Paper 191-29

A NEW STRATEGY OF MODEL BUILDING IN PROC LOGISTIC WITH AUTOMATIC VARIABLE SELECTION, VALIDATION, SHRINKAGE AND MODEL AVERAGING

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

This paper is a further development of our work presented at SUGI'26 and SUGI'28 in which we proposed an approach to model building for prediction based on the combination of stepwise logistic regression, information criteria, and the best subset selection. The approach inherited some strong features of the three components mentioned above, in particular it helped to avoid the agonizing process of choosing the "right" critical p-value in stepwise regression. At the same time automatic selection procedures are often criticized severely for instability and bias in regression coefficients estimates, their standard errors and confidence intervals. To avoid these disadvantages we propose to use some validation, shrinkage and averaging techniques. Note that these techniques are still not incorporated in SAS® PROC LOGISTIC. Meanwhile, variants of the techniques suggested in our presentation are based on the standard PROC LOGISTIC output. Together, they make a new robust model building strategy.

INTRODUCTION

Currently, stepwise selection methods (including best subset selection) are probably the most widely used in medical and other applications (see, for example, Steyerberg, et al. (2000)). The reason for their popularity is that stepwise methods automatically and with the appearance of objectivity select a limited number of predictors which may be considered important in a prediction problem. However, very important disadvantages are reported for these methods (for a more complete list of drawbacks, see, for example, Derksen & Keselman (1992), Harrell et al. (1996), Steyerberg et al. (2000), Steyerberg et al. (2001), Harrell (2001), Miller (2000), etc.) The basic drawbacks are:
(a) The selection is typically unstable, sensitive to small perturbations in the data. Addition or deletion of a small number of observations can change our chosen model markedly (see Breiman, (1995), and Steyerberg et al. (2000)).
(b) Standard errors of regression coefficients are biased low, confidence intervals are too narrow, p-values are too small, and R^2 or analogous measures are inflated.
(c) Regression coefficients are biased high in absolute value.
Drawbacks (a) – (c) can lead to a substantial decrease in predictive ability. As a result, Harrell et al. (1996), Steyerberg et al. (2000), Steyerberg et al. (2001), and Harrell (2001) recommend either not to use stepwise regression at all or to use it with a high critical P-value, e.g., 50%. This recommendation is hardly a realistic one, especially if we have a large number of covariates – a very typical situation in problems for prediction. For example, in microarray data analysis, a new and growing field of applications of logistic regression, where the number of predictors p (genes whose expression level is measured to predict a disease) is much greater than the number of observations N. Of course, this situation looks rather pathological from a conventional point of view in statistics. Logistic regression (or any other discriminant analysis technique) can be applied only after dimension reduction by principal component analysis or partial least squares (e.g., Nguyen & Rocke (2002) and supplementary SAS code therein). But more commonly, with $N >> p$, the best way to handle the situation with a large number of predictors is to use automatic selection methods in combination with validation techniques, shrinkage and averaging. See for example, in Simonoff (2000): "I would argue that automatic methods are crucial, and the key for

statisticians should be to get people to validate their models and correct for selection effects." Note that in Shtatland et al. (2003) we developed a three-step procedure, which incorporates the conventional stepwise logistic regression, information criteria, and finally best subsets regression (for a more detailed description see below). The approach softens some problems with stepwise selection, in particular it helps to avoid the agonizing process of choosing the "right" critical p-value in stepwise regression. A very important feature of our approach in Shtatland et al. (2003) is the fact that we used only standard elements of SAS/STAT®. In this presentation, we follow this approach and use validation techniques, shrinkage and averaging methods based on the already existing standard elements of SAS PROC LOGISTIC.

VALIDATION TECHNIQUES

Validation techniques are usually divided into two types: external and internal. When using external validation we apply the model we built to the completely new data deriving from the same source (i.e. having the same distribution). This will lead to a trustworthy evaluation of the prediction error and it is definitely the most stringent type of validation. But at the same time it is used in practice much less frequently because in most cases one is forced to use the same data for model building and its assessment. As a result, internal validation is much more ubiquitous. Among internal validation techniques the most popular are the bootstrap and cross-validation.

