Paper 504-2013

Estimating Harrell's Optimism on Predictive Indices Using Bootstrap Samples

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

In aging research, it is important to develop and validate accurate prognostic models whose predictive accuracy will not degrade when applied in external data sources. While the most common method of validation is split sample, alternative methods such as cross-validation and bootstrapping have some significant advantages. The macro that we present calculates Harrell's optimism for logistic and Cox regression models based on either the c-statistic (for logistic) or Harrell's c (for Cox). It allows for both stepwise and best subset variable selection methods (and can handle class variables for both methods), and uses both traditional and .632 bootstrapping methods. In addition, we present and discuss the advantages of using Best Subsets regression for model selection instead of stepwise procedures. The uses of Best Subsets regression and our Harrell_Optimism macro are demonstrated using data on post-hospitalization mortality.

INTRODUCTION

A common goal in medical research is to develop accurate prognostic models for health-related outcomes. Validation of the prognostic model is critically important to guarantee that its predictive accuracy will not degrade when applied in external data sources. By far the most common approach in the medical literature is split-sample validation; the model is developed in just one portion of the data and then validated in the remaining portion. Any discrepancy between the predictive accuracy in the development and validation sets is regarded as evidence of over-fitting or optimism. The statistical literature is in strong agreement that split-sample assessment of model optimism is extremely inefficient for two reasons: (i) there is a substantial loss of estimation precision from developing the model in a portion of the data, and (ii) unless sample sizes are extremely large, very little can actually be learned about the model optimism from a single split-sample [1].

Alternative approaches to assessing model optimism that make full use of the data include cross-validation and bootstrapping [2]. These methods have much in common in that they (i) replicate the development and validation cycle many times and (ii) use a full or nearly full version of the dataset for each cycle [3]. One difficulty in routinely implementing either cross-validation or bootstrapping has been a lack of user-friendly software to implement the computationally intensive calculations. Two notable exceptions are the validate.lrm function in the rms package for R [4] and a SAS® macro published earlier this year [5], both of which allow only for stepwise model selection. Many algorithms in the literature do not distinguish between optimism due to estimating the model coefficients in the development sample (i.e. maximum likelihood estimation means that the model coefficients are optimally chosen for the development sample) and optimism due to over-fitting in the model building process (e.g. selection of predictors; categorization of continuous predictors; choices related to functional form for continuous predictors).

We are developing a series of SAS macros to evaluate the optimism of the model selection process. Currently, our macro supports both stepwise and best subsets variable selection. Both of these variable selection methods have been strongly criticized in the literature for their potential for over-fitting, and it is thus of great interest to quantify the degree of over-fitting. Best subsets has one important advantage; whereas stepwise selection only shows one model and does not output comparisons to other potential models, best subset regression gives a tremendous amount of useful information for comparing models [6]. In practice, we find that a large number of models of reasonable parsimony are statistically nearly indistinguishable. It is very valuable for clinicians to be able to assess a lot of similarly performing prognostic models, in order to choose the one that is most practically applied. For these reasons and for clarity of presentation, we restrict our demonstration in this paper to the best subset selection method.

HARRELL'S ALGORITHM FOR CALCULATING OPTIMISM

Harrell et al [1] presented an algorithm for estimating the optimism, or over-fitting, in predictive models. Their method is based on using bootstrapped datasets to repeatedly quantify the degree of over-fitting in the model building process. Two common specific settings are: (a) logistic regression, with measure of discrimination the area under the ROC curve (the c-statistic) [3,6-8]; and (b) Cox proportional hazards regression, with measure of discrimination Harrell's c statistic [9-14]. The steps for estimating the optimism, as suggested by Harrell et al., are as follows:

1. Select the predictors and fit a model using the full dataset and a particular variable selection method. From that model, calculate the apparent discrimination (c_{app}).
2. Generate M=100 to 200 datasets of the same sample size (n) using bootstrap samples with replacement.
3. For each one of the new datasets m=1,…,M, select predictors and fit the model using the exact same algorithmic approach as in step 1 and calculate the discrimination ($c_{boot}^{(m)}$).
4. For each one of the new models, calculate its discrimination back in the original data set ($c_{orig}^{(m)}$). For this step, the regression coefficients can either be fixed to their values from step 3 to determine the joint degree of overfitting from both selection and estimation or can be reestimated to determine the degree of overfitting from selection only.
5. For each one of the bootstrap samples, the optimism in the fit is $o^{(m)} = c_{orig}^{(m)} - c_{boot}^{(m)}$. The average of these values is the optimism of the original model:

$$o = \frac{\sum_{m=1}^{M} o^{(m)}}{M}$$

6. The optimism corrected performance of the original model is then $c_{adj} = c_{app} - o$. This value is a nearly unbiased estimate of the expected values of the optimism that would be obtained in external validation.

