Paper 222-26

THE PERILS OF STEPWISE LOGISTIC REGRESSION AND HOW TO ESCAPE THEM USING INFORMATION CRITERIA AND THE OUTPUT DELIVERY SYSTEM

Ernest S. Shtatland, Emily Cain, and Mary B. Barton
Harvard Pilgrim Health Care, Harvard Medical School, Boston, MA

ABSTRACT

In this presentation, which is a continuation of our NESUG'2000 paper, we demonstrate that using SAS® stepwise logistic regression with the default and most typically used value of significance level for entry (SLENTRY) of 0.05 may be unreasonable and sometimes even dangerous because it results in the model that on one hand has usually too many variables for a reliable interpretation and on the other hand too few variables for a good prediction. Users who blindly rely on stepwise logistic regression will most likely get a rather poor choice for both purposes: interpretation and prediction. The recommendations of using critical p-values other than the default often look vague and even contradictory. We propose to resolve this problem by using the Akaike and Schwarz information criteria (which are standard components of the PROC LOGISTIC output), some elements of Bayesian reasoning, and capabilities of ODS (Output Delivery System) which are available in PROC LOGISTIC in SAS version 8. We also discuss the problem of improving the model selection process by taking into account model selection uncertainty.

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

THE PROBLEMS WITH MODEL SELECTION

Model selection is a fundamental task in data analysis, widely recognized as central to good inference. In SAS PROC LOGISTIC, we have 4 *automatic* model selection techniques: forward selection, backward elimination, stepwise selection which combines the elements of the previous two, and the best subset selection procedure. The first three methods are based on the same ideas and we will talk only about stepwise selection as more flexible and sophisticated selection procedure. This choice is subjective, some researchers prefer to work with backward selection. Typically, the final model selected by each of these procedures will be the same, but it is in no way guaranteed. Stepwise selection is intuitively appealing: it builds models in a sequential manner and it allows for the examination of a collection of models which might not otherwise have been examined. The best subsets selection method which is invoked with the statement SELECTION = SCORE is not as popular as forward, backward, and stepwise selections because it can compare only the models of the same size (with the same number of covariates). However, we will show how the best subset selection method can be very useful in the final step of our procedure in reducing model selection uncertainty.

Purposeful selection which combines subject measure knowledge with statistical significance considerations can be performed only when we have a *small* number of models to compare originally, or at some advanced step of selection when a small number of covariates has been left. It is worth noting that if we have 10 covariates , the number of all possible models is 2^{10} =1024. With 20 covariates we have more than 1,000, 000 possible models, and with 30 covariates the number of possible models is greater than 1,000,000,000. Thus, even with rather moderate numbers of covariates we cannot do without stepwise selection. The stepwise technique allows us to decrease drastically the total number of models under consideration and to produce the final model. The final result will depend substantially on the 2 parameters: SLENTRY (the significance level for entering) and SLSTAY (the significance level for stay). If the values of these parameters are not specified, the SAS system uses default values of 0.05 for both. This default value of the significance level is used more often than not without any grounds, just because of an unwritten statistical tradition which says: if you do not have strong personal opinions on this matter, then use 0.05. SLENTRY=0.05 does *not* mean that the overall significance level is 0.05, it is usually *much larger* than 5%. One way to deal with this problem is to specify a *very small* SLENTRY (see, for example,

SAS Institute Inc. (1995), p. 51). But how small? Is it enough to have SLENTRY=0.01? Or do we need SLENTRY=0.001?

On the other hand, in Hosmer and Lemeshow (1989), p. 108, the choice of SLENTRY = 0.05 is described as too stringent, often excluding important variables from the model. Hosmer and Lemeshow propose to use the range from 0.15 to 0.25 and even to 0.30. Thus, we have very substantial deviations in the *opposite* directions from the default value of 0.05. These recommendations to use on one hand values of SLENTRY much smaller than 0.05 and on the other hand much larger than 0.05, seem contradictory. This apparent contradiction can be resolved if we reject the idea of a *single* model choice as a dogma. In reality, there is no one "supermodel" which is good for all purposes, and even in the same study we might often need at least *two* types of models: one for description / interpretation and another for prediction. These two types of models are different and as such they require different ranges for SLENTRY. An apparent arbitrariness in specifying SLENTRY values from 0.001 to 0.01 to 0.05 for explanatory models and from 0.15 to 0.25 to 0.30 for predictive models reduces the degree of confidence in our process as governed by some rationale as opposed to a trial and error method. This arbitrariness adds much of uncertainty to the selection process and can reduce substantially the degree of automation in stepwise selection. This automation is perhaps the most important virtue of stepwise logistic regression. Thanks to this feature and in spite of all the criticism, stepwise logistic regression has been used and will be used widely just because there is no realistic alternative.

