Paper SAS4287-2020

A Survey of Methods in Variable Selection and Penalized Regression

Yingwei Wang, SAS Institute Inc.

ABSTRACT

Statistical learning often deals with the problem of finding a best predictive model from a set of possible models on the basis of the observed data. "Best" often means most parsimonious; thus a sparse model that is composed of a subset of variables is usually preferable to a full model that uses all input variables because of its better interpretability and higher prediction accuracy. To this extent, systematic approaches such as variable selection methods for choosing good interpretable and predictive models have been developed. This paper reviews variable selection methods in linear regression, grouped into two categories: sequential methods, such as forward selection, backward elimination, and stepwise regression; and penalized methods, also called shrinkage or regularization methods, including the LASSO, elastic net, and so on. In addition to covering mathematical properties of the methods, the paper presents practical examples using SAS/STAT® software and SAS® Viya®.

INTRODUCTION

The first two decades of this century have witnessed a technological revolution in data collection, leading to the point where high-dimensional and large-scale data provide the foundation for many business applications and fields of scientific research. Because of the increasing amount and complexity of data, greater computing power and highly reliable analysis are required for building statistical models that interpret the insights or predict the future responses accurately. Linear regression models are often preferred to more complicated statistical models because you can fit them relatively easily. Moreover, linearity with respect to fixed functions of the predictors is often an adequate first step in handling more complex behavior.

Suppose you observe the independent pairs $\{(y_i, \mathbf{x}_i)\} \in \mathbb{R} \times \mathbb{R}^p$ in the following linear regression model,

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots n$$

where $\mathbf{x}_i^T$ is a row vector that represents the predictors for the ith observation, y_i is the corresponding ith response variable, $\{\epsilon_j\}_{j=1}^p$ are centered iid noise terms with constant variance σ^2 and are independent of the predictors, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is the vector of linear regression coefficients. $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denotes the design matrix, with $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ as its ith row and with $\boldsymbol{X}_j = (x_{j1}, x_{j2}, \ldots, x_{jn})^T$ as its jth column, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$. Then the linear regression model can be written as the matrix-vector form

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The goal is to estimate the regression coefficients $\boldsymbol{\beta}$ on the basis of the observed sample $(\mathbf{y}, \boldsymbol{X})$. Meanwhile, it is assumed that only a few predictors among $\{X_j\}_{j=1}^p$ have a significant influence on the response $\mathbf{y}$ that warrants the use of model selection procedures.

Statistical variable selection involves taking full advantage of the observed data to infer relationships between the predictor variables, aiming at either discovering the insights into an existing phenomenon (interpretation) or making predictions based on partial information (prediction). Whether the goal is interpretation or prediction, the key task is to learn the important features and stable characteristics of the data. Indeed, quite frequently, the true solution to the coefficients $\boldsymbol{\beta}$ can be well approximated by a sparse vector, where only a few variables are truly important, whereas the remaining variables have values of either exactly zero or nearly zero. Therefore, the objective of variable selection is to figure out which variables (predictors, causes, effects, and so on) are the most relevant for explaining or predicting a phenomenon of interest.

Variable selection methods in linear regression are grouped into two categories: sequential selection methods, such as forward selection, backward elimination, and stepwise regression; and penalized regression methods, also known as shrinkage or regularization methods, including the LASSO, elastic net, and their modifications and combinations. Sequential selection methods are easy to interpret but are a discrete search process in which variables are either included in or excluded from the model. Penalization techniques for variable selection in regression models are

alternatives that are more continuous and do not suffer as much from this variability. They are becoming increasingly popular because they are able to perform variable selection while simultaneously estimating the coefficients in the model.

This paper describes the variable selection methods available in SAS/STAT and SAS Viya procedures. In addition to fitting models, these procedures provide modern approaches for building models by selecting variables and effects, such as classification effects and spline effects. These approaches rely on ordinary least squares and penalized least squares as theoretical frameworks for variable selection and feature extraction.

Ordinary Least Squares Regression and Sequential Selection Methods

Sequential selection methods estimate the regression coefficients for candidate models by solving the following ordinary least squares (OLS) problem:

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y} \tag{1}$$

In this paper, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is assumed to be invertible unless otherwise stated. The fitted values of the response $\mathbf{y}$ are

$$\hat{\mathbf{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{ols}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}$$

Usually, the OLS estimate $\hat{\boldsymbol{\beta}}^{\text{ols}}$ in Equation (1) is totally dense; that is, all values of the estimated coefficients are nonzero. However, as mentioned in the previous section, estimation and prediction that use the full model of all p covariates might not perform well or can even fail as a result of the accumulation of noise, high collinearity, spurious correlation, and lack of interpretability. The goal is to identify a smaller subset of these predictors that exhibit the strongest effects; this is the central issue in variable selection.

