

A macro of building predictive model in PROC LOGISTIC with AIC-optimal variable selection embedded in cross-validation

Hongmei Yang, Andréa Maslow, Carolinas Healthcare System.

ABSTRACT

Logistic regression leveraging stepwise selection has been widely utilized for variable selection in health care predictive modeling. However, due to the drawbacks of stepwise selection, new ideas of variable selection are emerging, including Akaike Information Criterion (AIC)-optimal stepwise selection which utilizes AIC as the criterion for variable importance and builds a model based on a combination of stepwise logistic regression and information criteria. As predictive factors selected over a single sample may over fit the sample and have poor prediction capability on independent test data, embedding variable selection in resampling techniques, such as cross-validation, is recommended to appropriately estimate expected prediction error, especially with a limited sample size. When processing the AIC-optimal selection through cross-validation, different lists of influential variables may be selected over the iterations. Simply averaging the coefficients would yield a final model with many more predictors than necessary, and therefore reduced predictive accuracy. This paper proposes additional steps to address this issue. Variables selected in the AIC-optimal stepwise process are ranked by their frequency appearing in the AIC-optimal lists obtained from cross-validation iterations. A final model is obtained by sequentially adding the variables with the same frequency until an optimal averaged area under the Receiver Operating Characteristic curve (AUC) is achieved. We present the algorithm and the macro used to achieve the selection in the context of cross-validation.

Intended audience: SAS users of all levels who work with SAS/STAT and PROC LOGISTIC in particular.

INTRODUCTION

In predictive modeling, researchers are interested in determining the “best” subset of predictors out of many covariates. Automatic stepwise selection allows researchers to select useful subsets of variables by evaluating the order of importance of variables. Since it was first developed in 1960s, stepwise selection has been widely used and remains the most commonly used approach for variable selection in academic and health care settings (Walter & Tiemeier, 2009). However, deficiencies of stepwise selection have been reported in the literature. The main drawbacks include (a) inflated statistical significance levels (i.e., standard errors of the model coefficients and p-values are biased downward) due to use of incorrect degrees of freedom; (b) default P-value ($\alpha=0.05$) used to determine a stopping rule; (c) lack of replicability due to its dependence on sampling error; and (d) reliance on the single best model, while ignoring model uncertainty in producing the estimates (Derksen & Keselman, 1992; Harrell, 2001; Rothman et al., 2008; Thompson, 1995; Wilkinson, 1979). The deficiencies of stepwise methods become more apparent when the number of covariates is large and multicollinearity exists.

To overcome some of the problems, new ideas of variable selection are emerging. Wang (2000) used Akaike information criteria (AIC) as a criterion of variable importance and built a model based on a combination of stepwise logistic regression and information criteria. Along the lines of AIC-optimal selection, Shtatland et al. (2000, 2002) proposed a three-step procedure in which a stepwise regression method was first used to obtain a full stepwise sequence, then AIC was used to find an AIC-optimal model in this stepwise sequence, and lastly, best subset selection was applied to model sizes that were in the neighborhoods of the optimal size to obtain a ‘confidence’ sets of models. Although the approaches avoid the agonizing process of choosing the ‘right’ critical p-value, the possible impact of sampling error was not adequately considered. It is recommended to embed variable selection in resampling techniques, such as cross-validation, to appropriately estimate expected prediction error, especially with a limited sample size (Fox, 1991; Harrell, Lee & Mark, 1996; Henderson & Valleman, 1981). When processing the AIC-optimal selection through cross-validation, different lists of influential variables may be selected over the iterations. Simply averaging the coefficients would account for model uncertainty but would also yield a final model with many more predictors than necessary, and therefore reduced predictive accuracy.

This paper proposes additional steps to address the issue of multiple lists of influential variables obtained through cross-validation. We rank variables selected in the AIC-optimal stepwise process by their frequency appearing in the AIC-optimal lists obtained from cross-validation iterations. We obtain a final model by sequentially adding the variables with the same frequency until an optimal averaged area under the Receiver Operating Characteristic curve (AUC) is achieved. We present the algorithm and the macros used to achieve the selection in the context of cross-validation.

ALGORITHM

We detail the algorithm of AIC-optimal variable selection embedded in cross-validation in Figure 1.

```

For each resampling iteration do
    Hold out specific samples
    Fit stepwise logistic regression on the remainder with SLENTRY and SLSTAY close to 1
    Identify the step with optimal AIC
    Obtain a list of variables achieving the optimal AIC (AIC-optimal variable list)
End
Merge all the AIC-optimal variable lists
Obtain frequency of each unique variable appearing in the merged data
Transpose data to show what variables have the same frequency and determine the number of variables at each frequency level
Concatenate variable names which are of the same frequency
Assign the concatenated variable names to macro variables and name the macro variables with a suffix equal to the frequency ( $S_F, S_{F-1}, S_{F-2}, \dots, S_{F-(F-1)}$ ); or  $S_{F-i}, i = 0 \text{ to } F - 1$ , where  $F$  is the total number of iterations)

For each cumulative sum of the macro variables ( $S_{F-i}, i = 0 \text{ to } F - 1$ ), do
    For each resampling iteration do
        Hold out specific samples
        Fit logistic regression model on the remainder
        Predict the hold-out samples
        Calculate AUC based on its relationship with Somer's D
    End
    Merge AUCs from all the iterations
    Calculate the average AUC across all the hold-out predictions
End
Determine the final list of predictors (i.e., the one with biggest average AUC) to be kept in the final model
Fit the final model to the original training data

