Estimating Harrell’s Optimism on Predictive Indices Using Bootstrap Samples

Irena Stijacic Cenzer, University of California at San Francisco, San Francisco, CA
Yinghui Miao, NCIRE, 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 for 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 functional decline.

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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting 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 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 overfitting or optimism. The statistical literature is in strong agreement that split-sample assessment of model optim...
HARRELL’S ALGORITHM FOR CALCULATING OPTIMISM

Harrell et al (Harrell, Lee, & Mark, 1996) presented an algorithm for estimating the optimism, or overfitting, in predictive models. Their method is based on using bootstrapped datasets to repeatedly quantify the degree of overfitting in the model building process. We consider the specific setting here of building a logistic regression model with interest centered around the area under the ROC curve (the c-statistic) which measures the model’s ability to discriminate between patients with disparate outcomes. 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 $m=1,\ldots,M$, select the predictor and fit the model using the exact same algorithmic approach as in step 1 and calculate the discrimination ($c_{\text{boot}}^{(m)}$).
4. For each one of the new models, calculate its discrimination back in the original data set ($c_{\text{orig}}^{(m)}$).
5. For each one of the bootstrap samples, the optimism in the fit is $o^{(m)} = c_{\text{orig}}^{(m)} - c_{\text{boot}}^{(m)}$. 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 $c_{\text{adj}} = c_{\text{app}} - o$. 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 (Efron & Tibshirani, 1997), 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 much more similar to the cross-validation idea of validating the model in repeated external samples.

OUR MACRO

Our macro calculates the optimism of predictive indices and can be used in a number of situations:

1. It works for logistic regression and survival analysis.
2. It works for best subsets regression and stepwise selection.
3. It works for regular bootstrapping methods and .632 bootstrapping.
4. It differentiates between optimism due to variable selection and optimism due to coefficient estimation.
5. In addition to estimating the optimism, the output of the macro also contains 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 logistic regression and the standard bootstrap method. The full macro can be obtained by contacting the authors.

DEFINING THE MACRO CALL

```macro
%MACRO HARRELL_OPTIMISM (ORIGDAT=, SEED=, EVENTNO=, OUTC=, FULLMODEL=, REPS=);
```

The macro HARRELL_OPTIMISM is specified in the following manner:

- **ORIGDAT** – the original full dataset.
- **SEED** – seed from which the bootstrapping should start. Allows for ability to replicate results.
Estimating Harrell’s Optimism on Predictive Indices Using Bootstrap Samples, continued

- EVENTNO – the value of outcome variable that specifies that outcome occurred.
- OUTC – the name of the outcome variable.
- FULLMODEL – list of all potential variables.
- REPS – desired number of bootstrap replications.

THE MACRO

Figure 1 shows the SAS code for shortened “Harrell Optimism” macro.

```sas
%MACRO Harrell_Optimism (ORIGDAT=, SEED=, EVENTNO=, OUTC=, FULLMODEL=, REPS=);

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

proc surveyselect data=&ORIGDAT out=BOOTDAT
   seed=&SEED method=URS samprate=1 outhits rep=&REPS;
run;
/*ORIGINAL DATASET - BEST MODEL*/
proc logistic data=&ORIGDAT;
model &OUTC (event="&EVENTNO")=&FULLMODEL / selection=score best=1;
odds output Bestsubsets=BSORIG1;
run;
proc sort data=BSORIG1; BY SCORECHISQ; RUN;
data BSORIG2; set BSORIG1; 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=VARIABLESINMODEL);
call symput ('BESTORIGM', trim(VARIABLESINMODEL));
run;
proc logistic data=&ORIGDAT;
model &OUTC (event="&EVENTNO")=&BESTORIGM;
odds output Association=C_OB1;
run;
data C_OB2 (keep=C_OB);
set C_OB1 (keep=LABEL2 NVALUE2 rename=(NVALUE2=C_OB));
   if LABEL2='c';
rundata C_OB2 (keep=C_OB);
set C_OB1 (keep=LABEL2 NVALUE2 rename=(NVALUE2=C_OB));
   if LABEL2='c';
run;
/*REPLICATED BOOTSTRAPPING DATASETS*/
%do M=1 %to &REPS;
/*BOOT DATASET - BEST MODEL*/
proc logistic data=BOOTDAT;
```

where REPLICATE=&M;
model &OUTC (event="&EVENTNO")=&FULLMODEL / selection=score best=1;
odds output Bestsubsets=BS1;
run;

