Paper 258-28

STEPWISE METHODS IN USING SAS® PROC LOGISTIC AND SAS® ENTERPISE MINER™ FOR PREDICTION

Ernest S. Shtatland, Ken Kleinman, and Emily M. Cain

Harvard Medical School, Harvard Pilgrim Health Care, Boston, MA

ABSTRACT

In this presentation, which is a sequel to our SUGI'26 paper, we demonstrate that if the goal of modeling is prediction, with a large number of covariates and little theoretical guidance for choosing among them, our approach based on the combination of stepwise logistic regression, information criteria, and best subset selection will result in fully automated procedure (due to ODS). The approach inherits the best features of the three components mentioned above and helps us avoid an agonizing process of choosing the "right" critical p-value. If we apply the approach to Enterprise Miner, we can strengthen the Regression node in comparison with other modeling nodes (the Neural Network and Tree).

The intended audience: SAS users of all levels who work with SAS/STAT® and PROC LOGISTIC in particular and Enterprise Miner.

THE EXISTING AUTOMATIC MODEL SELECTION TECHNIQUES IN SAS

Model selection is a fundamental task in data analysis. The process of selecting a subset of variables from a typically large number of variables, called model building, is particularly important in prediction. In SAS PROC LOGISTIC, there are three automatic model selection techniques: forward selection, backward elimination and stepwise selection which combines the elements of the previous two. All these procedures are intuitively appealing: they build models in a sequential manner and allow for examination of a collection of models *which might not otherwise have been examined.* Basically this is a matter of taste which one of the three sequential procedures to use. According to Everitt and Der (1996), "In the best of all worlds the final model selected by each of these procedures would be the same. This does often happen, but it is in no way guaranteed". Sequential procedures involve selection and stopping criteria. Selection criteria are based on the likelihood ratio statistics or their derivatives. The selection part is more or less straightforward. Choosing a critical P-value = α which determines a stopping rule is a real problem. A default P-value in SAS is $\alpha = 0.05$ and it has been used too often, intentionally or not. This choice has been criticized by many authors as absolutely inadequate for both prediction and interpretation purposes (see, for example, Shtatland, Cain, and Barton (2001) and references therein). By using Monte Carlo simulations Lee and Koval (1997) show that the best α varies between 0.05 and 0.40. At the same time, Steyerberg *et al.* (2000) recommend using $\alpha = 0.50$ to include all useful variables for a better prediction. Combining these recommendations we can conclude that the recommended interval for α should be $0.05 \le \alpha \le 0.50$. This is a very large interval and choosing the "right" value is an agonizing process that requires many trials, especially because there is no theory at all behind *any* choice of α. Derksen and Keselman (1992) describe this situation as follows: "If you torture the data for long enough, in the end they will confess… What more brutal torture can there be than subset selection? The data will always confess, and the confession will usually be wrong." To resolve this tormenting problem by putting it on a more theoretical basis we propose to use information criteria.

MODEL SELECTION AND INFORMATION CRITERIA

The basic idea behind the information criteria is penalizing the likelihood for the model complexity (the number of explanatory variables used in the model). The most popular in this family are the Akaike information criterion (AIC) and Schwarz information criterion (SIC). The most general form of information criteria is

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

where logL(M) and logL(0) are the maximized log likelihood for the fitted model and the "null" model containing only an intercept term, N is the sample size, K is the number of covariates (including the intercept) and c is known as a penalizing parameter. The AIC and SIC can be defined as information criteria with $c = 2$ and $c = \log N$

