Paper ST-157

SAS® Code for Variable Selection in Multiple Linear Regression Models Using Information Criteria Methods with Explicit Enumeration for a Large Number of Independent Regressors

Dennis J. Beal, Science Applications International Corporation, Oak Ridge, Tennessee

ABSTRACT

SAS® 9.1 calculates Akaike's Information Criteria (AIC), Sawa's Bayesian Information Criteria (BIC), and Schwarz Bayesian Criteria (SBC) for every possible $2^p - 1$ models for $p \leq 10$ independent variables. AIC, BIC, and SBC estimate a measure of the difference between a given model and the "true" underlying model. The model with the smallest AIC, BIC, or SBC among all competing models is deemed the best model. However, for large multivariate data sets where $p > 10$, SAS displays AIC, BIC, and SBC for only a fraction of all possible models. This paper provides SAS code that evaluates all possible linear regression models for any $p > 0$ to determine the best subset of variables that minimizes the information criteria among all possible models. Simulated multivariate data are used to compare the performance of the REG procedure to select the optimal model for $p > 10$ when the optimal model is determined through an explicit enumeration of all possible models using AIC, BIC, and SBC. This paper is for intermediate SAS users of SAS/STAT who understand multivariate data analysis and SAS macros.

Key words: Akaike's Information Criteria, Sawa's Bayesian Information Criteria, Schwarz Bayesian Criteria, multiple linear regression, model selection

INTRODUCTION

Multiple linear regression is one of the statistical tools used for discovering relationships between variables. It is used to find the linear model that best predicts the dependent variable Y from the independent X variables. A data set with p independent variables or regressors has $2^{p+1} - 1$ possible subset models to consider since each of the p variables and the intercept is either included or excluded from the model, not counting interaction terms. Information criteria statistics Akaike's Information Criteria (AIC), Sawa's Bayesian Information Criteria (BIC), and Schwarz Bayesian Criteria (SBC) are calculated for each model with the objective to identify the model that minimizes the information criteria. SAS 9.1 currently displays AIC, BIC, and SBC for all possible subset models only for $p \leq 10$ in the PROC REG procedure in SAS/STAT. For $p > 10$, SAS displays AIC, BIC, and SBC only for a subset of models.

This paper shows SAS code that will calculate AIC, BIC, and SBC for all possible linear regression models for any $p > 0$ so that the best model that minimizes AIC, BIC, or SBC is determined through explicit enumeration. This best model identified by explicit enumeration is then compared with the best model identified in PROC REG to determine how often these models are in agreement. A simulation is conducted 100 times using a varying combination of variables to determine the dependent variable Y for each simulation. The number of times PROC REG identifies the best model known from an explicit enumeration is summarized along with the number of models considered. Of course, the number of possible subset models grows exponentially as p gets larger, so p is limited by the speed and memory of the computer running the SAS code. However, the SAS code presented is general enough for any $p > 0$. The SAS code presented in this paper uses the SAS System for personal computers version 9.1.3 running on a Windows XP Professional platform with Service Pack 2.

INFORMATION CRITERIA

Information criteria are measures of goodness of fit or uncertainty for the range of values of the data. In the context of multiple linear regression, information criterion measures the difference between a given model and the "true" underlying model.

AKAIKE'S INFORMATION CRITERIA

Akaike (1973) introduced the concept of information criteria as a tool for optimal model selection. Akaike (1987) and Bozdogan (1987, 2000) discuss further developments of using information criteria for model selection. Akaike's Information Criteria (AIC) is a function of the number of observations n, the sum of squared errors (SSE), and the number of independent variables $k \leq p + 1$ where k includes the intercept, as shown in Eqn. (1).

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2k \tag{1}$$

The first term in Eqn. (1) is a measure of the model lack of fit while the second term ($2k$) is a penalty term for additional parameters in the model. Therefore, as the number of independent variables k included in the model increases, the lack of fit term decreases while the penalty term increases. Conversely, as variables are dropped from the model, the lack of fit term increases while the penalty term decreases. The model with the smallest AIC is deemed the "best" model since it minimizes the difference from the given model to the "true" model.

BAYESIAN INFORMATION CRITERIA

Sawa (1978) developed a model selection criterion that was derived from a Bayesian modification of the AIC criterion. Bayesian Information Criteria (BIC) is a function of the number of observations n, the SSE, the pure error variance fitting the full model (σ^2), and the number of independent variables $k \le p + 1$ where k includes the intercept, as shown in Eqn. (2).

