Paper SA01_05

SAS® Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria

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

ABSTRACT

Multiple linear regression is a standard statistical tool that regresses *p* independent variables against a single dependent variable. The objective is to find a linear model that best predicts the dependent variable from the independent variables. Information criteria uses the covariance matrix and the number of parameters in a model to calculate a statistic that summarizes the information represented by the model by balancing a trade-off between a lack of fit term and a penalty term. SAS® calculates Akaike's Information Criteria (AIC) for every possible 2^p models for $p \leq 10$ independent variables. AIC estimates a measure of the difference between a given model and the "true" model. The model with the smallest AIC among all competing models is deemed the best model. This paper provides SAS code that can be used to simultaneously evaluate up to 1024 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 to select the true model with standard statistical techniques such as minimizing RMSE, 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, multivariate 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 possible subset models to consider since each of the *p* variables is either included or excluded from the model, not counting interaction terms. Model diagnostics are calculated for each model to help determine which model is "best". These model diagnostics include the root mean square error (RMSE) and the coefficient of determination (R^2). A good linear model will have a low RMSE and a high R^2 close to 1. However, these model diagnostics alone are insufficient to determine the best model.

The usual techniques taught in statistics courses to find the best linear model include minimizing the RMSE, maximizing R^2, forward selection, backward elimination and stepwise regression. This paper will compare these techniques to minimizing the information criteria statistic on simulated data from several distributions. SAS code for determining the best linear model will be shown.

COMMON STATISTICAL TECHNIQUES

Five common statistical techniques taught in most statistics courses to determine the best linear model include minimizing the RMSE, maximizing R^2, forward selection, backward elimination and stepwise regression.

The RMSE is a function of the sum of squared errors (SSE), number of observations *n* and the number of parameters *p* and is shown in Eqn. (1).

$$RMSE = \sqrt{\frac{SSE}{n-p}} \tag{1}$$

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; so additional parameters will decrease both the numerator and denominator.

The coefficient of determination R^2 is the percentage of the variability of the dependent variable that is explained by the variation of the independent variables. Therefore, the R^2 value ranges from 0 to 1. R^2 is a function of the total sum of squares (SST) and the SSE and is shown in Eqn. (2).

$$R^2 = 1 - \frac{SSE}{SST} \tag{2}$$

The R^2 is calculated for all possible subset models. Using this technique, the model with the largest R^2 is declared the best linear model. However, this technique has several disadvantages. First, the R^2 increases with each variable included in the model. Therefore, this approach encourages including all variables in the best model although some variables may not significantly contribute to the model. This approach also contradicts the principal of parsimony that encourages as few parameters in a model as possible.

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 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 to the model again.

Stepwise regression is a modification of the forward selection technique in that variables already in the model do not necessarily stay there. 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.

Other model selection techniques not evaluated in this paper include adjusted R^2 and Mallow's C_p. Hocking (1976) and Sclove (1987) discuss the use of these and other statistical techniques in model selection.

INFORMATION CRITERIA

Information criteria is a measure of goodness of fit or uncertainty for the range of values of the data. In the context of multiple linear regression, information criteria measures the difference between a given model and the "true" underlying model. Akaike (1973) introduced the concept of information criteria as a tool for optimal model selection. Akaike's Information Criteria (AIC) is a function of the number of observations *n*, the SSE and the number of parameters *p*, as shown in Eqn. (3).

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2p \tag{3}$$

The first term in Eqn. (3) is a measure of the model lack of fit while the second term is a penalty term for additional parameters in the model. Therefore, as the number of parameters *p* 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.

Akaike (1973) forms the basis for the concept of information criteria. Other references that use AIC for model selection include Akaike (1987), Bozdogan (1987 and 2000) and Sawa (1978).

EXAMPLE DATA

A multivariate data set with 10 independent variables and one dependent variable was simulated from a known "true" model that is a linear function of a subset of the independent variables. The following SAS code simulates 1000 observations for these 10 independent X variables and one dependent Y variable. The 10 independent X variables come from normal, lognormal, exponential and uniform distributions with various means and variances. Variables X5, X6 and X9 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 = 100 + 50*rannor(0);              * Normal(100, 2500);
  x5 = x1 + 3*rannor(0);                * normal bimodal;
  x6 = 2*x2 + 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 = 200 + 90*rannor(0);             * normal(200, 8100);
  y = 3*x2 - 4*x8 + 5*x9 + 3*rannor(0); * true model with no intercept term;
  output;
  end;
```

SAS CODE FOR AIC

The following SAS code from SAS/STAT computes AIC for all possible subsets of multiple regression models for main effects. The `selection=adjrsq` option specifies the adjusted R^2 method will be used to select the model, 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` option displays the AIC statistic for each model. The first `proc reg` calculates AIC for all possible subsets of main effects using an intercept term. The second `proc reg` calculates AIC 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 to largest. The model with the smallest AIC value is deemed the "best" model. The SAS code presented in this paper uses the SAS System for personal computers version 8.2 (TS level 02M0) running on a Windows 2000 platform.

