Optimism of Best Subset Selection by AIC/BIC for Prognostic Model Building

Yinghui Miao, NCIRE, San Francisco, CA
Irena Stijacic Cenzer, University of California at San Francisco, San Francisco, CA
Katharine A. Kirby, University of California at San Francisco, San Francisco, CA
W. John Boscardin, University of California at San Francisco, San Francisco, CA

ABSTRACT

The widespread use of predictive models for health-related outcomes in life science research highlights the need for developing and validating accurate prognostic models. Over-optimism of model performance measures is a common phenomenon in risk prediction modeling in medical settings. In WUSS-2012 and SUGI-2013 presentations and proceedings, we proposed an approach to evaluate optimism due to variable selection and coefficient estimation and presented a SAS® macro that works for logistic and Cox regression models with both best subsets and stepwise selection by using the traditional and .632 bootstrapping methods per Harrell algorithm. This current paper is a further development of our work to find an optimal subset based on the Akaike information criterion (AIC) and the Schwarz Bayesian information criterion (BIC) and to estimate the optimism of this process. We demonstrate this approach and the related SAS® macro by developing a prognostic model on ten year mortality in older adults. This approach together with the SAS® macro should be useful for several aspects of prognostic model building.

INTRODUCTION

Accurate developing and validating prognostic models for health-related outcomes is important in life science research [1-3]. Multivariable regression modeling is one of the most commonly used statistical analysis techniques for developing prognostic models. The number of predictors is often quite large relative to the sample size, and thus some sort of model selection must be performed to limit the number of predictors in the model [3,4,6,22].

Regarding variable selection for modeling, an important, although somewhat artificial, distinction is to be drawn between (i) deductive model selection where the researcher interactively determines which explanatory variables are to be included in the regression model based on substantive considerations, and (ii) automated model selection where a computer is used to select the variables for inclusion. The drawbacks of stepwise method as the automated model selection such as instability and bias in regression coefficients estimates, their standard errors and confidence intervals lead to a substantial decrease in predictive ability [4,8,9]. In many situations, an adjuvant to deductive selection is desirable or even necessary. In these cases, alternative automated selection procedures, such as best subsets selection [6] can often substantially inform the deductive approach.

In order to validate multivariable regression models and correct for selection effects, some popular validation techniques, such as split-sample validation, cross-validation, and bootstrap are commonly used [1,2,4,7]. Also, some analytic procedures or approaches have been developed [5]. Shtatland and colleagues developed a three-step procedure, which incorporates the conventional stepwise logistic regression, information criteria, and best subsets regression. Shtatland’s procedure improves variable selection capabilities [8-10]. In particular, Shtatland’s approach attempts to mimic cross-validation and the bootstrapping results without performing both techniques.

Assessing the degree of over-fitting or optimism in regression models is a critical part of predictive modeling [14-16]. Harrell et al. presented modeling strategy for developing clinical multivariable prognostic models and for assessing their calibration and discrimination by using bootstrapped data sets to repeatedly quantify the degree of over-fitting in modeling process [1].

Based on Harrell’s algorithm [1], we present a macro for calculating Harrell’s optimism in the development of a predictive model by focusing on two common specific settings: (a) logistic regression, with measure of discrimination the area under the ROC curve (the c-statistic); and (b) Cox proportional hazards regression, with measure of discrimination Harrell’s c statistic. In WUSS-2012 and SUGI-2013, we proposed an approach to evaluate optimism due to variable selection and coefficient estimation per Harrell’s algorithm. In addition, we presented a SAS® macro that works for logistic and Cox regression models with both best subsets and stepwise selection by using the traditional and .632 bootstrapping methods [17,20,21]. This current paper is a further development of our work on strategies of model building to incorporate best AIC/best BIC subset selection and class variable functionality for best subset selection.
IDENTIFYING A SINGLE BEST MODEL FROM BEST SUBSETS SELECTION OUTPUT

Best subsets variable selection is a versatile technique because it allows researchers to compare models via summary statistics and then select one or more best sets of variables based on statistical and/or substantive grounds. However, in order to assess the optimism of best subsets regression [6], we need a method to (artificially/algorithmically) identify a single best model. Our macro has three alternatives: (i) a simple test of the point diminishing returns in the score statistic (i.e. we choose the largest number of predictors for which the drop in the score statistic from the next largest is still nominally statistically significant by a chi-squared test); (ii) the best of the best subset models by the AIC criterion; and (iii) the best of the best subset models by the BIC criteria. Finally, we apply “CLASS” functionality to supplement these best models with left-out dummy variables from grouped categorical predictors if there are incomplete grouped categorical predictors (e.g., if two of three non-reference levels for age group were chosen, we would automatically include the third level). This functionality gives a method of implementing class variables in the selection process [21].