$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \frac{2(k+2)n\sigma^2}{SSE} - \frac{2n^2\sigma^4}{SSE^2} \tag{2}$$

The penalty term for BIC is more complex than the AIC penalty term and is a function of n, the SSE, and σ^2 in addition to k.

SCHWARZ BAYESIAN CRITERIA

Schwarz (1978) developed a model selection criterion that was derived from a Bayesian modification of the AIC criterion. Schwarz Bayesian Criteria (SBC) is a function of the number of observations n, the SSE, and the number of independent variables $k \le p + 1$ where k includes the intercept, as shown in Eqn. (3).

$$SBC = n \cdot \ln\left(\frac{SSE}{n}\right) + k \ln n \tag{3}$$

The penalty term for SBC is similar to AIC in Eqn. (1), but uses a multiplier of ln n for k instead of a constant 2 by incorporating the sample size n.

SIMULATED DATA

A multivariate data set with 15 independent regressor X variables was simulated from normal, lognormal, exponential, gamma, and uniform distributions with various means and variances. The following SAS code simulates 100 observations for these 15 independent X variables. Variables X5, X6, X7, X8, X13, and X15 are correlated with other variables.

```
data dat;
  do i = 1 to 100;
    X1 = 10 + 3*rannor(0);
    X2 = 20 + 2*ranuni(0);
    X3 = 15 - exp(2*rannor(0));
    X4 = 35 + 2*ranexp(0);
    X5 = 3*x3 + 2*x1 - 2*rangam(0, 2);
    X6 = x1*x3 - x5 + 0.4*rannor(0);
    X7 = 2 + 3*x6 - 2*x1 + ranuni(0);
    X8 = x6 + ranuni(0)*ranexp(0);
    X9 = 10 + ranuni(0);
    X10 = 5 - 3*ranuni(0);
    X11 = 3 + 5*ranexp(0) + rannor(0);
    X12 = 5 + 2*exp(3*rannor(0));
    X13 = 53 + X5*X8 - 3*X10 + 2*ranuni(0);
    X14 = 50 + 7*rannor(0);
    X15 = X14 - 2*X1 + rangam(0, 3);
    output; end; drop i; run;
```

First, all possible subset models were generated for each p where $1 \le p \le 15$. For $p \le 6$ all possible subset models were evaluated with a dependent variable Y that was calculated as a linear combination of the independent X variables, constant, and error terms in each model. The signs and coefficients for each regressor X variable and constant term (if present in the model) were randomly assigned using the uniform random generator `ranuni`. The error term for each model was normally distributed using the normal random generator `rannor`.

For each p where $7 \le p \le 15$, 100 models were randomly selected from among all possible $2^p - 1$ subset models with at least one X regressor variable. The intercept or constant term was randomly assigned either as zero (i.e., not part of the model) or one with a nonzero coefficient (positive or negative). The dependent variable Y was then calculated as for $p \le 6$. Therefore, every possible subset model had an equally likely chance to be evaluated.

SAS CODE FOR EXPLICIT ENUMERATION OF ALL MODELS

The SAS code for explicitly enumerating all possible subset models begins with defining some SAS macros that are used within the code.

SAS MACROS

The SAS macro variable `P` is assigned the number of independent X regressors. The SAS macro variable `N_MODELS` calculates the number of possible subset models with at least one X regressor for any given p.

```
%let P = 11;  ** P = number of independent regressors;
%let N_MODELS = %eval(2**(&P.));
```

The SAS macro `OBSNVARS` calculates both the number of variables and the number of observations for a given input data set.

```
%macro obsnvars(ds);          ** DS = input data set name;
   %global dset nvars nobs;
   %let dset=&ds;
   %let dsid = %sysfunc(open(&dset));
   %if &dsid %then %do;
     %let nobs =%sysfunc(attrn(&dsid,NOBS));
     %let nvars=%sysfunc(attrn(&dsid,NVARS));
     %let rc = %sysfunc(close(&dsid));
     %end;
   %else
   %put Open for data set &dset failed - %sysfunc(sysmsg());
%mend obsnvars;
```

The SAS macro `STARTDO` initiates a `DO` loop into the SAS data step that assigns the presence (`X&num. = 1`) or absence (`X&num. = 0`) of independent regressor variable `X&num.` in the model. For example, if independent regressor variable X5 = 0, then X5 is not present in the model. If X5 = 1, then X5 is present in the model.

```
%macro startdo(num);
  do X&NUM. = 0 to 1;
%mend startdo;
```