```

Figure 1. Algorithm of AIC-Optimal Variable Selection in the Context of Cross-Validation

MACROS

We develop two macros to fulfil the above algorithm. The first macro (%AICoptSW) performs AIC-optimal stepwise logistic regression on each resampling iteration to obtain lists of variables achieving the optimal AIC. With some additional data steps, the macro creates a character variable with values of concatenated variable names which appear the same number of times in the AIC-optimal lists. The second macro (%cvAUC) performs repeated cross-validation of logistic regression and estimates model performance through averaged AUC over all hold-out predictions. The final model has the best performance based on the averaged AUC.

```

/*****
* Macro #1: AICoptSW
* Description: Perform AIC-optimal stepwise models in each iteration of K-fold cross-validation to obtain
lists of variables achieving the optimal AICs and their frequency appearing in the AIC-optimal lists
obtained from cross-validation iterations.
*
* Parameters:
*   The following parameters define the data used to fit the model.
*

```

```

* indat      SAS dataset containing all necessary variables.
* y          The response variable for the logistic regression model with '1' as the event of interest.
* x          The list of predictors that appear in the MODEL statement.
*
* The following parameters define the features of K-fold cross-validation.
*
* seed       A seed for reproducibility of random partition of the data into folds.
* fold       Specify the number of disjoint validation subsets.
* repeats    Number of times the cross-validation will be repeated
*
*****/

%macro AICoptSW(indat=, y=, x=, seed=, fold=, repeats=);

*Partition data into &fold folds and repeat &repeats times;
data _modif;
  set &indat;
  %do i=1 %to &repeats;
    unif_&i=&fold*ranuni(&seed+&i);
    fold_&i=ceil(unif_&i);
  %end;
run;

%do i=1 %to &repeats;
  %do j=1 %to &fold;
    *For each fold, run stepwise logistic regression on the remaining data with both SLENTRY and SLSTAY
    close to 1 to obtain the sequence of variables entering the model;
    proc logistic data=_modif (where=(fold_&i ne &j));
      model &y (event='1')= &x / selection=stepwise slentry=0.99 slstay=0.995;
      ods output ModelBuildingSummary=SUM;
      ods output FitStatistics=FIT;
    run;

    *For each selection sequence, identify the step with optimal AIC;
    proc sql;
      select Step into :nstep from FIT where Criterion="AIC"
        having InterceptAndCovariates=min(InterceptAndCovariates);
    quit;

    *Obtain a list of variables achieving the optimal AIC from the selection sequence;
    proc sql;
      create table sequence_&i&j as
        select EffectEntered, &i as rpts, &j as flds from SUM where Step<=&nstep;
    quit;
  %end;
%end;

*Merge all the AIC-optimal variable lists;
data seqData;
  set
    %do i=1 %to &repeats;
      %do j=1 %to &fold;
        sequence_&i&j
      %end;
    %end;
;

```

```

run;

*Get frequency of each unique variable appearing in the AIC-optimal lists;
proc sql;
  create table varFreq as
    select distinct EffectEntered, count(*) as counts from seqData group by EffectEntered
    order by counts,EffectEntered;
quit;

*Transpose data to show what variables have the same frequency;
proc transpose data=varFreq out=varFreq_wide;
  by counts;
  var EffectEntered;
run;

*Determine the number of variables at each frequency level;
proc sql;
  select nvar-3 into :nvar from dictionary.tables where libname='WORK' and
memname='VARFREQ_WIDE';
QUIT;

%let nvar=&nvar;

*Concatenate variable names which have the same frequency to create a character variable holding the
list of variable names for each frequency level;
*Suggest to save the output data set "varFreq_wide" from macro AICoptSW to a permanent library as
this data set is needed for following modeling;

data libName.varFreq_wide;
  length varlist $1000;
  set varFreq_wide;
  varlist= catx(" ", of COL1 - COL&nvar);
run;
%mend;

*Assign the concatenated variable names to macro variables and name the macro variables with a
suffix equal to the frequency;
data _null_;
  set libName.varfreq_wide;
  call symput('covar'||left(put(counts,3.)),varlist);
run;

*View value of user defined macro variables;
%put _user_;

/*****
* Macro #2: cvAUC
* Description: Perform repeated cross-validation of logistic regression and estimate model performance
by averaging AUCs over all fitted models.
*
* Parameters:
*   The following parameters define the feature of the data and K-fold cross-validation.
*
*   y           The response variable for the logistic regression with '1' as the event of interest.