THE BEST MODEL FOR AIC OR BIC

Model selection via AIC has been shown to be asymptotically equivalent to cross-validation and the bootstrap. Shtatland et al. proposed an approach to calculate AIC for a number of models near a model identified by stepwise selection. The models near to the given model are searched using best subsets techniques [8-10]. This process of manually calculating the AIC for a number of likely near-AIC-optimal models as an approximation to maximizing the AIC over all possible models is referred to as the AIC-blanket approach. We use a similar approach and calculate the AIC-blanket over all the models identified in the best subsets output, and we extend the idea to an BIC-blanket as well. BIC identifies more parsimonious models which can be appealing to substantive researchers in prognostic model building.

COMBINING AUTOMATED SELECTION AND LOGICAL MODEL BUILDING

Deductive models selection, also known as logical selection [18,19] is to determine potential predictors that can be directly included in the regression model based on substantive considerations, such as literature or a study-specific conceptual model. In it is usually preferable to combine aspects of automated selection and logical model building. To assess this modeling practice, we specify the “INCLUDE=” option which is incorporated in SAS® “SELECTION=SCORE” option to include the first \( n \) effects in the MODEL statement in every model. These first \( n \) effects are the selected predictors per the logical selection.

SAS® MACRO FOR MODEL BUILDING

We developed a SAS® macro that has the advantage of using multiple model performance indices for correcting over-optimism and model comparisons. The main features of this macro are:

- Identify an optimal prognostic using three best subsets methods: (i) diminishing returns; (ii) best AIC; (iii) best BIC as well as by stepwise selection.
- Calculate average Harrell’s optimism from bootstrapping validation data.
- Estimate Harrell’s optimism due to variable selection only or due to variable selection and coefficient estimation for the various selection methods.

The macro includes methods for both logistic regression (the macro HARRELL_OPTIMISM_LOGISTIC) and Cox regression (the macro HARRELL_OPTIMISM_COX); performs either best subsets selection or stepwise selection, and can handle grouping of categorical variables for both methods. It can differentiate between optimism due to variable selection only and optimism due to variable selection and coefficient estimation. Outputs of the macro provide the Best Subsets blanket in which the best subsets models, the best AIC models, and the best BIC models were chosen. Statistics regarding optimism are presented in outputs. Also, lists of variables in the full model and the percent of frequencies they were selected in the final model per best subsets selection and stepwise selection, respectively. To assess frequently selected models, the frequency along with percentage of each selected final model is summarized in tables. For Cox regression [11], Harrell’s \( c \) statistic is not directly available in PROC PHREG or other SAS® procedures. Therefore, our macro calls another macro, SURVCSTD, developed by WALTER Kremers to calculate Harrell’s \( c \) statistic [12,13]. In this paper we present code for a shortened version of the macro for logistic regression.

DEFINING THE MACRO CALL

The macro call for both the logistic regression and the Cox proportional hazards model is specified in the following manner:

ORIGDATA – the original full dataset.
SEED – specifies the initial seed for random number generation. Allows for ability to replicate results.
REPS – desired number of bootstrap replications.
ID – sampling unit (identifier)
ORIG_BESTNO – number of subsets of each size for best subsets to display for models fitted in the original data set.
BOOT_BESTNO – number of best subsets of each size to use in computing the AIC/BIC blanket.
FULLMODEL – a list of all predictors.
CLASSVAR – a list of all categorical variables (with all categories except the reference category entered as dummy variables) to be treated as a group in the selection process.
NONCVAR – a list of non-grouped variables among the potential predictor variables. See example for details on using the CLASSVAR and NONCVAR options.
STARTNO – the smallest subset size to consider for best subsets method.
STOPNO – the largest subset size to consider for best subsets method.
INCLUDENO – the number of predictors (first n effects in the MODEL statement in every model) to always include for best subset method.
SLENTRY – the significance level of the score chi-square for entering an effect (predictor) into the model.
SLSTAY – the significance level of the Wald chi-square for an effect (predictor) to stay in the model.

Specific macro call for the logistic regression:
OUTC – outcome variable
EVENTNO – the event category for the binary response model.

Specific macro call for the Cox proportional hazards model:
EVENT – the indicator for experiencing the event (EVENT=1) vs. censored (EVENT=0).
START/TIME – the entry time and event/censoring time variable.
TIES – specifies the method of handling ties in failure times.