CROSS-VALIDATION

Cross-validation is rooted in the well-known phenomenon that estimating prediction error on the same data as used for model building tends to give downward-biased values. The reason for this is that parameters estimates are optimized to reflect the peculiarities of the given data set. It was Stone (1974, 1977) who gave one of the first clearly stated accounts on this topic. The idea of cross-validation is to divide the data set into two parts. The first one is used for building the model, the second one is used for validating. This is done repeatedly for a (possibly large) number of divisions. The results of the test procedures are combined properly and used for subsequent model selection and assessment. Partitioning of the data set is the key element of cross-validation (CV). According to the type of partitioning there are:
(1) Leave-one-out cross-validation (LOOCV) a.k.a. Delete-1 CV; (2) Leave-many-out CV (LMOCV), or Leave-M-out, with $1 \le M < N$ where N is a sample size; (3) V-Fold Cross-Validation (where V is the number of disjoint validation subsets), etc. All of the most popular forms of CV are nearly unbiased, but can exhibit high variance, especially in small samples (Ripley (1994), Efron & Tibshirani (1995)). Perhaps LOOCV is an appealing alternative to LMOCV to many practitioners because (a) it eliminates the need to select the value of M or V; (b) it limits the number of splits (to a total number of N); (c) it maximizes the number of observations (i.e., $N - 1$) assigned to each model building set; (d) LOOCV often works well for continuous error functions such as the mean squared error. However, these reasons can be outweighed by the better performance of LMOCV or V-fold cross-validation. Some statisticians prefer to use V-fold CV with V in between 3 and 10. 10-fold CV often works well, but sometimes the result is very sensitive to the partitioning used. Though LOOCV works worse than V-fold CV in general, it is especially important to us because of the close relationship between LOOCV and AIC (see below). In summary, cross-validation is a very powerful method in model building and assessment. Unfortunately, cross-validation is *not* part of SAS PROC LOGISTIC or any other SAS regression procedure (see, for example, Potts, and Patetta (1999). There are examples of "home-made" macros for cross-validation offered by independent authors. See for example, SAS macro CVLR (Cross-Validation for Logistic Regression) written by Moore (2000). Another example is the SAS macro Logit cross-validation.sas that performs leave-one-out cross-validation for logistic regression, described in Peterson (2002). The last macro was written mostly for educational purposes. However, both macros mentioned above cannot be considered an adequate substitute for standard SAS/STAT software. In such a situation, it is of great importance that the cross-validation technique LOOCV is asymptotically equivalent to AIC, as shown in Stone (1974, 1977) and rediscovered in many other books and articles on cross-validation (see, for example, Shao (1993)). Since AIC simulates the cross-validation situation, AIC can be used as a substitute for CV (more exactly, COOCV) for the large enough data sets. So it is natural to prefer AIC to any other information criterion in the cross-validation context. And we recommend to use AIC as a "fast and cheap" version of cross-validation at *each* step of our process of model building and assessment: in the stepwise regression part, in the best subset regression step, and finally, in shrinkage. Concluding our section on cross-validation we can add that cross-validation

is *not* the ultimate answer to model selection. Cross-validation only works well in a few situations, because sometimes it has bad asymptotic properties (see Minutes of Euro-workshop on Statistical Modeling Model Building and Evaluation (2002)). That is why we need additional techniques such as the bootstrap, shrinkage, etc.

BOOTSTRAPPING

The bootstrap was developed 10 years later than its competitor, cross-validation. The literature on the bootstrap is extensive. See, for example, Efron (1982), Efron and Tibshirani (1993), Shao and Tu (1995), Venable and Ripley (1994), etc. Among many bootstrap techniques, we could mention as very popular the following ones: .632 bootstrap and .632+ bootstrap. The "magic" number .632 comes from the probability that an observation is in a bootstrap sample, and is nothing but $1 - 1/e$.
And though it was reported in Minutes of Euro-workshop on Statistical Modeling Model Building and Evaluation (2002)) that so far the bootstrap is rarely used for model validation, it is exactly the bootstrap methods which gain more and more popularity in model validation. Note that like cross-validation, the bootstrap is not part of SAS PROC LOGISTIC or any other SAS regression procedure. There exist a number of bootstrap-based SAS macros for model building. For example, bootlogit, a macro to calculate bootstrap estimates of probabilities in logistic regression submitted by Johnson (2001) in StatLib. Another example is Jackknife and Bootstrap Analyses Macros provided by SAS Institute Inc., which includes %BOOT macro for regression models resampling either observations or residuals, or %BOOTCI, computing several varieties of confidence intervals. Here we can only repeat our remark about macros for cross-validation written by independent authors: they are provided "as is", with no warranties, expressed or implied. By the way, we have the same situation with some macros provided by SAS Institute Inc. (see the previous example). In conclusion:
(1) There is a widespread myth that bootstrapping is a magic spell to perform valid statistical inference on *anything* and under any circumstances, but it *is* just a myth.
(2) A combination of bootstrapping and shrinkage seems to be close to serving the practical needs in model building and validation, and is highly desirable.
(3) Bootstrapping and AIC (and also LOOCV) have been proven to be roughly asymptotically equivalent (see, for example, Shibata (1997), Simon et al. (2003), etc.) This is additional kudos to AIC. Below, we will discuss this topic in more detail.

