Paper 206-30

MODEL BUILDING IN PROC PHREG WITH AUTOMATIC VARIABLE SELECTION AND INFORMATION CRITERIA

Ernest S. Shtatland, Harvard Medical School, Harvard Pilgrim Health Care, Boston, MA
Ken Kleinman, Harvard Medical School, Harvard Pilgrim Health Care, Boston, MA
Emily M. Cain, Harvard Medical School, Harvard Pilgrim Health Care, Boston, MA

ABSTRACT

In SUGI'29 presentation, we suggested that our strategy of model building in PROC LOGISTIC (see also our SUGI'26 and SUGI'28 papers) could work for PROC PHREG as well. Our suggestion was based on the close similarity between logistic and Cox's regressions, including information criteria and stepwise, forward, backward and score options. Here we elaborate on this suggestion. As in logistic regression, we propose an approach to model building for prediction in survival analysis based on the combination of stepwise regression, Akaike information criteria, and the best subset selection. As in the case of PROC LOGISTIC, the approach inherits some strong features of the three components mentioned above. In particular, the approach helps to avoid the agonizing process of choosing the "right" critical p-value in stepwise regression.

INTRODUCTION

Like any regression procedure, PROC PHREG inevitably requires stepwise selection methods (including best subset selection) if the number of predictors is large. Currently, these methods are probably most widely used in medical and other applications (see, for example, Steyerberg et al. (2000)). At the same time automatic (stepwise) selection methods have serious drawbacks and are often criticized severely for instability and bias in regression coefficients estimates, their standard errors and confidence intervals (Derksen & Keselman (1992), Harrell et al. (1996), Steyerberg et al. (2000), Harrell (2001), etc.). For details, see also Shtatland et al. (2004**).** Basically, this criticism is related to logistic regression. But PROC PHREG shares so many properties with PROC LOGISTIC: same techniques for model building, in particular, stepwise, forward, backward and best subsets options, with the same confusing SLE and SLS default values of 0.05, same information criteria AIC and Schwarz that PROC PHREG can also "plead guilty" of the same drawbacks. (It is interesting that in PROC PHREG we can find only the *values* of AIC and SBC in printouts without definition and any theoretical background.) Besides this formal similarity between PROC PHREG and PROC LOGISTIC, some examples of close relationship between logistic regression and Cox regression can be found in Allison (1995), pp 211-222, Altman and Royston (2000), Efron (1988), Hosmer and Lemeshow (1989), pp. 238 - 245, SAS Institute Inc. (1995), pp 119-124, and Mayo, Kimler, and Fabian (2001). In Allison (1995) it is discussed how to use PROC LOGISTIC to analyze survival data in which many events occur at the same

point in time. The use of logistic regression with survival data has been also studied in Efron (1988) and Hosmer & Lemeshow (1989). According to Altman and Royston (2000) there is no need to distinguish *conceptually* between models for survival time (Cox regression) and binary regression models (logistic regression), because survival models can generate predicted probabilities at any given time-point within the follow-up period of the study. In SAS Institute Inc. (1995) it is shown how to use PROC PHREG to fit a conditional logistic regression model in matched case-control studies. Mayo, Kimler, and Fabian (2001) compare both logistic regression and Cox proportional hazards regression models for the prediction of breast cancer in the same cohort. In summary, we can conclude that such close relationship between PROC LOGISTIC and PROC PHREG is not surprising because both procedures are *event-oriented*. In PROC LOGISTIC we are interested in whether or not some events like hospitalizations, deaths, bankruptcies, mergers, residence changes, consumer purchases, etc., happened. In PROC PHREG we are interested in *when* these events happened. Thus, we can suggest that in PROC PHREG we have the same elements that allow us to develop an approach similar to the one developed for PROC LOGISTIC in Shtatland et al. (2003) and Shtatland et al. (2004). This approach is *absolutely necessary* when the number of predictors is large (see for example, Lu (2003), where this number is above 300).