MACRO (SHORTENED VERSION)
Select the predictors and fit a model using the full data set and the best subsets selection method, keep all optimal models in a blanket.
/*Overall Macro call*/
%MACRO Harrell_Optimism_Logistic (ORIGDATA=, SEED=, REPS=, ID=, OUTC=, EVENTNO=, ORIG_BESTNO=, BOOT_BESTNO=, FULLMODEL=, CLASSVAR=, NONCVAR=, STARTNO=, STOPNO=, INCLUDENO=, SLENTRY=, SLSTAY=);
/*Generate bootstrapped data sets with replacement*/
proc surveyslect data=&ORIGDATA out=BOOT seed=&SEED method=URS samprate=1 outhits rep=&REPS;
run;
/*Logistic regression: fit a model using the full data set with the best subsets selection method and deductive variables selection*/
proc logistic data=&ORIGDATA;
model &OUTC (event="&EVENTNO")=&FULLMODEL / selec selection=score best=&ORIG_BESTNO start=&STARTNO stop=&STOPNO include=&INCLUDENO;
ods output BESTSUBSETS=BSORIG;
run;
/*Call macro of Best subsets selection procedure: identify the best subset model, the best AIC model, the best SC (BIC) model based on the full data set*/
%BESTSUBSET (ORIGDATA=&ORIGDATA, INDATA=BSORIG, CLASSVAR=&CLASSVAR, NONCVAR=&NONCVAR, MODELDATA=&ORIGDATA, LEFTDATA=&ORIGDATA, OUTC=&OUTC, EVENTNO=&EVENTNO, OUTDATA=ORIGBS);

Select the predictors and fit a model using the bootstrapped data set and the best subsets selection method, keep all optimal models in a blanket.
%do M=1 %to &REPS;
/*Generate individual bootstrapped data set and its corresponding left data set*/
data BOOT_A; set BOOT;
where REPLICATE=&M;
proc sort; by &ID;
run;
Estimating Harrell’s Optimism on Predictive Indices Using Bootstrap Samples, continued

data BOOT_B; set BOOT_A (KEEP=&ID REPLICATE); by &ID;
if first.&ID; proc sort; by &ID; run;
proc sort data=&ORIGDATA; by &ID; run;
data LEFT;
merge &ORIGDATA (in=A) BOOT_B (keep=&ID in=B);
by &ID; if A and not B; proc sort; by &ID; run;
/*/logistic regression: fit a model using the full data set with the best subsets
selection method and deductive variables selection*/
proc logistic data=BOOT_A;
model &OUTC (event="&EVENTNO")=&FULLMODEL / selection=score best=&BOOT_BESTNO
start=&STARTNO stop=&STOPNO include=&INCLUDENO;
ods output BESTSUBSETS=BS_BOOT&M; run;
data BSBOOT; set BS_BOOT&M; run;
/*Call macro of Best subsets selection procedure: identify the best subset model, the
best AIC model, the best BIC (BIC) model based on the full data set*/
%MACRO BESTSUBSET (ORIGDATA=, INDATA=, CLASSVAR=, NONCVAR=, MODELDATA=, LEFTDATA=,
OUTC=, EVENTNO=, OUTDATA=);
/*Identify the best Best Subsets Model*/
data INDATA0; set &INDATA;
if _N_=1 or CONTROL_VAR="1"; proc sort; by ScoreChiSq; run;
data INDATA1; set INDATA0; by ScoreChiSq;
DIFF_SCORECHISQ=dif(SCORECHISQ);
if DIFF_SCORECHISQ^=0 and DIFF_SCORECHISQ<3.841459 then delete; run;
data INDATA2; set INDATA1 end=last; if last then output; run;
data INDATA2; set INDATA2; BESTMODEL=1;
proc sort; by CONTROL_VAR NUMBEROFVARIABLES; run;
data INDATA3; merge &INDATA INDATA2; by CONTROL_VAR NUMBEROFVARIABLES; run;
/*Extract each of the individual optimal models from the blanket of models from best
subsets selection*/
%let DSID=%sysfunc(open(&INDATA));
%let NOBS=%sysfunc(ATTRN(&DSID,NOBS));
%do I=1 %to &NOBS;
data A; set INDATA3; if _N_=&I; run;
/*CLASS functionality*/
data _NULL_;
CLASSVNUMBER=countw("&CLASSVAR");
NONCVNUMBER=countw("&NONCVAR");
call symputx ('CLASSVARNO',CLASSVNUMBER);
call symputx ('NONCLASSVARNO',NONCVNUMBER);
CVAR=compress("&CLASSVAR","0123456789");
call symputx ('CLASSVAR2',CVAR);
run;
%if &CLASSVAR^=" " and &NONCVAR^=" " %then %do;
data BS0; set A (keep=NUMBEROFVARIABLES VARIABLESINMODEL SCORECHISQ DIFF_SCORECHISQ BESTMODEL);
length CLASSVNUMBER NOCLASSVNUMBER CVAR CVARSINMODEL $1024;$;
array CVAR1[&CLASSVARNO] $32 _temporary_; array CVAR2[&CLASSVARNO] $32 _temporary_;
Estimating Harrell’s Optimism on Predictive Indices Using Bootstrap Samples, continued