CROSS-VALIDATION vs. BOOTSTRAPPING

A fundamental difference between cross-validation and bootstrapping is that the bootstrap re-samples the available data at random *with* replacement, whereas cross-validation does it *without* replacement. However, now there exist a number of hybrid methods – some combinations of cross-validation and bootstrapping. See, for example, a hybrid method realized in macro bcv.sas by Johnson (2001) in StatLib. As a result, fundamental distinctions between CV and the bootstrap are becoming blurred by recent hybrid methods that exploit advantages of both techniques. Cross-validation is simpler, more intuitively appealing and provides nearly unbiased estimate of the future error. The bootstrap is commonly believed to be more biased than cross-validation, but to have lower variance at least in small samples. It is interesting that the extent to which stepwise methods drawbacks a) - d) are widespread became clear only after developing the bootstrap and cross-validation. And the idea that it is desirable (if not necessary) that any data-driven model building procedure (including stepwise) be accompanied with validation step through bootstrap or cross-validation becomes common now. However, the bootstrap and cross-validation are expensive compared to information criteria (for example, AIC). According to Efron & Tibshirani (1995), .632+ bootstrap substantially outperforms cross-validation in a catalog of 24 simulation experiments.

AIC AND OTHER INFORMATION CRITERIA IN MODEL SELECTION

In Shtatland et al. (2003), we propose a three-step procedure. Firstly, we apply stepwise logistic regression with critical p-values SLENTRY and SLSTAY close to 1 (for example, 0.95 or 0.99, the exact value does not matter) to build a complete stepwise sequence, a trajectory in the model space, starting with the null model without predictors and finishing with the full model containing all K potential predictors. In doing so, we reduce the total number of $M=2^K$ potential candidate models to the manageable number of K

models. For example, if K = 20, we have 2^{20} = 1048576 vs. 20. A huge reduction in the model space! At the same time Step #1 helps us to avoid a painful process of choosing the "right" critical p-value. Only a partial list of the "right" critical p-values recommended in books and articles on logistic regression looks like this: 0.01, 0.05, 0.15, 0.25, 0.3, 0.4 or 0.5. Also, 0.05 is the SAS default value. Neither value is theoretically justified. But if we torture the data long enough, in the end they will confess, and the confession will usually be wrong. After our model space reduction we are able to apply any information criterion and compare the models. Secondly, we find the optimal model in this stepwise sequence with respect to a chosen information criterion. In doing so we use Output Delivery System statements in combination with PROC MEANS and PROC PLOT. Both steps 1 and 2 are automated. A general form of information criteria (IC) is