The SAS macro `ENDDO` closes each `DO` loop in the SAS data step.

```
%macro enddo(num);
  end;
%mend enddo;
```

The SAS macro `DSNAMES` is used in the `SET` statement for any number of data sets.

```
%macro dsnames(name, startno, endno);
  %do z = &STARTNO %to &ENDNO;
  &NAME&z.
  %end;
%mend dsnames;
```

The SAS macro GETALLX generates all possible subset models where X&i. = 1 if variable X&i. is included in the model, and X&i. = 0 if variable X&i. is not included in the model. The data set A creates *p* nested DO loops that are used to generate all possible subset models for the *p* X variables. The data sets B and C then create the macro variable VARS that stores all X variables in the model as a character field. The VARS&i. macro variable will be used in each PROC REG. The next DO loop calculates the AIC, BIC, and SBC for each possible subset model assuming an intercept term and stores the calculations in the withint&i. data sets. Similarly, the last DO loop calculates the AIC, BIC, and SBC for each possible subset model assuming no intercept term and stores the calculations in the noint&i. data sets. Note this last DO loop must start at 2 since the first subset model has no X variables and no constant term.

```
%macro getallx;
 data A;
  %do i = 1 %to &P;
    %startdo(&i.)
       %end;
    output;
  %do i = 1 %to &P;
    %enddo(&i.)
       %end;
  run;

  data B;
    length VARS $100;
    set A;
       VARS = ' ';
       %do i = 1 %to &P;
         if x&i. = 1 then VARS = trim(left(vars)) || "X&i.,";
         %end;
       VARS = trim(left(vars));
       run;

 data C;
   set B;
   vars = translate(vars, ' ', ','); ** replace each comma with a space;
   call symput('VARS'||left(_N_), trim(left(vars))); run;

  %do i = 1 %to &N_MODELS;
       PROC REG data=dat outest=withint&i. noprint;
         model y = &&VARS&i. / aic sbc bic;  ** assumes intercept term;
         run; quit;
   %end;

  %do i = 2 %to &N_MODELS;
       PROC REG data=dat outest=noint&i. noprint;
         model y = &&VARS&i. / noint aic sbc bic; ** noint = no intercept;
         run; quit;
   %end;
%mend getallx;
```

SAS CODE FOR SELECTING THE BEST MODEL USING EXPLICIT ENUMERATION

The data set ALL combines the 2^p subset models including an intercept term with the $2^p - 1$ models excluding an intercept term. Therefore, the data set ALL has all $2^{p+1} - 1$ possible subset models. Note that the data set noint1 is not created since there are no independent regressor X variables and no intercept term for that model.

```
data ALL;
  set %dsnames(withint, 1, &N_MODELS) %dsnames(noint, 2, &N_MODELS);
  rename _aic_=AIC _sbc_=SBC _bic_=BIC _rmse_=RMSE _rsq_=RSQ;
  drop _model_ _type_ _depvar_ y; run;
```

The data are sorted by AIC, and the model with the smallest AIC is deemed best according to the AIC criterion.

```
  proc sort data=ALL; by aic;
data BEST_AIC;  ** identify the best model according to AIC;
  set ALL;
  if _N_=1 then BEST_AIC='AIC'; run;
```

The data are then sorted by SBC, and the model with the smallest SBC is deemed best according to the SBC criterion.

```
  proc sort data=BEST_AIC; by sbc;
data BEST_SBC;  ** identify the best model according to SBC;
  set BEST_AIC;
  if _N_=1 then BEST_SBC='SBC'; run;
```

Last, the data are then sorted by BIC, and the model with the smallest BIC is deemed best according to the BIC criterion.

```
  proc sort data=BEST_SBC; by bic;
data BEST_BIC;  ** identify the best model according to BIC;
  set BEST_SBC;
  if _N_=1 then BEST_BIC='BIC'; run;
```

The macro variable `N_MODELS2` stores the total number of models evaluated ($2^{p+1} - 1$).

```
%obsnvars(best_bic); %let N_MODELS2 = &nobs;
```

Last, the best models as determined by AIC, BIC, or SBC are printed.

```
proc print data=BEST_BIC;
  where best_aic='AIC' or best_sbc='SBC' or best_bic='BIC';
run;
```

Note that Beal (2007) showed that of the three information criteria calculated within SAS (AIC, BIC, and SBC), SBC performed better than either AIC or BIC for selecting the correct model in simulations for both large (n = 1000) and small sample sizes (n = 100). Beal (2005) compared information criteria with both heuristic methods and common model diagnostics to show that information criteria are superior for selecting the best model.

