SESUG Paper 143-2023

%modelselect: A macro for automatically selecting statistical models based on fit statistics

William B. Smith, Auburn University

ABSTRACT

When analyzing data from an experiment, it is not uncommon to have multiple candidate models from which to choose. In basic statistics courses, we are taught that selection of the single "appropriate" model should be done based on fit statistics, such as Akaike information criterion (AIC) or Bayesian information criterion (BIC). However, manual selection of models for each response variable is a tedious undertaking that can be burdensome in the process of data analysis. In an effort streamline this task, %modelselect was developed as a macro to automate model selection. The macro uses PROC SQL in SAS® v. 9.4 to extract fit statistics from FitStatistics table of PROC MIXED or PROC GLIMMIX, then generates a full summary table for easy comparison. In addition to traditional fit statistics, such as AIC and BIC, the macro was also programmed to calculate derivations of fit statistics such as the Akaike weight (Burnham and Anderson 2004). This paper will demonstrate the functionality of the %modelselect macro and its application in research data analysis.

INTRODUCTION

Any student that has taken the most basic of science courses is likely familiar with the classical scientific method and the statistical implications of such a process. In this method, the researcher is to observe a phenomenon, develop a hypothesis, test the hypothesis, and then choose to accept or reject the hypothesis based on data collected in the experiment. Such notions are reinforced in most introductory statistics courses in which students are taught that statistical analysis of experimental data revolves around null hypothesis testing. This thought process requires that the researcher has a single working hypothesis (or theory of the underlying phenomenon) and is testing this against a “straw man.” However, even as early as 1890, it was proposed that researchers entertain multiple working hypotheses (Chamberlin 1890). Such a notion requires a method by which these hypotheses may be tested for selection of a “best fit” model. The need for such a method led to the information theory and, by extension, information criteria.

The use of information criteria has long been used as the method by which candidate models are tested and selected. However, coding for all working hypotheses is often tedious and requires recording of information criteria for multiple models, followed manual selection of the “best” candidate model for each of the variables of interest. Automation of this process would greatly reduce burden (and chance of error) on the part of the researcher. Thus, this paper will present a SAS macro used to collect and compare information criteria from multiple candidate models for ease of model selection.

INFORMATION CRITERIA

To understand the purpose of the %modelselect macro, it is first necessary to understand the concepts of statistical information and information criteria. These topics are more exhaustively reviewed both by Burnham and Anderson (2004) and by Christensen (2018). This section will briefly summarize these concepts.

The concept of statistical information, at least as we know it today, can be traced back to Kullback and Leibler (1951 . In their note, these authors equate statistical information to the “divergence” or “distance” between two populations; in essence, the Kullback-Leibler (K-L) information is an assessment of population discrimination (Kullback and Leibler 1951). The “distance” described as the K-L information is understood to be the lost information when a candidate model $g(x)$ is used to estimate the true model $f(x)$, though Burnham and Anderson (2004) note that information theorists do not believe in the existence of a “true” model. Information criteria generally use this same concept of statistical information (or, rather, information loss) to discriminate between candidate models. When a candidate model sufficiently

approximates the essential elements of the true model, information loss is minimized. In practice, the true model is never known and, thus, relative distances are used to rank candidate models.

Akaike (1974) presented the first information criterion (IC) for use in model selection. The critical finding behind this IC was the formal relationship that existed between K-L information and likelihood theory (Akaike 1974; Burnham and Anderson 2004). The resulting information criterion, known as Akaike Information Criterion (AIC), uses the maximum likelihood as an estimator of the expected, relative K-L information (Burnham and Anderson 2004). It was later found, however, that AIC is asymptotically efficient in infinite dimensions, or a large sample size, but does not provide consistent selection with finite dimensionality, or small sample sizes (Hurvich and Tsai 1989). To correct this problem, Hurvich and Tsai (1989) proposed the corrected AIC (AICC) which resulted in an unbiased estimator of K-L information for small sample sizes (Burnham and Anderson 2004).

One of the strongest critiques of AIC is that it is dimension inconsistent; in other words, there is a small (but non-zero) probability that AIC will select an overly complex, or overparameterized, model (Christensen 2018). Thus, others attempted to solve this problem by proposing equations that are consistent. In this context, consistency means that, as sample size increases, a finite model will be chosen as long as it exists among the candidate models (Christensen 2018). One such correction, the Hannan and Quinn Information Criterion (HQIC), achieved consistency by applying a natural logarithm to the equation (Hannan and Quinn 1979). Others proposed an increased penalty for additional parameters to achieve consistency, resulting in the Consistent Akaike Information Criterion, or CAIC (Bozdogan 1987). Another suggestion for consistency was to base selection on the Bayes Theorem rather than information theory (the basis of K-L information); this resulted in what is now known as the Bayesian Information Criterion, or BIC (Schwarz 1978).