The choice of SLENTRY is perhaps the most difficult and crucial aspect of using stepwise logistic regression. We propose to resolve the problem of the apparent arbitrariness in specifying SLENTRY values by using the Akaike and Schwarz information criteria (which are the standard components of the PROC LOGISTIC output), some elements of Bayesian reasoning, and capabilities of ODS (Output Delivery System) which are available in PROC LOGISTIC in SAS version 8.

In addition, we propose to use the best subsets selection method (SELECTION = SCORE) which helps us to overcome, at least partly, the model uncertainty problem on the purely frequentist grounds.

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 (see, for example, Akaike (1983)). The most popular in this family are the Akaike information criterion (AIC) and Schwarz information criterion (SC). The AIC and SC can be defined by the equations:

$$AIC = -2\log L(M) + 2*K$$
$$SIC = -2\log L(M) + (\log N)*K \qquad (1)$$

where logL(M) is the maximized log likelihood for the fitted model, N is the sample size and K is the number of covariates including an intercept. It can be seen that AIC and SC have some important optimal properties, often complementary, which justify choosing precisely these information criteria out of the entire family.

As shown in Stone (1977), AIC is asymptotically *equivalent* to the cross-validation criterion which is based on predictive ideas. Moreover, AIC is considered a cornerstone of the modern approach to prediction. Striving predominantly for good prediction, AIC sometimes tends to select more covariates than it seems necessary. And last but not least, model comparisons based on AIC are asymptotically *equivalent* to those based on Bayes factors if the prior information is comparable to the information in the likelihood (i.e. in the data), see Kass and Raftery (1995).

It is interesting that the *implied* significance level varies for AIC from 30% to 15-16% as the sample size increases. More exactly, the use of AIC is *equivalent* asymptotically to the stepwise procedure with a critical level of 15.7% (see for example, Atkinson (1981) or Lindsey & Jones (1998)). Thus, we can see that this interval of the implied significance level for AIC closely corresponds to the interval of SLENTRY recommended for predictive models (15% - 30%). The asymptotic equivalence of AIC and stepwise logistic regression with the critical level of 15.7% is especially important. It can be interpreted as a proper *theoretical* justification for using this interval for SLENTRY and especially the critical level of 15 - 16% for large sample sizes. Note that these results on critical values were derived originally from simulation studies. It can be suggested that using the AIC implied significance levels is much more valid and convincing than working with more or less arbitrary levels of SLENTRY.

Compare the AIC-based approach with the following strategy recommended in Hosmer and Lemeshow (1999), p.184: "In many applications it may make sense to use 25-50 percent to allow more variables to enter than will ultimately be used and then narrow the field of selected variables using $p < 0.15$ to obtain a multivariable model for further analysis." Some uncertainty can be seen in this recommendation. So, is not it better from the very beginning to rely on the AIC-implied significance level?

Unlike AIC, SC is consistent: the probability of choosing incorrectly a bigger model converges to 0 as the sample size increases. Also, asymptotically SC provides the *shortest code length* data description (see Dawid (1992) and references therein). And most importantly, when the prior information is small relative to the information in the data (more exactly, when the amount of information in the prior is equal to that in one observation), exp(-SC / 2) provides a surprisingly good approximation to the Bayes factor. According to Kass and Wasserman (1995), SC may be competitive with and even preferable to the more sophisticated Bayesian methods. SC is a part of the standard output of PROC LOGISTIC. Note that although SC (and AIC) optimal properties related to the Bayes factor approximations are asymptotic, the sample sizes needed to provide accuracy of the approximation are not prohibitively large (Kass and Wasserman (1995)). According to Kass and Raftery (1995) and Kass and Wasserman (1995), SC functions as a "fully automatic Occam's razor", which is very useful for reporting results in scientific communication.

