

Paper SA05

Information Criteria Methods in SAS® for Multiple Linear Regression Models

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

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. This paper provides the SAS code that can be used to simultaneously evaluate up to 1023 models to determine the best subset of variables that minimizes the information criteria among all possible subsets. Simulated multivariate data are used to compare the performance of AIC, BIC, and SBC with model diagnostics root mean square error (RMSE), Mallows' Cp, and adjusted R^2 , and the three heuristic methods forward selection, backward elimination, and stepwise regression. This paper is for intermediate SAS users of SAS/STAT who understand multivariate data analysis.

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 from the independent variables. A data set with p independent variables has $2^p - 1$ possible subset models to consider since each of the p variables is either included or excluded from the model, not counting interaction terms or the intercept. Diagnostics are calculated for each model to help determine which model is "best". These model diagnostics include the root mean square error (RMSE), the adjusted coefficient of determination (R^2), and Mallows' Cp. A good linear model will have small RMSE and Cp and a high adjusted R^2 close to 1. The three common heuristic methods include forward selection, backward selection, and stepwise regression. However, these model diagnostics and heuristic methods alone are insufficient to determine the best model.

This paper will compare these model diagnostic and heuristic techniques to minimizing the information criteria statistics Akaike's Information Criteria (AIC), Sawa's Bayesian Information Criteria (BIC), and Schwarz Bayesian Criteria (SBC) on simulated data from several distributions. The SAS code for determining the best linear model will be shown. 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.

MODEL DIAGNOSTICS

Three model diagnostic statistical techniques often taught in statistics courses to determine the best linear model include minimizing the RMSE, maximizing the adjusted R^2 , and minimizing Mallows' Cp.

ROOT MEAN SQUARE ERROR

The RMSE is a function of the sum of squared errors (SSE), number of observations n , and the number of independent variables $k \leq p + 1$ where k includes the intercept and is shown in Eqn. (1).

$$RMSE = \sqrt{\frac{SSE}{n-k}} \quad (1)$$

The RMSE is calculated for all possible subset models. Using this technique, the model with the smallest RMSE is declared the best linear model. This approach does include the number of parameters in the model. Additional parameters will decrease the numerator since the SSE decreases as additional variables are included in the model and the denominator decreases as k increases.

ADJUSTED R²

The adjusted coefficient of determination R² is the percentage of the variability of the dependent variable that is explained by the variation of the independent variables after accounting for the intercept and number of independent variables $k \leq p$. Therefore, the adjusted R² value ranges from 0 to 1 and is a function of the total sum of squares (SST), the SSE, number of observations n , number of independent variables $k \leq p + 1$ where k may include the intercept, and the binomial variable i ($i = 1$ if the intercept is included in the model, $i = 0$ otherwise). The equation for the adjusted R² is shown in Eqn. (2).

$$adjR^2 = 1 - \frac{(n-i) \cdot SSE}{(n-k) \cdot SST} \quad (2)$$

The adjusted R² is calculated for all possible subset models. Using this technique, the model with the largest adjusted R² is declared the best linear model. This approach also includes the number of variables k in the model; thus, additional parameters will decrease both the numerator and denominator. However, several models often will have an adjusted R² = 1, so determining the best model among tied values is problematic.

MALLOWS CP

Mallows' Cp (Mallows 1973) is another model diagnostic that is a function of the SSE, the full model pure error estimate σ^2 , number of observations n , and the number of independent variables $k \leq p + 1$ where k includes the intercept. Mallows' Cp is shown in Eqn. (3).

$$Cp = \frac{SSE}{\sigma^2} + 2k - n \quad (3)$$

Mallows' Cp is calculated for all possible subset models. Using this technique, the model with the smallest Cp is declared the best linear model. As the number of independent variables k increases, an increased penalty term ($2k$) is offset with a decreased SSE.

HEURISTIC STATISTICAL TECHNIQUES

Three heuristic statistical techniques often taught in statistics courses to determine the best linear model include forward selection, backward elimination, and stepwise regression.

FORWARD SELECTION

Forward selection begins with only the intercept term in the model. For each of the independent variables, the F statistic is calculated to determine each variable's contribution to the model. The variable with the smallest p -value below a specified α cutoff value (e.g., 0.15) indicating statistical significance is kept in the model. The model is rerun keeping this variable and recalculating F statistics on the remaining $p - 1$ independent variables. This process continues until no remaining variables have F statistic p -values below the specified α . Once a variable is in the model, it remains in the model.