A variant on this method, known as .632 bootstrapping [3,15], changes step 4 to calculating a weighted average of the discrimination in the original dataset and the discrimination in the observations that were not included in the m^{th} bootstrapped sample. This method is more similar to the cross-validation idea of validating the model in repeated external samples.

THE "BEST" MODEL FOR BEST SUBSET SELECTION

Although a major advantage of best subset selection is that it highlights a range of candidate models, for the purposes of evaluating optimism we need to mimic the process of a researcher choosing a single model. Our process is to first choose an optimal model size followed by an optional step to supplement the chosen model with left-out dummy variables from grouped categorical predictors. There are a number of ways to determine what size model we should choose, including AIC and BIC values [16-17] ; we are currently using a simplistic test of the point of diminishing returns in growing the model. We look at the best model for each number of possible predictors and find the point at which the difference in score statistics (approximating the likelihood ratio test) between the best models of size k and size k+1 is no longer statistically significant at a level of 0.05. We are thus acting as if the two models are nested, which is not always the case. Once this model is identified, we then (optionally) check to see if any of its predictors are incomplete grouped categorical predictors and then put the missing dummy variables into the final model (e.g. if two of three non-reference levels for age group were chosen we would automatically include the third level). This simple step reflects common practice in using best subsets regression. Furthermore, it gives a method of implementing class variables in the selection that, to our knowledge, is not directly available in any software implementation of best subsets regression.

OUR MACRO

Our full macro calculates the optimism of predictive indices and has the following features:

1. Includes methods for both logistic regression and Cox regression.
2. Performs either best subsets regression or stepwise selection, and can handle grouping of categorical variables for both methods.
3. Uses either regular bootstrapping methods or .632 bootstrapping.
4. Differentiates between optimism due to variable selection only and optimism due to variable selection and coefficient estimation.
5. Outputs the list of variables in the full model and the percent of time they were selected in the final model.

In this paper we present only a shortened version of the macro. It performs best subset Cox regression and the standard bootstrap method. The shortened version and output of the macro performing best subsets logistic regression was published in the WUSS 2012 proceedings [18]. The full macro can be obtained by contacting the authors.

DEFINING THE MACRO CALL

```
%Macro Harrell_Optimism_Cox
   (ORIGDAT=,
    SEED=,
    REPS=,
    ID=,
    EVENT=,
    FULLMODEL=,
    CLASSVAR=,
    NONCVAR=,
    START=,
    TIME=,
    TIES=);
```

The macro HARRELL_OPTIMISM_COX is specified in the following manner:

- ORIGDAT – 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 (patient) identifier.
- EVENT – the indicator for experiencing the event (EVENT=1) vs. censored (EVENT=0).
- FULLMODEL – a list of all potential predictor variables.
- 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.
- START/TIME – the entry time and event/censoring time variable.
- TIES – specifies the method of handling ties in failure times.

THE MACRO

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 [11] to calculate Harrell's c.