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

where logL(M) is the maximized 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)). But we use IC(3/2) and IC(1) in Shtatland et al. (2003) basically because we would like to have the implied critical p-value for stepwise regression up to 50%, the value recommended in Harrell et al. (1996), Steyerberg et al. (2000), Steyerberg et al. (2001), and Harrell (2001). 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 (1974, 1977**),** Shibata (1997), Simon et al. (2003), etc. And thirdly, we use the best subset selection in the neighborhood of the optimal model found in the step 2. Thus, instead of the single optimum we have a " bouquet" of good candidate-models, the output of the best subsets selection. It is interesting and somewhat surprising that AIC-optimal stepwise model may be better (in terms of AIC, not the SCORE statistic, of course) than any best subset selections model with the parameter BEST = 5, 7 or 10, for example. A natural question arises whether we can combine the first two steps and build a stepwise regression based completely on AIC. Thus, it is proposed to use as a criterion of "importance" of a variable the AIC statistic instead of the log-likelihood one, which is used in Hosmer and Lemeshow (2000) and SAS/STAT PROC LOGISTIC. Several authors, for example, Venables and Ripley (1999), Wang (2000), and Moore (2000), realized this idea in S-Plus, Stata, and SAS, correspondingly. Nevertheless, we prefer the conventional stepwise procedure which utilizes log-likelihood statistic as a criterion of importance (see SAS/STAT User's Guide, Version 8 (1999), PROC LOGISTIC and Hosmer and Lemeshow (2000)). This choice is "neutral" with respect to all information criteria used in Shtatland et al. (2003): AIC, IC(1) and I(3/2). We apply AIC only to the output of the stepwise procedure – the stepwise sequence, and to the output of the best subsets selection.

AIC CORRECTIONS FOR SMALL SAMPLES

When the sample size is small (the case of special interest in Harrell et al. (1996), Steyerberg et al. (2000) and Steyerberg et al. (2001)), AIC can lose its asymptotic optimal properties and become strongly biased. To overcome this weakness, two approaches have been taken. One approach is to use a corrected AIC, called AIC_C (Hurvich & Tsai (1989, 1991))

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

where N is the sample size and K is the number of predictors in the model. AIC_C usually outperforms AIC as a selection criterion in small-sample studies. AIC_C is generally recommended when $N / K < 40$ (Burnham and Anderson (2002), p.445). This AIC_C superiority over AIC is one more reason not to build stepwise regression based *exclusively* on AIC. Nevertheless, AIC is unique, mostly due to the close relationship with bootstrapping and cross-validating.

AIC vs. CROSS-VALIDATION and BOOTSTRAPPING

As mentioned above, neither bootstrapping nor cross-validation are part of SAS PROC LOGISTIC (see, for example, Potts, and Patetta (1999)). Fortunately, AIC is asymptotically equivalent to cross-validation and bootstrapping and can be considered a "fast and cheap" substitute for both validation methods. Thus, not having cross-validation or bootstrapping directly in SAS PROC LOGISTIC we should use AIC in model building as a substitute for both. In doing so we are on a safer side as to validation. If the sample size is not large enough to apply an asymptotic approach, the small-sample correction AIC_C should be used. Another approach is based on the bootstrap method. The idea of the bootstrap bias correction, called EIC (an extended information criterion), was formalized as a model selection criterion by Ishiguro, Sakamoto & Kitagawa (1997) and Shibata (1997). Simulation results show that EIC largely overcomes AIC's overfitting problem in model selection with small samples. Also Shibata (1997) proved that EIC is asymptotically equivalent to AIC (though EIC was developed for a small-sample situation). At the same time, simulation results in Liquet et al. (2003) demonstrate that differences between EIC and AIC_C are very small in many situations. Thus, a combination of AIC_C and AIC can be viewed equivalent to the bootstrap-based EIC. In Johnson (2001), we can find SAS macro for calculating the AIC, CP, PRESS, a bootstrap-based extension to AIC (EIC), a bootstrap-smoothed likelihood cross-validation (BCV) and its modification (632CV). This macro, called bcv.sas is *not* incorporated into SAS/STAT standard software. Concluding this section we can say that when we work with AIC / AIC_C, we are trying to mimic cross-validation / bootstrapping results *without* performing both techniques. And the combination AIC / AIC_C will allow the assessment of whether any step up or down in an automatic model selection procedure would be expected to improve prediction (the approach discussed in Clayton and Hills (1993)). After using automatic model selection procedures in combination with AIC / AIC_C, we have to decide if we need shrinkage.

