HOW TO BE A BAYESIAN IN SAS: MODEL SELECTION UNCERTAINTY IN PROC LOGISTIC AND PROC GENMOD

Ernest S. Shtatland, Sara Moore, Inna Dashevsky, Irina Miroshnik,
Emily Cain, Mary B. Barton

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

ABSTRACT

The SAS system is known not to support any more or less developed Bayesian method. At the same time a Bayesian framework is the ideal environment for resolving the problem of model selection uncertainty (which is important for getting proper inference based on the model), though at a price of very complex and time-consuming algorithms. In this presentation, which is a continuation of our SUGI'2000 paper, the possibility of avoiding the complexities of fully developed Bayesian methods is discussed. A Bayesian-like approach to resolving the problem of model selection uncertainty in PROC LOGISTIC and PROC GENMOD is developed, while staying completely within the maximum-likelihood methodology. Only standard elements of the output are used, such as the likelihood, the Akaike information criterion, and the Schwarz information criterion, etc., or some equivalent R^2 measures discussed in the above mentioned Shtatland, Moore & Barton (2000). The proposed approach uses some averaging and improves the model selection process by taking model uncertainty into account. The average of a (usually small) number of 'good' models is often better than any one model alone. The improvement is seen in terms of the quality of predictions, more realistic confidence intervals, etc. Applications to some medical studies are discussed.

MODEL SELECTION AND INFERENCE: FREQUENTIST APPROACH

Model selection is a fundamental task in data analysis, widely recognized as central to good inference. It is also a very complex matter, so it is not surprising that there is no definitive breakthrough in this field and still there is no widely accepted model building strategy. At the same time in the research community there is a clear need for such a strategy. Many researchers come to the conclusion that the appropriate model selection criteria should be specified in the protocol for any study, including clinical trials (Lindsey & Jones (1998)), and that model selection should be considered an integral part of inference. Until recently the relationship between model selection and inference was a one-way street: first, we search for a reasonable model (optimal or sub-optimal in some meaning) and then, conditioning on a *single* choice, we make statistical inference (confidence intervals, etc.). In other words, we proceed in our inference as if our chosen model were the true one, which is almost always incorrect. As a result, we ignore the model uncertainty uncovered in the search, underestimate the total uncertainty, and work with too narrow confidence intervals. In short, we can be overly optimistic, which is dangerous.

MODEL SELECTION: LIKELIHOOD RATIO TEST AND STEPWISE REGRESSION

Currently, there are two basic approaches to model selection in SAS PROC LOGISTIC: the classical approach based primarily on the likelihood ratio test (LRT) and the approach based on the family of information criteria such as the Akaike information criterion (AIC), Schwarz or Bayesian information criterion (SIC or BIC). The classical approach through LRT, though still being widely used, is unsatisfactory because of three basic disadvantages:

a) It works only if *nested* models are compared;
b) Asymptotic Chi-Square approximation may be poor for small sample sizes;
c) LRT is inconsistent: inherently it favors larger models unduly.

The most popular implementation of the LRT idea is stepwise selection which is realized in both multiple linear regression and logistic regression cases and which, in principle, can be implemented in generalized linear modeling as a whole.

In SAS PROC LOGISTIC, the most commonly used model selection methods are three automatic procedures: forward selection, backward elimination, and stepwise regression which is, essentially, a combination of the previous two. Ideally, we expect that the final model selected by each of these procedures would be the same. This does often happen, but it is not 100% guaranteed. All of them are based on the 'importance' of a covariate defined in terms of the statistical significance of the coefficient of the variable (Hosmer & Lemeshow, pp. 106-107). Significance is assessed via the likelihood ratio chi-square test, and at any step in the stepwise procedure the most important covariate will be the one that produces the largest change in the log-likelihood relative to model without the covariate (in other words, the one that would result in the largest likelihood ratio statistic). Also, the most important explanatory variable is the one with the smallest *P*-value. However, it is well known that the *P*-values used in stepwise selection procedures are not *P*-values in the traditional hypotheses testing context. They should be rather thought of as indicators of relative importance among explanatory variables.

