

## A SAS® MACRO FOR VALIDATING A LOGISTIC MODEL WITH SPLIT SAMPLE AND BOOTSTRAP METHODS

Kwan Hur, Charles A. Oprian, William G. Henderson, Bharat Thakkar, and Sharon Urbanski  
Center for Cooperative Studies in Health Services,  
Department of Veterans Affairs Hospital, Hines, IL

### ABSTRACT

A SAS macro for testing the validity of a model developed using a stepwise logistic regression by split sample and/or a bootstrap approach is presented. For the split sample method, the macro will split a sample into two random halves: a learning and a testing sample. A stepwise model is constructed on the learning half and a c-index is calculated. The model is then applied to the testing sample and the c-index is calculated and compared to the c-index of the learning sample. For the bootstrap method, the macro can sample with replacement the population and build a corresponding model. This can then be replicated to produce repeated logistic models with their corresponding c-indexes.

### INTRODUCTION

An issue that often arises in clinical research is the identification of important risk factors that are predictive of outcomes, such as development of a disease or resultant morbidity or mortality once a patient develops the disease. Regression techniques, with or without variable selection methods such as stepwise procedures, are often used for this purpose. After a model is built it is necessary to evaluate the model through a model validation procedure. The investigation of the stability and predictability of a selected model is a significant part of the modeling process.

There are two basic questions related to model validation. Is the representation or inclusion of a particular set of variables truly a valid portrayal of the actual model or a mere artifact of the sample represented by the data at hand? Is the model an accurate and reliable predictor of the outcome being studied? In order to answer these two questions a split sample<sup>1</sup> and a bootstrap approach<sup>2,3</sup>, may be used.

Logistic regression is a widely accepted technique for developing a model for binary outcomes, such as mortality or morbidity. As an example, this macro uses a stepwise logistic regression (PROC LOGISTIC) to identify prognostic risk factors for the 30-day mortality associated with major pulmonary resections (lobectomy or pneumonectomy) performed in the VA medical care system. Both split sample and bootstrap methods will be demonstrated in this example.

### METHODS

#### Split sample method

The most preferred method to validate a selected regression model is through the collection of new data. Often, however, the collection of new data is not practical. A reasonable alternative when the sample is large enough is to randomly split the data into two halves: a learning and a testing sample. The first half or learning sample is used to identify risk factors which will be included in the model. The second half or testing sample is used to evaluate the predictability of the model. Figure 1 displays the flow chart of the split sample method.

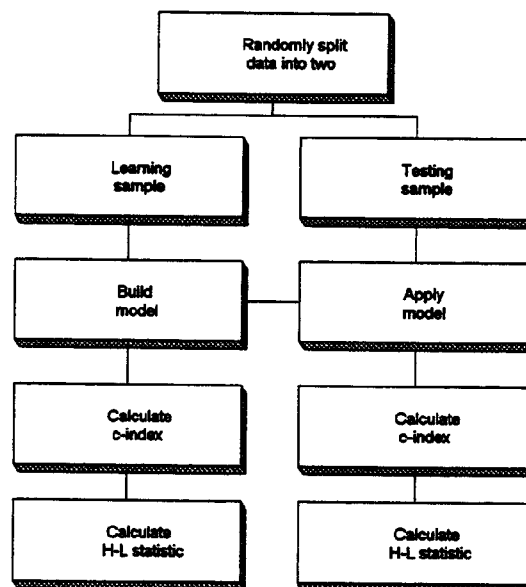


Figure 1. Split sample method

This method is replicated several times to assess the stability of the model by measuring the frequency of occurrence of each risk factor in the model and by examining the variability of the beta coefficients of the risk factors. The predictive ability of the model can be examined by calculating and comparing c-indexes<sup>4</sup> for the learning and testing samples. In general, a c-index of 0.5 indicates no predictive ability and a c-index of 1.0

indicates a perfect predictive ability.

The goodness-of-fit of the model can be measured with the Hosmer-Lemeshow statistic<sup>5</sup>. This statistic provides a measure of the calibration ability of the model for patients classified into the deciles of risk.

**Bootstrap method**

In the bootstrap approach replicated samples of *n* observations are selected with replacement from the original data set of *n* observations. The stability of the model can be examined similarly to the procedure described above by recording the frequency of occurrence of the variable in the model as shown in Figure 2. If a variable is truly representative of the model it will occur in the majority of the models. The c-index is calculated for each replication in order to examine the predictive ability of each model.