HOW TO PERFORM SHRINKAGE IN SAS PROC LOGISTIC

As we have seen above one of the drawbacks of the stepwise methods is that regression coefficients are biased high in absolute value. According to Copas (1983), shrinkage is particularly marked when stepwise fitting is used. To cope with this problem, shrinkage methods were proposed as a companion to the subset selection. The shrinkage seems also to soften instability of the stepwise methods (Sauerbrei (1999)). In general, shrinkage of regression coefficients may improve the performance of a prognostic model substantially. According to Steyerberg & Harrell (2000), when the sample size is large, *no shrinkage will be required*, in contrast to a small data set, where a substantial shrinkage may be needed. More exactly, if the number of predictors over the number of observations is less than 1/10, shrinkage is necessary, if this ratio is between 1/10 and 1/20, shrinkage is advisable, and if the ratio is smaller than 1/20, shrinkage is not necessary. Of course this rule is empirical. All shrinkage methods rely on a parameter: the ridge parameter for ridge regression, the garrote parameter for the nonnegative garrote (Breiman (1995)), and the tuning parameter for the lasso (Tibshirani (1996)) which was inspired by Breiman (1995). In each case the parameter value significantly influences the result. Breiman's garrote and Tibshirani's lasso are not available in SAS/STAT. Luckily, we have two other shrinkage methods that are "ready-to-go" in SAS PROC LOGISTIC. The first one is based on an heuristic linear shrinkage factor

$$(\text{Model } \chi^2 - K) / \text{Model } \chi^2 \qquad (3)$$

where Model χ^2 is Model chi-square and K the number of parameters in the model M (see Copas (1983) and Van Houwelingen & Le Cessie (1990)). Formula (3) can be re-written in the form

$$\lambda=(2\log L(M)-2\log L(0)-K)/(2\log L(M)-2\log L(0)) \qquad (3a)$$

The shrinkage factor (3, 3a) 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

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

Note that the AIC-based shrinkage factor (4) is smaller than (3a), which means that (4) provides more effective shrinking. It is also interesting that (4) is closely related to the R^2 measures discussed in Shtatland et al. (2002):

$$R^2 = 1 - \log L(M) / \log L(0) \qquad (5)$$

and

$$\text{Adj-}R^2 = 1 - (\log L(M) - K - 1)/(\log L(0) - 1) \qquad (6)$$

The R^2 measure (5) is known as McFadden or deviance or entropy R^2. (6) is its adjustment for the number of predictors. From (4), (5) and (6) it is not difficult to show the following approximate equation

$$\lambda_{AIC} \approx \text{Adj-}R^2 / R^2 \qquad (7)$$

i.e., the heuristic shrinkage factor λ_{AIC} has the meaning of the ratio of adjusted and unadjusted R^2 measures.

Shrunken regression coefficients are calculated by multiplication by the shrinkage factors (4) and (7). Note that 2logL(M) and 2logL(0) are the elements of the standard output of the SAS PROC LOGISTIC. So the heuristic linear shrinkage discussed above can be easily implemented in SAS PROC LOGISTIC by using ODS. In Steyerberg et al. (2000), the authors found no major differences between application of a linear shrinkage factor and a more advanced shrinkage technique, Tibshirani's lasso. In one of our models with 2629 observations and 45 predictors, the shrinkage factor λ_{AIC} is approximately equal to 0.81, which is rather substantial. It means that roughly 19% of the model fit is noise. According to Harrell et al. (2000), if the estimated shrinkage falls below 0.85, for example, we should be thinking about applying shrinkage. At the same time, according to Steyerberg & Harrell (2000), shrinkage is not necessary since the ratio $45/2629 < 1/50$. From this, we can conclude that the final decision about shrinkage is up to the researcher. The second shrinkage technique, which can be used in SAS PROC LOGISTIC, is based on averaging. See the next section about averaging in more detail, and in a more general context.

AVERAGING AND GETTING MORE ROBUST PREDICTIONS

It is well-known that such modeling procedures as neural networks, classification and regression trees (CART), and variable selection for regression are unstable, that is sensitive to small changes in data. To overcome such a disadvantage, Breiman (1996a) introduced a special method, named bagging (bagging is short for "Bootstrap AGGregatING"). The idea of the method is straightforward. Instead of making predictions from a single model fit to the observed data, bootstrap samples of the data are taken, and predictions are averaged over all of the fitted models to get the bagged prediction. According to Perlich et al. (2001), bagging has been applied widely to machine learning techniques, but it has rarely been applied to statistical tools such as logistic regression. Still, bagging in logistic regression is straightforward. That is, one creates B random sub-samples with replacement from the original data set and estimates for each of them the logistic model. The prediction for an observation is the mean of the B predictions. Perlich et al. (2001) provides an example of 10-fold bagged logistic regression in which they use 10 subsamples with replacement of the original data set, estimate 10 logistic regression models and take the mean of the probability predictions as the final probability prediction. It is important that bagging may become necessary and quite effective if we use stepwise selection (including best subsets selection) methods in logistic regression. As mentioned above, bootstrapping is not incorporated in SAS PROC LOGISTIC. That is why we propose a different approach to averaging. We average the prediction probabilities for the bouquet of candidate-models obtained as an output of best subsets selection, plus the AIC-optimal model in the stepwise sequence. According to Tukey (1995) in discussion of the paper by Draper (1995), p.78, we can only afford a *small* number of separate alternatives, i.e. our bouquet of models should contain a reasonably small number of models. The proposed formula for averaging predictions is