Burnham and Anderson (2004) noted that the information criteria, alone, are not interpretable because they are unitless and contain arbitrary values related to sample size. Instead, they proposed scaling the values of AIC, resulting in relative deviations known as Δ_i. These values could more easily be interpreted as the best-fit model would have a $\Delta_i = 0$ (Burnham and Anderson 2004). This relative value could then be normalized such that the sum of all likelihoods is equal to 1. The resulting parameter, known as the Akaike weight (w_i), would serve as the “weight of evidence” in support of a candidate model (Burnham and Anderson 2004).

CHOOSING AN INFORMATION CRITERION

As shown above, there are a bevy of IC from which to choose in selecting best-fit models. Each IC performs best in certain situations, and each has trade-offs. For instance, AIC is best suited for datasets with infinitely large sample sizes (Hurvich and Tsai 1989). However, AIC has also been shown, mathematically, to favor models with a greater number of parameters (Christensen 2018). Occam’s razor would dictate that the most parsimonious (i.e., the simplest) model is the best model (Burnham and Anderson 2004). Thus, it would stand to reason that a consistent information criterion (such as BIC, CAIC, or HQIC) be selected. Others, though, have criticized BIC for selecting overly underparameterized models (Downs and Cheng 2013). In contrast, if the goal is not to select a single model but, rather, approach the data from the perspective of model averaging, it may be that the Akaike weight offers the most insight (Burnham and Anderson 2004).

There is no one correct answer in designating an IC for model selection. It is strictly a choice of the researcher in terms of subject-matter knowledge and usefulness of explanation and prediction that results in favoring one criterion over another (Mac Nally et al. 2018).

MACRO IMPLEMENTATION

In order to combat the tediousness of coding and computing each candidate model, transcribing the information criteria, and manually comparing criteria to select the best-fit model, a macro was developed to automate the process. It should be noted that this macro does not choose the best information criterion; that is still at the discretion of the researcher and must be specified in the macro. The macro is simply a tool to aid in coding for model comparison and selection.

MACRO REQUIREMENT

The SAS macro, `%modelselect`, requires two parameters. The first parameter, `merge`, contains the list of FitStatistics tables that were output by the preceding analysis procedures. The second parameter, `fitstat`, specifies which fit statistic (i.e., IC) will be used for model comparison and selection. The options for `fitstat` are AIC, BIC, CAIC, HQIC, or AKAIKE_WT.

The macro is called as follows:

```
%modelselect(merge=, fitstat=)
```

MACRO PROGRAM

The `%modelselect` macro is based on a number of iterative steps to compile a table of fit statistics for model comparison. It is designed to work with output from either PROC MIXED or PROC GLIMMIX.

Compile FitStatistics tables into a single table

The first step in the macro uses a data step to merge all of the tables specified in the macro call into a single table titled `fit`.

```
%macro modelselect(merge=, fitstat=);
   data fit;
      merge &merge;
   run;
```

Transpose dataset

In the merged table, each dataset has its own column, and fit statistics are listed in a single row. This step uses PROC TRANSPOSE to collect fit statistics into a single column for concise presentation of the data.

```
proc transpose
   data = fit
   out = fit_transpose
   name = Model;
   id descr;
run;
```

Identify the selected fit statistic

The FitStatistics table uses a lengthy description (`descr`) to identify the fit statistics and tell the user that a smaller number is preferred. However, for purposes of this macro, it is more concise to only have the fit statistic of interest as the column header. Using the `%upcase` function (to prevent errors based on capitalization), this step pairs `fitstat` with d `descr`.

```
%if %upcase(&fitstat) = AIC %then %let descr = AIC__Smaller_is_Better_;
%else %if %upcase(&fitstat) = AICC %then %let descr =
                               AICC__Smaller_is_Better_;
%else %if %upcase(&fitstat) = BIC %then %let descr =
                               BIC__Smaller_is_Better_;
%else %if %upcase(&fitstat) = CAIC %then %let descr =
                               CAIC__Smaller_is_Better_;
%else %if %upcase(&fitstat) = HQIC %then %let descr =
                               HQIC__Smaller_is_Better_;
```

Create table with only the selected fit statistic

Using PROC SQL, this step of the macro creates a simplified table (`fit_compare_%upcase(&fitstat)`) that shows only the models being compared and the fit statistic of interest.