```
proc reg data=a outest=est;
  model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / selection=adjrsq sse aic ;
  output out=out p=p r=r; run; quit;

proc reg data=a outest=est0;
  model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 / noint selection=adjrsq sse aic ;
  output out=out0 p=p r=r; run; quit;

data estout;
  set est est0; run;

  proc sort data=estout; by _aic_;
  proc print data=estout(obs=8); run;
```

COMPARISON OF AIC RESULTS WITH HEURISTIC METHODS

SAS will calculate the AIC for every possible subset of variables for models with up to 10 independent variables. SAS confirmed the minimum AIC for all possible subsets of variables is 2239.73 with only the X2, X8 and X9 variables in the model and no intercept term.

Independent variables X1 through X10 were regressed against the no intercept dependent variable Y using forward selection, backward selection and stepwise regression with an assumed entry and exit significance level of 0.15. An entry significance level of 0.15, specified in the `slentry=0.15` option, means a variable must have a *p*-value < 0.15 in order 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 in order 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 x2 x3 x4 x5 x6 x7 x8 x9 x10 / slstay=0.15 slentry=0.15
    selection=forward ss2 sse aic;
  output out=out1 p=p r=r; 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 x2 x3 x4 x5 x6 x7 x8 x9 x10 / slstay=0.15 slentry=0.15
    selection=backward ss2 sse aic;
  output out=out1 p=p r=r; 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 x2 x3 x4 x5 x6 x7 x8 x9 x10 / slstay=0.15 slentry=0.15
    selection=stepwise ss2 sse aic;
  output out=out3 p=p r=r; run; quit;
```

The following SAS code calculates the RMSE for each possible subset model, sorts the models from smallest to largest RMSE and then prints the best 10 models. Specifying `adjrsq` in the option `selection=adjrsq` is not crucial since the goal is to minimize RMSE. Other choices for the `selection` option are `rsquare` or `CP`. The model diagnostics are output into the data sets `est4` and `est5`.

```
proc reg data=a outest=est4;
  model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
    selection=adjrsq sse aic adjrsq;
  output out=out p=p r=r;          run; quit;

proc reg data=a outest=est5;
  model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
    noint selection=adjrsq sse aic adjrsq;
  output out=out p=p r=r;          run; quit;

data both; set est4 est5;  run;
    proc sort data=both; by _rmse_; run;
    proc print data=both(obs=10); run;
```

Table 1 shows that all four heuristic methods include variables X2, X3, X8, X9 and X10 along with an intercept term. A heuristic method is an approximate method that does not guarantee convergence to the optimal model. The AIC for these models is 2240.3, which is higher than the AIC for the true model. Forward, backward and stepwise regression methods selected the same model that includes a nonzero intercept and variables X3 and X10 that are not part of the true underlying model. The minimized RMSE method includes variables X3, X7 and X10 in addition to the nonzero intercept term that are not part of the true underlying model and has the largest AIC of the methods in Table 1. The model that minimized AIC is the true underlying model. Therefore, Table 1 illustrates how the forward, backward, stepwise regression and minimizing RMSE heuristic methods fail to identify the underlying model which minimizes AIC.

Table 1. A comparison of AIC with four heuristic methods

	Parameter estimates					
Variable	True Model	Minimized AIC	Forward selection	Backward selection	Stepwise	Minimized RMSE
Intercept			1.30456	1.30456	1.30456	1.30058
X1						
X2	3	3.0212	3.01216	3.01216	3.01216	3.0118
X3			-0.04905	-0.04905	-0.04905	-0.04722
X4						
X5						
X6						
X7						-6.415E-06
X8	-4	-3.98233	-4.02578	-4.02578	-4.02578	-4.02631
X9	5	4.97887	4.98791	4.98791	4.98791	4.98827
X10			-0.00166	-0.00166	-0.00166	-0.0016849
Model Diagnostics						
R^2		1	1	1	1	1
Adjusted R^2		1	1	1	1	1
RMSE		3.05985	3.05615	3.05615	3.05615	3.05613
AIC		2239.73	2240.3	2240.3	2240.3	2241.27
F		1.07E+10	6.44E+09	6.44E+09	6.44E+09	5.36E+09
Pr > F		<.0001	<.0001	<.0001	<.0001	<.0001

CONCLUSION

SAS is a powerful tool that utilizes AIC 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 such as forward selection, backward elimination, stepwise regression and minimizing RMSE using simulated data with a known underlying model.

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.

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.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32, 1-49.

Sclove, S. L. (1987). Application of model selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343.

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

CONTACT INFORMATION

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

Dennis J. Beal
Statistician / Risk Scientist
Science Applications International Corporation
P.O. Box 2501
151 Lafayette Drive
Oak Ridge, Tennessee 37831
phone: 865-481-8736
fax: 865-481-8714
e-mail: dennis.j.beal@saic.com