SAS CODE WITHOUT EXPLICIT ENUMERATION

SAS calculates the AIC, BIC, and SBC for every possible subset of variables for models with up to p = 10 independent variables except for the intercept only model. The following SAS code from SAS/STAT computes AIC, BIC, and SBC for all possible subsets of multiple regression models for main effects with at least one regressor X variable. Note that `X1-X&P` in the `model` statement indicates that all variables `X1`, `X2`, …, `X&P` are included. The `selection=adjrsq` option specifies the adjusted R^2 method will be used to select the model so that all possible 2^p models are considered with an intercept term for $p \leq 10$ and a subset of models for $p > 10$, although other `selection` options may also be used such as `selection=rsquare`. The `AIC, BIC,` and `SBC` options display the AIC, BIC, and SBC statistics for each model, respectively. The first `PROC REG` calculates AIC, BIC, and SBC for all possible subsets of main effects using an intercept term. The second `PROC REG` calculates AIC, BIC, and SBC for all possible subsets of main effects without an intercept term by specifying the `noint` option. The output data sets `EST` and `EST0` are combined, sorted, and printed from smallest AIC, BIC, and SBC to largest. The model with the smallest AIC, BIC, or SBC value is deemed the "best" model. Note that for $p > 10$, only a small fraction of all possible models are displayed.

```
proc reg data=A outest=EST;
  model y = x1-x&P / selection=adjrsq aic bic sbc;  ** with intercept;
  run; quit;

proc reg data=A outest=EST0;
  model y = x1-x&P / noint selection=adjrsq aic bic sbc; ** no intercept;
  run; quit;
```

```
data ESTOUT;
  set EST EST0; run;

  proc sort data=ESTOUT; by _aic_;
  proc print data=ESTOUT(obs=1); title 'best model by AIC from proc reg'; run;

  proc sort data=ESTOUT; by _bic_;
  proc print data=ESTOUT(obs=1); title 'best model by BIC from proc reg'; run;

  proc sort data=ESTOUT; by _sbc_;
  proc print data=ESTOUT(obs=1); title 'best model by SBC from proc reg'; run;
```

COMPARISON OF PROC REG WITH EXPLICIT ENUMERATION

The "best" models (i.e., models with the lowest AIC, BIC, or SBC values) identified by PROC REG with the selection=adjrsq option were compared with the "best" models identified through explicit enumeration of all possible subset models. Since PROC REG displays only a fraction of all possible models for $p > 10$, it is unknown how often PROC REG with the selection=adjrsq option identifies the same best model as found by explicit enumeration of all possible subset models.

Simulations were run for each p where $1 \le p \le 15$. For $p \le 6$ all possible subset models were evaluated using the selection=adjrsq option. For $7 \le p \le 15$, 100 randomly selected models from among all possible subset models were evaluated for each p due to the exponential growth of the number of all subset models for larger values of p. For $p \le 10$, PROC REG with the selection=adjrsq option displays all subset models except the intercept only model. For $p > 10$ only a small fraction of all possible models are displayed using the selection=adjrsq option.

Table 1 summarizes the results of the comparison between an explicit enumeration of all possible subset models with the selection=adjrsq option using AIC, BIC, and SBC for $1 \le p \le 15$. The second column of Table 1 shows the total number of possible subset models for each p is $2^{p+1} - 1$ since the intercept term is either included or excluded along with the p independent regressor X variables. The third column of Table 1 shows the number of models displayed using the selection=adjrsq option along with the percent (column 3 divided by column 2) of all possible subset models for each p. The fourth column of Table 1 shows the number of models evaluated for this paper using the selection=adjrsq option. The last three columns of Table 1 show the number of best models evaluated using the selection=adjrsq option that agree with the best models identified using explicit enumeration for AIC, BIC, and SBC. Note that the last three columns of Table 1 are not quantifying the number of models that correctly identify the true underlying model, only the number of models where both explicit enumeration and selection=adjrsq option are in agreement as to the best model. Table 1 shows that the best models identified by AIC agree more often than BIC or SBC with the best models identified by explicit enumeration.