None of these automatic procedures are foolproof and numerous warnings are issued to use them with care. When using stepwise selection techniques we capitalize on chance because we perform many significance tests to compare different combinations of explanatory variables. As a result, completely unrelated variables can be chosen by chance alone, and a thorough analysis is needed that examines the substantive importance of the variables in addition to their statistical significance. Any stepwise selection identifies candidates for a model based solely on statistical grounds. A common modification of the stepwise selection procedure is to begin with a model which already contains some known important covariates (option=INCLUDE) irrespective of their statistical significance.

As a whole, the stepwise selection approach is a very convenient and powerful technique. Unfortunately, it is too often misused especially when researchers rely heavily on the result of the stepwise search as a *single* choice without further exploratory analysis. According to Breiman (1992) such a usage has long been " a quiet scandal in the statistical community". Note that stepwise techniques and LRT as a whole do not address overfitting or underfitting problems. As a result, a stepwise model choice could be biased. It does not provide confidence intervals with the proper coverage. Confidence intervals produced with a stepwise procedure are falsely narrow, and cross-validation or bootstrapping techniques are usually needed to get more realistic results. Unfortunately, these techniques are not implemented in SAS.

MODEL SELECTION AND INFORMATION CRITERIA

To overcome the disadvantages of LRT mentioned above, the information criteria family was introduced. 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). AIC must be credited with being the first widely known measure that attempts to address the divergent requirements of model complexity and estimation accuracy (fit, likelihood). AIC and SIC can be defined in two different forms. The "smaller-is-better" form is defined by the equations:

$$\begin{aligned} AIC &= -\ 2\log L(M) + 2*K \\ SIC &= -\ 2\log L(M) + (\log N)*K \end{aligned} \qquad (1)$$

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 and K is the number of covariates (including an intercept). The "larger-is-better" form uses the equations:

$$\begin{aligned} AIC &= \log L(M) – K \\ SIC &= \log L(M) – (\log N/2)*K \end{aligned} \qquad (2)$$

Having these two forms seems confusing especially if they are implemented simultaneously in the same procedure (SAS PROC MIXED, for example). Using both forms can be explained by statistical tradition. In any case, there is no problem in it since (1) and (2) are related to each other by a one-to-one mapping. We can add that the most general form of AIC-type information criteria is

$$IC(c) = -\ 2\log L(M) + c*K \qquad (3)$$

If *c* = 0, (3) is equivalent to the classical likelihood statistic. If *c* = 1, (3) is equivalent to the GLIM goodness-of-fit procedure based on plotting the deviance against degrees of freedom (Smith and Spiegelhalter (1980)). If *c* = 2, IC is identical to AIC. And lastly, if *c* = logN, (3) is equivalent to SIC. The question of what value of parameter *c* to choose is not easy. Atkinson (1981) suggests that the range between 2 and 5 or 6 may provide "a set of plausible initial models for further analysis".

AIC and SIC have some optimal properties providing certain justification for choosing these information criteria out of the entire family (3). AIC is based on the errors of prediction ground and as such has some minimax properties for prediction over the experimental region, but larger values of *c* may be required for extrapolation (Atkinson (1981)). Striving predominantly for good prediction, AIC may tend to select too many covariates. From the prediction standpoint, occasionally retaining an unnecessary covariate is preferable to occasionally omitting a necessary one. It is known (Stone (1977)) that *c* = 2 is asymptotically equivalent to a cross-validation criterion – a very important property of AIC. Unlike AIC, the Bayesian Information Criterion is consistent: the probability of choosing incorrectly the bigger model converges to 0. SIC arises automatically out of the minimum coding approach (Dawid (1992)). Also important is that AIC and SIC can be used in model comparison of nested as well as non-nested models (unlike LRT)). Some applied statisticians strongly believe that in the future the emphasis will shift from studying the effect of a single covariate (after correction for confounders) to building prognostic models for individual cases. See in Van Houwelingen (1997): "Maybe Akaike's information criterion will take over completely from the *P*-values… This asks for a different view on statistical model building." Thus, we see that strong properties of AIC and SIC are often mutually complementary: SIC is consistent - AIC is not, AIC is good in prediction - SIC is better in extrapolation, SIC often performs better when the true model is very simple - for relatively complex models AIC is consistently more accurate. The philosophy underlying AIC is that "truth" is high-dimensional, requiring many (possibly infinitely many) explanatory variables. By contrast, working with SIC we assume that a true model is low-dimensional (Buckland, Burnham, and Augustin, (1997)). An applied researcher has to be capable of maneuvering between "AIC-SIC truths" and reconciling them. Below we give an example of such heuristic reconciling. If we want to avoid overfitting a model, we should use more conservative criteria, such as SIC, sometimes at the cost of underfitting a model for finite samples, which leads to a significant increase in bias. If we want to avoid uderfitting a model, then we should use more liberal AIC. We will see below that AIC and SIC are also mutually complementary from a different, Bayesian analysis standpoint.

