Bootstrap Cox regression using SAS procedures

Chikuma Hamada, Faculty of Medicine, University of Tokyo,
Junji Kishimoto, SAS Institute Japan Ltd.

Abstract

In the analysis of biomedical studies variable selection is often necessary because of too many explanatory variables that are highly correlated. The stability of model selection should be evaluated because of variability across repeated experiments. With Cox regression, an analytical approach to checking stability is very difficult; however, the jackknife or bootstrap method implemented with a high-speed computer can be a practical alternative.

We illustrate how to do bootstrap Cox regression using SAS(R) procesures PROC MULTTEST and PROC PHREG, and apply the method to pancreatic carcinoma data. Although we focus on the Cox proportional hazards model, this methodology can be applied to other regression models with slight modification.

1. Introduction and purpose

Cox proportional hazard regression model is a standard tool for the analysis of censored survival data. Many applications for biomedical data , particularly evaluation of cancer clinical trials, were already reported. SAS PHGLM(Version 5) and PHREG(Version 6) procedure are available for this purpose. In the analysis of medical data, the selection of variables in the framework of a regression model which might influence the response variables is a traditional and important problem.

In prognostic studies the selection of variables is an essential problem , since the main goal is to distinguish between risk and less risk factors. Ignoring important variables may result in bias and low power for evaluation of the effect of treatment. Risk factors may also characterize patient groups and influence the adapted therapies. In the analysis of randomized clinical trials it is often necessary to evaluate the influence of factors other than treatment. Even if randomization was done, the distribution of important prognostic factors may be biased among treatment groups when sample size is relatively small, such as 100 or interim analysis is performed on the early stage of the study. Adjustment for effective covariates also strengthen statistical power to detect effect of treatment. Chapters about variable selection techniques can be found in many standard textbooks on regression analysis. PHREG procedure allows four types of variable selection, best subset, stepwise, forward and backward.

The evaluation of the stability of selected variables in the model plays a very important role in the model selection. Because it is well known that the choice of variables for inclusion in the regression model will vary across repeated samples or sub samples from unit population. One reason for this unstability may be the high interrelationship(correlation) among the explanatory covariates. Sample sizes of prognostic studies are usually too small to apply the well-known and often recommended method of data splitting , such as cross validation. Computer intensive-methods such as the jackknife or especially the bootstrap method can be a practical approach.

Bootstrap Cox regression is now available with PHREG and MULTTEST procedures. We will illustrate this method and show the result of application to pancreatic carcinoma data in the process of model building.

2.Data

Nishimura et al. analyze effect of intraoperative radiotherapy for carcinoma of the pancreas. We illustrate bootstrap Cox regression using sub part of this data.

The sample data (SAS data set name is PCANCER) comes from the 83 patients with

carcinoma of the pancreas. Of those, 82 died and 1 survived. The variables consist of TIME(survival time in months from surgery) ,CENSOR(dead=0 alive=1) and 8 covariates, AGE (age at the surgery) SEX(male=0 female=1) TREAT(intraoperative radiotherapy no=0 yes=1) LOC(location of carcinoma head=0 other=1) CH(involvement of the intrapancreatic bile duct) P(peritoneal spreading no=0 yes=1) STAGE(stage of TNM classification) and PS(performance status) .

3.Method

The basic idea of the bootstrap method is that if independent identically distributed observations with sample size N are obtained, then the variability of estimated characteristics of the distribution can be assessed by studying the variability of the estimate across a large number of bootstrap samples. The bootstrap samples are produced by taking samples from original data using random sampling with replacement.

MULTTEST procedure is a new procedure in release 6.07 SAS/STAT software. The procedure originated from MBIN and MTEST procedures which were produced by Westfall, Lin and Young. It offers p-value adjustment for multiplicity of statistical test by resampling method. This procedure also produces bootstrap sample from raw data set and outputs to new data set using outsamp=SASdata set option in the PROC MULTTEST statement. Each bootstrap sample produced by MULTTEST procedure is analyzed by PHREG procedure.

The example of SAS coding of Bootstrap Cox regression is as follows.

Example of program for fixed bootstrap regression

```
DATA PCANCER;SET PCANCER;CLASS=1;
PROC MULTTEST DATA=PCANCER
     NSAMPLE=1000 OUTSAMP=OUT   SEED=4989
     NOCENTER NOPRINT BOOTSTRAP;
     TEST MEAN(TIME CENSOR AGE SEX
         TREAT LOC CH P STAGE PS) ;
     CLASS CLASS;
DATA OUT;SET OUT;
     NO=SUBSTR(_R_, 1, 4);
PROC PHREG DATA=OUT  OUTEST=EST;
     MODEL TIME*CENSOR(1)=AGE SEX TREAT LOC
     CH P STAGE PS  ;
     BY NO;
```