BACKWARD ELIMINATION

Backward elimination begins by including all variables in the model and calculating F statistics for each variable. The variable with the largest p -value exceeding the specified α cutoff value is then removed from the model. This process continues until no remaining variables have F statistic p -values above the specified α . Once a variable is removed from the model, it cannot be added back into the model.

STEPWISE REGRESSION

Stepwise regression is a modification of the forward selection technique in that variables already in the model do not necessarily stay in. As in the forward selection technique, variables are added one at a time to the model, as long as the F statistic p -value is below the specified α . After a variable is added, however, the stepwise technique evaluates all of the variables already included in the model and removes any variable that has an insignificant F statistic p -value exceeding the specified α . Only after this check is made and the identified variables have been removed can another variable be added to the model. The stepwise process ends when none of the variables excluded from the model

has an F statistic significant at the specified α and every variable included in the model is significant at the specified α .

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 criteria measure 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. Other authors who use AIC for model selection include Akaike (1987) and Bozdogan (1987, 2000). Akaike's Information Criteria (AIC) is a function of the number of observations n , the SSE, and the number of independent variables $k \leq p + 1$ where k includes the intercept, as shown in Eqn. (4).

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2k \quad (4)$$

The first term in Eqn. (4) 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 $k \leq p + 1$ where k includes the intercept, as shown in Eqn. (5).

$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \frac{2(k+2)n\sigma^2}{SSE} - \frac{2n^2\sigma^4}{SSE^2} \quad (5)$$

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 $k \leq p + 1$ where k includes the intercept, as shown in Eqn. (6).

$$SBC = n \cdot \ln\left(\frac{SSE}{n}\right) + k \ln n \quad (6)$$

The penalty term for SBC is similar to AIC in Eqn. (4), 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 10 independent variables and one dependent variable was simulated from known “true” models that are linear functions of a subset of the independent variables. The following SAS code simulates 1000 observations for these 10 independent X variables. The 10 independent X variables come from normal, lognormal, exponential, and uniform distributions with various means and variances. Variables X5, X6, X9, and X10 are correlated with other variables.

```

data a;
do i = 1 to 1000;
  x1 = 10 + 5*rannor(0);      * Normal(10, 25);
  x2 = exp(3*rannor(0));    * Lognormal;
  x3 = 5 + 10*ranuni(0);    * Uniform;
  x4 = 50 + 10*rannor(0);   * Normal(100, 2500);
  x5 = x1 + x4 + rannor(0); * Normal bimodal;
  x6 = 5 + 2*x2 + 3*ranexp(0); * Lognormal and exponential mixture;
  x7 = 0.5*exp(4*rannor(0)); * Lognormal;
  x8 = 10 + 8*ranuni(0);    * Uniform;
  x9 = x2 + x8 + 2*rannor(0); * Lognormal, uniform and normal mix;
  x10 = 20 + x7 + 9*rannor(0); * Lognormal and normal mix;
output;
end;
drop i; run;

```

The dependent variable Y was calculated using the simulated data for all $2^{10} - 1 = 1023$ possible subset “true” models with an intercept and another 1023 subset models without an intercept. The coefficient terms for the true models were randomly assigned for all 10 variables for each true model. A pure random error term $\varepsilon = 2 * \text{rannor}(0)$ was added to each simulated Y observation. AIC, BIC, SBC, the three heuristic statistical techniques, and the three model diagnostics discussed earlier were calculated for every possible model. The “best” model determined by each method was compared with the true simulated model.

SAS CODE FOR INFORMATION CRITERIA

SAS calculates the AIC, BIC, and SBC for every possible subset of variables for models with up to 10 independent variables. The following SAS code from SAS/STAT computes AIC, BIC, and SBC for all possible subsets of multiple regression models for main effects. Note that “x1-x10” in the `model` statement indicates that all variables x1, x2, x3, x4, x5, x6, x7, x8, x9, and x10 are included. The `selection=adjrsq` option specifies the adjusted R^2 method will be used to select the model so that all possible 1023 models are considered, although other `selection` options may also be used such as `selection=rsquare`. The `SSE` option displays the sum of squared errors for each model, while the `AIC`, `BIC`, and `SBC` options display the AIC, BIC, and SBC statistics for each model, respectively. The `adjrsq`, `cp`, and `rmse` options display the adjusted R^2 , Mallows’ Cp, and the RMSE statistics for each model, respectively. The first `PROC REG` calculates AIC, BIC, SBC, adjusted R^2 , Mallows’ Cp, and the RMSE for all possible subsets of main effects using an intercept term. The second `PROC REG` calculates AIC, BIC, SBC, adjusted R^2 , Mallows’ Cp, and the RMSE 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.

```

proc reg data=a outest=est;
  model y = x1-x10 / selection=adjrsq sse aic bic sbc adjrsq cp rmse;
run; quit;

proc reg data=a outest=est0;
  model y = x1-x10 / noint selection=adjrsq sse aic bic sbc adjrsq cp rmse;
run; quit;

data estout;
  set est est0; run;

proc sort data=estout; by _aic_;
proc print data=estout(obs=8); title 'best model by AIC'; run;

proc sort data=estout; by _bic_;
proc print data=estout(obs=8); title 'best model by BIC'; run;

proc sort data=estout; by _sbc_;
proc print data=estout(obs=8); title 'best model by SBC'; run;