correspondingly. If $c = 0$ (no penalty), (1) is equivalent to the classical likelihood statistic on which the "importance" of variables is based in stepwise logistic regression (Hosmer and Lemeshow (2000), pp 116-128). If $c = 1$, (1) is equivalent to the GLIM goodness-of-fit procedure based on plotting the deviance against degrees of freedom (see Nelder and Wedderburn (1974) and Smith and Spiegelhalter (1980)). See also Box and Kanemasu (1973) whose entropy-based criterion is also equivalent to IC(1). Smith and Spiegelhalter (1980) show that their version of the local Bayes factors corresponds to IC(3/2). The question of which value of the parameter c to choose is not easy. Atkinson (1981) suggests the range between 2 and 6. Most likely, this is the right suggestion if we build the model for interpretation purposes. Since we are interested in prediction, as always the case when we work with Enterprise Miner, we should use $c \le 2$ (AIC and more liberal criteria). It is known (Smith and Spiegelhalter (1980)) that values of $c < 1$ tend to favor complex models unduly. At the same time, if $c > 1$ then smaller models are favored over complex models (Laud and Ibrahim (1995)). Thus, a borderline $c = 1$ is a sensible choice, especially for prediction problems and the interval of interest for prediction becomes more narrow: $1 \le c \le 2$. From the interval $1 \le c \le 2$, we recommend to use the following landmarks: $c = 1$ (GLIM-value), $c = 3/2$ (the local Bayes factor value) and $c = 2$ (AIC-value). Among these three candidates, AIC is undoubtedly choice #1. Though we will see below that $c = 1$ (and also $c = 3/2$) is also very useful when we would like to keep more variables than AIC does. AIC has some minimax properties for prediction over the experimental region (Atkinson (1981)). Also, as shown in Stone (1977), AIC is asymptotically *equivalent* to the cross-validation criterion, which is a very important property. And last but not least, model comparisons based on AIC are asymptotically *equivalent* to those based on Bayes factors under the assumption that the prior information is comparable to the information in the likelihood (i.e. in the data), see Kass and Raftery (1995). At the same time, $c = 1$ is unique as a border between favoring complex models and simple ones. As we will see below, I(3/2) and especially IC(1) are very useful and should be added to our model selection kit.

MODEL BUILDING, STEP 1: CONSTRUCTING A FULL STEPWISE SEQUENCE

Whatever IC criterion we have decided to work with we encounter a very serious problem: the process is not automated. The method of calculating AIC = IC(2) or IC(1) or IC(3/2) for *every possible* sub-model with subsequent direct comparison is absolutely impractical even with a moderate or moderately large number of variables. 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*. 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 entire sequence. In doing so, we reduce the total number of $K=2^P$ potential candidate models to the manageable number of P models. In one of our examples
(not the largest one) 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.

MODEL BUILDING, STEP 2: MINIMIZING INFORMATION CRITERIA - IC(1), IC(3/2) AND IC(2) ON THE FULL STEPWISE SEQUENCE

If we use PROC LOGISTIC with the following ODS statements:

```
ods output ModelBuildingSummary=SUM;
ods output FitStatistics=FIT;
```

we get the data sets SUM and FIT that contain Summary of Stepwise Procedure and by default the values of -2log L and AIC statistics for each member of the stepwise sequence. We also obtain the values of Schwarz information criterion (SIC), but we are not interested in them since our sole goal is prediction. By using statistic –2logL we can calculate any information criterion, including IC(3/2) and IC(1). Then using PROC MEANS and the MERGE statement, it is easy to find the minimum of AIC, IC(3/2) and IC(1) and the corresponding AIC-, IC(3/2)-

and IC(1)- optimal submodels. Also, it is strongly recommended to apply PROC PLOT and visualize the behavior of AIC, IC(3/2) and IC(1) *vs.* the number of predictors in the model, i.e. the step number in stepwise regression. In all our examples below, we can see from the plot that AIC has a unique distinct minimum, which clearly indicates the AIC-optimal model. The situation with IC(3/2) is very similar. As to IC(1), this criterion has usually a more or less large plateau with no distinctive (clear) minimum. In this situation, it is reasonable to address the R^2 measure corresponding to IC(1)

$$\text{Adj-}R^2_{IC(1)} = 1-(2\ \text{logL}(M) - k - 1)\ /(2\text{logL}(0) - 1) \quad (2)$$

where k is the number of covariates without an intercept. About this adjustment see Mittlbock and Schemper (1996), Menard (1995), p. 22, and Shtatland, Kleinman and Cain (2002) (in a similar way we can define Adj-R^2_{AIC}). Adj-$R^2_{IC(1)}$ is equivalent to IC(1): their graphs mirror each other. Nevertheless, we prefer to work with Adj-$R^2_{IC(1)}$ rather than IC(1). The reason for this is that IC(1), like AIC, takes rather arbitrary values: from very large positive to very large negative, and these values are hard to interpret. At the same time adjustment (9) takes values between 0 and 1 which are easier to interpret. This is why we suggest to use Adj-$R^2_{IC(1)}$ and Adj-R^2_{AIC} at least as a supplement to IC(1) and AIC (if not instead of IC(1) and AIC). Working with Adj-$R^2_{IC(1)}$ it is much easier to find the boundaries of the around-the-maximum plateau than the boundaries of around-the-minimum IC(1) plateau (actually, we need only the right boundary) This procedure can be automated by rounding the values of Adj-$R^2_{IC(1)}$ e.g., up to 0.001 and finding the *last* maximum of the rounded Adj-$R^2_{IC(1)}$.