Summarizing, we come to the following conclusions. First, AIC and SIC have some optimal mutually complementary properties, and on this ground should be chosen out of the entire family of information criteria (this choice is implemented in SAS PROC REG, PROC LOGISTIC and PROC MIXED). Second, there is no *single* information criterion which will play the role of a panacea in model selection. As a whole, information criteria resolve (at least partly) some problems related to the classical LRT approach:

(a) Information criteria work in both nested and non-nested cases;
(b) Information criteria are not tests of significance. As such, they do not indicate that the better of two models is "significantly better". But at the same time, they do not depend on asymptotic Chi-Square approximations which may be poor for small sample sizes. Although, asymptotically, the use of AIC is equivalent to a stepwise procedure with a critical level of 15.7% (Lindsey & Jones (1997));
(c) At least some of the information criteria (SIC, for example) are consistent.

Information criteria, originally designed specifically for prediction purposes in time series, are much more wildly used now. Regarding their use in biostatistics and health care applications see Van Houwelingen (1997) and Lindsey & Jones (1998).

There are two serious problems related to the information criteria. First, they are *not* automated. Second, it is still assumed that we will come to a *single* model: AIC-optimal or SIC-optimal, etc. The first problem is technical. 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. Thus, finding the best AIC or SIC model by complete enumeration is usually impractical, and we need some shortcuts. One of the possible solutions to this problem is to use the stepwise selection method with the level of significance for entry SLENTRY close to one, for example SLENTRY=0.95. In this case we will get most likely the sequence of models starting with the null model (with the intercept only), and ending with the full model with all explanatory variables included. The models in this sequence will be ordered in the way maximizing the increment in likelihood at any step. Note that we use the stepwise procedure in a way different from the one typically used. Instead of getting a *single* stepwise pick for a small SLENTRY, say, 0.05, we are planning to work with the whole sequence of K models and calculate AIC and SIC for them. Thus, instead of comparing the values of AIC or SIC for 1024 models we have to do this for 10 models (10 vs. 1000 and 20 vs. 1000000). This is a huge gain in efficiency. Moreover, the behavior of AIC or SIC on this sequence is very simple and easy to interpret: when the number of covariates grows, the values of both AIC and SIC decrease then increase with one minimum. And this minimum corresponds exactly to the AIC or SIC-pick.

The second problem related to the information criteria is much deeper because it is connected to our tradition of selecting a *single* model, optimal or sub-optimal according to some unique chosen criterion. To overcome this limitation, we have to turn to the Bayesian approach.

MODEL SELECTION AND THE BAYESIAN APPROACH

It is well known (see Draper (1995), Chatfield (1995), Kass & Raftery (1995)) that the Bayesian approach with averaging across the models is the most natural environment for resolving the problem of model selection uncertainty, superior to bootstrapping. Note that neither the Bayesian approach nor bootstrapping are implemented in SAS. It is also known that the fully developed Bayesian approach has two disadvantages. First, it is assumed that we know the prior distributions while they are usually *unknown*, and any assumptions about these distributions "are not checkable, however many data are collected" (according to Nelder (1999)). This is a very important disadvantage that deters many statisticians from becoming Bayesians. The second disadvantage is a technical one: the difficulty of calculating the Bayes factors (the Bayes factor in the simple vs. simple hypothesis testing setting can be defined as the odds in favor of one model over the competing model) and the number of terms for Bayes averaging which can be enormous (Kass and Raftery (1995)). Many methods were proposed to overcome these problems: the Occam's window approach, to minimize the number of models for averaging; the Markov Chain Monte Carlo method, to average all the models; the intrinsic Bayes factor approach of Berger and Pericchi; the fractional Bayes factor method of O'Hagan, etc. (see Kass and Raftery (1995)). All these methods are very complex technically, and have not yet resolved the problem. Also, all these techniques are unavailable for SAS users.