```

SAS CODE FOR MODEL DIAGNOSTICS

The following SAS code determines the best models for the model diagnostics adjusted R^2 , Mallows' Cp, and RMSE.

```
proc sort data=estout; by _rmse_;
proc print data=estout(obs=8); title 'best model by RMSE'; run;

proc sort data=estout; by _cp_;
proc print data=estout(obs=8); title 'best model by CP'; run;

proc sort data=estout; by descending _adjrsq_; ** want largest adjusted R2;
proc print data=estout(obs=8); title 'best model by adjusted R2'; run;
```

SAS CODE FOR HEURISTIC METHODS

Independent variables X1 through X10 were regressed against the dependent variable Y using forward selection, backward selection, and stepwise regression with an assumed entry and exit significance level $\alpha = 0.15$. An entry significance level $\alpha = 0.15$, specified in the `slentry=0.15` option, means a variable must have a p -value ≤ 0.15 to enter the model during forward selection and stepwise regression. An exit significance level of 0.15, specified in the `slstay=0.15` option, means a variable must have a p -value ≥ 0.15 to leave the model during backward selection and stepwise regression.

The following SAS code performs the forward selection method by specifying the option `selection=forward`. The model diagnostics are output into the data set `est1`.

```
proc reg data=a outest=est1;
  model y=x1-x10 / slentry=0.15 selection=forward sse aic bic sbc;
run; quit;
```

The following SAS code performs the backward elimination method by specifying the option `selection=backward`. The model diagnostics are output into the data set `est2`.

```
proc reg data=a outest=est2;
  model y=x1-x10 / slstay=0.15 selection=backward sse aic bic sbc;
run; quit;
```

The following SAS code performs stepwise regression by specifying the option `selection=stepwise`. The model diagnostics are output into the data set `est3`.

```
proc reg data=a outest=est3;
  model y=x1-x10 / slstay=0.15 slentry=0.15 selection=stepwise sse aic bic sbc;
run; quit;
```

COMPARISON OF METHODS

SAMPLE SIZE $n = 1000$

Fig. 1 shows the frequency each method group correctly selected the true models from all 2046 possible models (including the intercept term) using a sample size $n = 1000$. In general, the information criteria methods (AIC, BIC and SBC) performed best by correctly selecting 63% of the true models. The heuristic methods (stepwise, backward, and forward selection) correctly selected 43% of the true models. The model diagnostics (adjusted R^2 , Cp, and RMSE) performed the worst by correctly selecting only 28% of the true models.

Fig. 1 also shows the percent each method correctly selected the true models. Clearly, the information criterion SBC easily outperformed the other methods by correctly selecting 96% of the true models. Information criteria AIC and BIC correctly selected 45% and 46% of the true models, respectively. Across all methods stepwise regression was a distant second to SBC by correctly selecting 54% of the true models. Backward selection correctly selected 46% of the true models, while forward selection correctly selected only 29% of the true models. For the model diagnostics, Mallows' Cp correctly selected 47% of the true models. However, the adjusted R^2 and RMSE methods performed the worst by selecting only 20% and 18% of the true models, respectively.