MODEL BUILDING, STEP 1: CONSTRUCTING A FULL STEPWISE SEQUENCE

To understand all the difficulties with model building in each regression procedure even with a moderate or moderately large number of variables let us consider a couple of examples. For instance, if we have p=10 possible explanatory variables (which is a comparatively small number), then there are $K = 2^{10} = 1024$ possible models to compare. If p=20 (which is rather moderate), then the number of possible models is about one million. With p=34 we have more than 16 billion candidate models. Finding the best model by *direct* comparison is an unrealistic task. One of the possible ways, a reasonable and cheap one, to resolve the problem is to use the stepwise selection method with SLENTRY and SLSTAY close to 1 (e.g., SLENTRY = 0.99 and SLSTAY = 0.995). As a result, we will get the sequence of models starting with the null model and ending with the full model (all the explanatory variables included). The models in this sequence will be ordered in the way maximizing the increment in likelihood at every step. It is natural to call this sequence the *stepwise sequence*. Our choice: SLENTRY = 0.99 and SLSTAY = 0.995 is absolutely arbitrary. We can use any other values if only they result in the (full) stepwise sequence. It is important that we use the stepwise procedure in a way different from the one typically used. Instead of getting a *single* stepwise pick for some specific SLENTRY value (for example, 0.05, or 0.15, or 0.30, or 0.50, etc.) we obtain the *full sequence*. In doing so, we reduce the total number of $K=2^P$ potential candidate models to the manageable number of P models. Thus with 34 potential explanatory variables, we reduce the number of candidate models from 2^{34} (more than 16,000,000,000) to just 34. After this reduction we are able to apply any information criterion.

STEP 1 is performed by using the following code:

```
ods output ModelBuildingSummary=Summary;
ods output FitStatistics=Fit;

proc phreg data=MYDATA;
   model Time*Status(0)= X1 X2 X3 … XP
             / selection=stepwise slentry=0.99 slstay=0.995;
   run;
```

MODEL BUILDING, STEP 2:
MINIMIZING AIC ON THE FULL STEPWISE SEQUENCE

```
ODS statements are used above to get AIC values for each model in the
stepwise sequence. These values reside in the output data set FIT.

data AIC;
   set FIT;
   if CRITERION = 'AIC     ';
run;

proc print data=AIC;
run;
```

It is easy to find the minimum of AIC by using PROC MEANS and the MERGE statement. Also, it is strongly recommended to apply PROC PLOT and visualize the behavior of AIC *vs.* the number of predictors in the model, i.e. the step number in stepwise regression. Typically (but not always) AIC has a unique distinct minimum, which clearly indicates the AIC-optimal model. We will explain later why we use AIC and not any other information criterion.

STEP 3: SHOPPING AROUND AIC-OPTIMAL

Obviously, it would be too simplistic to recommend AIC-optimal models as the best models for prediction. First of all, there could be a number of nearly optimal models in the vicinity of the AIC-optimal choice. Second, and maybe most important, we have screened the *stepwise sequence only*, not *all possible models*. Up to now this limitation has been considered a clear advantage and the only practical way to use stepwise regression with a very large number of predictors. But at this point we have to face the fact that we work with the stepwise sequence only, not all possible models. The problem can be resolved more or less satisfactory by using the best subset selection procedure and a macro below. We will apply PROC PHREG with selection = SCORE to the *neighborhood* of the AIC-optimal model with the model size k_{AIC}, the parameter START smaller than k_{AIC}: $k_{AIC} - L$, and the parameter STOP larger than k_{AIC}: $k_{AIC} + M$. The parameters L, M and BEST determine the size of the two-dimensional subset of models resulted from the PROC PHREG run. It is natural to call this two-dimensional subset the *stepwise-AIC-best subset* "*blanket*". It is very likely that this "blanket" covers the really optimal model.

```
ods output BestSubsets=Best_subsets;

   proc phreg data=MYDATA;
```

```
      model Time*Status(0)= X1 X2 X3 … XP
                / selection=score START= k_AIC – L STOP= k_AIC + U best = K;
   run;

   proc print data=Best_subsets;run;
```