Thus, AIC and SC that originated within a purely frequentist approach, serve as a bridge between Bayesian and frequentist methods. They can emulate the Bayesian approach in two extreme and opposite situations: when the priors are as important as the likelihood (i.e. the data), and when the priors are of little importance. This is another example of AIC and SC being mutually complementary (this time from a Bayesian standpoint). It emphasizes a particular significance of AIC and SC in the family of information criteria. We have to remember that for large sample sizes AIC and SC are supported from the Bayesian standpoint. Note also that we can use AIC and SC absolutely "free": they are components of the standard output of PROC LOGISTIC. Summarizing the properties of AIC and SC, we can suggest using the AIC-optimal model for prediction, and the SC-optimal model for description and interpretation.

THE OUTPUT DELIVERY SYSTEM (ODS) IN MODEL SELECTION

The most serious problem in using AIC and SC for model selection is that this process is not automated. The method of calculating AIC and SC for *every possible* submodel with the following direct comparison is impractical (see above in the section on stepwise regression), and we obviously need some shortcuts. One of the possible shortcuts is to use the stepwise selection method with SLENTRY = 1 and SLSTAY = 1 (we can use here any number sufficiently close to 1). 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). It is natural to call this sequence the *stepwise sequence*. Note that the stepwise sequence is optimal from a maximum likelihood standpoint. It is important also that we use here the stepwise procedure in a way different from the one used typically. Instead of getting a *single* stepwise pick for some specific SLENTRY value (for example, 0.05 or 0.15, etc.), we propose to work with 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 our example with 34 explanatory variables, we reduce the number of candidate models from 2^{34} (more than 16,000,000,000) to just 34. It is worth noting that the stepwise sequence contains (in a very condensed form) *any* stepwise model for *any* possible significance level.

Then, 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 the values of AIC and SC for the members of the stepwise sequence. For convenience, we separate the data set FIT into two data sets: AIC and SC (34 observations in each) which contain the values of criteria AIC and SC correspondingly. By using PROC MEANS and the MERGE statement, it is easy now to find the minimum of AIC and SC and the corresponding AIC- and SC-optimal submodels. Also, it is highly recommended to apply PROC PLOT and visualize the behavior of AIC and SC *vs.* the model size. Usually, the AIC plot has a 'U' shape with a plateau around the minimum. On the other hand, the SC graph has typically a 'V' shape with a substantially smaller number of 'nearly' optimal

models. In our example with 34 explanatory variables, the SC-optimal model contains 7 covariates, the default stepwise pick has 13 variables, and the AIC-optimal model contains 19 predictors. So the default stepwise choice is located literally halfway between SC- and AIC- optimal models. As a result, the default stepwise model (which is used much more often than it deserves) contains 6 more variables than we need for reliable description/interpretation of the data, and 6 less predictors to be a good predictive model. It is unlikely that starting with the default stepwise model, in the course of purposeful selection we would easily arrive at SC-optimal model if we are interested in interpretation or at AIC-optimal model if we are interested in prediction. Note also that the SC-optimal model corresponds to SLENTRY = 0.0028 and AIC-pick to SLENTRY = 0.1271, which is reasonably close to 0.1571 taking into account that the number of observations in our example is equal to 2629. It is also important that when building a model for description / interpretation and working with SC (which is consistent) we do not have such a yardstick as 15-16% for AIC. So we should rely only on the concrete values of SC and their implied significance levels. In our example we can see that SC sets a very low p-value of 0.0028 *automatically*, based on the data and functioning as a "fully automatic Occam's razor".

INCLUDING THE BEST SUBSET SELECTION PROCEDURE INTO MODEL SELECTION

In the sections above we have discussed the procedure which is the combination of stepwise selection of covariates with information criteria AIC and SC. The first step of this procedure is automated stepwise selection with SLENTRY = 1 that reduces the number of potential candidate models from usually enormous to manageable. The second step is finding SC-optimal and AIC-optimal models. This step can be also made automatic thanks to ODS. Let k_{AIC} and k_{SC} be the numbers of covariates in AIC- and SC-optimal models correspondingly. Obviously, it would be too simplistic to recommend SC- and AIC-optimal models of the stepwise sequence as the best models for description / interpretation and prediction correspondingly.

First of all, there is a number of *nearly optimal* models in the vicinity of AIC- and SC-optimal choices. These nearly optimal models make a plateau around AIC and SC picks. The number of nearly optimal models is usually larger for AIC, and smaller for SC. These sub-optimal models can be added to the optimal ones to create a pool of potential candidates.