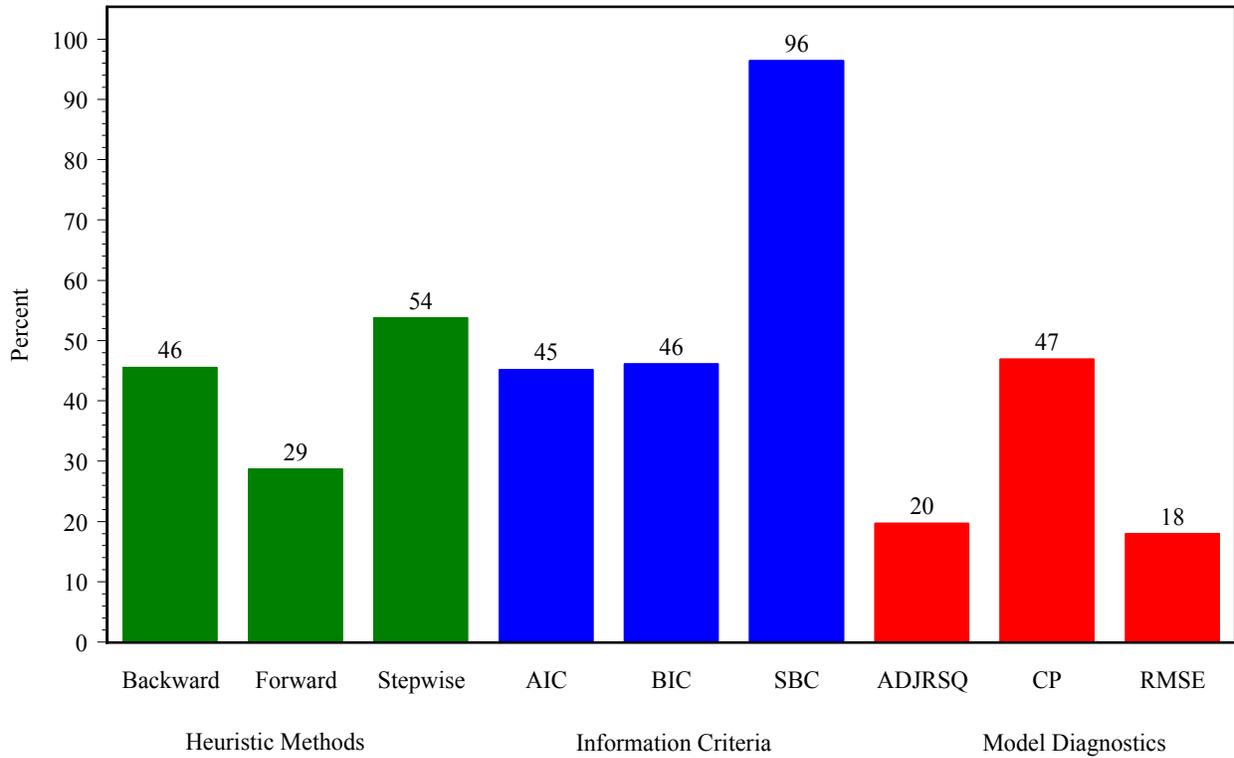
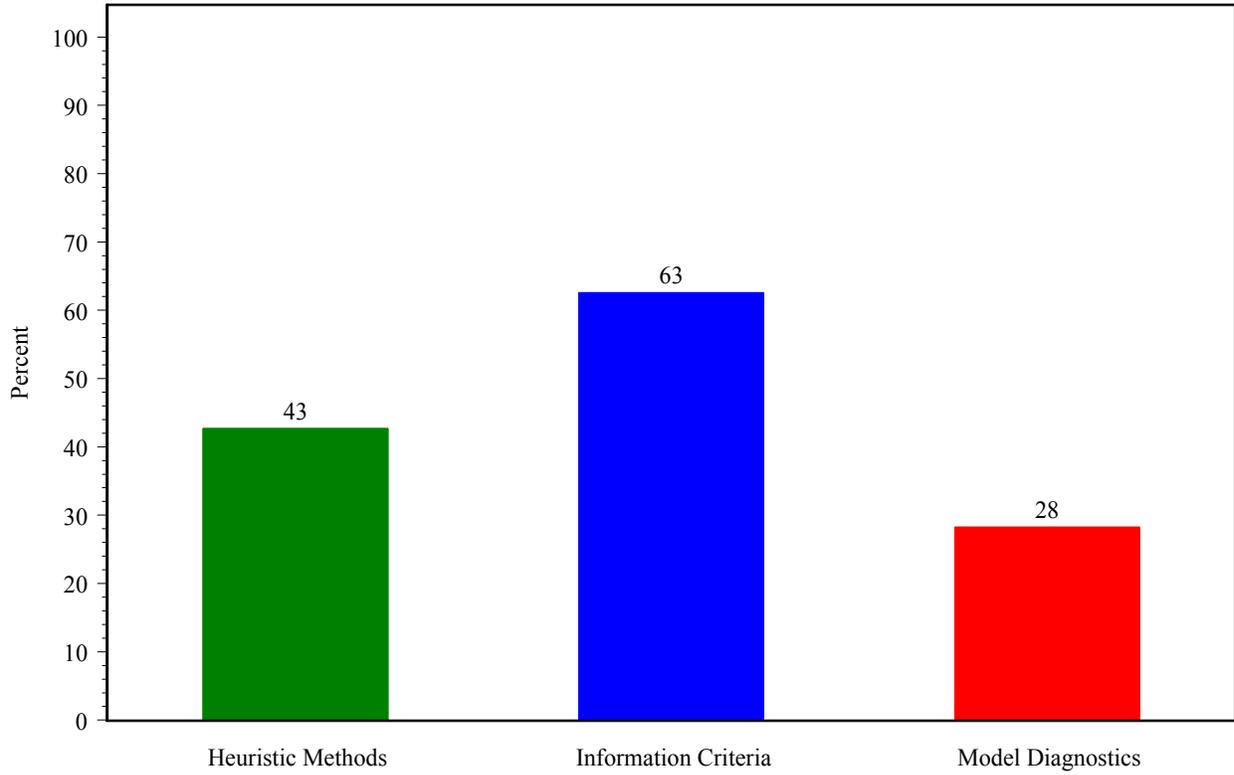