Figure 1. The shortened version of the Harrell_Optimism_Cox macro

```
%MACRO Harrell_Optimism_Cox (ORIGDAT=, SEED=, REPS=, ID=, EVENT=, FULLMODEL=, CLASSVAR=,
                             NONCVAR=, START=, TIME=, TIES=);

proc datasets lib=work details;
delete C_STAT_LIST (memtype=data);
run;

proc surveyselect data=&ORIGDAT out=BOOT
     seed=&SEED method=URS samprate=1 outhits rep=&REPS;
run;

/*ORIGINAL DATASET - FULL MODEL*/
proc phreg data=&ORIGDAT;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&FULLMODEL / ties=&TIES;
output out=ORIGFULL xbeta=SCORE_OF;
```

```
RUN;
%SURVCSTD(DATA=ORIGFULL, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=SCORE_OF, PRINT=0,
          ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_OF);
/*ORGINAL DATASET - BEST MODEL*/
proc phreg data=&ORIGDAT;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&FULLMODEL / ties=&TIES selection=SCORE best=1;
ods output BESTSUBSETS=BSORIG1;
run;

%if &CLASSVAR^=' ' and &NONCVAR^=' ' %then %do;
data _NULL_;
CLASSVNUMBER=countw("&CLASSVAR");
NONCVNUMBER=countw("&NONCVAR");
call symputx ('CLASSVARNO',CLASSVNUMBER);
call symputx ('NOCLASSVNO',NONCVNUMBER);
CVAR=compress("&CLASSVAR","0123456789");
call symputx ('CLASSVAR2',CVAR);
run;

data BSORIG1B; set BSORIG1;
length CLASSVINMODEL NCVARINMODEL CNVARSINMODEL $256;
NCVARINMODEL=' ';
array CVAR1[&CLASSVARNO] $32 _temporary_;
array CVAR2[&CLASSVARNO] $32 _temporary_;
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 CLASSVINMODEL=strip(CLASSVINMODEL)!!" "!!strip(CVAR1[K])!!"1"!!"-"!!strip(CVAR2[K]);
end; drop K;
array NVAR1[&NOCLASSVNO] $32 _temporary_;
do L=1 to dim(NVAR1);
  NVAR1[L]=scan("&NONCVAR",L,' ');
  if find(VARIABLESINMODEL,strip(NVAR1[L]),'I')>0
    then NCVARINMODEL=strip(NCVARINMODEL)!!" "!!(strip(NVAR1[L])); end; drop L;
call catx (" ", CNVARSINMODEL, NCVARINMODEL, CLASSVINMODEL);
proc sort; by SCORECHISQ;
run; %end;
%else %if &CLASSVAR=' ' and &NONCVAR=' ' %then %do;
  data BSORIG1B; set BSORIG1;
  CNVARSINMODEL=VARIABLESINMODEL;
  proc sort; by SCORECHISQ;
  run; %end;

data BSORIG2; set BSORIG1B; by SCORECHISQ;
DIFF_SCORECHISQ=DIF(SCORECHISQ);
if DIFF_SCORECHISQ^=. and DIFF_SCORECHISQ<3.841459 then delete;
run;
data BSORIG3; set BSORIG2 end=last;
if last then output;
run;
data _NULL_; set BSORIG3 (keep=CNVARSINMODEL);
call symputx ('BESTORIGM', CNVARSINMODEL);
run;
proc phreg data=&ORIGDAT;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&BESTORIGM / ties=&TIES;
output out=ORIGBEST xbeta=SCORE_OB;
run;
%SURVCSTD(DATA=ORIGBEST, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=SCORE_OB, PRINT=0,
                 ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_OB);
/*REPLICATED BOOTSTRAPPING DATASETS*/
%do M=1 %to &REPS;
/*CREATE LEFT-OUT DTASET FOR .632-VALIDATION*/
data BOOT_A; set BOOT (keep=&ID REPLICATE);
where REPLICATE=&M;
proc sort; by &ID; run;
data BOOT_B; set BOOT_A; by &ID;
```

```
if first.&ID;
proc sort; by &ID; run;
proc sort data=&ORIGDAT; by &ID; run;
data LEFT;
merge &ORIGDAT (in=A) BOOT_B (keep=&ID in=B);
by &ID; if A and not B;
proc sort; by &ID; run;
/*BOOT DATASET - FULL MODEL*/
proc phreg data=BOOT outest=BF1;
where REPLICATE=&M;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&FULLMODEL / ties=&TIES;
output out=BOOTFULL xbeta=SCORE_BF;
run;
%SURVCSTD(DATA=BOOTFULL, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=SCORE_BF, PRINT=0,
                ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_BF);
/*LEFT DATASET - FULL MODEL*/