"NSAMPLE=1000" in PROC MULTTEST statement specifies the number of resamples 1000 、 "OUTSAMP=OUT" specifies to make bootstrap resamples output to SAS data set OUT、"SEED=4989" specifies the initial seed for the random number generator. "NOPRINT" suppresses all printed outputs, because those have no meaning. "Bootstrap" specifies resampling with replacement from the original data set. Substituting " PERMUTATION" for "BOOTSTRAP", MULTTEST resamples without replacement. Since the purpose is not to compare among groups, but to produce the data set, CLASS statement has no meaning. TEST statement specifies the survival time , the state of censoring and covariates which are necessary for Cox Regression. If stratified variable, such as center or study exists, stratified bootstrap Cox regression can be done using STRATA statment. MULTTEST procedure performs resampling stratified the level of strata variable(specified by STRATA statement).

In this article , the number of resampling is set to 1000. In release 6.10, the variable _R_ is not present but is split into two numeric variables _SAMPLE_ and _OBS_ this allows direct manipulation "by _sample_ instead of "NO=SUBSTR(_R_,1,4)";

Fixed model

A Cox regression model with the same explanatory variables fitted to each bootstrap sample. If B*ik is the estimated regression coefficient for the kth variables in the ith bootstrap sample(where * indicates a bootstrap estimate). Then bootstrap estimates of the regression coefficient and its standard error are the mean and standard deviation of B*ik(i=1 to resampling number), say B*k and SE(B*k). Ninety-five percent confidence limits of B*k(quantiles) were obtained from the empirical cumulative distribution function of the B*ik.

Stepwise model

Each bootstrap sample was analyzed by stepwise selection using STEPWISE and SLS=0.20 SLE=0.20 option.

4.Result

4.1 The analysis of original data

Table 1 shows the analysis of full model. TREAT is only significant at 0.05 level and intraoperative radiotherapy decreases the risk of death by almost half.

Table 2 shows the analysis of stepwise regression. TREAT, STAGE and LOC are selected by this method.

Table 1 result of full model(original data)

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
AGE	0.016192	0.01301	1.54939	0.2132
SEX	-0.156399	0.24251	0.41591	0.5190
TREAT	-0.733026	0.32100	5.21466	0.0224
LOC	0.504438	0.29915	2.84342	0.0917
CH	0.095407	0.11696	0.66540	0.4147
P	0.468512	0.29339	2.55012	0.1103
STAGE	0.388938	0.26535	2.14845	0.1427
PS	-0.205395	0.18938	1.17627	0.2781

Table 2 result of stepwise regression(original data)

Step	Variable Entered	Removed	Number In	Score Chi-Square	Pr > Chi-Square
1	TREAT		1	5.6713	0.0172
2	STAGE		2	3.7617	0.0524
3	LOC		3	2.1742	0.1403

Table 3 shows the analysis of best subset regression. The best model including three variables consists of TREAT, LOC and STAGE. This result is consistent with stepwise regression.

Table 3 result of best subset regression using BEST=1 option (original data)

Number of Variables	Score Value	Variables Included in Model
1	5.6713	TREAT
2	9.4148	TREAT LOC
3	11.5226	TREAT LOC STAGE
4	12.8568	AGE TREAT LOC STAGE
5	14.0891	AGE TREAT LOC P STAGE
6	15.2356	AGE TREAT LOC P STAGE PS
7	15.8855	AGE TREAT LOC CH P STAGE PS
8	16.3798	AGE SEX TREAT LOC CH P STAGE PS

4.2 The analysis of fixed bootstrap model

Table 4 shows the coefficient of the full model and bootstrap estimates. Standard errors of bootstrap estimates are higher than those of original estimates.

Table 5 shows the coefficients of TREAT in the best subset models including from 1 to 8 variables specified selection=score option and those of bootstrap estimates corresponding to each model. As the number of variables in the model increases, standard errors of bootstrap estimates are larger. Otherhand those of original analysis are almost stable. Bootstrap estimates show that the model including more than five variables increases standard error of TREAT.

Table 4 comparison bootstrap estimates with original analysis

Variable	original analysis Bk	SE(Bk)	bootstrap estimates B*k	SE(B*k)
AGE	0.016	0.013	0.018	0.014
SEX	-0.156	0.242	-0.169	0.315
TREAT	-0.733	0.321	-0.764	0.450
LOC	0.504	0.299	0.580	0.338
CH	0.095	0.116	0.110	0.157
P	0.469	0.293	0.519	0.441
STAGE	0.389	0.265	0.428	0.302
PS	-0.205	0.189	-0.172	0.257

Table 5 comparison original coefficient of TREAT with bootstrap estimates.