Fig. 1. Histograms of percent each method correctly selected the true model of all possible models ($n = 1000$)

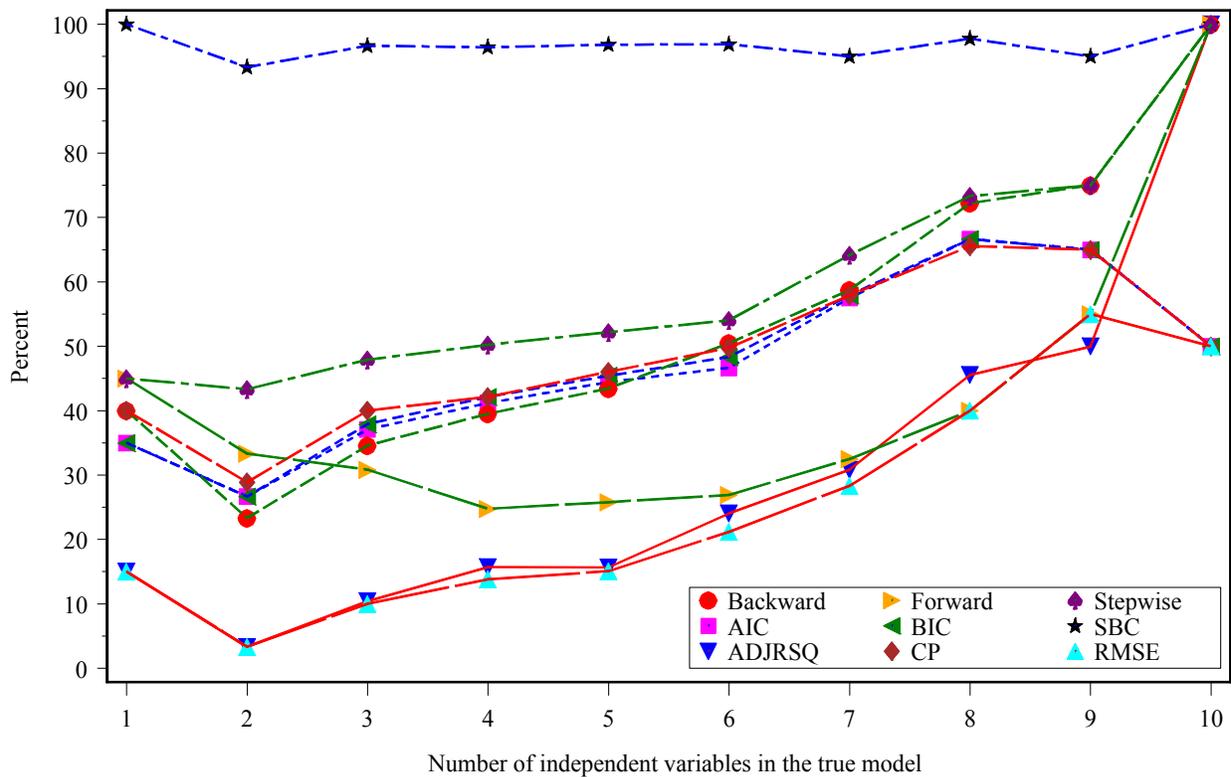


Fig. 2. Line plot of percent of models correctly selected for each method by the number of independent variables in the true model ($n = 1000$)