CVARNOSUM=0; length CVARCATE $32;
do K=1 to dim(CVAR1);
 CVAR1[K]=scan("&CLASSVAR2",K,' '); CVAR2[K]=scan("&CLASSVAR",K,' ');
 if find(VARIABLESINMODEL,strip(CVAR1[K]),'I')>0 then do;
 CLASSVINMODEL=strip(CLASSVINMODEL)!!""!!strip(CVAR1[K])!!""""!!strip(CVAR2[K]);
 CVARCATE=compress(CVAR2[K],"ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz_");
 CVARNOSUM=CVARNOSUM+(input(CVARCATE,8.)); CVARCATE=' ';
end; drop K;

array NVAR1[&NOCLASSVNO] $32 _temporary_;
NVAROSUM=0;
do L=1 to dim(NVAR1);
 NVAR1[L]=scan("&NONCVAR",L,' '); if find(VARIABLESINMODEL,strip(NVAR1[L]),'I')>0 then do;
 NCVARINMODEL=strip(NCVARINMODEL)!!strip(NVAR1[L]);
 NCVARNOSUM+1; end; drop L;
call catx (" ", CNVARSINMODEL, NCVARINMODEL, CLASSVINMODEL);
NOINCOMPMODEL=SUM(CVARNOSUM,NCVARNOSUM);
label NUMBEROFVARIABLES='Number in original model'
 NOINCOMPMODEL='Number in complete model'; run;
%end;
%else %if &CLASSVAR=' ' and &NONCVAR=' ' %then %do;
data BS0;
 set A (keep=NUMBEROFVARIABLES VARIABLESINMODEL SCORECHISQ DIFF_SCORECHISQ BESTMODEL);
 length CLASSVINMODEL NCVARINMODEL CVARNOSUM $1024;
 CLASSVINMODEL=' '; NCVARINMODEL=' '; CVARNOSUM=0; NCVARNOSUM=0; NOINCOMPMODEL=NUMBEROFVARIABLES;
 label NUMBEROFVARIABLES='Number in original model'
 NOINCOMPMODEL='Number in complete model'; run;
%end;

/*Logistic regression procedure by using the complete grouped categorical predictors; output information criteria (AIC, SC) for the model, variable selection and coefficient estimation; identify OUTMODEL statement for saving Variable Selection and Coefficient Estimation*/
proc logistic data=&MODELDATA outmodel=BB1;
 model &OUTC (event="&EVENTNO")=&COVARIATE;
ods output ASSOCIATION=C_1 FITSTATISTICS=FITS1; run;
/*Output c statistic*/
data C_2 (keep=LABEL2 NVALUE2 rename=(NVALUE2=HARRELL_C));
 set C_1 (keep=LABEL2 NVALUE2 rename=(NVALUE2=HARRELL_C)) where=(LABEL2='c'); run;
data FITS2 (keep=AIC_COVAR SC_COVAR);
 set FITS1 end=EOF;
 retain AIC_COVAR SC_COVAR;
 format AIC_COVAR SC_COVAR 10.4;
 if _N_=1 then do;
 AIC_COVAR=; SC_COVAR=. end;
 if CRITERION='AIC' then AIC_COVAR=INTERCEPTANDCOVARIATES;
 if CRITERION='SC' then SC_COVAR=INTERCEPTANDCOVARIATES;
 if EOF; run;

/*For each one the new models, calculate its discrimination back in the original/full data set; output c statistic [per Variable Selection]*/
proc logistic data=&ORIGDATA;
 model &OUTC (event="&EVENTNO")=&COVARIATE;
 ods output ASSOCIATION=C_ORIG_BB (keep=LABEL2 NVALUE2 rename=(NVALUE2=HARRELL_C_VARS) where=(LABEL2='c')); run;
/*For each one the new models, calculate its discrimination back in the original/full data set; output c statistic [per Variable Selection and Coefficient Estimation]*/
proc logistic inmodel=BB1;
score data=&ORIGDATA out=ORIG_BBSCORE; run;

proc logistic data=ORIG_BBSCORE;
model &OUTC (event="&EVENTNO")=P_1;
ods output ASSOCIATION=C_ORIG_BBS (keep=LABEL2 NVALUE2 rename=(NVALUE2=HARRELL_C_EST) where=(LABEL2='c')); run;

/*For each one the new models, calculate its discrimination in the left data set, then output c statistic [per Variable Selection]*/
proc logistic data=&LEFTDATA;
model &OUTC (event="&EVENTNO")=&COVariate;
ods output ASSOCIATION=C_LEFT_BB (keep=LABEL2 NVALUE2 rename=(NVALUE2=C_LEFT_BB) where=(LABEL2='c')); run;