No. vars	original analysis B	SE(B)	B2.5	B97.5
8	-0.733	0.321	-1.362	-0.104
7	-0.766	0.315	-1.379	-0.144
6	-0.847	0.292	-1.419	-0.276
5	-0.749	0.273	-1.284	-0.214
4	-0.766	0.271	-1.297	-0.234
3	-0.703	0.265	-1.223	-0.183
2	-0.716	0.263	-1.232	-0.199
1	-0.610	0.260	-1.120	-0.101

vars	bootstrap estimates B*	SE(B*)	B*2.5	B*97.5
8	-0.764	0.450	-1.737	0.083
7	-0.780	0.424	-1.631	0.056
6	-0.874	0.357	-1.630	-0.221
5	-0.788	0.306	-1.442	-0.245
4	-0.806	0.298	-1.420	-0.273
3	-0.742	0.282	-1.346	-0.248
2	-0.754	0.288	-1.367	-0.252
1	-0.621	0.238	-1.137	-0.198

B2.5=B−1.96×SE(B) B97.5=B+1.96×SE(B)

4.3 The analysis of stepwise model

Table 6 shows how often each of the 8 variables appeared in the 1000 bootstrap resamples. 6 times do not meet with convergence criterion. Although three

variables TREAT, LOC and STAGE are selected in the original stepwise model, LOC and STAGE are present in less than two-third of the bootstrap models. Variable P which is not selected in the original model, is present in almost half of the bootstrap models. These results show unstability of selected models.

Table 6 Frequency of selection of variables in stepwise regression

Variable	Selection frequency	proportion (%)
AGE	456	45.9
SEX	279	28.1
TREAT	822	82.7
LOC	625	62.9
CH	261	26.3
P	512	51.5
STAGE	556	55.9
PS	307	30.9

Table 7 shows the distribution of the number of variables in bootstrap samples and the number of variables in TREAT STAGE and LOC which are selected in the original stepwise regression. The original model including variables TREAT, STAGE and LOC reappears only 128 times in bootstrap samples.

Table 7 Distribution of the number of variables in TREAT STAGE and LOC

	Number of variables in bootstrap models				
Frequency	0	1	2	3	Total
0	1	0	0	0	1
1	4	22	0	0	26
2	5	38	85	0	128
3	4	49	174	24	251
4	2	48	160	75	285
5	2	24	99	77	202
6	0	1	33	43	77
7	0	0	10	11	21
8	0	0	0	3	3
Total	18	182	561	233	994

Discussion

The bootstrap approach is used for the estimation of parameters and their variability in a particular model. We evaluate the relationship between variability of regression coefficient and the number of variables included in the model. This analysis gives us the suggestion until how many variables should be included in the model. In our paper we also adopt this approach to investigate of stability of selected models. The result of a stepwise selection is just a single model without any information about its stability. The variable may be selected due to small highly influential observations. The information from bootstrap approach is useful to build robust model which is less sensitive to small observations. We may feel more confident about the importance of a variable when it was selected in nearly all the bootstrap sample. The changeability among variables may also be evaluated from this analysis. if the following tendency exist "When variable A is selected in the model, variable B is not selected and in opposite case variable B is selected' variable B could be candidate of surrogate variable A.

Since an analytical approach of these problems is almost impossible at present. bootstrap Cox regression using SAS PHREG and MULTTEST turn on the light for building appropriate model. It should be noted that although this article focused on Cox proportional hazard regression model, it can also be applied to other regression, such as multiple regression (REG), logistic regression (LOGISTIC), repeated measurement analysis(MIXED) and acceralated survival model (LIFEREG), with slightly straightforward modification.

References

Westfall, P.H. and Young, S.S. (1992), Resampling-Based Multiple Testing. New York:John Wiley & Sons

SAS Institute Inc. (1992), SAS Technical Report P-229, SAS/STAT Software:Changes and Enhancements, Release 6.07, 369-405, 435-479, Cary, NC:SAS Institute Inc.

Sauerbrei, W. and Schumacher, M. (1992). "A Bootstrap Resampling Procedure for Model Building:Application to the Cox Regression Model." Statistics in medicine, 11, 2093-2109

Nishimura, A et al. (1988). "Evaluation of Intraoperative Radiotherapy for Carcinoma of the Pancreas: Prognostic Factors and Survival Analyses." Radiation Medicine, 6-2, 85-91