$$P_{AVE} = \Sigma w_j \, (1/(1 + \exp(-b_{0j} - b_{1j} X_1 - b_{2j} X_2 - \ldots - b_K j X_K)) \qquad (8)$$

where weights w_j may be either equal or AIC-related: $w_j = \exp(-AIC/2)$ (see, for example, Shtatland (2001) and references therein). Using averaging by (8) makes our predictions more reliable. Another approach is to average the coefficients themselves in (8). The reason is that the coefficients in logistic regression have very simple interpretation as a logarithm of odds-ratios, and odds-ratios are very popular in such applications as clinical studies, health care research, etc.

ALTERNATIVES TO AVERAGING

Alternative variants to averaging are:
(a) Choosing the AIC optimal model from the best subsets regression bouquet (based on *statistical* consideration only);
(b) Keeping all the models from the bouquet in order to make our pick later based on considerations other than the *statistical* one (for example, medical, financial, etc.);
(c) Building the final model including *all the covariates* from the bouquet models. This model may contain many more predictors than the final models in the two previous cases.

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 SAS PROC LOGISTIC.
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 develop a shrinkage technique that requires neither cross-validation nor bootstrapping and is based on elements of the standard SAS PROC LOGISTIC output. In particular, we have obtained the formula for the shrinkage factor in terms of both adjusted and unadjusted R-Square measures in logistic regression (Shtatland et al. (2002)). Shrinkage is very important in getting realistic prediction.
3) We have proposed a method that provides us with more robust prediction: averaging of good models (built through stepwise and best subsets) with some weights, either equal or AIC-related. Averaging almost always helps performance. Some alternatives to averaging are proposed as well.

IS OUR APPROACH APPLICABLE TO SURVIVAL ANALYSIS (PROC PHREG)?

The answer is YES. PROC PHREG shares 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 criterion. However, in PROC PHREG we can find only the *values* of AIC and SBC in printouts without any theoretical background and definition. Also, some information about close relationship between logistic regression and Cox regression can be found in Allison (1995), p. 6, Altman and Royston (2000), Efron (1988), Hosmer and Lemeshow (1989), pp. 238 - 245, and Mayo, Kimler, and Fabian (2001). Such similarity 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, in PROC PHREG we have the same elements that allow us to develop an approach similar to the one discussed above for PROC LOGISTIC. This work is currently in progress. This approach is *absolutely necessary* when the number of predictors is large (see for example, Lu (2003), where this number is above 300).

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.

Breiman, L. (1995). Better subset regression using the nonnegative Garotte. *Technometrics*, **37**, 373-384.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 26, 123-140.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition. New York: Springer-Verlag.

Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology.* Oxford: Oxford University Press.

Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society, Series B* **45**, 311-354.

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. (1982). *The Jacknife, the Bootstrap, and other Resampling Plans*. SIAM, Philadelphia.

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

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Efron, B. & Tibshirani, R. (1995). Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Technical report 176, Department of Statistics, Stanford University.

George, E. I. (2000). The variable selection problem. Technical Report. Dept. of MSIS, University of Texas, Austin. September 2000.

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.

Harrell, F. E. (2003). *Regression Modeling Strategies*. Pfizer Global Research and Development, Statistical Research Center • Clinical Technology / Computational Medicine, New London CT, 13-14 January 2003.

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

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression,* 2nd edition. New York: John Wiley & Sons, Inc.

Hurvich, C.M. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297 - 307.