Second, we should remember that AIC- and SC-optimal and nearly optimal models have been chosen from the *stepwise sequence only,* not from *all* possible models. It has been to our advantage because we can afford looking for optimum only for much smaller subset of models. At the same time, it is an obvious disadvantage because we screen only a small portion of all possible models. The problem can be resolved by using best subset regression. The method of best subset selection can provide a computationally efficient way to screen many more possible models. Note that this method becomes especially efficient only if we limit ourselves to a small number of model sizes of interest. We propose to use only the model sizes of the AIC- and SC-optimal and nearly optimal members of the stepwise sequence. For example, sizes: $k_{AIC} - 2$, $k_{AIC} - 1$, k_{AIC}, $k_{AIC} + 1$,and $k_{AIC} + 2$ for predictive models, and $k_{SC} - 2$, $k_{SC} - 1$, k_{SC}, $k_{SC} + 1$,and $k_{SC} + 2$ for models for description /interpretation. In short, we can say: 2 models to the left and 2 models to the right from the optimal model; or 1 model to the left and 1 model to the right, etc.

It is worth noting that the OUTPUT of the original best models selection procedure provides only score statistics and the list of covariates with no coefficient estimates, odds ratios, AIC, SC and other statistics. That is why the procedure is not *fully automated* and models of different sizes cannot be compared directly. By using an ODS statement

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

and a macro we can simultaneously run logistic regressions for *all* selected model sizes of interest (around k_{AIC} and k_{SC}) , and for a specified value of the BEST option (for example, BEST = 3, 4, 5, etc.). As a result we can suggest a number of separate alternative models (for interpretation or prediction). It contradicts common practice: to search for a reasonable model and then to settle on a *single* choice, ignoring the model uncertainty uncovered in the search. In doing so, we most likely underestimate the total uncertainty, work with too narrow confidence intervals, and make poor predictions. Model selection uncertainty can be an order of magnitude worse than parameter uncertainty (see Draper (1995), pp. 84-85). Precisely this common practice mentioned above, is characterized in (Breiman (1992), p. 738) as "a quiet scandal in the statistical community". That is why it is recommended in Harrell, Lee and Mark (1996): "If doing stepwise variable selection, present a summary

table depicting the variability of the list of 'important factors' selected over the bootstrap samples or cross-validations". The problem is that both computer-intensive methods, cross-validation and bootstrapping, are unavailable in SAS, version 8.

CONCLUSIONS

We propose a three-step procedure:

1) Using stepwise regression with SLENTRY = 1 and building the stepwise sequence;
2) Finding AIC / SC- optimal and sub-optimal models for the stepwise sequence;
3) Applying best subset selection to the sample sizes that correspond to AIC/SC-optimal and nearly optimal models.

All three steps are *automated* by using the capabilities of ODS and the macro language.

The resulting, usually *small* number of separate alternatives can either be used by an investigator in further purposeful selection (a purely frequentist approach), or can be averaged in a Bayesian manner by using the weights based on AIC (see , for example, Shtatland *et al* (2000)). In both cases, model selection uncertainty is taken into account at least partly.

ACKNOWLEDGMENTS

The authors would like to thank Irina L. Miroshnik and Inna Dashevsky 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. (1981). Likelihood ratios, posterior odds and information criteria. *Journal of Econometrics*, **16**, 15-20.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, **87**, 738-754.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In *Bayesian Statistics 4*, eds. J. M. Bernardo et al. Oxford: Oxford Science Publications, 109-125.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 45-97.

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

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

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

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

Kass, R. E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928-934.

Lindsey, J. K. & Jones, B. (1998) Choosing among generalized linear models applied to medical data. *Statistics in Medicine*, **17**, 59-68.

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

Shtatland, E. S., Moore, S., Dashevsky, I., Miroshnik, I., Cain, E. & Barton, M. B. (2000). How to be a Bayesian in SAS: Model selection in PROC LOGISTIC and PROC GENMOD. *NESUG'2000 Proceedings*, Northeast SAS Users Group, Inc., 724-732.

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
126 Brookline Avenue, Suite 200
Boston, MA 02215
tel: (617) 421-2671

email: ernest_shtatland@hphc.org