proc phreg data=LEFT outest=LF1;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&FULLMODEL / ties=&TIES;
output out=LEFTFULL xbeta=SCORE_LF;
RUN;
%SURVCSTD(DATA=LEFTFULL, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=SCORE_LF, PRINT=0,
                ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_LF);
/*BOOT DATASET - BEST MODEL*/
proc phreg data=BOOT;
where REPLICATE=&M;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&FULLMODEL / ties=&TIES selection=SCORE best=1;
ods output BESTSUBSETS=BS1;
run;

%if &CLASSVAR^=' ' and &NONCVAR^=' ' %then %do;
data BS1B; set BS1;
length CLASSVINMODEL NCVARINMODEL CNVARSINMODEL $256;
NCVARINMODEL=' ';
array CVAR1[&CLASSVARNO] $32 _temporary_;
array CVAR2[&CLASSVARNO] $32 _temporary_;
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 CLASSVINMODEL=strip(CLASSVINMODEL)!!" "!!strip(CVAR1[K])!!"1"!!"-"!!strip(CVAR2[K]); end; drop K;
array NVAR1[&NOCLASSVNO] $32 _temporary_;
do L=1 to dim(NVAR1);
  NVAR1[L]=scan("&NONCVAR",L,' ');
  if find(VARIABLESINMODEL,strip(NVAR1[L]),'I')>0
    then NCVARINMODEL=strip(NCVARINMODEL)!!" "!!(strip(NVAR1[L])); end;
drop L;
call catx (" ", CNVARSINMODEL, NCVARINMODEL, CLASSVINMODEL);
proc sort; by SCORECHISQ;
run; %end;
%else %if &CLASSVAR=' ' and &NONCVAR=' ' %then %do;
  data BS1B; set BS1;
  CNVARSINMODEL=VARIABLESINMODEL;
  proc sort; by SCORECHISQ;
  run; %end;

data BS2; set BS1B; by SCORECHISQ;
DIFF_SCORECHISQ=DIF(SCORECHISQ);
if DIFF_SCORECHISQ^=. and DIFF_SCORECHISQ<3.841459 then delete; run;
data BS3; set BS2 end=last;
if last then output; run;

data _NULL_; set BS3 (keep=NUMBERINMODEL CNVARSINMODEL DIFF_SCORECHISQ);
call symput ('DIF_CHISQ',put(DIFF_SCORECHISQ,8.6));
call symput ('BESTMODNO',put(NUMBERINMODEL,4.0));
call symputx ('BESTMODEL',CNVARSINMODEL);
run;
```

```
proc phreg data=BOOT outest=BB1;
where REPLICATE=&M;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&BESTMODEL / ties=&TIES;
output out=BOOTBEST xbeta=SCORE_BB;
run;
%SURVCSTD(DATA=BOOTBEST, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=SCORE_BB, PRINT=0,
                ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_BB);
/*ORIGINAL DATASET - BOOT-BEST MODEL*/
proc phreg data=&ORIGDAT;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&BESTMODEL / ties=&TIES;
output out=ORIG_BOOTBEST xbeta=SCORE_ORIG_BB;
run;
%SURVCSTD(DATA=ORIG_BOOTBEST, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT,
              SCORE=SCORE_ORIG_BB, PRINT=0, ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_ORIG_BB);
/*LEFT DATASET - BOOT-BEST MODEL*/
proc phreg data=LEFT;
ID &ID;
model (&START,&TIME)*&EVENT(0)=&BESTMODEL / ties=&TIES;
output out=LEFT_BOOTBEST xbeta=SCORE_LEFT_BB;
run;
%SURVCSTD(DATA=LEFT_BOOTBEST, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT,
          SCORE=SCORE_LEFT_BB, PRINT=0, ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_LEFT_BB);
/*ORIGINAL DATASET - SCORE BY BETA FROM BOOT-FULL MODEL*/
proc score data=&ORIGDAT score=BF1 out=ORIG_BF_SCORE type=PARMS;
var &FULLMODEL;
run;
%SURVCSTD(DATA=ORIG_BF_SCORE, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=&TIME.2,
              PRINT=0, ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_ORIG_BFSCORE);
/*ORIGINAL DATASET - SCORE BY BETA FROM BOOT-BEST MODEL*/
proc score data=&ORIGDAT score=BB1 out=ORIG_BB_SCORE type=PARMS;
var &BESTMODEL;
run;
%SURVCSTD(DATA=ORIG_BB_SCORE, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=&TIME.2,
                PRINT=0, ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_ORIG_BBSCORE);