Risk factors	Replications						f(x <sub>i</sub> )
	1	2	3	.....	n-1	n	
x <sub>1</sub>	x	x				x	64
x <sub>2</sub>	x	x					x 34
x <sub>3</sub>			x			x	45
x <sub>4</sub>	x					x	x 37
.							
.							
x <sub>p-1</sub>	x					x	52
x <sub>p</sub>		x	x				x 86

Figure 2. Frequency of risk factors  
 x indicates that a risk factor is included in the model for each replication.  
 f(x<sub>i</sub>) is the frequency of occurrence for each risk factor.

**MACRO MODELVLD**

The macro MODELVLD allows the users to perform a model validation using a split sample approach as well as a bootstrap approach. At the end of the specified number of replications it creates two output files, one for the beta coefficients and the other for the c-indexes. The number of replications can be specified through a macro parameter.

**Macro Parameters:**

**indsn** is the name of the input data set. Input data should contain only one dependent and all the independent variables that will be introduced in the stepwise logistic model.

**method** is an indicator for the method being used (*split* for split sample and *boot* for bootstrap).

**N** is the counter for the number of replications.

**outdsn1** is the name of the output data set that contains the parameter estimates.

**outdsn2** is the name of the output data set that contains the c-indexes.

An example of the macro call statement should be:

```
%modelvld (indsn,split,200,beta,cindex);
```

**Output:**

This macro generates 3-4 tables depending on the specified method. If the user uses the split sample method, for the learning sample, it generates the frequency table of the risk factors that are included in the model, the actual parameter estimates (beta coefficients), and the c-index. For the testing sample it produces a table for the c-indexes and an additional table for the Hosmer-Lemeshow goodness-of-fit statistics. Likewise the same tables except for the Hosmer-Lemeshow statistics are generated for the bootstrap approach.

**EXAMPLE**

A total of 1,655 cases that had either a lobectomy or a pneumonectomy were identified from the National Surgical Quality Improvement Program database which contains 87,078 major surgery cases performed in the VA system. The primary objective of the pulmonary resection study was to identify the pre-operative patient risk factors that are related to 30-day mortality following the procedure.

In the first step a univariate test was done on each of 66 risk factors to see which are significantly related to the outcome variable, mortality. A chi-square test was used for the discrete variables and a t-test was performed for the continuous variables. Those variables with a significant level (p-value) less than 0.2 were then entered in a stepwise logistic regression. A dummy variable was introduced for the type of the operation (dummy = 0 for lobectomy and dummy = 1 for pneumonectomy). The final model (p-value < 0.05) from a stepwise logistic regression is shown in table 1.

**Table 1 Result of a stepwise regression**

<b>Variable</b>	<b>Parameter Estimate</b>	<b>Standard Error</b>	<b>p-value</b>
Intercept	-1.0107	0.7824	0.8647
cvaneuro	1.0076	0.3959	0.0109
discancr	0.9440	0.3030	0.0018
albumin	-0.5947	0.2080	0.0043
RBC	0.2798	0.0830	0.0007
dummy	0.8594	0.2296	0.0002

cvaneuro - residual neuro deficit, discancr - disseminated cancer, albumin - pre-operative serum albumin, RBC - red blood cell units transfused, dummy - type of operation

Both split sample and bootstrap methods were applied to the pulmonary resection data with 200 replications. The first output generated from the macro will be a table similar to Figure 2 which indicates the frequency and the selection of the risk factors in each replication. Table 2 is a summary of the frequency of occurrence of each of the pre-operative risk factors in the replication of 200 from a bootstrap method. The mean and standard deviation of the beta coefficients of the risk factors are also listed in Table 2. A total of 21 risk factors were included in the model at least once, but only those risk factors that were included in the model more than 50% of the time are presented in this table. The risk factors that were included in the final model listed in Table 1 occurred in a majority of the 200 bootstrap replications. This shows the stability of the final model, i.e. the selection of the variables is not an artifact of the particular data set used.