The values of the parameters L, U, and K are more or less arbitrary and usually depend on the situation. It is worth noting that the output of the standard best subset selection procedure provides only score statistics and the *list* of predictors with no coefficient estimates and other statistics. The problem with using score statistics is that it is difficult to compare models of different sizes since the score statistic tend to increase with the number of variables in the model. By using ODS statement with PROC PHREG and the following macro we can simultaneously run Cox regressions for all selected model sizes of interest around k_{AIC} and for a specified value of the BEST option:

```
OPTIONS MPRINT SYMBOLGEN MLOGIC;

%MACRO SCORE;

proc sql noprint;
  select (nobs -delobs) into: num
    from dictionary.tables
      where libname ='WORK'
        and memname = "BEST_SUBSETS";
   %let num=&num;
quit;


%do i=1 %to &num;

   data _null_ ;
      set Best_Subsets;
       if _N_ = &i;
      call symput('list', VariablesInModel);
   run;

   proc PHREG data=MYDATA;
   model OUTCOME = &list;
   run;

%end;

%MEND;

%SCORE;
```

AIC *vs.* OTHER INFORMATION CRITERIA

A general form of information criteria (IC) is

$$IC(c) = -2\log L(M) + c*K \qquad (1)$$

where logL(M) is the maximized partial log-likelihood, K is the number of covariates and *c* is a penalty parameter. In Shtatland et al. (2003), we use IC with c=1, 3/2 and 2. IC(2) is the famous Akaike Information Criterion, AIC. IC(1) and IC(3/2) have merits of their own (see, for example, in Shtatland et al. (2003)). AIC is still viewed as our key information criterion, a key player in the "information field". AIC has a number of optimal properties related to prediction, including asymptotical equivalency to Bayesian methods (Bayes factors). But what is especially important in the context of our presentation, AIC is asymptotically equivalent to the cross-validation criterion LOOCV and the bootstrap, as shown in Stone (1977**),** Shibata (1997), etc. When the sample size is small, AIC can lose its asymptotic optimal properties and become severely biased. To overcome this disadvantage a corrected AIC, AICC is proposed (see Shtatland et al. (2004) and references therein)

$$AICC = AIC + 2*K*(K+1)/(N-K-1), \qquad (2)$$

where N is the sample size and K is the number of predictors in the model.

SHRINKAGE

The predictive value of a Cox regression can be increased if the regression coefficients related to the K covariates in the model are multiplied by the shrinkage factor

$$\gamma = (\chi^2_k - K) / \chi^2_k \qquad (3a)$$

where χ^2_k is the model χ^2 . (Heinze and Schemper (2001)). This formula (3) can be re-written as

$$\gamma = (2\log L(M) - 2\log L(0) - K)/(2\log L(M) - 2\log L(0)) \qquad (3b)$$

The shrinkage factor (3a, 3b) is based on the Information Criterion IC(1). By the reasons discussed above AIC is viewed as our key information criterion. That is why we prefer to work with an AIC-based shrinkage factor

$$\gamma_{AIC} = (\log L(M) - \log L(0) - K)/(\log L(M) - \log L(0)) \qquad (4)$$

WHAT TO DO WITH THE STEPWISE-AIC-BEST SUBSETS BLANKET

As in case of logistic regression (Shtatland et al. (2004)) we can use the Stepwise-AIC-Best Subsets blanket in one of the following ways:

(a) Averaging all the models from the blanket which results in more robust but less interpretable model;
(b) Choosing the AIC optimal model from the blanket (based on *statistical* consideration only, in particular on the fact that AIC is asymptotically equivalent to cross-validation and the bootstrap, two most popular validation methods. When we work with AIC we are

trying to mimic cross-validation / bootstrapping results without performing both techniques.);
(c) Keeping all the models from the blanket in order to make the final pick later based on considerations other than the *statistical* one (for example, medical, biological, financial, etc.);
(d) Building the final model including *all the covariates* from the blanket models. This model may contain many more predictors than the final models in the two previous cases.