Fortunately, there exists a "shortcut" method that allows us to resolve the problem of *unknown* priors on one hand and the formidable calculations on the other hand. It can be done by using AIC and SIC simultaneously. As shown in Kass and Raftery (1995) (see also Akaike (1983)), model comparisons based on AIC are asymptotically equivalent to those based on Bayes factors under the assumption that the precision of the priors is comparable to that of the likelihood (in other words, only if the information in the prior increases at the same rate as the information in the likelihood). This situation is considered not very typical, though not impossible (Kass and Raftery (1995), Carlin and Louis (1996), pp.48-49). Much more usual is the situation when the prior information is small relative to the information provided by the data. In this case SIC should be used. According to Kass and Wassermam (1995), exp(-SIC / 2) provides a surprisingly good approximation to the Bayes factor when the amount of information in the prior is equal to that in one observation (at least when comparing nested models). Thus, AIC and SIC can emulate the Bayesian approach in the two extreme and opposite situations: when the priors are as important as the likelihood (i.e. the data), and when the priors are almost of no importance at all. This is one more example of AIC and SIC being mutually complementary, this time from a Bayesian standpoint. It emphasizes a particular significance of AIC and SIC in the family of information criteria, and suggests that we have to pay special attention to the segment of the stepwise sequence between SIC and AIC. We call this segment the AIC-SIC window. By the way, we can always add to this segment some models that are *substantially important*. It should be reminded that both stepwise regression and information criteria are based solely on the statistical grounds. By using the AIC-SIC window we get some important benefits of the Bayesian approach without its disadvantages. There is a variety of ways to use the AIC-SIC window. If we need a quick and reasonably reliable prediction, we can use the largest model in this window: the AIC-optimal model. If we are interested in the best model for description and interpretation, we can use the smallest model: the SIC-pick. According to Kass and Wasserman (1995), SIC may be preferable to the fully Bayesian techniques: the intrinsic Bayes factors of Berger and Pericchi and the fractional Bayes factors of O’Hagan. Kass and Raftery (1995) think that SIC may be used for reporting scientific results with other analyses omitted but serving as background support. But we can also use AIC-SIC window in a purely Bayesian way – through averaging. The promising Bayesian model averaging approach to coping with model uncertainty should appeal not only to Bayesians, but also to any “broad-minded” statistician. The key to the success of this approach lies in not having to choose the single best model but rather in averaging over a variety of plausible competing models (Chatfield (1995)). We can average the models from the AIC-SIC window with weights

$$w_k=\exp(-AIC_k/2) \,/\, \Sigma \exp(-AIC_i/2) \qquad (4)$$

or

$$w_k = \exp(-SIC_k/2) \,/\, \Sigma \exp(-SIC_i/2) \qquad (5)$$

We would like to emphasize one more time that working with the AIC-SIC approach has important advantages over the fully developed Bayesian approach. Prior distributions are usually unknown and can only be hypothesized. This is a major problem. Bayesian factors are inherently sensitive to errors of specification of prior distributions. This is another major and very serious limitation. The AIC-SIC window approach works *without reference to any prior distribution.* Still another difficulty is that Bayesian methods may lead to intractable computational problems. All widely available statistical packages (including SAS) use classical frequentist methods. Besides the fact that the fully developed Bayesian approach is unavailable in SAS, we can add that some areas of research such as clinical trials and epidemiology are especially resistant to Bayesian methods (Freedman (1996)). The AIC-SIC window approach is much simpler in terms of computation. It uses only standard elements of the PROC LOGISTIC, such as likelihood, stepwise selection, AIC, SIC, etc. augmented by some relatively simple calculations.

EXAMPLES OF USING THE AIC-SIC WINDOW

We will give two examples of applying the AIC-SIC window approach. In Barton *et al* (2000), the authors study the dependence of mammography utilization in a prepaid health care system on socioeconomic factors. PROC LOGISTIC is used with the number of cases N = 1667 and the number of models in the stepwise sequence K = 6. In this example, AIC and SIC demonstrate a very uncommon consent and choose as optimal the model with covariates age, age-squared, and income. Thus, the AIC-SIC window contains only one model which makes unnecessary farther work on choosing a submodel in or averaging over the window. This is a very atypical situation.