**Table 2 Bootstrap results**

<b>Risk Factors</b>	<b># of times Selected</b>	<b>% Selected</b>	<b>Beta Mean</b>	<b>Beta S.D</b>
dummy	196	98.0	1.013	0.234
discancr	156	78.0	1.075	0.253
cvaneuro	120	60.0	1.227	0.258
age	118	59.0	0.042	0.010
RBC	113	57.0	0.344	0.089
albumin	112	56.0	-0.713	0.186
wt. loss	111	56.0	0.832	0.194

The predictive ability of the model can be shown by comparing the c-indexes of the learning group and the testing group in the split sample approach. Table 4 shows the c-indexes from the learning and testing group for each replication. As expected, in general, the c-index in the testing

group is lower than the c-index in the learning group. The degradation of the c-index in the testing group was about 9 points on average (0.73 to 0.64). The mean, standard deviation and the ranges of c-indexes for each group are displayed at the bottom of the table.

In addition to the c-indexes, the Hosmer-Lemeshow goodness-of-fit statistic is computed from each sample for each replication. An example of this computation for a particular testing sample is obtained from the frequencies in Table 5. The H-L statistic equals 8.558 and the corresponding p-value is 0.4790 with 8 degrees of freedom. This suggests that the model fit is adequate for the testing sample.

**Table 4 Comparison of c-indexes**

<b>Replication</b>	<b>Learning Sample</b>	<b>Testing Sample</b>
1	0.727	0.554
2	0.716	0.703
3	0.721	0.583
4	0.735	0.654
5	0.783	0.631
.		
.		
200	0.763	0.671
Average	0.728	0.636
S.D.	0.034	0.036
Minimum	0.635	0.455
Maximum	0.812	0.708

**Table 5 H-L goodness-of-fit statistic**

<b>Decile</b>	<b>Observed Deaths</b>	<b>Expected Deaths</b>	<b>Observed Survivors</b>	<b>Expected Survivors</b>
1	3	1	77	79
2	2	0	54	56
3	5	3	109	111
4	2	2	56	56
5	5	8	112	109
6	4	3	71	72
7	5	7	88	89
8	5	6	69	68
9	9	12	72	69
10	16	14	69	71
Total	56	56	778	778

Note: H-L statistic = 8.558, p-value = 0.4790

## DISCUSSION

Both the split sample and bootstrap methods examine the stability of the model by looking at the frequency of the selection of the variables in the model as well as the variability in the coefficients of the variables in the model. The predictability of a model may be more fairly judged through the split sample approach because the model is generated from the learning sample and then applied to the testing sample which is independent from the learning sample. The bootstrap method, however, is better in building the final model in so far as it maximizes the utilization of the data by using the entire data set rather than half of the data as is the case with the split sample approach.

The CPU time required for the above example (1,655 observations and 22 independent variables) run on VAX/VMS V6.1 Model 4000-700 was approximately 97 minutes. For better assessment of the model a larger number of replications is recommended within the limitations of the system being used.

## REFERENCES

1. Picard, RR and Berk KN. "Data Splitting," American Statistician, 44:104-7, 1990.
2. Efron, B. and Tibshirani, R. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," Statistical Science, 1:1:54-77, 1986.
3. Sauerbrei, W. and Schumacher, M. "A Bootstrap Resampling Procedure for Model Building: Applications to the Cox Regression Model," Statistics in Medicine, 1:2093-2109, 1992.
4. SAS Institute Inc. (1990), "The LOGISTIC Procedure," Chapter 27 of SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2. Cary, NC: SAS Institute Inc.
5. Hosmer, DW and Lemeshow, S. Applied Logistic Regression. New York: John Wiley & Sons, 1989.

SAS is a registered trademark or trademarks of SAS institute Inc. in the USA and other countries.

Other brand and product names and registered trademarks or trademarks of their respective companies.

## ACKNOWLEDGMENT

The authors wish to thank Dr. Khuri and Dr. Daley, chairpersons of the National Surgical Quality Improvement Program, for allowing us to use a portion of the NSQIP data. This material is based upon work supported by the Office of Research and Development, Health Services Research, Department of Veterans Affairs.

## CONTACT

Any suggestions or comments are welcome. A copy of the macro can be obtained from:

Kwan Hur  
CCSHS (151A)  
Edward Hines VA Hospital  
Hines, IL 60141  
(708)343-7200 Ext. 5823