USING STEPWISE-AIC-BEST SUBSETS APPROACH IN COX REGRESSION FOR GENE EXPRESSION DATA

This is one of "hottest" applications of our Stepwise-AIC-Best Subsets approach to Cox regression.
MICROARRAYS. We describe our approach within gene expression data analysis (complementary DNA or cDNA microarrays, oligonucleotide microarrays and the "gold standard" of gene expression measurement, real-time quantitative PCR). cDNA and Affymetrix microarrays, two basic microarray platforms, have become a common research technique for the *genome-wide* measurement of expression levels. Global gene expression profiles are used for diagnosis of many diseases, especially cancer. According to Pawitan et al. (2004), it is now widely accepted that microarrays have the *potential* to revolutionize medicine (through new understanding of disease etiology, potentially better diagnosis, and new targets for therapy). Extremely rich information from this new technology has raised the hope of individualized therapy: specific treatment for each patient depending on patient's characteristics. While cDNA microarrays have the potential of developing into powerful molecular diagnostic tools, more work needs to be done before this potential is reached. From the biological perspective, the challenge is how to interpret a long list of genes resulting from microarray analysis (hundreds of candidate genes after filtering with some adjusted multiple t-test). Although high-density arrays may provide much more raw data on gene expression level, this is not necessarily a requirement, or even a desirable attribute, for meaningful molecular classification and tumor stratification studies. Even a few genes that exhibit highly-differential expression patterns may serve as robust separators for class distinction and class discovery experiments. Even after adjusted t-test filtering the resulting candidate genes may or may not be associated with prognosis. From the data analysis perspective, the results of microarray analysis are usually very unstable. Microarray measurements often appear to be systematically biased and the numerous contributing factors are poorly understood. This makes interpretation of gene expression data using microarrays very difficult and comparison of data generated by different users across different microarray systems almost impossible. For example, Lossos et al. (2004) refer to the two well known and widely cited studies on diffuse large-B-cell lymphoma with *no overlap* at all among the candidate genes in both models. Of course this disparity can be explained by technical differences, the composition of the microarrays used, and different algorithms used for constructing predictive models. We can conclude that though microarrays contain huge amount of biological information they are far from being the instrument which can been used *directly* in clinical practice, in particular clinical testing. We should also add that the microarray data/price ratio is too high for using in clinical practice. The price of a

microarray chip is related to how many genes it measures. Given the cost of the technology, the situation is not likely to change anytime soon. What microarrays can do is to produce a list of candidate genes in human cancer research to be analyzed in the next step.

RT-PCR. To overcome this instability and other disadvantages of microarrays and also to decrease the number of the candidate genes from hundreds to tens the following approach is recommended. Once significant results have been obtained from the analysis of microarray data, they are validated using techniques such as quantitative reverse-transcription polymerase chain reaction (RT-PCR) or Northern blots on the key components. Though this requires time and effort, these "gold standard" techniques are seen as more valid than the variable results of the microarray. Most journals will not publish microarray data without some validation of the candidate genes identified in microarray studies, and the emerging "gold-standard" validation technique RT-PCR has provided a simple, reliable, and rapid method of validating gene expression results obtained by microarray analysis. More importantly, this is also the first step for translating microarray results from the laboratory to the clinic. Thus, the future of gene expression research lies in the combination of microarrays and RT-PCR. RT-PCR is not suited for simultaneous analysis of large numbers of genes. Generally, it validates one gene at a time. At the same time, RT-PCR tests that involve simultaneous analysis of tens to hundreds of markers within a clinical sample will soon be required to allow maximum translation of post-genomic research to patient care. There already exist RT-PCR techniques that can process from hundreds to low thousands of genes (Iwao, Matoba, Ueno et al (2002)). For example, the authors mentioned above use high throughput RT-PCR (ATAC-PCR) that allows to accurately measure expression of 2412 genes and produce 21 candidate genes. Typically, array analysis is used as a tool to screen for target genes that are differentially expressed between biological samples, and RT-PCR then provides precise quantification and validation of the microarray results. Note that a gene-by-gene PCR processing is slightly similar to the univariate Cox regression analysis with the significance level of 0.20 – 0.25 aiming at the selection of variables for the multivariate Cox regression. The important difference is that RT-PCR processing is a knowledge-based, biological filtering and the gene-variables selection based on the 0.20 – 0.25 significance level is an empirical technique, which is often used, but also often questioned.