```

```

*          Same as in macro %AICoptSW.
* covars   The macro variables obtained from above which hold predictors in each frequency
*          level. Sequentially add the macro variables from most frequent to least and
*          assess model performance till an optimal averaged AUC is achieved.
* fold     Specify the number of disjoint validation subsets. Same as in macro %AICoptSW.
* repeats  Number of times the cross-validation will be repeated. Same as in macro %AICoptSW.
*
*****/

%macro cvAUC(y=, covars=, repeats=, fold=);
  %do i=1 %to &repeats;
    %do j=1 %to &fold;
      *For each fold, perform logistic regression on the remaining data to train the model;
      proc logistic data=_modif (where=(fold_&i ne &j)) outmodel=_mod&i&j;
        model &y (event='1')=&covars /firth lackfit;
        ods output ParameterEstimates=coeff&i&j;
      run;
      %if print^=0 %then %do; proc printto file='junk.txt'; %end;

      *For each fold, apply the trained model to the fold data to predict;
      proc logistic inmodel=_mod&i&j;
        score data=_modif(where=(fold_&i=&j)) out=out&i&j fitstat;
      run;

      *For each fold, obtain Somers' D;
      proc freq data=out&i&j;
        tables p_1*&y/noprint measures;
        ods output measures=measure&i&j;
      run;

      *For each fold, calculate AUC based on its relationship with Somers' D;
      data measure&i&j (keep= AUC AUC_95LL AUC_95UL rpts flds);
        set measure&i&j (keep= statistic value ase);
        where statistic="Somers' D R|C";
        AUC=(value+1)/2;
        AUC_95LL= AUC-1.96*(ase/2);
        AUC_95UL= AUC+1.96*(ase/2);
        rpts=&i;
        flds=&j;
      run;

      %if print^=0 %then %do; proc printto; run; %end;
    %end;
  %end;

  *Merge all AUC measures over the fitted models;
  data auc;
    set
      %do i=1 %to &repeats;
        %do j=1 %to &fold;
          measure&i&j
        %end;
      %end;
  ;
run;

```

```

*Obtain averaged AUC;
proc means data=auc;
  class rpts;
  var auc;
run;

proc means data=auc ;
  var auc;
run;

%mend;

%cvAUC (y=y, covars=&covar30, repeats=3, fold=10);
%cvAUC (y=y, covars=&covar30 &covar29 &covar28 &covar27, repeats=3, fold=10);
%cvAUC (y=y, covars=&covar30 &covar29 &covar28 &covar27 &covar26 &covar25 &covar24,
repeats=3, fold=10);

```

CONCLUSION

Along the lines of AIC-optimal stepwise selection, this paper proposes additional steps to address the challenge of variable selection in the context of cross-validation. Conclusions suggest to select variables based on a combination of stepwise sequence, AIC, and frequency appearing in the AIC-optimal variable lists over all the cross-validation iterations. The steps are automated by using the macro language.

REFERENCES

- Rothman KJ, Greenland S, Lash TL. 2008. Modern Epidemiology. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.
- Walter S, Tiemeier H. 2009. Variable selection: current practice in epidemiological studies. *Eur J Epidemiol*, 24:733–736. doi: 10.1007/s10654-009-9411-2.
- Derksen S, Keselman HJ. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *Br J Math Stat Psychol*, 45:265–282. doi: 10.1111/j.2044-8317.1992.tb00992.x.
- Harrell FE. 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer.
- Thompson B. 1995. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55: 525-534.
- Wilkinson L. 1979. Tests of significance in stepwise regression. *Psychological Bulletin*, 86: 168-174.
- Wang Z. 2000. Model selection using Akaike information criterion. *STATA Technical Bulletin*, 54: 47-49.
- Fox J. 1991. Regression diagnostics: An introduction. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-079. Newbury Park, CA: Sage.
- Harrell FE, Lee K, Mark DB. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15: 361-387.
- Henderson HV, Velleman PF. 1981. Building multiple regression models interactively. *Biometrics*, 37: 391 – 411.
- Shtatland ES, Kleinman K, Cain EM. 2003. Stepwise methods in using SAS PROC LOGISTIC and SAS enterprise miner for prediction. SUGI'28 Proceedings, Paper 258-28. Cary, NC: SAS Institute Inc.
- Shtatland ES, Cain E, Barton MB. 2001. The perils of stepwise logistic regression and how to escape them using information criteria and the Output Delivery System. SUGI'26 Proceedings, Paper 222-26. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Hongmei Yang

Care Delivery and Population Health Analytics, Carolinas HealthCare System

Tel: 704-446-7745

Email: Hongmei.yang@carolinashealthcare.org