proc sort data=BS1; by SCORECHISQ; run;
data BS2; set BS1 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=NUMBEROFVARIABLES VARIABLEsinMODEL);
call symput ('BESTMODNO', put(NUMBEROFVARIABLES, 4.0));
call symput ('BESTMODEL', trim(VARIABLEsinMODEL));
run;

proc logistic data=BOOTDAT;
where REPLICATE=&M;
model &OUTC (event="&EVENTNO")=&BESTMODEL;
odds output Association=C_BB1;
run;

data C_BB2 (keep=C_BB);
set C_BB1 (keep=LABEL2 NVALUE2 rename=(NVALUE2=C_BB));
if LABEL2 = 'c';
run;

/*ORIGINAL DATASET - BOOT-BEST MODEL*/
proc logistic data=&ORIGDAT;
model &OUTC (event="&EVENTNO")=&BESTMODEL;
odds output Association=C_ORIG_BB1;
run;

data C_ORIG_BB2 (keep=C_ORIG_BB);
set C_ORIG_BB1 (keep=LABEL2 NVALUE2 rename=(NVALUE2=C_ORIG_BB));
if LABEL2 = 'c';
run;

/*MERGE RESULTS - 2 C-STATISTICS*/
data C_STAT (keep=RESAMPLE V_RETAIN BESTNO C_BB C_ORIG_BB OPTIMISM_BB_OBB);
length RESAMPLE $8 V_RETAIN $256 BESTNO $8;
merge C_BB2 (KEEP=C_BB) C_ORIG_BB2 (KEEP=C_ORIG_BB);
RESAMPLE=&M;
V_RETAIN="&BESTMODEL";
BESTNO=STRIP("&BESTMODNO");
OPTIMISM_BB_OBB=C_BB-C_ORIG_BB;
label RESAMPLE='Re-Sample' V_RETAIN='Variable in BestModel' BESTNO='Variable no. in Best Model' C_BB='C statistic from best subset model fitted in the bootstrap dataset' C_ORIG_BB='C statistic from best subset model fitted in the bootstrap dataset, applied to original dataset'
Estimating Harrell's Optimism on Predictive Indices Using Bootstrap Samples, continued

```sas
OPTIMISM_BB_OBB='C-diff: BootBest-Orig_BootBest';
run;

proc append base=C_STAT_LIST data=C_STAT force; run;

proc means data=C_STAT_LIST mean;
var OPTIMISM_BB_OBB;
output out=OPT_MEAN mean(OPTIMISM_BB_OBB)=OPTIMISM_BB_OBBMEAN;
run;

data C_VALIDATION (keep=C_OB OPTIMISM_BB_OBBMEAN C_VALD_OBB);
merge C_OB2 (keep=C_OB) OPT_MEAN (keep=OPTIMISM_BB_OBBMEAN);
C_VALD_OBB=C_OB-OPTIMISM_BB_OBBMEAN;
label C_OB='C statistic from best subset model fitted in the original dataset';
run;

data _null_; VNUMBER=countw("&FULLMODEL");
call symput ('VARNO',put(VNUMBER,4.0)); 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;
%MEND Harrell_Optimism;
```

Figure 1. The shortened version of the Harrell_Optimism macro
EXAMPLE - ACE

Data Description

We use data collected in the Unit for Acute Care of the Elderly (ACE Unit) at University Hospitals of Cleveland to illustrate the use of best subsets models and our “Harrell Optimism” macro. Participants were drawn from two prospective studies of an intervention to improve functional outcomes of hospitalized elders in two teaching hospitals in Ohio. Patients aged 70 years or older who had emergency admissions to the general medical services between 1993 and 1998 were eligible. Exclusion criteria were admission to an intensive care unit or oncology ward, elective admission, or length of stay less than two days.

Briefly, the cohort data consisted of 1638 patients who had emergency admission. Patients’ 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. Out outcome was decline in functional status (dependent in one or more ADLs or death) at the time of discharge from ICU. Out of 1638 patients in our cohort, 498 (30.4%) experienced decline in functional status and/or death during the follow up period.

We looked at 26 potential risk factors from the domains of sociodemographics, functional status two weeks before and at admission, clinical conditions and laboratory results. The demonstration below is based on the actual, more detailed model selection processed that was used by our research group to develop a published (Mehta, et al., 2011) predictive index for new onset disability in hospitalized elders.