/*For each one the new models, calculate its discrimination back in the left data set; output c statistic [per Variable Selection and Coefficient Estimation]*/
proc logistic inmodel=BB1;
score data=&LEFTDATA out=LEFT_BBSCORE; run;
proc logistic data=LEFT_BBSCORE;
model &OUTC (event="&EVENTNO")=P_1;
ods output ASSOCIATION=C_LEFT_BBS (keep=LABEL2 NVALUE2 rename=(NVALUE2=C_LEFT_BBS) where=(LABEL2='c')); run;

/*Merge data of calculated discrimination*/
data MODELTAB1;
merge C_2 C_ORIG_BB (keep=HARRELL_C_VARS)
  C_ORIG_BBS (keep=HARRELL_C_EST)
  C_LEFT_BB (keep=C_LEFT_BB)
  C_LEFT_BBS (keep=C_LEFT_BBS)
    FITS2; run;

data MODELTAB2;
length VAR_MODEL $1024;
set MODELTAB1;
VAR_MODEL="&COVariate";
if _n_=1 then set BS0 (keep=NUMBEROFVARIABLES VARIABLESINMODEL NOINCOMPMODEL SCORECHISQ DIFF_SCORECHISQ BESTMODEL); run;
proc append base=MODELTAB3 data=MODELTAB2 force; run; %end;
%let RC=%SYSFUNC(CLOSE(&DSID));

/*Identify the best AIC Model*/
data MODELTAB4 (DROP=VAR_MODELLAG); set MODELTAB3;
by VAR_MODELLAG ascending BESTMODEL; length VAR_MODELLAG $1024;
VAR_MODELLAG=lag(VAR_MODELLAG);
if _n_=1 then DUPMODEL=0;
else if VAR_MODELLAG=VAR_MODELLAG then DUPMODEL=1; else DUPMODEL=0;
label NUMBERINMODEL='Number of variables in original model'
  VARIABLESINMODEL='Variables in original model'
  NOINCOMPMODEL='Number of variables in complete model'
  VAR_MODEL='Variables in complete model'
  AIC_COVAR='AIC with covariates in complete model'
  SC_COVAR='SC with covariates in complete model'
  DUPMODEL='Duplicated complete model 0:No 1:Duplicated';
if DUPMODEL=0; proc sort; by AIC_COVAR; run;
data AICORDER; set MODELTAB4; by AIC_COVAR;
if _n_=1 then AIC_BEST=1; else AIC_BEST=0; proc sort; by SC_COVAR; run;

/*Identify the best BIC (SC) Model*/
data &OUTDATA; set AICORDER; by SC_COVAR; if _n_=1 then SC_BEST=1; else SC_BEST=0;
Estimating Harrell’s Optimism on Predictive Indices Using Bootstrap Samples, continued

```
proc sort; by AIC_COVAR; run;

proc datasets lib=work details;
delete A BS0 C_1 C_2 FITS1 FITS2 C_ORIG_BB ORIG_BBBSCORE C_ORIG_BBS C_LEFT_BB
LEFT_BBBSCORE C_LEFT_BBS MODELTAB1-MODELTAB4 &INDATA INDATA0 INDATA1 INDATA2 INDATA3
(memtype=data); run;
%MEND BESTSUBSET;
```

**EXAMPLE – TEN YEAR MORTALITY IN HEALTH AND RETIREMENT STUDY**

**Data Description**

We use a subset of data collected from the Health and Retirement Survey (HRS) to illustrate the use of best subsets models and our “Harrell Optimism” macro. HRS is a representative sample of all persons in the contiguous United States aged 50 years and above, and data are collected primarily through phone interviews with a response rate of 81%. Participants who were enrolled in 1998 were eligible and their information was cross-referenced with the National Center for Health Statistics National Death Index to determine vital status. Exclusion criteria were nursing home residents and indeterminate vital status.

Briefly, the cohort data consisted of 19710 community-dwelling participants, and we use a random subsample of 1000 participants to demonstrate the macro. Subjects’ functional status was defined as ability to perform five Activities of Daily Living (ADL) and it was measured at initial admission to hospital and at discharge from ICU. We demonstrate using logistic regression to model a binary indicator of ten-year mortality with 13% dying during that time frame. We looked at 23 potential risk factors from the domains of sociodemographics, functional measures, and comorbidities and behaviors. The demonstration below is based on the actual, more detailed model selection process that was used by our research group to develop a predictive index for ten-year mortality [19].