The following sequential selection methods are among the most popular and widely used techniques in variable selection. They provide systematic ways to search through models and fit a sequence of regression models. At each step, new models are obtained by adding or deleting one predictor variable from the models at the previous stages.

- Forward selection starts with the null model. In the first step, it fits all the single variable models and selects the predictor variable that makes the best individual contribution. Here "best" means the best selection criterion shown in the section "Model Selection Criteria," such as lowest AIC, lowest cross validation error, and so on. At each step, the procedure continues in this way, adding a candidate variable that improves model fitting the most, given the variables already in the model. The algorithm stops when the stop criterion is satisfied or when there are no more candidate variables. The stop criteria that are supported in variable selection procedures are also shown in the section "Model Selection Criteria."
- Backward elimination is similar to forward selection, but it moves in the opposite direction. That is, starting with the full model, at each step you consider eliminating the variable that has the least impact on the model, given the other variables already included. Again, you can use a predetermined threshold for dropping variables from the model to decide whether you can indeed remove the candidate. When no more candidates meet the criterion for removal, the algorithm stops.
- In both forward selection and backward elimination, after a variable has been acted on, that action cannot be reversed. Hence, a variable that was eliminated at some earlier point during a backward elimination step, for example, is never allowed back into the model. This lack of flexibility is remedied in the stepwise approach to variable selection. Here, at each step, the algorithm considers either adding or deleting each variable, until it reaches a point where no inclusion or elimination improves the model. In other words, a variable might be included in an early stage but removed later, or a variable that was removed from the model might be allowed back in.

Although these techniques were originated for linear regression models to help solve the variable selection problem, they can also be applied in settings that extend the basic linear models, such as polynomial regression, generalized linear models, and Cox proportional hazards models. For these other model types, the residual sum of squares in Equation (1) would be replaced by deviance or other relevant measures. Here the discussion is restricted to the linear regression context, with the understanding that you can use the search philosophy in other settings as well.

Model Selection Criteria

The total sum of squares (SST) and the sum of squared estimate of errors (SSE) are defined by

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2 \quad \text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Table 1 provides the criteria for the CHOOSE=, SELECT=, and STOP= options in the SELECTION statement in the REGSELECT procedure.

Table 1 Criteria Used in SAS/STAT and SAS Viya Procedures

Options	Statistics	Formula
ADJRSQ	R^2_{adj}	$1 - \frac{(n-1)(1-R^2)}{n-p}$
AIC	Akaike's information criterion	$n \log\left(\frac{\text{SSE}}{n}\right) + 2p + n + 2$
AICC	Corrected Akaike's information criterion	$n \log\left(\frac{\text{SSE}}{n}\right) + \frac{n(n+p)}{n-p-2}$
BIC | SBC	Schwarz Bayesian information criterion	$n \ln\left(\frac{\text{SSE}}{n}\right) + p \ln(n)$
CP	Mallows' C_p	$\frac{\text{SSE}}{\hat{\sigma}^2} + 2p - n$
PRESS	Predicted residual sum of squares	$\sum_{i=1}^{n} \frac{r_i^2}{(1-h_i)^2}$ where r_i = residual at observation i and h_i = leverage of observation $i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$
RSQUARE	R^2	$1 - \frac{\text{SSE}}{\text{SST}}$
SL	Significance level used to assess the contribution of an effect to the fit when it is added to or removed from a model	
VALIDATE	Average square error over the validation data	

Penalized Least Squares Regression and Shrinkage Selection Methods

A penalization technique can be described as follows. In general, a shrinkage method solves the penalized least squares (PLS) problem in Lagrangian form,

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \boldsymbol{X\beta}\|_2^2 + P_\lambda(\boldsymbol{\beta}) \right\} \tag{2}$$

where $P_\lambda(\cdot)$ is the sparsity-inducing penalty function on the coefficient vector $\boldsymbol{\beta}$, and nonnegative λ is also called the decay/tuning/regularization parameter that controls the trade-off between data fitting and regularization. In contrast to the OLS solution, the penalty has the effect of shrinking the coefficients, even setting some to zero. This approach produces a spectrum of solutions, depending on the value of λ; such methods are often referred to as regularization or shrinkage methods.