Best Subsets Output

In order to identify the best predictors of decline in functional status we show how to run a single best subsets regression. Best subsets regression is fitted using the option “selection=score” in the proc logistic. We determine how many subsets of each size we want to include in the output in option “best=”, and what is the maximum number of variables in the model in option “stop=”. Best subsets regression is fitted using the following call:

```
proc logistic data=testdat descending;
model depdeadvsvind = AGEC_2 AGEC_3 female MARITAL white myocard chf chronlung pvd cvd diabet CHARCAT APACHET ALBUMLOWC_1 SPMSQSEVE_1 STEADYC malnut walkblockless walkuphill deprcat IADLBASE2_1 adladm23 adladm45 HLTHADM livalonen baddx /best=3 stop=15 selection = score;
run;
```

Best Subsets regression gives us the following output:

<table>
<thead>
<tr>
<th>Number of Variables</th>
<th>Score</th>
<th>Chi-Square</th>
<th>Variables Included in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>194.4078</td>
<td></td>
<td>SPMSQSEVE_1</td>
</tr>
<tr>
<td>1</td>
<td>134.4664</td>
<td></td>
<td>adladm45</td>
</tr>
<tr>
<td>1</td>
<td>100.2635</td>
<td></td>
<td>IADLBASE2_1</td>
</tr>
<tr>
<td>2</td>
<td>261.1653</td>
<td></td>
<td>SPMSQSEVE_1 adladm45</td>
</tr>
<tr>
<td>2</td>
<td>255.2301</td>
<td></td>
<td>SPMSQSEVE_1 STEADYC</td>
</tr>
<tr>
<td>2</td>
<td>227.3459</td>
<td></td>
<td>SPMSQSEVE_1 IADLBASE2_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...some output omitted</td>
</tr>
<tr>
<td>15</td>
<td>428.4826</td>
<td></td>
<td>AGEC_2 AGEC_3 CHRONLUNG ALBUMLOWC_1 SPMSQSEVE_1 STEADYC MALNUT walkblockless walkuphill DEPRCAT IADLBASE2_1 adladm23 adladm45 LIVALONEN baddx</td>
</tr>
<tr>
<td>15</td>
<td>428.2701</td>
<td></td>
<td>AGEC_2 AGEC_3 CHF CHRONLUNG ALBUMLOWC_1 SPMSQSEVE_1 STEADYC MALNUT walkblockless walkuphill IADLBASE2_1 adladm23 adladm45 LIVALONEN baddx</td>
</tr>
<tr>
<td>15</td>
<td>428.2520</td>
<td></td>
<td>AGEC_2 AGEC_3 CHRONLUNG PVD ALBUMLOWC_1 SPMSQSEVE_1 STEADYC MALNUT walkblockless walkuphill IADLBASE2_1 adladm23 adladm45 LIVALONEN baddx</td>
</tr>
</tbody>
</table>

Output 1. Output of PROC LOGISTIC for Best Subsets regression
There are a number of ways to determine what size model we should choose, including AIC and BIC values (Shtatland ES, 2003). We chose to look if the difference in score statistics between the best models of two different sizes is statistically significant. (We are acting as if the two models are nested which is not always the case.) That is, if score statistics of the best model of size n+1 is more than 3.84 higher than the score statistics of the best model of size n, we select the model of size n+1.

Once we select the correct model size, we have some freedom in selecting the final model of that size. One of the advantages of best subsets regression is that we can look at similar models that are indistinguishable statistically, but might vary practically. For example, two models of same size can have almost identical score statistic, but can have one or more different variables. Since the models are statistically similar, a researcher can decide which model to use based on clinical importance of variables, ease of data collection, etc.

Using just this algorithm, the selected model in ACE example had 12 variables: age (AGEC_2 AGEC_3), serum albumin (ALBUMLOWC_1), cognitive impairment (SPMSQSEVE_1), gait (STEADYC), mobility (walkblockless walkuphill), IADL (IADLBASE2_1), ADLs at admission (adladm23 adladm45), living alone (LIVALONEN) and diagnosis with cancer and/or stroke (baddx). Clinical considerations would probably necessitate some changes to this model but we do not consider this aspect further in this paper. The c statistic for the final model selected is 0.797.

**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 200 datasets using bootstrapping procedures. In each one of those datasets, we fit a new best subsets model and we calculate the 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:
### Table 1: A sample of ten best subset models selected in first 10 bootstrap samples.