The second example is related to the application of Poisson regression (Barton, Moore, Polk et al (2000)). In this case, the number of cases N = 992 and the number of the models in the sequence K = 10. The AIC-optimal model contains 7 covariates (including the intercept). The SIC-optimal model is a submodel of the AIC pick with 4 covariates. The AIC-SIC window contains 4 models. Note that in this case SIC may over-penalize, placing too much weight on parsimony. As a result, the SIC-optimal model does not contain some substantially significant explanatory variables.

SAS MACRO FOR AIC-SIC WINDOWS IN LOGISTIC AND POISSON REGRESSION

With the enhanced capabilities in Version 8 to output the resulting statistics for many SAS statistical procedures, it is not difficult to write a SAS macro for model selection based on the AIC-SIC window approach which combines the stepwise selection, information criteria and the Bayesian averaging approach. The macro will build the model in three steps:

(1) A stepwise regression (or its PROC GENMOD analog) with the probability of entry high enough to allow construction of a sequence of models starting with the null model (the intercept only) to the full model with all explanatory variables of interest;
(2) Comparing the values of AIC and SIC for each model of the stepwise sequence, finding the AIC and SIC-optimal models in this sequence and constructing an AIC-SIC window;
(3) Using the AIC-SIC window. If we are interested in description and interpretation, then we should use the simplest model and run the SIC-optimal regression. If we need a quick but more or less reliable prediction, we can use the largest model in this window and run AIC-optimal regression. If we need a more accurate prediction, we might use averaging over the AIC-SIC window. Kass and Raftery (1995) show that model averaging by the Occam's window and Markov Chain Monte Carlo methods gives consistently and substantially better predictions than predictions based on any one model alone, for several data sets. We can expect similar effects when using averaging over the AIC-SIC window. Note that we can apply the averaging approach only for large enough data sets because the Bayesian interpretation of AIC and SIC is justified only asymptotically. Note also that Bayesian averaging does not lead to a simple model. According to Chatfield (1995), though "even a simple average is often as good as anything, the user does not receive a simple model to describe the data." It is interesting to note that the stepwise pick with the default SLENTRY of 0.05 (which is typically used) is usually located half-way between AIC-optimal and SIC-optimal models. Thus, the default pick is usually *too large for interpretation and too small for prediction*. Users who trust stepwise regression blindly have a rather bad choice for both purposes. It is one more manifestation of "a quiet scandal in the statistical community" (Breiman (1992)).

Thus, before using the macro, the user needs to clarify the objectives of using the model: either for data description and interpretation, or for a quick and simple prediction, or maybe for a more accurate prediction. All of these goals can be achieved using the AIC-SIC window and the macro based on it.

CONCLUSIONS

In this paper, we propose a model selection approach that combines the advantages of stepwise regression, information criteria and Bayesian averaging. The basic message of the paper is that we should not ignore the model uncertainty uncovered in the search and underestimate the total uncertainty, and that we can afford to do this by taking into consideration *only a small number* of candidate models.

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.

Barton, M. B., Moore, S., Shtatland, E. S. & Bright, R. (2000). The relationship of household income to mammography utilization in a prepaid health care system. (submitted).

Barton, M. B., Moore, S., Polk, S., Shtatland, E.S., Elmore, J. G., & Fletcher, S. W. (1999). Anxiety and health care utilization after false positive mammograms (Abstract). *Journal of General Internal Medicine*, **14**, 9

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.

Buckland, S. T., Burnham, K. P. & Augustin. N. H. (1997). Model selection: an integral part of inference. *Biometrics*, 53, 603-618.

Carlin, B. P. & Louis, T. A. (1998). *Bayes and Empirical Bayes Methods for Data Analysis*, New York: Chapman & Hall/CRC.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A* **158**, 419-466.

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.

Freedman, L. (1996). Bayesian statistical methods (A natural way to assess clinical evidence). *British Medical Journal*, 313, 569-570.

Hosmer, D. W. & Lemeshow, S. (1989). *Applied Logistic Regression*, 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.

Nelder, J. A. (1999). Statistics for the millennium (from statistics to statistical science). *The Statistician*, **48**, 257-269.

SAS Institute Inc. (1997). *SAS/STAT Software Changes and Enhancements Through Release 6.12*, Cary, NC: SAS Institute Inc.

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

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.

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, H. C. (1997). The future of biostatistics: expecting the unexpected. *Statistics in Medicine*, **16**, 2773-2784.

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