Hurvich, C.M. and Tsai, C. (1991). Bias of the corrected AIC criterion for uderfitted regression and time series models. *Biometrika*, **78**, 499 - 510.

Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, 49, 411-434.

Jackknife and Bootstrap Analyses Macros, SAS Institute Inc. www.asu.edu/it/fyi/dst/helpdocs/statistics/sas/tips/stats/jackboot.html

Johnson, P. (2001). StatLib. http://lib.stat.cmu.edu/general/.

Liquet, B., Sakarovitch, C. and Commenges, D. (2003). Bootstrap choice of estimators in non-parametric families: an extension of EIC. *Biometrics* (in press).

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, vol. 1, Issue 1, 1 – 22.

Miller, A. (2002). *Subset Selection in Regression,* 2nd Edition. CRC Press LLC.

Minutes of Euro-workshop on Statistical Modeling Model Building and Evaluation (2002). http://www.stats.gla.ac.uk/~goeran/euroworkshop/webpages/2002/minutes.html

Moore, C. (2000). http://www.arches.uga.edu/~ctmoore/prof/CVLRnotes.htm

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39 - 50.

Pan, W. (1999). Bootstrapping likelihood for model selection with small samples. *Journal of Computational and Graphical Statistics*, **8**, 687- 968.

Peterson, J. T. (2002). http://coopunit.forestry.uga.edu/unit_homepage/Peterson/fors8360/labs/lab2

Potts, W.E. and Patetta, M.J. (1999). *Logistic Regression Modeling*. Cary, N.C.: SAS Institute Inc.

Ripley, B.D. (1994). *.Address to the Conference on Neural Networks in the Capital Markets*. Pasadena, California, USA. November, 1994.

Ripley, B.D. (2002). *Model Choice. Oxford Graduate Lecture*. Oxford University, Dept of Applied Statistics, December 9, 2002.

SAS/STAT *User's Guide, Version 8.* Cary, NC: SAS Institute Inc., 1999.

Sauerbrei, W. (1999). The use resampling methods to simplify regression models in medical statistics. *Applied Statistics*, **48**, 313-329.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486 - 494.

Shao, J., & Tu, D. (1995). *Jackknife and Bootstrap*. Springer-Verlag New York Inc.

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

Shtatland, E. S., Moore, S. & Barton, M. B. (2000). Why we need R^2 measure of fit (and not only one) in PROC LOGISTIC and PROC GENMOD. SUGI 2000 Proceeding, 1338-1343, Cary, NC, SAS Institute, Inc.

Shtatland, E. S., Kleinman K., and Cain E. M. (2002). One more time about R^2 measures of fit in logistic regression. NESUG 2002 Proceeding, 742 - 747, NorthEast SAS Users Group, Inc.

Shtatland, E. S., Moore, S., Dashevsky I., Miroshnik, I., Cain, E. M. and Barton, M. B. (2000). How to be Bayesian in SAS: model selection uncertainty in PROC LOGISTIC and PROC GENMOD. NESUG 2000 Proceeding, 724 -732, NorthEast SAS Users Group, Inc.

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.

Simon, G., Lendasse, A., Wertz, V. & Verleysen, M. (2003). *ESANN'2003 – European Symposium on Artificial Neural Networks*. Springer-Verlag, Berlin, Heidelberg.

Simonoff, J. S. (2000). http://www.biostat. wustl.edu /archives/html/s-news /2000-11/msg00184.html

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.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell Jr, F. E., and Habbema, J. D. F (2001). Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets.
Medical Decision Making, **21**, 45 -56.

Steyerberg, E. W. and Harrell Jr, F. E. (2001). Chapter 8: Statistical Models for Prognostication http://symptomresearch.com/chapter_8/cesauthorbio.html

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, *Series B*, **36**, 111-147.

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.

van Houwelingen, J.C. and le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine*, 9, 1303-1325

Venables, W.N. and Ripley, B.D. (1999). *Modern Applied Statistics with S-Plus*, 3rd ed. Springer, New York.

Voit, E. O.and Knapp, R. G. (1997). Derivation of the linear-logistic model and Cox's proportional hazard model from a canonical system description. *Statistics in Medicine*, **16**, 1705 -1729.

Wang, Z. (2000). Model selection using Akaike information criterion. *STATA Technical Bulletin*, **54**, 47-49.

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