**Macro Call**

For our example, the macro will be called as follows:

```
%Harrell_Optimism_Logistic
  (ORIGDATA=samplehrs,
   SEED=12345, REPS=100, ID=ID,
   OUTC=findead, EVENTNO=1,
   ORIG_BESTNO=5, BOOT_BESTNO=5,
   FULLMODEL=AGECAT1 AGECAT2 AGECAT3 AGECAT4 AGECAT5
       AGECAT6 RACEETH1 EDUCATION MALE SMOKE EAT
       BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY
       STROKE DEMENTIA INCONT WALKROOM,
   CLASSVAR=AGECAT6,
   NONCVAR=RACEETH1 EDUCATION MALE SMOKE EAT
       BMI HYPERTEN DIABETES CANCER CHF LUNG ARTERY
       STROKE DEMENTIA INCONT WALKROOM,
   STARTNO=1, STOPNO=23, INCLUDENO=0,
   SLENTRY=0.2, SLSTAY=0.05);
```

The CLASSVAR option instructs the macro to keep all or none of the 6 age category dummy variables. The final model selected by the best AIC in our macro (see Table 1) includes the following 12 variables:

```
AGECAT3 AGECAT5 AGECAT6 RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM
```

which is then completed by adding AGECAT1, AGECAT2, and AGECAT4 back into the model as AGECAT was specified as a CLASSVAR to give a 15 predictor model with an apparent c-statistic of 0.848. The more parsimonious best completed BIC model has 4 fewer predictors (dropping RACEETH1, SMOKE, EAT, and CHF) with an apparent c-statistic of 0.829.

**Best Subsets models on Bootstrap datasets**

Once the final model is selected and the corresponding c-statistic is calculated, we estimate the optimism associated with the predictive index. The first step in the optimism algorithm is to generate 100 datasets using bootstrapping procedures. In each one of those datasets, we find new ‘best’ models by the various criteria. After fitting each model, we correct for missing levels of categorical variables if necessary. Finally, we calculate the c-statistic for both: (i) the bootstrap sample and (ii) back in the original sample.
Table 1. Best Models Generated in the Original/Full Data Set by Best Subsets Procedure

<table>
<thead>
<tr>
<th>Number of Variables in Original Model</th>
<th>Variables in Original Model</th>
<th>Number of Variables in Complete Model</th>
<th>Variables in Complete Model</th>
<th>AIC with Covariates in Complete Model</th>
<th>SC with Covariates in Complete Model</th>
<th>Harrell’s c Statistic</th>
<th>Score Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>15</td>
<td>RACEETH1 MALE SMOKE EAT DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>548.5637 [Best AIC Model]</td>
<td>626.9754</td>
<td>0.848265</td>
<td>219.7859</td>
</tr>
<tr>
<td>13</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>16</td>
<td>RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>549.5971</td>
<td>632.9095</td>
<td>0.847923</td>
<td>221.6129</td>
</tr>
<tr>
<td>14</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>16</td>
<td>RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>549.6694</td>
<td>632.9018</td>
<td>0.847757</td>
<td>222.7761</td>
</tr>
<tr>
<td>11</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>14</td>
<td>MALE SMOKE DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>549.9432</td>
<td>623.4542</td>
<td>0.843986</td>
<td>217.4989</td>
</tr>
<tr>
<td>15</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>17</td>
<td>RACEETH1 MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.4291</td>
<td>638.6423</td>
<td>0.850518 [Best Harrell’s c]</td>
<td>223.6737</td>
</tr>
<tr>
<td>11</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>14</td>
<td>RACEETH1 MALE SMOKE DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.6212</td>
<td>624.1774</td>
<td>0.846005</td>
<td>218.9489</td>
</tr>
<tr>
<td>15</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>17</td>
<td>RACEETH1 EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.8061</td>
<td>639.0011</td>
<td>0.847096</td>
<td>223.7079</td>
</tr>
<tr>
<td>15</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>17</td>
<td>RACEETH1 MALE SMOKE EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>550.8468</td>
<td>639.0399</td>
<td>0.848321</td>
<td>224.0256</td>
</tr>
<tr>
<td>16</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>18</td>
<td>RACEETH1 EDUCATION MALE SMOKE DRESS EAT HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.5204</td>
<td>644.6152</td>
<td>0.849145</td>
<td>224.2317</td>
</tr>
<tr>
<td>16</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>18</td>
<td>RACEETH1 MALE SMOKE DRESS EAT BMI HYPERTEN DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.6417</td>
<td>644.7556</td>
<td>0.849305</td>
<td>224.7451</td>
</tr>
<tr>
<td>16</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>18</td>
<td>RACEETH1 EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.6501</td>
<td>644.7448</td>
<td>0.849034</td>
<td>224.3984</td>
</tr>
<tr>
<td>12</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>15</td>
<td>RACEETH1 MALE SMOKE EAT HYPERTEN DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.6758</td>
<td>630.1178</td>
<td>0.847998</td>
<td>219.9215</td>
</tr>
<tr>
<td>12</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>15</td>
<td>RACEETH1 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.7328</td>
<td>630.1327</td>
<td>0.846240</td>
<td>220.7631</td>
</tr>
<tr>
<td>10</td>
<td>AGECAT3 AGECAT5 AGECAT6</td>
<td>13</td>
<td>MALE SMOKE DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
<td>551.8301</td>
<td>620.4825</td>
<td>0.841289</td>
<td>216.6517</td>
</tr>
</tbody>
</table>
Estimating Harrell’s Optimism on Predictive Indices Using Bootstrap Samples, continued