Here the coordinate-wise separable penalty functions are considered; in other words, $P_\lambda(\boldsymbol{\beta})$ can be decomposed as

$$P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_\lambda(\beta_j)$$

There are many choices for the penalty function $p_\lambda(\cdot)$. Let's focus on two of them:

- LASSO: $p_\lambda(t) = \lambda|t|$
- Elastic net: $p_\lambda(t) = \lambda_1|t| + \lambda_2 t^2$

For each fixed λ, you can find an optimal solution by solving the PLS problem (Expression 2). After the solution paths have been obtained by solving the PLS problem on a grid of λ's, it is important to decide on a rule to choose the optimal solution, or equivalently, the best tuning parameter λ. There are several ways to do this. In a data-rich environment, you could use a validation set to compute the average square error. If no validation set is available, you can use other techniques, such as cross validation or an information criterion shown in the section "Model Selection Criteria" to either stop the process or choose the best model.

Next, four kinds of penalized selection are briefly introduced: LASSO, adaptive LASSO, elastic net, and adaptive elastic net. For more information about these methods, see, for example, Hastie, Tibshirani, and Friedman (2009).

LASSO Selection

Tibshirani (1996) proposed the least absolute selection and shrinkage operator (LASSO), which minimizes the residual sum of squares under a constraint on the ℓ_1 norm of the coefficient vector $\boldsymbol{\beta}$. The LASSO solves the optimization problem

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

The reason for using the ℓ_1 penalty is twofold: the geometry of the ℓ_1 norm induces sparse solutions, and the convexity greatly simplifies the computation. Furthermore, the use of the ℓ_1 penalty in the LASSO has been the foundation of many other penalization techniques for variable selection.

The introduction of the LASSO opened the doors to simultaneously performing variable selection and providing regularized estimated coefficients via penalization. For small values of λ, the LASSO method produces ordinary least squares estimates. Increasing λ in discrete steps leads to a sequence of coefficient estimates, where some are exactly zero and the rest, which correspond to selected effects, are shrunk toward zero.

The LASSO solution path for linear regression, which constitutes the trajectory of coefficient estimates as a function of λ in Equation (3), is piecewise linear with changes in slope where variables enter or leave the active set. The LASSO solution path can be efficiently computed by using the least angle regression (LARS) algorithm (Efron et al. 2004). This algorithm provides the complete solution path for the LASSO problem by taking advantage of the fact that the solutions are piecewise linear with respect to λ, and it builds a connection between the LASSO and the forward stepwise selection. It has the same order of computational efforts as a single OLS fit $O(np^2)$.

Adaptive LASSO Selection

One potential drawback of the LASSO is that the same shrinkage effect that sets many estimated coefficients exactly to zero also shrinks all nonzero estimated coefficients toward zero. One possible solution is to use the weighted penalty approach.

Zou (2006) proposed the adaptive LASSO to permit different weights for different parameters; that is,

$$\hat{\boldsymbol{\beta}}^{\text{a-lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |w_j \beta_j| \right\} \tag{4}$$

where the $\{w_j\}_{j=1}^{p}$ are the weights to ensure good sampling performance of the adaptive LASSO estimator. This adaptivity can permit larger penalties to be imposed on unimportant covariates and smaller penalties to be imposed on important variables.

The most commonly applied adaptive LASSO takes $w_j = 1/\hat{\beta}_j^{\text{ols}}$ with the convention that when $\hat{\beta}_j^{\text{ols}} = 0$ (that is, $w_j = \infty$), the jth variable is excluded in this second stage. Furthermore, if $|\hat{\beta}_j^{\text{ols}}|$ is large, the adaptive LASSO uses a small penalty (that is, a little shrinkage) for the jth coefficient β_j, which implies less bias. Thus, the adaptive LASSO yields a sparse solution.

You can obtain the adaptive LASSO estimator (Equation 4) by solving a LASSO-type problem; that is,

$$\hat{\boldsymbol{\beta}}_*^{\text{a-lasso}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \boldsymbol{X}_*\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

$$\hat{\boldsymbol{\beta}}_j^{\text{a-lasso}} = w_j \hat{\boldsymbol{\beta}}_{*,j}^{\text{a-lasso}}$$

where $\boldsymbol{X}_* = \boldsymbol{X}\boldsymbol{W}^{-1}$ and the diagonal matrix $\boldsymbol{W} = \text{diag}\{1/w_j\}$.