/*LEFT DATASET - SCORE BY BETA FROM BOOT-FULL MODEL*/
proc score data=LEFT score=BF1 out=LEFT_BF_SCORE type=PARMS;
var &FULLMODEL;
run;
%SURVCSTD(DATA=LEFT_BF_SCORE, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=&TIME.2,
                PRINT=0, ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_LEFT_BFSCORE);
/*LEFT DATASET - SCORE BY BETA FROM BOOT-BEST MODEL*/
proc score data=LEFT score=BB1 out=LEFT_BB_SCORE type=PARMS;
var &BESTMODEL;
run;
%SURVCSTD(DATA=LEFT_BB_SCORE, ID=&ID, START=&START, TIME=&TIME, EVENT=&EVENT, SCORE=&TIME.2,
                PRINT=0, ALPHA=0.05, SEMETHOD=I, NADJ=0, OUT=SURVCS_LEFT_BBSCORE);
/*MERGE RESULTS - C-STATISTICS (RELATED TO BOOTSTRAPPING DATASETS)*/
data C_STAT (keep=RESAMPLE V_RETAIN BESTNO C_BB C_BF C_ORIG_BB C_ORIG_BFS C_ORIG_BBS
                  OPTIMISM_BB_OBB OPTIMISM_BF_OBFS OPTIMISM_BB_OBBS
                  C_LF C_LEFT_BB C_LEFT_BFS C_LEFT_BBS);
length RESAMPLE $8 V_RETAIN $256 BESTNO $8;
merge SURVCS_BF (keep=CN rename=(CN=C_BF))
      SURVCS_LF (keep=CN rename=(CN=C_LF))
      SURVCS_BB (keep=CN rename=(CN=C_BB))
      SURVCS_ORIG_BB (keep=CN rename=(CN=C_ORIG_BB))
      SURVCS_LEFT_BB (keep=CN rename=(CN=C_LEFT_BB))
      SURVCS_ORIG_BFSCORE (keep=CN rename=(CN=C_ORIG_BFS))
      SURVCS_ORIG_BBSCORE (keep=CN rename=(CN=C_ORIG_BBS))
      SURVCS_LEFT_BFSCORE (keep=CN rename=(CN=C_LEFT_BFS))
      SURVCS_LEFT_BBSCORE (keep=CN rename=(CN=C_LEFT_BBS));
RESAMPLE=&M;
V_RETAIN="&BESTMODEL";
BESTNO=strip("&BESTMODNO");
OPTIMISM_BB_OBB=C_BB-C_ORIG_BB;
OPTIMISM_BS_OBS=C_BS-C_ORIG_BS;
OPTIMISM_BF_OBFS=C_BF-C_ORIG_BFS;
```

```
OPTIMISM_BB_OBBS=C_BB-C_ORIG_BBS;
label RESAMPLE='Re-Sample'
 V_RETAIN='Variable in BestModel'
 BESTNO='Variable no. in Best Model'
 C_BF="Harrell's C-statistic: from full model fitted in the bootstrap dataset"
 C_BB="Harrell's c: from best subset model fitted in the bootstrap dataset"
 C_ORIG_BB="Harrell's c: from best subset model fitted in the bootstrap dataset, applied to original
dataset"
 C_ORIG_BFS="Harrell's C-statistic: score by Beta from full model fitted in the bootstrap dataset, applied
to original dataset"
 C_ORIG_BBS="Harrell's c: score by Beta from best subset model fitted in the bootstrap dataset, applied to
original dataset"
 OPTIMISM_BB_OBB="Harrell's c difference: BootBst-Orig_BootBest"
 OPTIMISM_BF_OBFS="Harrell's c difference: BootFull-Orig_BootFullScore"
 OPTIMISM_BB_OBBS="Harrell's c difference: BootBest-Orig_BootBestScore"
 C_LF="Harrell's C-statistic: from full model fitted in the dataset of .632 bootstrapping method"
 C_LEFT_BB="Harrell's C-statistic: from best subset model fitted in the bootstrap dataset, applied to
dataset of .632 bootstrapping method"
 C_LEFT_BFS="Harrell's C-statistic: score by Beta from full model fitted in the bootstrap dataset, applied
to dataset of .632 bootstrapping method"
 C_LEFT_BBS="Harrell's C-statistic: score by Beta from best subset model fitted in the bootstrap dataset,
applied to dataset of .632 bootstrapping method";
run;
proc append base=C_STAT_LIST data=C_STAT force;
run; %end;

proc means data=C_STAT_LIST mean;
var OPTIMISM_BB_OBB OPTIMISM_BF_OBFS OPTIMISM_BB_OBBS C_LF C_LEFT_BB C_LEFT_BFS C_LEFT_BBS;
output out=OPT_MEAN
  mean(OPTIMISM_BB_OBB OPTIMISM_BF_OBFS OPTIMISM_BB_OBBS C_LF C_LEFT_BB C_LEFT_BFS C_LEFT_BBS)
  =OPTIMISM_BB_OBBMEAN OPTIMISM_BF_OBFSMEAN OPTIMISM_BB_OBBSMEAN C_LFMEAN C_LEFT_BBMEAN C_LEFT_BFSMEAN
C_LEFT_BBSMEAN;
run;

data C_VALIDATION (keep=C_OF C_OB OPTIMISM_BB_OBBMEAN OPTIMISM_BF_OBFSMEAN OPTIMISM_BB_OBBSMEAN