<table>
<thead>
<tr>
<th>16</th>
<th>AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1</th>
<th>EDUCATION MALE SMOKE EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>RACETH1 EDUCATION MALE SMOKE EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM AGECAT1 AGECAT6</td>
<td></td>
</tr>
<tr>
<td>552.9752</td>
<td>645.0499</td>
<td>0.847745</td>
</tr>
</tbody>
</table>

| 13 | AGECAT3 AGECAT4 AGECAT5 raceeth1 MALE SMOKE DRESS EAT BMI DIABETES CANCER LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 16 | RACETH1 MALE SMOKE DRESS EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 552.1751 | 635.5388 | 0.848200 | 221,5102 |

| 17 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG INCONT WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 19 | RACETH1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG INCONT WALKROOM AGECAT1 AGECAT6 |
| 552.6967 | 650.6710 | 0.850258 | 224,7541 |

| 17 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 19 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM AGECAT1 AGECAT6 |
| 552.7690 | 650.7835 | 0.848772 | 225,6325 |

| 14 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 16 | RACETH1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 552.9063 | 636.2700 | 0.844406 | 223,1476 |

| 14 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 16 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 552.9353 | 636.2820 | 0.846076 | 222,8307 |

| 11 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 14 | MALE SMOKE EAT BMI DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 553.2233 | 626.7795 | 0.842387 | 218,2163 |

| 15 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 17 | RACETH1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 553.3024 | 641.5698 | 0.847663 | 223,9518 |

| 9  | AGECAT3 AGECAT4 AGECAT5 AGECAT6 MALE SMOKE DIABETES CANCER LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 12 | MALE SMOKE DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 553.3385 | 617.2003 | 0.838857 | 210,0891 |

| 17 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG DEMENTIA WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 19 | RACETH1 MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG DEMENTIA WALKROOM AGECAT1 AGECAT6 |
| 553.6636 | 651.6509 | 0.849565 | 224,8665 |

| 18 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG INCONT WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 20 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG INCONT WALKROOM AGECAT1 AGECAT6 |
| 553.9194 | 656.7712 | 0.849305 | 225,4347 |

| 15 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 17 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 554.0322 | 642.2816 | 0.846808 | 223,7465 |

| 18 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM AGECAT1 AGECAT6 |
|----|------------------------------------------|---------------------------------------------------------------------|
| 20 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM AGECAT1 AGECAT6 |
| 554.1450 | 657.0270 | 0.847539 | 225,4893 |

| 18 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG ARTERY WALKROOM |
|----|------------------------------------------|---------------------------------------------------------------------|
| 20 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG ARTERY WALKROOM AGECAT1 AGECAT6 |
| 554.3911 | 657.2641 | 0.848827 | 225,4327 |

| 16 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
|----|------------------------------------------|---------------------------------------------------------------------|
| 18 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 554.4426 | 647.5547 | 0.848188 | 224,6044 |

| 18 | AGECAT3 AGECAT4 AGECAT5 AGECAT6 raceeth1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM AGECAT1 AGECAT6 |
|----|------------------------------------------|---------------------------------------------------------------------|
| 20 | RACETH1 EDUCATION MALE SMOKE DRESS EAT BMI HYPERP DIABETES CANCER CHF LUNG WALKROOM AGECAT1 AGECAT6 |
| 554.7838 | 657.6780 | 0.848858 | 225,5425 |

| 10 | AGECAT3 AGECAT4 AGECAT6 raceeth1 MALE EAT DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
|----|------------------------------------------|---------------------------------------------------------------------|
| 13 | RACETH1 MALE EAT DIABETES CANCER LUNG WALKROOM AGECAT1 AGECAT6 |
| 554.8705 | 623.5230 | 0.837170 | 215,9501 |

| 11 | AGECAT3 AGECAT4 AGECAT6 raceeth1 MALE EAT BMI |
|----|------------------------------------------|---------------------------------------------------------------------|
| 14 | RACETH1 MALE EAT BMI DIABETES CANCER LUNG |
| 555.2910 | 628.8472 | 0.837751 | 218,0153 |
Estimating Harrell’s Optimism on Predictive Indices Using Bootstrap Samples, continued