Fig. 2 examines further the results from Fig. 1 by plotting the percent each method correctly selected the true models by the number of independent variables in the true model. The success rate for SBC consistently exceeded 90% regardless of the number of variables in the true model. All methods generally improved as the number of independent variables increased except for forward selection. Forward selection performed better when either one or two independent variables or at least nine independent variables were in the true model. Adjusted R^2 and RMSE performed the worst of all the methods, particularly for up to seven independent variables in the true model.

Given these results for a large sample size of $n = 1000$, a natural extension to consider is whether these results still hold for a smaller sample size. The first 100 observations were selected from the simulated data set of 1000 observations. The SAS code was rerun using this smaller data set of $n = 100$ observations so the performance of these methods for selecting the true models can be compared with the full data set.

SAMPLE SIZE $n = 100$

The frequency each method group correctly selected the true models from all 2046 possible models using a sample size $n = 100$ is shown in Fig. 3. In general, the information criteria methods again performed best by correctly selecting 59% of the true models. The heuristic methods (stepwise, backward, and forward selection) correctly selected 41% of the true models. The model diagnostics (adjusted R^2 , Cp, and RMSE) performed the worst by correctly selecting only 26% of the true models. These results agree well with Fig. 1.

Fig. 3 also shows the percent each method correctly selected the true models for $n = 100$. The information criterion SBC again easily outperformed the other methods by correctly selecting 83% of the true models. Information criteria AIC and BIC correctly selected 42% and 52% of the true models, respectively. Across all methods stepwise regression was a distant second to SBC by correctly selecting 53% of the true models. Backward selection correctly selected 45% of the true models, while forward selection correctly selected only 26% of the true models. For the model diagnostics, Mallows' Cp correctly selected 43% of the true models. However, the adjusted R^2 and RMSE methods performed the worst by selecting only 17% and 18% of the true models, respectively. Note with a smaller sample size, less information from the data is available to correctly select the true models. Thus, the percentages in Fig. 3 are typically smaller than in Fig. 1, but still consistent.