The number of best models identified by BIC using PROC REG with the selection=adjrsq option in agreement with the explicit enumeration of all models was considerably smaller than for either AIC or SBC. This is because SAS sometimes calculates BIC differently using PROC REG with the selection=adjrsq option and PROC REG without the selection=adjrsq option for the same model. It is unknown why the calculation for BIC differs between these methods for the same model using identical data. The author suspects a different estimate for the pure error variance fitting the full model (σ^2), since this is the only other parameter used in the calculation of BIC that is not used for AIC or SBC.

The reason the number of models in the last three columns may not match the number of all subset models evaluated for $p \le 6$ in column four is because PROC REG without the selection=adjrsq option calculates information criteria for the intercept only model, whereas PROC REG with the selection=adjrsq option does not. Therefore, when the intercept only model is the best model having the lowest AIC, BIC, or SBC value, then the PROC REG with the selection=adjrsq option cannot identify it as the model with the lowest information criteria since it does not consider it.

Figure 1 is a graph of the percent of all possible subset models shown using PROC REG with the selection=adjrsq option from the third column of Table 1. Clearly, the percent monotonically increases from 66.7% (two out of three models) for $p = 1$ to 99.95% for $p = 10$. However, for $p = 11$, the percent plummets to 5.4% and then decreases monotonically for $p > 11$. While the number of models shown for each p increases linearly for $p > 11$, the number of all possible subset models increases exponentially.

Table 1. Summary table of results for $1 \le p \le 15$

Number of regressors (p)	Total number of subset models ($2^{p+1} - 1$)	Number of models shown with selection=adjrsq option with percent of all possible subset models shown	Number of models evaluated	Number of models in agreement		
				AIC	BIC	SBC
1	3	2 (66.7%)	3	2	2	2
2	7	6 (85.7%)	7	7	7	6
3	15	14 (93.3%)	15	14	14	14
4	31	30 (96.8%)	31	30	30	30
5	63	62 (98.4%)	63	63	62	62
6	127	126 (99.2%)	127	127	121	127
7	255	254 (99.6%)	100	99	95	99
8	511	510 (99.8%)	100	100	95	100
9	1023	1022 (99.9%)	100	100	92	100
10	2047	2046 (99.95%)	100	100	96	100
11	4095	222 (5.4%)	100	100	92	97
12	8191	266 (3.2%)	100	100	90	95
13	16,383	314 (1.9%)	100	100	83	98
14	32,767	366 (1.1%)	100	95	83	89
15	65,535	422 (0.6%)	100	100	79	90