MODEL BUILDING, STEP 3: INFORMATION CRITERIA AND IMPLIED CRITICAL P-VALUES FOR STEPWISE PROCEDURES: WHY WE NEED INFORMATION CRITERIA OTHER THAN AIC

It is of paramount interest to know the relationship between AIC and stepwise logistic regression. It has been shown in Hosmer and Lemeshow (1989), p.184 that if one has to pick a *unique* critical P-value, one should choose it around 0.15. This value $\alpha = 0.15$ (0.157 to be exact) is mentioned in Steyerberg *et al*. (2001). Also it appears in Lee and Koval (1997) as the end of a narrow interval $0.15 \le \alpha \le 0.20$ recommended to use. All these suggestions are empirical, based either on Monte Carlo simulations or some particular examples. Thus the "magic" number of 0.15 or 0.157 cannot be explained within the theory of stepwise logistic regression, but it is naturally understood within the information criteria theory. Atkinson (1980) and Lindsey & Jones (1998) show that *asymptotically* AIC is equivalent to a stepwise procedure with a critical P-value of 0.1573. Steyerberg *et al.* (2001) also associate $\alpha = 0.157$ with AIC (though it is an improper association since the authors consider only small data sets and asymptotic considerations needed for $\alpha = 0.157$ are hardly applicable). The implied significance level α is known to vary for AIC from 0.30 to 0.157 as the sample size increases (Sawa, (1978)). Consequently, to cover the combined critical P-value interval $0.05 \le \alpha \le 0.50$ we have to involve information criteria other than AIC. It seems that subinterval $0.05 \le \alpha < 0.15$ is less important since in problems for predicting it is better to over-parameterize the model than to under-parameterize it. To cover subinterval $0.30 \le \alpha \le 0.50$ we suggest to use another popular information criteria, IC(1), which is much more liberal than AIC.

EXAMPLES

We have considered how our procedure works in 5 cases with the same sample of 2629 and the numbers of original variables: 30, 34, 36, 41 and 46. The implied AIC_{MIN} critical p-values vary between 0.1142 and 0.1472, which is reasonably close to the asymptotic value of 0.1572. It means that sample size of 2629 is large enough for asymptotic considerations. The implied $IC(1)_{MIN}$ critical p-values vary between 0.3023 and 0.3166, which is very close to the landmark p-value of 0.30. And the implied critical values for $IC(1)_{RB}$ (the *right boundary* of the around-the-minimum IC(1) plateau) vary between 0.4903 and 0.5541, which is close to another landmark of 0.50. Thus, we see that AIC and IC(1) combined, over-cover subinterval $0.15 \le \alpha \le 0.50$. And the most important is that according to our methodology we are not using *the prescribed* α of 0.15, or 0.30, or 0.50 in stepwise regression, but rather we are guided by the information criteria minimum principle and *our data*. It is interesting that when we move from AIC_{MIN} model to $IC(1)_{MIN}$ one, we add a number of potentially important predictors. The same we do when moving from $IC(1)_{MIN}$ model to $IC(1)_{RB}$ one. It is investigator's job to determine which of the mentioned above added variables are not only statistically, but also substantially important.

SHOPPING AROUND IC(1)- AND AIC- OPTIMAL MODELS BY USING BEST SUBSET SELECTION

Obviously, it would be too simplistic to recommend AIC- or IC(1)- optimal models as the best models for prediction. First of all, there could be a number of nearly optimal

models in the vicinity of AIC- and IC – optimal choices. 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. These problems can be resolved by using the best subset selection procedure and a macro below.

We will apply PROC LOGISTIC with selection = SCORE to the neighborhood of AIC_{MIN}-, $IC(1)_{MIN}$-and $IC(1)_{RB}$ – optimal models, with the model sizes: k_{AIC}, $k_{IC(1)}$ and $k_{IC(1)RB}$ correspondingly. Choosing the basic parameters START, STOP and BEST in the best subset selection procedure is more or less arbitrary and depends on the situation. But it is reasonable to /expect/ select a smaller number of models in the vicinity of AIC_{MIN} -and $IC(1)_{RB}$ - optimal models, and a larger number of models in $IC(1)_{MIN}$ neighborhood. It is worth noting that the output of the ordinary best subset selection procedure provides only score statistics and the list of predictors with no coefficient estimates, odds ratios, AIC, IC(1), 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

```
ods output BestSubsets= Best_Subsets;
```

and the following macro we can simultaneously run logistic regressions for all selected model sizes of interest (around k_{AIC}, $k_{IC(1)}$ and $k_{IC(1)RB}$) 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 logistic data=MYDATA
descending;
   model OUTCOME = &list;
   run;