STATISTICAL MODELING WITH GENE EXPRESSIONS. In addition to problems discussed above on microarrays analysis there is a very serious statistical challenge related to this high dimensionality of the gene space. In microarray data analysis, when using logistic or Cox regressions, the number of predictors p (genes whose expression level is measured to predict a disease) is much larger than the number of observations N (patients): $p >> N$. Typically p is from 10,000 to 30,000, and N is around 100. This situation looks rather pathological from a conventional point of view in statistics, and traditional statistical techniques fail. That is why before using logistic or Cox regression, Nguyen & Rocke (2002) recommend to perform some dimension reduction by principal component analysis or partial least squares. The problem with this approach is how to interpret the components that includes hundreds and thousands of genes. Another approach (Ernest Shtatland & Timur Shtatland, unpublished (2004)) builds an optimal regression model in the following stepwise manner. First, we find AIC-optimal one-gene

regression model (among tens of thousands). For each next step *k* we add to the previous model with *k-1* genes the next gene to get AIC-optimal model with *k* genes. Each step is performed by using a simple macro, and we proceed until we get the model size we need. The problem with this method of model building is that the resulting model may be suboptimal. For other approaches see also Li & Gui (2004) and Pawitan et al. (2004).
A reasonable approach is:
(A) Creating a list of candidate genes (sources: cancer literature, microarray data, genomic databases and molecular biology information) by narrowing initial pool of about 30,000 to about 250. In spite of the results obtained in Nguyen & Rocke (2002), Park et al. (2002), Li & Gui (2004) and Pawitan et al. (2004), we think it is questionable to use Cox regression with thousands (up to 30, 000) of genes. Also we think that working with such a large number of genes as 250 or so in Cox regression is questionable too since the number of observations (patients) is usually too small in human cancer research problems. (B) The next step should be validating candidate genes by using RT-PCR. As a result we can decrease the number of genes from hundreds (for example, 250) to tens (for example, 20 or even less). In addition to validation, RT-PCR provides precise quantification of the microarray results which is important for the next step.
(C) With these 20 gene variables plus clinical variables we can use Cox regression, in particular apply our Stepwise-AIC-Best Subsets approach. It is worth noting that there is a well-known rule of thumb when working with logistic or Cox regression. This is the **1 in 10 rule**: approximately 1 candidate predictor can be studied for every 10 events (death, relapse, etc). The number of events is usually smaller than the number of patients, thus the 1: 10 rule is a much stricter limitation than 1 predictor for every 10 patients. Also there are more rigorous limits: the **1: 20** and **1: 50 rules**. The 1: 10 rule is actually the minimum for reliable modeling. When the 1:10 rule is violated, the number of candidate predictors is in fact too large for the data set, and overfitting will occur. Note that the 1: 10 rule is kept very rarely by the researches in the gene expressions and human cancer field. A typical example is a paper by Jenssen et al. (2002), which is solid with the exception of the model they built. The authors of this article build a Cox regression model for 56 patients and 11 predictors (genes plus clinical variables). Taking into account that the number of events should be smaller than 56, we can see that even the most liberal rule 1: 10 is severely violated. Thus the model built in Jenssen et al. (2002) can be absolutely unreliable. It is worth adding that the usually low numbers of patients (and consequently, events) in combination with comparatively large numbers of candidate gene variables make difficult or even impossible using stepwise Cox regression. This makes model building for Cox regression in gene expression analysis even more necessary and urgent. We suggest a new approach to model building.

SUMMARY OF OUR APPROACH: BUILDING A MODEL

In spite of some disadvantages of variable selection procedures, including stepwise procedures mentioned above, the demand for variable selection will be strong and it will continue to be a basic strategy for data analysis. In our presentation, we show how to improve variable selection capabilities within PROC PHREG.
1) We propose to use AIC at all steps of building the final model, since AIC is asymptotically equivalent to cross-validation and the bootstrap, two most popular

validation methods. When we work with AIC we are trying to mimic cross-validation / bootstrapping results without performing both techniques.

2) We demonstrate a shrinkage technique that requires neither cross-validation nor bootstrapping and is based on elements of the standard PROC PHREG output. Shrinkage is very important in getting realistic prediction.

3) Averaging of the models from the blanket is also possible as in Shtatland et al. (2004) for PROC LOGISTIC. Averaging almost always helps performance.