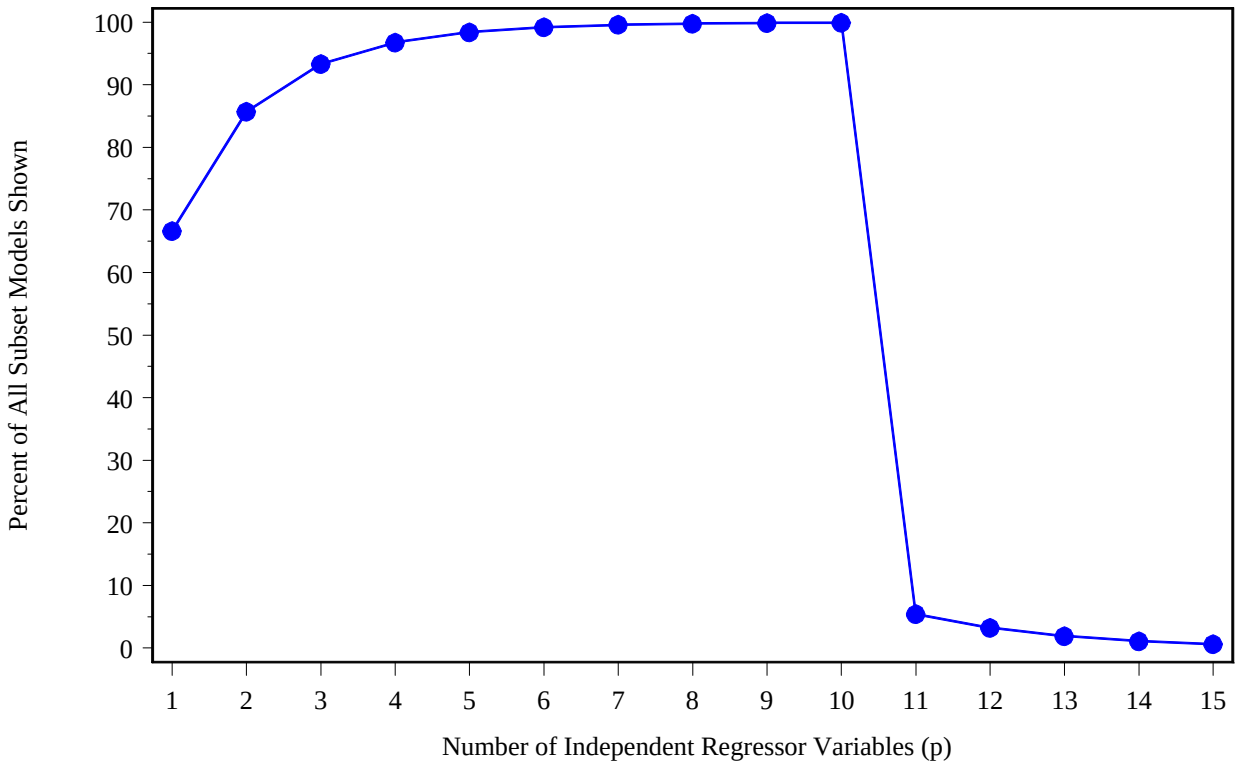

Figure 1. Line plot of the percent of all possible subset models shown using `PROC REG` with the `selection=adjrsq` option for the number of independent regressor variables $1 \le p \le 15$.

Figure 2 shows the percent of all models evaluated where PROC REG with the selection=adjrsq option agreed with the explicit enumeration of all possible subset models for AIC, BIC, and SBC using the ratios of the last three columns to the fourth column of Table 1. AIC more consistently selects the same best model using PROC REG with the selection=adjrsq option that agrees with an explicit enumeration of all models over BIC and SBC. BIC performs less consistently than AIC or SBC because SAS sometimes calculates BIC differently using PROC REG with the selection=adjrsq option than PROC REG without the selection=adjrsq option for the same model with identical data. BIC also is trending lower as the number of independent regressor variables *p* increases.