Appendix Data set PCANCER

```
DATA PCANCER;
  INPUT CASENO TIME CENSOR  AGE SEX TREAT
        LOC CH P STAGE PS @@;
CARDS;
   1  2.4  0  66  0  0  1  4  1  4  3
   2  1.7  0  69  0  0  1  4  1  4  3
   3  0.1  0  48  0  0  1  1  0  3  2
   4  1.0  0  73  0  0  1  4  0  3  4
   5  4.8  0  65  0  0  1  4  0  4  2
   6  6.4  0  38  0  0  0  4  0  3  3
   7 10.8  0  62  1  0  1  4  0  3  3
   9  5.1  0  59  1  0  1  1  1  4  3
  10  1.1  0  53  0  0  1  1  1  4  3
  11  0.5  0  70  0  0  1  1  1  4  3
  12  0.8  0  71  0  0  0  3  0  4  3
  14  4.0  0  61  1  0  1  4  0  4  3
  15  4.0  0  69  0  0  0  4  0  3  3
  16  4.0  0  41  1  0  1  4  0  4  4
  17  8.5  0  49  0  0  1  4  0  3  2
  18  3.6  0  56  0  0  0  4  0  3  2
  19  6.9  0  59  1  0  0  4  1  3  4
  20  6.2  0  53  0  0  0  3  1  4  4
  21  1.0  0  72  1  0  0  3  0  4  2
  22  6.2  0  57  1  0  0  3  0  3  2
  23  4.3  0  49  0  0  0  3  0  4  2
  24  3.1  0  74  0  0  1  4  0  3  3
  25  8.3  0  43  1  1  1  4  0  4  2
  26 12.7  0  60  1  1  1  4  1  3  3
  28  4.9  0  55  1  1  1  4  0  3  3
  30  2.7  0  70  0  1  1  3  0  3  3
  33 10.6  0  63  0  1  1  1  0  3  2
  35 18.2  0  69  1  1  0  4  0  3  2
  36  1.4  0  66  1  1  1  1  1  4  2
  37  5.8  0  58  1  1  1  1  0  4  2
  38  3.0  0  67  1  1  1  1  1  4  3
  39  1.5  0  74  0  1  1  1  1  4  2
  40  2.4  0  77  1  1  1  1  1  3  2
  41  2.0  0  70  1  1  1  4  0  3  4
  42  1.1  0  75  0  1  1  4  0  4  2
  43  2.5  0  65  1  1  1  2  0  3  2
  44  5.4  0  71  1  1  1  4  0  3  2
  45  4.4  0  50  0  1  1  1  1  4  2
  46  4.8  0  56  0  1  1  3  0  3  2
  47  3.1  0  68  1  1  1  1  0  3  1
  48  5.6  0  65  1  1  1  3  0  3  3
  49  3.1  0  65  0  1  0  3  0  3  2
  50  1.3  0  43  1  1  1  3  0  3  2
  51 11.5  0  83  0  1  0  3  0  3  2
  53  3.8  0  65  0  1  1  2  0  3  2
  54  2.9  0  63  0  1  1  1  0  4  3
  55  2.2  0  47  1  1  1  2  0  4  2
  56  1.7  0  75  1  1  1  2  1  4  2
  58  3.5  0  63  1  1  1  1  1  4  2
  62 11.3  0  54  0  1  1  2  0  3  2
  63  9.0  0  56  1  1  0  4  0  4  3
  64 12.5  0  50  0  1  0  1  0  3  1
  65  6.8  0  62  0  1  1  1  0  3  1
  66 10.8  0  53  1  1  1  1  0  3  3
  67  3.0  0  63  0  1  1  1  1  4  4
  68  1.8  0  59  0  1  1  4  0  4  4
  69  5.0  0  66  1  1  1  4  0  4  2
  70  8.0  0  62  0  1  0  3  0  3  2
  71  6.8  0  72  0  1  1  1  0  3  2
  72 11.1  1  54  0  1  0  3  0  3  1
  73  9.4  0  68  1  1  0  3  0  3  2
  74  3.9  0  63  1  1  0  4  0  3  2
  75  2.1  0  68  1  1  1  2  0  3  2
  76  4.3  0  48  1  1  1  4  0  3  3
  77  9.3  0  68  0  1  1  1  0  4  2
  78  8.8  0  75  1  1  1  1  0  4  2
  79  2.4  0  49  0  1  1  1  1  4  3
  80 21.6  0  62  0  1  1  1  0  4  3
  81  5.6  0  56  1  1  1  1  0  4  2
  83 11.4  0  56  0  1  1  1  0  3  2
  84 18.3  0  59  1  1  1  1  1  4  3
  85  9.2  0  59  0  1  1  1  1  3  2
  86  4.5  0  48  0  1  1  2  1  3  3
  87  8.2  0  64  0  1  1  2  0  3  2
  88 15.0  0  60  0  1  1  4  0  3  4
  89  6.9  0  75  0  1  1  1  0  4  2
  90  3.5  0  46  0  1  1  1  1  4  2
  91  2.1  0  53  1  1  1  1  1  4  2
  92  3.1  0  62  0  1  1  3  0  3  3
  93  3.2  0  47  0  1  1  3  0  4  3
  94  1.9  0  62  0  1  1  1  0  4  3
  95  2.1  0  55  0  1  1  1  1  3  2
  96  7.0  0  80  0  1  1  1  0  4  3
;
```