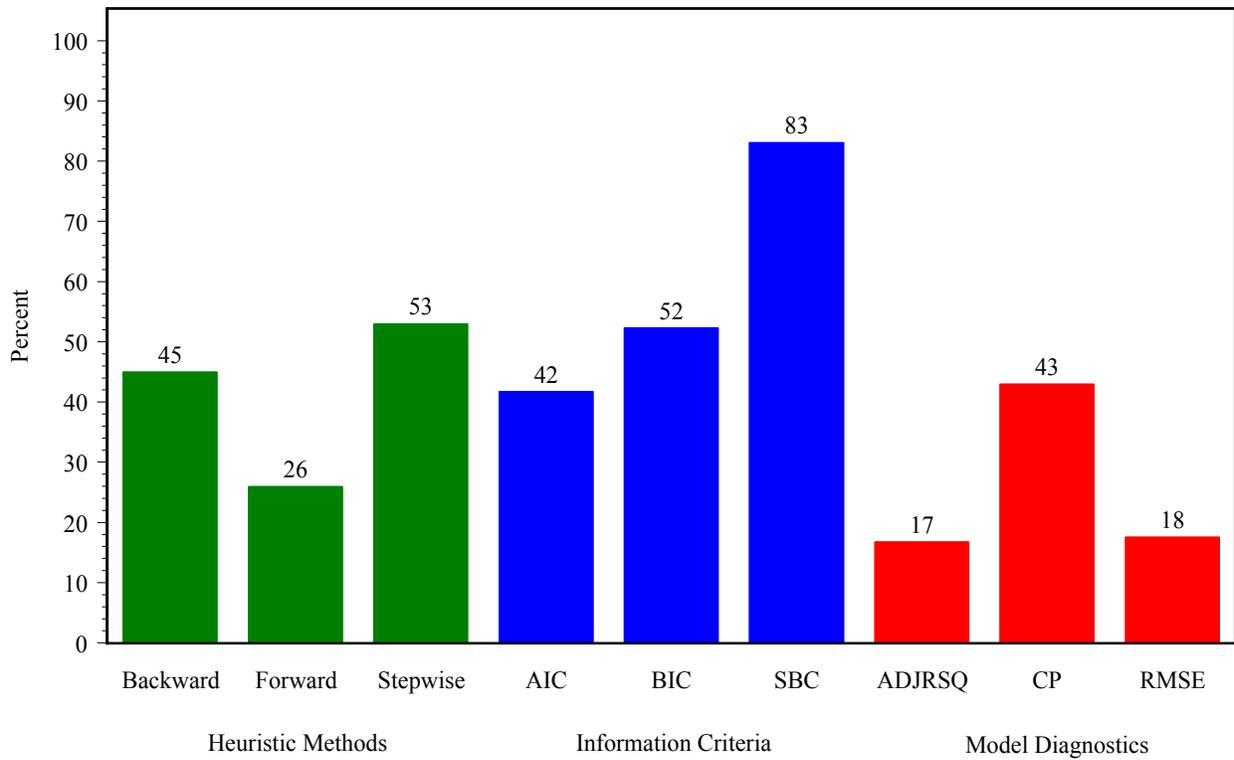
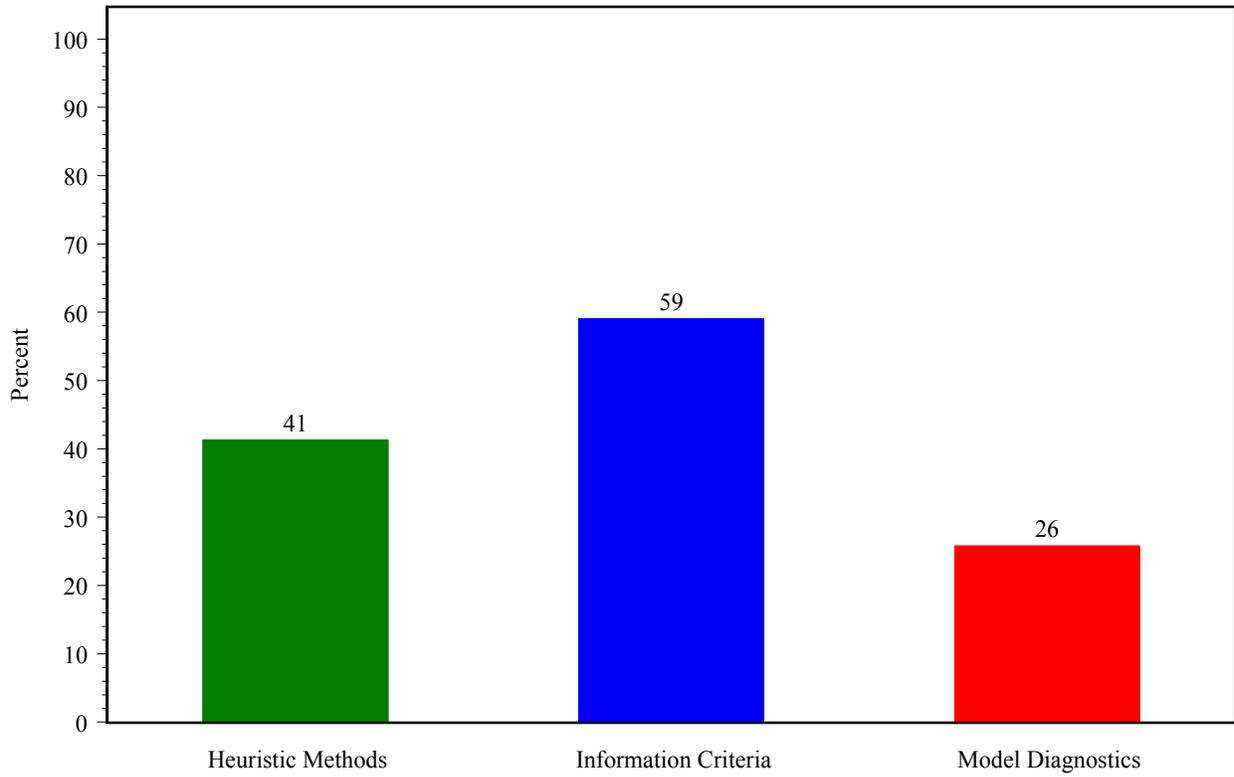


Fig. 3. Histograms of percent each method correctly selected the true model of all possible 2046 models ($n = 100$)