ACKNOWLEDGEMENTS

The authors are grateful to Irina Miroshnik for her help in writing the macro, and to Timur Shtatland for his discussion of bioinformatics topics.

REFERENCES:

Allison, P. D. (1995). *Survival Analysis Using the SAS® System: A Practical Guide*. Cary, NC: SAS Institute Inc.

Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, **19**, 453-473.

Derksen, S. & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noisy variables. *British Journal of Mathematical and Statistical Psychology,* **45**, 265-282.

Efron, B. (1988). Logistic regression, survival analysis and the Kaplan-Meier curve. *Journal of the American Statistical Association*, **83**, 414 - 425.

Harrell, F.E., Lee, K.L., and Mark, D.B. (1996). Multivariate prognostic models: issues in developing models, evaluating assumptions and accuracy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361-387.

Harrell, F. E. (2001). *Regression modeling strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer-Verlag New York Inc.

Heinze, G. and Schemper, M. (2003). Comparing the importance of prognostic factors in Cox and logistic regression using SAS. *Computer Methods and Programs in Biomedicine*, **71,** 155-163.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression,* 1st edition. New York: John Wiley & Sons, Inc.

Hosmer, D. W. & Lemeshow, S. (1999). *Applied Survival Analysis,* New York: John Wiley & Sons, Inc.

Iwao, K., Matoba, R., Ueno, N., et al (2002). Molecular classification of primary breast tumors possessing distinct prognostic properties. *Human Molecular Genetics*, **11**, No 2, 199-206.

Jenssen, T., Kuo, W. P., Stokke, T., et al. (2002), Associations between gene expressions in breast cancer and patient survival. *Human Genetics*, 111, 411-420.

Li, H. and Gui, J. (2004). Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, **20**, 0000 - 0000.

Lossos, I. S., Czerwinski, D. K., Alizadeh, A. A., et al. (2004). Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *The New England Journal of Medicine*, **350**, 25, 1828-1837.

Lu, J. (2003). Modeling customer lifetime value using survival analysis – an application in the telecommunication industry. SUGI '28 Proceeding, Paper 120-28, Cary, NC: SAS Institute, Inc.

Mayo, M. S., Kimler, B. F. and Fabian, C. J. (2001). Evaluation of models for the prediction breast cancer development in women at high risk. *The Journal of Applied Research in Clinical and Experimental Therapeutics*, **1**, Issue 1, 1 – 22.

Nguyen, D. V. and Rocke, D. M. (2002). Partial least squares proportional hazard regression for application to DNA microarray data. *Bioinformatics*, **18**, 1625 - 1632.

Park, P. J., Tian L. and Kohane, I. S. (2002). Linking expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, S120 – S127.

Pawitan, Y., Bjohle, J., Wedren, S., et al. (2004) Gene expression profiling for prognosis using Cox regression. *Statistics in Medicine*, **23**, 1767 -1780.

SAS Institute Inc. (1995). *Logistic Regression Examples Using the SAS® System*, Version 6, First Edition, Cary, NC: SAS Institute Inc.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, **7**, 375 - 394.

Shtatland, E. S., Barton, M. B., and Cain E. M. (2001). The perils of stepwise logistic regression and how to escape them using information criteria and the Output Delivery System. SUGI '26 Proceeding, Paper 222-26, Cary, NC: SAS Institute, Inc.

Shtatland, E. S., Kleinman K., and Cain E. M. (2003). Stepwise methods in using SAS® PROC LOGISTIC and SAS® ENTERPRISE MINER for prediction. SUGI '28 Proceeding, Paper 258-28, Cary, NC: SAS Institute, Inc.

Shtatland, E. S., Kleinman K., and 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.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell Jr, F. E., and Habbema, J. D. F. (2000). Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, **19**, 1059 -1079.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B,* **39**, 44 -47.

CONTACT INFORMATION:

Ernest S. Shtatland
Department of Ambulatory Care and Prevention
Harvard Pilgrim Health Care & Harvard Medical School
133 Brookline Avenue, 6th floor
Boston, MA 02115
tel: (617) 509-9936
email: ernest_shtatland@hphc.org