If the user selected Akaike weights as the fit statistic of choice, these are not automatically computed in SAS. Thus, a calculation is required to output the values. In this case, using an `%if`-`%then`-`%else`

sequence, the AICC is obtained from the dataset fit and is used to compute Akaike weights according to the equations presented by Burnham and Anderson (2004).

```
%if %upcase(&fitstat) = AKAIKE_WT %then %do;
   proc sql;
      create table fit_comp_aicc as
      select Model, AICC__Smaller_is_Better_ as AICC,
             min(AICC) as minAICC format 8.2,
             AICC - calculated minAICC as Delta_i format 8.2,
             exp(-0.5* calculated Delta_i) as RL format 6.4
      from fit_transpose;

      create table fit_compare_%upcase(&fitstat) as
      select Model, AICC, minAICC, Delta_i, RL,
             sum(RL) as sumRL format 6.4,
             RL/calculated sumRL as Akaike_wt format percent8.2
             from fit_comp_aicc;
   quit;
%end;

%else %do;
   proc sql;
      create table fit_compare_%upcase(&fitstat) as
      select Model, &descr as %upcase(&fitstat)
      from fit_transpose;
   quit;
%end;
```

Sort table in order of fit statistic value

For ease of comparing models, this step uses PROC SORT to sort the fit statistics table from least to greatest.

```
proc sort data=fit_compare_%upcase(&fitstat);
   by &fitstat;
run;
```

Print compiled table

Finally, the model comparison table is printed to the output using PROC PRINT.

```
      proc print data=fit_compare_%upcase(&fitstat) noobs;
         var Model %upcase(&fitstat);
      run;
%mend;
```

DEMONSTRATION OF %MODELSELECT

DESCRIPTION OF THE EXPERIMENT

An experiment was conducted at Auburn University to determine the methane production potential of four bermudagrass (*Cynodon dactylon* [L.] Pers.) cultivars. Data from this experiment were reported at the XXV International Grassland Congress (the proceedings have not been published at the time of this paper). Ruminally-fistulated heifers (n = 4) were assigned randomly to one of four bermudagrass cultivars (Coastal [COS], Russell [RUS], Tifton 44 [T44], or Tifton 85 [T85]) for four 30-d *in vivo* periods in a Latin square design. On d 28 of each period, rumen fluid was collected from each heifer for use in CH_4 production evaluation. Samples of each bermudagrass, corresponding to the cultivar fed, were weighed into duplicate 10-mL serum bottles and incubated at 39°C for 0, 2, 4, and 24 h. Following incubation, headspace samples were assayed for CH_4 concentrations by gas chromatography.

EXPERIMENTAL DATA

The DATA step (excerpted) for demonstration of the procedure was as follows:

```
data methane;
input period animal_ID treatment$ time rep CH4_conc;
cards;
2  1432  TIF44 0     1     0.24
2  1432  TIF44 0     2     0.00
2  1432  TIF44 2     1     0.57
2  1432  TIF44 2     2     0.60
2  1432  TIF44 4     1     0.86
. . .
5  1510  Russell     2     2     0.14
5  1510  Russell     4     1     0.25
5  1510  Russell     4     2     0.29
5  1510  Russell     24    1     0.90
5  1510  Russell     24    2     0.72
;
```

DATA ANALYSIS

These data were analyzed using PROC GLIMMIX. The response variable was methane concentration. The fixed effects were treatment, incubation time, and their interaction. Denominator degrees of freedom were adjusted using the second-order Kenward-Roger approximation (Kenward and Roger 2009). Random effects included period (the blocking factor), animal_ID (the experimental unit), and rep within animal_ID × period × incubation time (the observational unit nested within the experimental unit). Incubation time was treated as a repeated measurement on the observational unit across incubation times.

When using the `%modelselect` macro, it is necessary to output the FitStatistics table for each of the candidate models. It is useful to have SAS suppress all results files while assessing the candidate models so that the only output that exists is the comparison table generated by `%modelselect`. This is achieved by wrapping the code in an ODS EXCLUDE option. The FitStatistics table is then output using the ODS OUPUT option, and the table is renamed with a descriptive title.