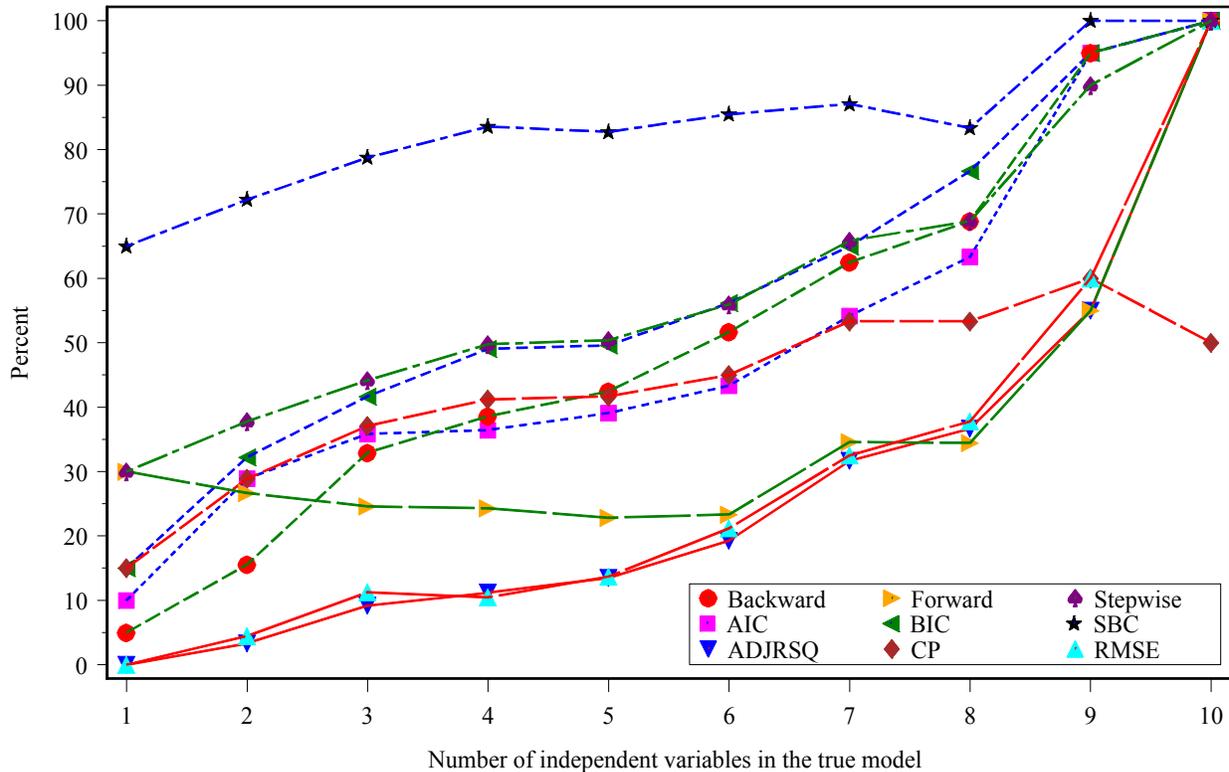


Fig. 4. Line plot of percent of models correctly selected for each method by the number of independent variables in the true model ($n = 100$)

Fig. 4 examines further the results from Fig. 3 by plotting the percent each method correctly selected the true models by the number of independent variables in the true model. The success rate for SBC consistently exceeded 65% regardless of the number of variables in the true model. All methods generally improved as the number of independent variables increased except for forward selection. Forward selection performed better when either one or two independent variables or at least nine independent variables were in the true model. Adjusted R^2 and RMSE again performed the worst of all the methods, particularly for up to seven independent variables in the true model.

CONCLUSION

SAS is a powerful tool that utilizes AIC, BIC, and SBC to simultaneously evaluate all possible subsets of multiple regression models to determine the best model for up to 10 independent variables. Using information criteria for multivariate model selection has been shown to be superior to heuristic methods (forward selection, backward elimination, stepwise regression) and model diagnostics (adjusted R^2 , Mallows' Cp and RMSE) using simulated data with a known underlying model. Among the three information criteria methods studied, SBC performed best by correctly selecting the true model the most consistently after enumerating all possible true models from the simulated data. SBC consistently performed best for both sample sizes $n = 1000$ and $n = 100$.

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: Akademiai Kiado.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52, 317-332.
- Beal, D J. (2005). SAS code to select the best multiple linear regression model for multivariate data using information criteria, *Proceedings of the 13th Annual Conference of the SouthEast SAS Users Group*.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychometrika*, 52, No. 3, 345-370.

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

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, 15, 661-675.

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
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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.