<table>
<thead>
<tr>
<th>Re-Sample</th>
<th>Variable no. in Best Model</th>
<th>Variable in Best Model</th>
<th>C statistic from best subset model fitted in the bootstrap dataset</th>
<th>C statistic from best subset model fitted in the bootstrap dataset, applied to original dataset</th>
<th>C-diff: BootBest-Orig_BootBest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>AGEC_2 AGEC_3 ALBUMLOWC_1 SPMSQSEVE_1 STEADYC walkblockless IADLBASE2_1 adladm23 adladm45 LIVALONEN baddx</td>
<td>0.785228</td>
<td>0.791934</td>
<td>-0.006705</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>AGEC_2 AGEC_3 MARITAL CHRONLUNG ALBUMLOWC_1 SPMSQSEVE_1 STEADYC walkblockless walkuphill IADLBASE2_1 adladm23 adladm45 LIVALONEN baddx</td>
<td>0.806941</td>
<td>0.797685</td>
<td>0.009257</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>AGEC_2 AGEC_3 ALBUMLOWC_1 SPMSQSEVE_1 STEADYC MALNUT walkblockless walkuphill IADLBASE2_1 adladm23 adladm45 baddx</td>
<td>0.806589</td>
<td>0.794096</td>
<td>0.012494</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>AGEC_2 AGEC_3 CHRONLUNG ALBUMLOWC_1 SPMSQSEVE_1 walkblockless walkuphill DEPRCAT IADLBASE2_1 adladm23 adladm45 baddx</td>
<td>0.802921</td>
<td>0.792168</td>
<td>0.010753</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>AGEC_2 AGEC_3 CHRONLUNG ALBUMLOWC_1 SPMSQSEVE_1 walkblockless walkuphill IADLBASE2_1 adladm23 adladm45 HLTHADM baddx</td>
<td>0.813756</td>
<td>0.791365</td>
<td>0.022392</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>AGEC_2 AGEC_3 MYOCARD CHRONLUNG ALBUMLOWC_1 SPMSQSEVE_1 STEADYC walkblockless walkuphill IADLBASE2_1 adladm23 adladm45</td>
<td>0.798810</td>
<td>0.787386</td>
<td>0.011424</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>AGEC_2 AGEC_3 ALBUMLOWC_1 SPMSQSEVE_1 STEADYC IADLBASE2_1 adladm23 adladm45 LIVALONEN baddx</td>
<td>0.780781</td>
<td>0.789747</td>
<td>-0.008966</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>AGEC_2 AGEC_3 CHRONLUNG ALBUMLOWC_1 SPMSQSEVE_1 MALNUT walkblockless walkuphill adladm23 adladm45 baddx</td>
<td>0.803696</td>
<td>0.791141</td>
<td>0.012555</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>AGEC_2 AGEC_3 ALBUMLOWC_1 SPMSQSEVE_1 STEADYC walkblockless walkuphill IADLBASE2_1 adladm23 adladm45 baddx</td>
<td>0.763751</td>
<td>0.793322</td>
<td>-0.029570</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>AGEC_2 AGEC_3 WHITE CHRONLUNG APACHE2 ALBUMLOWC_1 SPMSQSEVE_1 walkblockless walkuphill IADLBASE2_1 adladm23 adladm45 LIVALONEN baddx</td>
<td>0.805697</td>
<td>0.793829</td>
<td>0.011868</td>
</tr>
</tbody>
</table>
Average Optimism of the Models

The average optimism of the C statistics is the average of the 200 difference values between the two C statistics calculated above. In ACE example the optimism is 0.009. Therefore, the unbiased estimate of C statistics in the ACE study is 0.797 – 0.009 = 0.788. The final output of the macro is shown in Table 2:

<table>
<thead>
<tr>
<th>C statistic from best subset model fitted in the original dataset</th>
<th>Mean_C-Diff: BootBest-Orig_BootBest</th>
<th>C Validation: OrigBest-[Mean_C-Diff: BootBest-Orig_BootBest]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.796889</td>
<td>.008523115</td>
<td>0.78837</td>
</tr>
</tbody>
</table>

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 the c-statistic of a model selected by logistic regression using best subsets. The full version of the macro can also estimate the optimism of the c-statistic for stepwise regression, as well as optimism of Harrell’s c for Cox regression models. Additionally, our macro can implement both standard bootstrapping and the .632 bootstrap method. The .632 bootstrap method is preferable by some because it mimics more closely cross validation procedures. 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


CONTACT INFORMATION

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

Irena Stijacic Cenzer
University of California, San Francisco
4150 Clement St.
San Francisco, CA 94121
(415) 221-4810 x2707
Irena.Stijacic@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.