The code for this analysis is presented as follows:

```
%macro methane(struc=);
   ods exclude all;
   proc glimmix data=methane;
      class period animal_id treatment time rep;
      model CH4_conc = treatment|time/ ddfm=kr2;
      random period animal_id rep(animal_id*period*time);
      random _residual_/
         subject=period*animal_id*treatment*rep type=&struc;
      ods output
         FitStatistics=Fit%upcase(&struc)(rename=(value=%upcase(&struc)));
   run;
   ods exclude none;
%mend;
```

Repeated measurements offer a prime example for the suitability of the `%modelselect` macro. In this example, the candidate models are identical in their fixed and random effects but differ in the modeling of the residuals. It is not uncommon to specify a covariance structure to account for the inter-relatedness of residuals in repeated measurements. Thus, the macro was called as follows:

```
%methane(struc=cs)
%methane(struc=csh)
%methane(struc=toep)
```

```
%methane(struc=vc)
```

The %modelselect macro is then called to compile the four output tables as follows:

```
%modelselect(merge=FitCS FitCSH FitTOEP FitVC,fitstat=AICC)
```

Running this code results in an output as given in Output 1.

```
                    The SAS System

Model          AICC

CSH          133.77
VC           500.48
CS           502.16
TOEP         505.70
```

Output 1 Output from the %modelselect macro

In this example, AICC was chosen as the IC of interest; results may have differed had BIC or Akaike weight been chosen for model comparison. From this analysis, the most appropriate model for these data is that in which the residuals have a heterogenous compound symmetry covariance structure. Using the selected model, there was no interaction of treatment and incubation time ($P = 0.0946$), nor was there an effect of treatment ($P = 0.4151$). There was, however, an effect of incubation time as a main effect ($P < 0.0001$), and the results are presented in Table 1.

Hours of Incubation	Methane concentration, mmol/L
0	0.0569[d]
2	0.3822[c]
4	0.8050[b]
24	5.0981[a]
[a-d]Means that do not share superscript letters are different ($P < 0.05$).	

Table 1 Methane production from four bermudagrass cultivars are affected by incubation time

CONCLUSION

In analysis of experimental data, it is useful to develop and test multiple competing hypotheses. Coding of these analyses and transcription of the resulting fit statistics for model selection can become cumbersome. This macro helps to solve this problem by automating the process of model comparison and selection. This does not free the researcher from using subject matter expertise in evaluating the most appropriate model (Mac Nally et al. 2018), but it does streamline analysis efforts.

REFERENCES

Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19 (6), 716-723. DOI: 10.1109/TAC.1974.1100705.

Bozdogan, H. (1987), "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, 52 (3), 345-370. DOI: 10.1007/BF02294361.

Burnham, K. P., and Anderson, D. R. (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33 (2), 261-304. DOI: 10.1177/0049124104268644.

Chamberlin, T. C. (1890), "The Method of Multiple Working Hypotheses," *Science*, ns-15 (366), 92-96. DOI: doi:10.1126/science.ns-15.366.92.

Christensen, W. 2018. Model Selection Using Information Criteria (Made Easy in SAS). In *SAS Global Forum*. Denver, CO, USA.

Downs, D. E., and Cheng, Y. W. (2013), "Length–Length and Width–Length Conversion of Longnose Skate and Big Skate Off the Pacific Coast: Implications for the Choice of Alternative Measurement Units in Fisheries Stock Assessment," *North American Journal of Fisheries Management*, 33 (5), 887-893. DOI: 10.1080/02755947.2013.818080.

Hannan, E. J., and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society: Series B (Methodological)*, 41 (2), 190-195. DOI: 10.1111/j.2517-6161.1979.tb01072.x.

Hurvich, C. M., and Tsai, C.-L. (1989), "Regression and time series model selection in small samples," *Biometrika*, 76 (2), 297-307. DOI: 10.1093/biomet/76.2.297.

Kenward, M. G., and Roger, J. H. (2009), "An improved approximation to the precision of fixed effects from restricted maximum likelihood," *Computational Statistics & Data Analysis*, 53 (7), 2583-2595. DOI: 10.1016/j.csda.2008.12.013.

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22 (1), 79-86.

Mac Nally, R., Duncan, R. P., Thomson, J. R., and Yen, J. D. L. (2018), "Model selection using information criteria, but is the “best” model any good?," *Journal of Applied Ecology*, 55 (3), 1441-1444. DOI: 10.1111/1365-2664.13060.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6 (2), 461-464.

ACKNOWLEDGMENTS

This macro, in its conceptual form, was developed as part of the author’s Ph.D. dissertation with input from Drs. F. M. “Monte” Rouquette, Jr., Jamie Foster, and Jason Banta (Texas A&M AgriLife Research), and Drs. Luis Tedeschi and Larry Redmon (Texas A&M University).

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

W. Brandon Smith
Assistant Professor
Department of Animal Sciences
Auburn University
wbs0001@auburn.edu