<table>
<thead>
<tr>
<th>AGECAT1</th>
<th>AGECAT2</th>
<th>AGECAT3</th>
<th>AGECAT4</th>
<th>AGECAT5</th>
<th>AGECAT6</th>
<th>AGEETH1</th>
<th>MALE</th>
<th>EA</th>
<th>INCONT</th>
<th>DIABETES</th>
<th>CANCER</th>
<th>LUNG</th>
<th>WALKROOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGECAT1</td>
<td>AGECAT2</td>
<td>AGECAT3</td>
<td>AGECAT4</td>
<td>AGECAT5</td>
<td>AGECAT6</td>
<td>AGECAT6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGECAT1</td>
<td>AGECAT2</td>
<td>AGECAT3</td>
<td>AGECAT4</td>
<td>AGECAT5</td>
<td>AGECAT6</td>
<td>AGECAT6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGECAT1</td>
<td>AGECAT2</td>
<td>AGECAT3</td>
<td>AGECAT4</td>
<td>AGECAT5</td>
<td>AGECAT6</td>
<td>AGECAT6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Best Model</th>
<th>[Best BIC Model]</th>
</tr>
</thead>
<tbody>
<tr>
<td>616.8285</td>
<td>0.828947</td>
</tr>
<tr>
<td>199.2249</td>
<td>0.821430</td>
</tr>
<tr>
<td>209.8783</td>
<td>0.823021</td>
</tr>
<tr>
<td>211.1422</td>
<td>0.823055</td>
</tr>
<tr>
<td>204.5870</td>
<td>0.823084</td>
</tr>
<tr>
<td></td>
<td>AGECAT5</td>
</tr>
<tr>
<td>----</td>
<td>---------</td>
</tr>
<tr>
<td>11</td>
<td>EAT DIABETES CANCER LUNG WALKROOM AGECAT1-AGECAT6</td>
</tr>
<tr>
<td>198.7040</td>
<td>[Best diminishing returns]</td>
</tr>
</tbody>
</table>
Average Optimism of the Models

The average optimism of the c-statistics is the average of the 100 difference values between the two c-statistics calculated above. The final output of the macro is shown in Table 2. In the example, the optimism due to variable selection from the diminishing returns method is 0.026 and optimism due to variable selection and coefficient estimation is 0.033. Therefore, the optimism-corrected value of Harrell’s c statistics per variable selection is 0.827-0.026 = 0.801; and the unbiased estimate of Harrell’s c statistics per variable selection and coefficient estimation is 0.827-0.033 = 0.794. The degree of optimism is virtually identical for the four selection methods.

Table 2. Final Harrell Optimism macro output

<table>
<thead>
<tr>
<th>Method</th>
<th>Apparent c-Statistic</th>
<th>Optimism from Selection Only</th>
<th>Optimism from Selection and Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference in Score</td>
<td>0.827</td>
<td>0.026</td>
<td>0.033</td>
</tr>
<tr>
<td>Best AIC</td>
<td>0.848</td>
<td>0.023</td>
<td>0.035</td>
</tr>
<tr>
<td>Best SC (BIC)</td>
<td>0.829</td>
<td>0.027</td>
<td>0.034</td>
</tr>
<tr>
<td>Stepwise</td>
<td>0.839</td>
<td>0.025</td>
<td>0.033</td>
</tr>
</tbody>
</table>

CONCLUSION

In this paper we present a macro for calculating Harrell’s bootstrap optimism in the development of a predictive model. The paper focuses on the optimism of the c-statistic of a model selected by Logistic regression using best subsets selection. The full version of the macro can also estimate the optimism of the c-statistic for stepwise regression, as well as optimism of the optimism of the Harrell’s c-statistic for Cox regression models. We allow for a variant of best subset selection that augments the selected model to include any pieces of grouped categorical predictors that were not chosen by the algorithm. Additionally, our macro can implement both standard bootstrapping and the .632 bootstrap method. The macro can also estimate what portion of the total optimism is due to variable selection and what portion is due to coefficient estimation by both scoring and refitting the coefficients for the model in the validation set. The application demonstrated here (which is consistent with our broader experience) is that the degree of optimism is nearly identical for stepwise selection and for the three best subset selection methods.

REFERENCES


**CONTACT INFORMATION**

We are grateful for support from a number of grant sources including the UCSF Pepper Center Statistical Analysis Core (NIH/NIA P30 AG044281), a Methodology grant from the UCSF CTSI (NIH/NCATS UL1 RR024130), and two NIH/NIA research project grants (R01 AG029233, PI Landefeld; R01 AG028481, PI Covinsky).

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

John Boscardin
University of California, San Francisco
4150 Clement St., Mailstop 151R
San Francisco, CA 94121
E-mail: john.boscardin@ucsf.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.