%end;

%MEND;

%SCORE;
```

LOGISTIC REGRESSION *vs.* NEURAL NETWORKS AND DECISION TREES

As shown in Steyerberg *et al.* (2001), Fedenczuk (2002), and Lajiness (2001) a predictive model obtained with logistic regression analysis is no black box in contrast to, for example, a neural network or a decision tree. Also, using logistic regression, we have a naturally interpretable model because the regression coefficients represent odds ratios and in addition we have a number of very useful statistics such as Classification Table (including Sensitivity and Specificity), Receiver Operating Characteristic Curves, very powerful regression diagnostic statistics, etc. Thus not only do we have a predictive model, but we can evaluate the importance of individual predictors, and can figure out how to improve the model. In (With) this regard logistic regression is superior to its competitors in Enterprise Miner. This is applicable to a conventional stepwise logistic regression usage. Moreover it is true for our thee-step procedure.

THE THREE-STEP PROCEDURE *vs.* LOGISTIC REGRESSION BASED ON AIC

Our three-step procedure is based on the consecutive application of the traditional stepwise logistic regression to build a complete stepwise sequence, then finding an optimal model in this stepwise sequence with regard to some information criteria (for example, AIC, IC(1) or IC(3/2)), and then constructing neighborhoods of the optimal models by using best subset selection to have "confidence" sets of models instead of single optimums. A natural question arises whether we can combine the first two steps and build a stepwise regression based completely on AIC, for example. A basic idea is to use as a criterion of "importance" of a variable the AIC statistic instead of the log-likelihood one. This idea was realized in Wang (2000). But from the discussion above it can be seen that one has to work with the information criteria, other than AIC. That is why we should prefer the usual stepwise procedure, which is "neutral" with respect to AIC, IC(1)

and I(3/2).

CONCLUSIONS

The proposed technique can be used for predictive purposes with SAS PROC LOGISTIC or within Enterprise Miner in solving such problems as drug discovery, database marketing, credit risk evaluation, fraud detection, and other predictive modeling applications in banking, financial services, insurance, telecommunication industry, etc.

ACKNOWLEDGMENTS

The authors would like to thank Irina L. Miroshnik for helpful discussions and assistance in writing the macro.

REFERENCES

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute*, **50**, 277-290.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413-418.

Atkinson, A. C. (1981). Likelihood ratios, posterior odds and information criteria. *Journal of Econometrics*, **16**, 15-20.

Box, G. E. P. and Kanemasu, H. (1973). Posterior probabilities of candidate models in model discrimination. Technical Report 322, University of Wisconsin.

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.

Everitt, B. S. & Der, G. (1996). *A handbook of Statistical Analyses Using SAS.* New York: Chapman & Hall / CRC.

Fedenczuk, L. L. (2002). To neural or not to neural? - This is the question. System. *SUGI'27 Proceedings, Paper 113-27*. Cary, NC: SAS Institute Inc.

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

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

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Lajiness, M. S. (2001). Using Enterprise Miner to explore and exploit drug discovery data. System. *SUGI'26 Proceedings, Paper 266-25*. Cary, NC: SAS Institute Inc.

Laud, P. W. & Ibrahim, J. G. (1995). Predictive model selection. *J. R. Statist. Soc*. **B**, **57**, 247-262

Lee, K. and Koval, J. J. (1997). Determination of the best significance level in forward logistic regression. *Communications in Statistics - Simulations*, **26**, 559-575
Lindsey, J. K. & Jones, B. (1998) Choosing among generalized linear models applied to medical data. *Statistics in Medicine*, **17**, 59-68.

Nelder, J. A. and Wedderburn, R. W. M. (1974). Generalized linear models. *Journal of the Royal Statistical Society, Series A,* **135,** 370-384.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, 46,1273-1282.

Smith, A. F. M. & Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear model. *Journal of the Royal Statistical Society, Series B* **42**, 213-220.

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

Shtatland, E.S, Kleinman K, and Cain, E. M. (2002). One more time on R^2 measures of fit in logistic regression. *NESUG'15 Proceedings, Paper 222-26*. 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.

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.

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.

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 02215
tel: (617) 509-9936
email: ernest_shtatland@hphc.org