                        C_VALD_OBB C_VALD_OBFS C_VALD_OBBS C_LFMEAN C_LEFT_BBMEAN C_LEFT_BFSMEAN
                        C_LEFT_BBSMEAN C_EST_FULL C_EST_BEST C_EST_BFS C_EST_BBS);
merge SURVCS_OF (keep=CN rename=(CN=C_OF))
      SURVCS_OB (keep=CN rename=(CN=C_OB))
      OPT_MEAN (keep=OPTIMISM_BB_OBBMEAN OPTIMISM_BF_OBFSMEAN OPTIMISM_BB_OBBSMEAN
                     C_LFMEAN C_LEFT_BBMEAN C_LEFT_BFSMEAN C_LEFT_BBSMEAN);
C_VALD_OBB=C_OB-OPTIMISM_BB_OBBMEAN;
C_VALD_OBFS=C_OF-OPTIMISM_BF_OBFSMEAN;
C_VALD_OBBS=C_OB-OPTIMISM_BB_OBBSMEAN;
C_EST_FULL=0.368*C_OF+0.632*C_LFMEAN;
C_EST_BEST=0.368*C_OB+0.632*C_LEFT_BBMEAN;
C_EST_BFS=0.368*C_OF+0.632*C_LEFT_BFSMEAN;
C_EST_BBS=0.368*C_OB+0.632*C_LEFT_BBSMEAN;
label C_OF="Apparent C-statistic [full model]"
C_OB="Apparent Harrell's C"
OPTIMISM_BB_OBBMEAN="Optimism from selection only"
OPTIMISM_BF_OBFSMEAN="Optimism from estimation only [full model]"
OPTIMISM_BB_OBBSMEAN="Optimism from selection and estimatin"
C_VALD_OBB="Harrell's C-statistic corrected for selection optimism"
C_VALD_OBFS="Harrell's C-statistic corrected for estimation optimism [full model]"
C_VALD_OBBS="Harrell's C-statistic corrected for selection and estimatin optimism"
C_LFMEAN="Harrell's C-statistic [mean]: from full model fitted in the dataset of .632 bootstrapping
method"
C_LEFT_BBMEAN="Harrell's C-statistic: from best subset model fitted in the bootstrap dataset applied to
dataset of .632 bootstrapping method"
C_LEFT_BFSMEAN="Harrell's C-statistic: score by Beta from full model fitted in the bootstrap dataset
applied to dataset of .632 bootstrappong method"
C_LEFT_BBSMEAN="Harrell's C-statistic [mean]: score by Beta from best subset model fitted in the bootstrap
dataset applied to dataset of .632 bootstrappong method"
C_EST_BEST="Harrell's C-statistic corrected for selection only [.632 bootstrap]"
C_EST_BFS="Harrell's C-statistic corrected for estimation only [.632 bootstrap]"
C_EST_BBS="Harrell's C-statistic corrected for selection and estimation [.632 bootstrap]";
run;
```

```
data _null_;
VNUMBER=countw("&FULLMODEL");
call symputx ('VARNO', VNUMBER);
run;
data BSVAR1; set C_STAT_LIST (keep=V_RETAIN);
array VAR1[&VARNO] $32 _TEMPORARY_;
array VAR2[&VARNO] &FULLMODEL;
do I=1 to dim(VAR1);
  VAR1[I]=scan("&FULLMODEL",I,' ');
  if find(V_RETAIN,strip(VAR1[I]),'I')>0 then VAR2[I]=1; else VAR2[I]=0; end; drop I; run;
proc summary data=BSVAR1; var &FULLMODEL; output out=BSVAR2 mean(&FULLMODEL)=&FULLMODEL; run;
proc transpose data=BSVAR2 (drop=_TYPE_ _FREQ_) out=BSVAR3 prefix=PERCENT; run;

data BSVAR4; set BSVAR3;
PERCENT=PERCENT1*100;
label PERCENT='Percent (%) [BESTSUBSET]';
proc sort; by descending PERCENT;
run;

%if &CLASSVAR=' ' and &NONCVAR=' ' %then %do;
  data BESTSUBSET_MODEL; set BSVAR4;
  by descending PERCENT;
  run; %end;
%else %if &CLASSVAR^=' ' and &NONCVAR^=' ' %then %do;
data BSVAR5; set BSVAR4;
length VARNAME $70;
array CVAR1[&CLASSVARNO] $32 _temporary_;
array CVAR2[&CLASSVARNO] $32 _temporary_;
do O=1 to dim(CVAR1);
  CVAR1[O]=scan("&CLASSVAR2",O,' ');
  CVAR2[O]=scan("&CLASSVAR",O,' ');
  if find(_NAME_,strip(CVAR1[O]),'I')>0
    then VARNAME=STRIP(CVAR1[O])!!"1"!!"-"!!STRIP(CVAR2[O]); end; drop O;
if VARNAME=' ' then VARNAME=_NAME_;
proc sort; by VARNAME descending PERCENT; run;
data BSVAR6; set BSVAR5 (keep=VARNAME PERCENT); by VARNAME; if first.VARNAME;
proc sort; by descending PERCENT; run;
data BESTSUBSET_MODEL; set BSVAR6; by descending PERCENT; run; %end;

%MEND Harrell_Optimism_Cox;
```

