reduced models, the *DF*s for the deviance tests are larger than 92 and the corresponding $\chi^2 > 115.39$. Therefore, all models here have no significant differences from the saturated model at the significance level 0.05.

The result of the likelihood test was the same as that of the deviation test. The difference of *D*s between any two adjacent columns in Table 5 was much smaller than 3.84, the critical value of the Chi-Square statistic with $DF = 1$. Therefore, removing any covariates except $x7$ did not change the overall model significantly at $p = 0.05$ level.

If it were important to keep covariate $x4$, for example, in the model, the iteration should be started

from the model containing $x2$, $x4$ and $x7$ for the model selection process.

Finally, the result of the best subsets regression (not shown) verified that among all possible models with the same number of covariates, all reduced models generated by the macro with the hypothetical data had the smallest Mallows' *Cp* [1,10].

The method and macro were used for the project to detect factors associated with recommendations for rare disease drug coverage in Canada [9]. The real data consist of 92 observation and 15 covariates. All variables were binary. Based on the proposed method, five covariates were remained in the model. Although they were the same as using stepwise method with that the significance levels for entry and stay were set at 0.2 in chance, with the proposed method, the coefficient changes (< 20%) of significant covariates in the model were derived. The overall model did not change significantly at $p = 0.05$ level after 10 covariates were removed [9].

Discussion

Why a new method?

The SAS LOGISTIC procedure is a convenient tool for us to perform stepwise regressions. Statisticians only need to specify significance levels for variable entry and/or stay, then LOGISTIC will automatically select covariates to build the model for us. However, the significance levels are set arbitrarily. The arbitrariness may result in bias for the following analyses. Because of different significance levels, different sets of covariates will remain in the fixed models. If the significance level were too low, important effects would be missed. In contrast, if the significance level were set too high, some covariates, which have little effect on the outcome would be included in the model, diluting the important relationships between other covariates and the outcome.

A remedial strategy is using a higher significance level to run the LOGISTIC first, and then manually removing one covariate with the largest *p*-value and running the LOGISTIC again. This process will repeat until there are no covariates with *p*-values larger than the pre-set significance level in the model. Every time after removing a covariate, statisticians have to evaluate the characteristic changes of the remaining covariates and the overall model. The higher the significance level pre-set, the more covariates will be included. Therefore, the manual workload to eliminate redundant covariates will increase. While the significance level increases to 1.0, regardless of forward, backward, combined stepwise or purposeful selection methods, the selection process becomes a backward selection without a significant level. The

idea of the proposed method came from this strategy.

What is new?

Traditional backward elimination regression needs a pre-set significance level to control the covariate selection process. The process stops when there are no covariates that meet the criteria for removal. The proposed method does not need a pre-set significance level. Therefore, the process will not stop until only one covariate remains in the model.

The delta-beta-hat-percent $\Delta\hat{\beta}\%$ proposed by Hosmer et al. [1] was calculated for each covariate in each reduced model. The numerators are the differences of the coefficient estimates between the full model and each reduced model [11]. As we have seen from Table 5, the estimates of the coefficients vary depending on the presence or absence of other covariates included in the model. However, the variations are not monotonic. Removing a covariate may increase or decrease other covariates' effects on outcome. During the process of the backward selection, even if at a stage a coefficient changed a lot, the direction of the change might toward to its initial value in the full model. In contrast, even if the change were minor at a stage of the process, the estimated coefficient might quite differ to its original value. Therefore comparing reduced models with the full model appears more reasonable.

A user-oriented macro was developed. To validate the proposed method and the macro, a hypothetical dataset was generated. For simplicity, all variables are binary. Referring to the three output tables, statisticians can decide their final models quickly and confidently. The decision is more flexible than with the traditional stepwise method. Statisticians can change their criteria and build different final models to meet their special requirements. Same as the best subsets method, the output of the macro provides multiple models for further selections. However, with the proposed method, the number of covariates included in the final model is easy to determine.

The results of deviance and likelihood tests are the same. If observation number is larger, the latter is more convenient. Instead of checking the Chi-Squared distribution table with larger *DF*, statisticians only need to compare the difference between each pair of adjacent deviances *D*s with 3.84, the critical value of Chi-Square statistic with *DF* =1.

How to use the macro?

As with traditional logistic regressions, the data should not contain any variables with all "0"or all "1" values. Complete or quasi-complete separation should not occur. In addition, the data should not contain any variables with missing values. Because the LOGISTIC in the macro runs iteratively, the dataset should

always be the same during the iteration. If the data contained a variable with missing values, the effective observation number will increase while this variable removes.

After data preparation, users need to change their variable names to *y*, *x*1, *x*2, *x*3, and so on, exclude variables not for the regression, and then input their file names to invoke the macro. The number of variables and observations are arbitrary.

Readers can also use the macro in Table 1 to generate a dataset with a different seed to validate the macro in Table 2. I chose seed = "2" for convenience. If you choose another seed number, you may need to deal with complete or quasi-complete separation before running the macro. For illustrative purposes, I set the covariate number to "8". Because the number was relatively small, the characteristic changes of the overall model and the coefficient estimates during the process were not obvious.

After running the macro, statisticians can screen the output tables to select their final models. The macro focuses on the main effect model building. If statisticians want to identify interactions, they can add corresponding interactions to the fixed model and run the LOGISTIC procedure to build their final models with interactions easily.

In practice, one or more important covariates may have to stay in the model. Suppose that the purpose of a medical study is to compare two treatment effects. If the covariate represented the treatment were removed, the analysis would be meaningless. In this case, the statistician should select a final model containing the treatment from the macro outputs.

Conclusions

This paper proposed a novel method for stepwise logistic regression. A user-oriented SAS macro was included. With this method, model selection is much easier than with purposeful or the best subsets method. This method improved the stepwise covariate selection process. Broad applications are expected.

Acknowledgment

The author thanks Dr. Tania Stafinski, School of Public Health, University of Alberta, for the initial manuscript revision.

References

[1]. D. W. Hosmer, S. Lemeshow and R. X. Sturdivant. Applied logistic regression. 3rd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2013.

[2]. A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71, 1, 1-10, 1984.

[3]. C. Rainey. Dealing with separation in logistic regression model. Political Analysis, 24, 339-355, 2016.

[4]. W. H. Greene. Econometric analysis, 6th ed., Prentice Hall, Boston, 2008.

[5]. P. McCullaph and J. A. Nelder. Generalized linear model, 2nd ed., Chapman and Hall, London, 1989.

[6]. A. Agresti. Categorical data Analysis, 3rd ed., John Wiley & Sons, Inc., Hoboken, New Jersey, 2013.

[7]. D. G. Kleinbaum. Logistic regression: a self-learning text, Springer-Verlag, 1994.

[8]. J. E. King. Running a best-subsets logistic regression: an alternative to stepwise methods. Educational and Psychological Measurement, 63, 392-403, 2003.

[9]. F. N. I. Nagase, T. Stafinski, Sun J, G. Jhangri and D. Menon. Factors associated with positive and negative recommendations for cancer and non-cancer drugs for rare diseases in Canada. Orphanet Journal of Rare Diseases, 2019; https://doi.org/10.1186/s13023-019-1104-7.

[10]. C. L. Mallows. Some comments on Cp. Technometrics, 15, 661-675, 1973.

[11]. Z. Bursac, C. H. Gauss, D. K. Williams and D. W. Hosmer. Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine, 3, 17, 2008, (no page numbers).