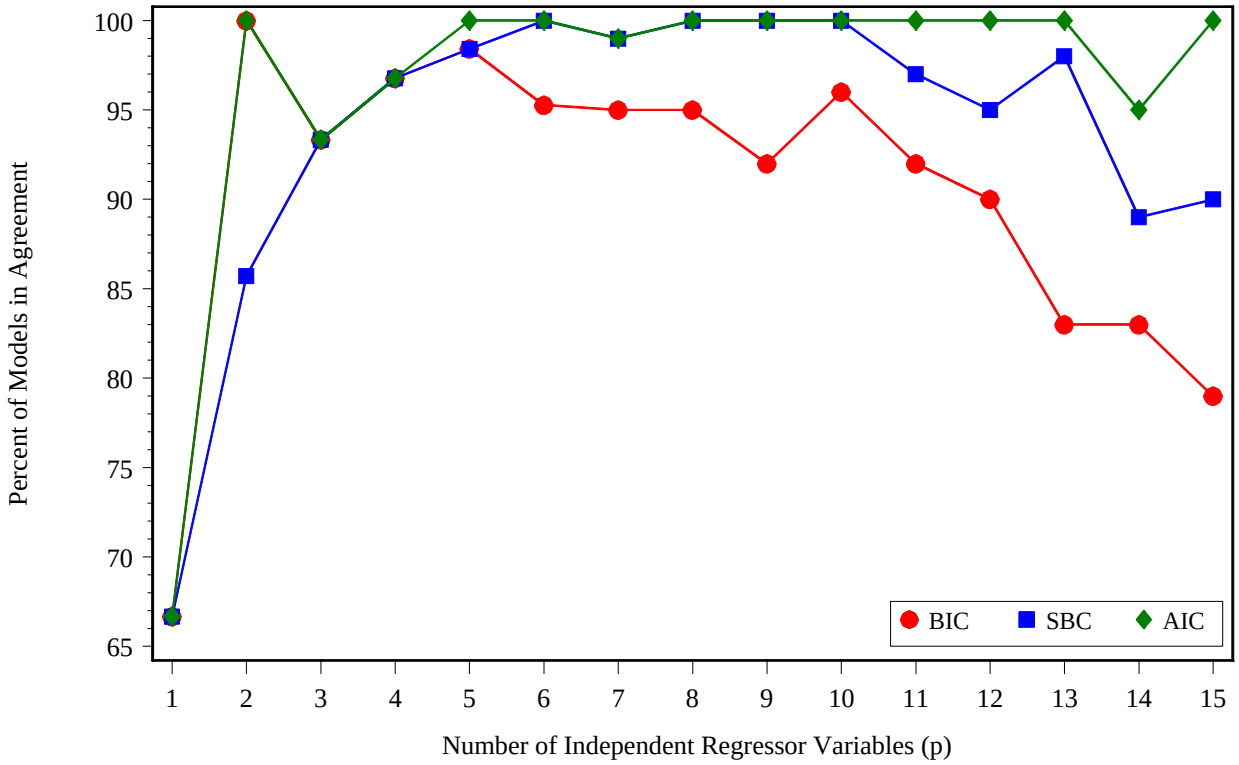

Figure 2. Line plot of the percent of all models evaluated where PROC REG with the selection=adjrsq option agreed with explicit enumeration using AIC, BIC, and SBC for $1 \le p \le 15$.

CONCLUSION

SAS is a powerful tool that utilizes AIC, BIC, and SBC to simultaneously evaluate all possible subsets of multiple linear regression models to determine the best model for up to $p = 10$ independent variables. However, for $p > 10$ variables, only a fraction of all possible models are shown using the selection=adjrsq option when identifying the best subset model with the smallest information criterion. SAS code was presented using explicit enumeration of all possible subset models for any $p > 0$. This code was used to compare the best model among all possible subsets with the best model determined by PROC REG displaying only a fraction of all possible models using the selection=adjrsq option.

Simulations were run for $1 \le p \le 15$ using AIC, BIC, and SBC to show that PROC REG with the selection=adjrsq option agreed on the same best model as explicit enumeration most consistently using AIC. BIC performed less consistently than AIC or SBC because SAS sometimes calculates BIC differently for PROC REG using the selection=adjrsq option than it does for PROC REG without using the selection=adjrsq option for the same model parameters using identical data.

The data analyst can identify the best model using any of the three information criterion for any $p > 0$ by using the explicit enumeration SAS code presented in this paper. Since the number of models grows exponentially as *p* increases, the SAS code provided can be used up to the limits of available computer speed and memory. For small to

moderate *p*, an explicit enumeration of all possible subset models can be run using the SAS code presented to ensure the best model is found for the selected information criterion. For large *p*, explicit enumeration likely will be computationally intractable. In this case, using `PROC REG` with the `selection=adjrsq` option will likely identify the best model using AIC as it considers a subset of all possible models. However, Beal (2007) showed that SBC more consistently selected the true model over AIC using large and small simulated data sets. In the case of very large *p*, an implicit enumeration algorithm should identify the best model with the lowest information criterion, as discussed in Bao (2005), with a minimum of iterations.

REFERENCES

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle". In B.N. Petrov and F. Csaki (Eds.), *Second international symposium on information theory*, 267-281. Budapest: Academiai Kiado.

Akaike, H. 1987. "Factor analysis and AIC". *Psychometrika* 52:317-332.

Bao, X. 2005. "An implicit enumeration algorithm for mining high dimensional data". *International Journal of Operational Research* 1:123-143.

Beal, D. J. 2007. "Information criteria methods in SAS® for multiple linear regression models". *Proceedings of the Fifteenth Annual Conference of the SouthEast SAS Users Group*, Hilton Head, SC.

Beal, D. J. 2005. "SAS code to select the best multiple linear regression model for multivariate data using information criteria". *Proceedings of the Thirteenth Annual Conference of the SouthEast SAS Users Group*, Portsmouth, VA.

Bozdogan, H. 1987. "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions". *Psychometrika* 52:345-370.

Bozdogan, H. 2000. "Akaike's information criterion and recent developments in informational complexity". *Journal of Mathematical Psychology* 44:62-91.

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

Schwarz, G. 1978. "Estimating the dimension of a model". *Annals of Statistics* 6:461-464.

CONTACT INFORMATION

The author welcomes and encourages any questions, corrections, feedback, and remarks. Contact the author at:

Dennis J. Beal, Ph.D. candidate
Senior Statistician / Risk Scientist
Science Applications International Corporation
P.O. Box 2501
151 Lafayette Drive
Oak Ridge, Tennessee 37831
phone: 865-481-8736
fax: 865-481-4757
e-mail: dennis.j.beal@saic.com