EXAMPLE – TEN YEAR MORTALITY IN HEALTH AND RETIREMENT STUDY

Data Description

We use data collected in 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. Our outcome was four-year mortality with 13% dying during that time frame.

We looked at 29 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 four-year mortality [19].

Macro Call

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

```
%Harrell_Optimism_Cox (ORIGDAT=TESTDAT,
                       SEED=130211,
                       ID=ID,
                       EVENT=findead,
                       START=START,
                       TIME=survtime4,
                       TIES=BRESLOW,
                       FULLMODEL= AGECAT1 AGECAT2 AGECAT3  AGECAT4 AGECAT5
                                  AGECAT6 RACEETH1 RACEETH2 RACEETH3
                                  EDUCATION MALE SMOKE DRESS EAT BMI
                                  HYPERTEN DIABETES CANCER CHF LUNG ARTERY
                                  STROKE DEMENTIA INCONT WALKROOM,
                       REPS=200,
                       CLASSVAR= AGECAT6 RACEETH3 ,
                       NONCVAR= EDUCATION MALE SMOKE DRESS EAT BMI HYPERTEN
                                DIABETES CANCER CHF LUNG ARTERY STROKE
                                DEMENTIA INCONT WALKROOM);
```

The CLASSVAR option instructs the macro to keep all or none of the 6 age category dummy variables and all or none of the 3 for race/ethnicity. The final model selected by the best subsets regression in our macro includes the following 19 variables:

```
EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG ARTERY STROKE DEMENTIA

INCONT WALKROOM AGECAT1 AGECAT2 AGECAT3 AGECAT4 AGECAT5 AGECAT6
```

The model has an apparent Harrell's C statistic of 0.8005.

Best Subsets models on Bootstrap datasets

Once the final model is selected and the corresponding Harrell's C statistic is calculated, we estimate the optimism associated with the predictive index. The first step in the optimism algorithm is to generate 200 datasets using bootstrapping procedures. In each one of those datasets, we fit a new best subsets model. After fitting each model, we correct for missing levels of categorical variables if necessary. Finally, we calculate the Harrell's C statistic for both: (i) the bootstrap sample and (ii) back in the original sample. A dataset generated by the macro contains all that information. The first 10 bootstrap samples for the example are shown in Table 1. We note that all the bootstrapped best models contain the categorical age predictor, but only some include the race/ethnicity predictor.

Re-Sample	Variable no. in Best Model	Variable in BestModel	Harrell's c: from best subset model fitted in the bootstrap dataset	Harrell's c: from best subset model fitted in the bootstrap dataset, applied to original dataset	Harrell's c-difference: BootBest-Orig_BootBest	Harrell's c: score by Beta from best subset model fitted in the bootstrap dataset, applied to original dataset	Harrell's c-difference: BootBest-Orig_BootBestScore
1	19	EDUCATION MALE DRESS EAT BMI DIABETES CANCER CHF LUNG ARTERY STROKE INCONT WALKROOM AGECAT1-AGECAT6 RACEETH1-RACEETH3	0.8133	0.7984	0.014945	0.7954	0.017909
2	15	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG WALKROOM AGECAT1-AGECAT6	0.8043	0.7989	0.005395	0.7973	0.006979
3	15	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG ARTERY DEMENTIA WALKROOM AGECAT1-AGECAT6	0.794	0.8	-0.005969	0.7991	-0.005024
4	17	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG ARTERY DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6	0.8087	0.7997	0.009014	0.7973	0.011398
5	19	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG ARTERY DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6 RACEETH1-RACEETH3	0.8058	0.8006	0.005203	0.7995	0.006355
6	17	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG STROKE DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6	0.7969	0.799	-0.002101	0.7985	-0.001602
7	19	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG ARTERY STROKE DEMENTIA WALKROOM AGECAT1-AGECAT6 RACEETH1-RACEETH3	0.803	0.8017	0.001385	0.8006	0.002432
8	17	EDUCATION MALE SMOKE DRESS EAT BMI DIABETES CANCER CHF LUNG ARTERY WALKROOM AGECAT1-AGECAT6	0.7956	0.8003	-0.004722	0.7996	-0.004029
9	16	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG ARTERY DEMENTIA WALKROOM AGECAT1-AGECAT6	0.7992	0.8	-0.000853	0.8	-0.000862
10	19	EDUCATION MALE SMOKE EAT BMI DIABETES CANCER CHF LUNG ARTERY STROKE DEMENTIA INCONT WALKROOM AGECAT1-AGECAT6	0.8131	0.8005	0.012611	0.7987	0.014417

Table 1. A sample of ten best subset models selected in first 10 bootstrap samples.

Average Optimism of the Models

The average optimism of the Harrell's C statistics is the average of the 200 difference values between the two C statistics calculated above. In HRS example, the optimism due to variable selection on is 0.0035 and optimism due to variable selection and coefficient estimation is 0.0048.Therefore, the unbiased estimate of Harrell's C statistics in the ACE study is 0.8005-0.0048 = 0.7957. The final output of the macro is shown in Table 2:

Apparent Harrell's C	Optimism from selection only	Harrell's C-statistic corrected for selection optimism	Optimism from selection and estimation	Harrell's C-statistic corrected for selection and estimation optimism
0.8005	.003491025	0.79699	.004797234	0.79569

Table 2. Final Harrell Optimism macro output

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 Harrell's c-statistic of a model selected by Cox regression using best subset selection. The full version of the macro can also estimate the optimism of the c-statistic for stepwise regression, as well as optimism of logistic 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. This capability is similar to that provided in SAS for CLASS variables in stepwise selection. 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.

REFERENCES

1. Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Tutorial in Biostatistics: Multivariable prognostic models. *Statistics in Medicine*, 15:361-387.
2. Efron B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78:316-331.
3. Steyerberg, E. W., Harrell, F. E., Borsboom, G. J. J. M., Eijkemans, M. J., Vergouwe, Y. & Habbema, J. D. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54:774-781.
4. Harrell, F. (2001). *Regression Modeling Strategies: With applications to linear models, logistic regression, and survival analysis.* NY: Springer.
5. Mannan, H.R., McNeil, J.J. (2012). Computer programs to estimate overoptimism in measures of discrimination for predicting the risk of cardiovascular diseases. *Journal of Evaluation in Clinical Practice.* ISSN 1365-2753.
6. King, J. (2003). Running a best-subsets logistic regression: an alternative to stepwise methods. *Educational and Psychological Measurement*, 63:392-403.
7. Gong, G. (1986). Cross-validation, the Jacknife, and the bootstrap: excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, 81:108-113.
8. Shtatland, E. S., Kleinman, K., & Cain, E. M. (2003). Stepwise methods in using SAS® PROC LOGISTIC and SAS® Enterpise Miner for prediction. *SUGI '28 Proceeding, Paper* 258-28, Cary, NC: SAS Institute, Inc.
9. Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B*, 34:187-220
10. Kremers, W. K. (2007). Technical Report Series No. 80, Concordance for survival time data: Fixed and time-

dependent covariates and possible ties in predictor and time. Department of Health Science Research, Mayo Clinical, Rochester, Minnesota, 2007.

11. Kremers, W. K. (2008). Calculates the C-statistic (concordance, descrimination index) for survival data with time dependent covariates and corresponding SE and 100(1-alpha)% CI. http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm
12. Pencina, M. J. & D' Agostino R. B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23:2109-2123
13. D' Agostino, R. B. & Nam, B. H. (2004). Evaluation of the performance of survival analysis models: discriination and calibration measures. *Handbook of Statistics*, 23:1-25. Amsterdam: Elsevier.
14. Steyerberg, E. W. (2009). *Clinical prediction models: a practical approach to development, validation, and updating*. New York: Springer.
15. Efron, B. & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92:548-560.
16. Shtatland, E. S., Kleinman, K., & Cain, E. M. (2004). A new strategy of model building in PROC LOGISTIC with automatic variable selection, validation, shrinkage and model averaging. *SUGI '29 Proceeding, Paper* 191-29, Cary, NC: SAS Institute, Inc.
17. Shtatland, E. S., Kleinman, K., & Cain, E. M. (2005). Model building in PROC PHREG with automatic variable selection and information criteria. *SUGI '30 Proceeding, Paper* 206-30, Cary, NC: SAS Institute, Inc.
18. Stijacic Cenzer I, Miao Y, Kirby K, Boscardin WJ. (2012) Estimating Harrell's optimism on predictive indices using bootstrap samples. *Proceedings of the Western Users of Sas Software Conference*, 74-12, 2012.
19. Lee SJ, Lindquist K, Segal MR, Covinsky KE.(2006). Development and validation of a prognostic index for 4-year mortality in older adults. *Journal of American Medical Association*. 295(7):801-8. Erratum in: *Journal in American Medical Association*. 295(16):1900.

CONTACT INFORMATION

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