Elastic Net Selection

Although the LASSO is widely used in variable selection, it has several drawbacks:

- If the number of variables is greater than the number of observations (that is, $n < p$), or if the number of informative variables (variables that are relevant for the model) is expected to be greater than n, the LASSO can select at most n variables before it stops, and the model might perform poorly.
- When there are groups of correlated variables, the LASSO tends to randomly select only one variable from a group and ignore all the others.
- For the usual $n > p$ scenarios, if there are high correlations between predictors, the prediction performance of the LASSO is dominated by ridge regression (Tibshirani 1996).

To address these drawbacks, Zou and Hastie (2005) proposed the elastic net method to combine the benefits of ℓ_1 and ℓ_2 regularizations; that is,

$$\hat{\boldsymbol{\beta}}^{\text{enet}}_{\text{naïve}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \tag{5}$$

The elastic net method strikes a balance between having a parsimonious model and borrowing strength from correlated regressors, by solving the least squares regression problem with constraints on both the sum of the absolute coefficients and the sum of the squared coefficients. The ℓ_1 part of the penalty performs variable selection by setting some coefficients to exactly 0, and the ℓ_2 part of the penalty encourages the group selection by shrinking the coefficients of correlated variables toward each other.

Elastic net can be treated as a convex combination of the LASSO and ridge penalty, with pure LASSO and pure ridge as two limiting cases. If λ_1 is set to 0, then the elastic net method reduces to the ridge regression. If λ_2 is set to 0, then the elastic net method reduces to the LASSO. If λ_1 and λ_2 are both set to 0, then the elastic net method reduces to OLS regression.

The elastic net method has several advantages. First, it can enforce sparsity. Second, it has no limitation on the number of selected variables. Third, it encourages a grouping effect in the presence of highly correlated predictors.

The naïve elastic net (Equation 5) could be transformed into a LASSO-type problem in an augmented space:

$$\hat{\boldsymbol{\beta}}^{\text{enet}}_{\text{naïve}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\}$$

where

$$\boldsymbol{X}^* = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I} \end{pmatrix}, \quad \mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \tag{6}$$

Adaptive Elastic Net Selection

The adaptive elastic net selection method, proposed by Zou and Zhang (2009), is an improved version of the elastic net and adaptive LASSO selection methods. It penalizes the squared error loss by using a combination of the ℓ_2 penalty and the adaptive ℓ_1 penalty; that is,

$$\hat{\boldsymbol{\beta}}^{\text{a-enet}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} |w_j\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

Like the naïve elastic net, the adaptive elastic net can also be transformed into an adaptive LASSO-type problem in some augmented space

$$\hat{\boldsymbol{\beta}}^{\text{a-enet}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} |w_j\beta_j| \right\}$$

where $\boldsymbol{X}^*$ and $\mathbf{y}^*$ are as defined in Equation (6).

This implies that for each fixed λ_2, you can efficiently solve both elastic net and adaptive elastic net problems by using the LARS algorithm.

Variable Selection Procedures in SAS/STAT and SAS Viya

Both SAS/STAT and SAS Viya provide a rich set of tools for performing variable selection by using sequential and penalized methods. Table 2 summarizes the variable selection methods that the SAS/STAT and SAS Viya procedures support. The methods are listed in increasing order of complexity.

Table 2 Methods of Variable Selection in SAS/STAT and SAS Viya Procedures

Method	PROC REG	PROC GLMSELECT	PROC HPREG	PROC REGSELECT
Forward	Yes	Yes	Yes	Yes
Forward swap	No	No	Yes	Yes
Backward	Yes	Yes	Yes	Yes
Stepwise	Yes	Yes	Yes	Yes
LAR	No	Yes	Yes	Yes
LASSO	No	Yes	Yes	Yes
Adaptive LASSO	No	Yes	Yes	Yes
Elastic net	No	Yes	No	Yes
Adaptive elastic net	No	No	No	Yes
Group LASSO	No	Yes	No	No

There are many useful references in the proceedings of previous SAS® Global Forum conferences that discuss how to perform variable selection by using the procedures shown in Table 2. For a discussion of penalized regression and the GLMSELECT procedure, see Günes (2015). For guidance on high-performance statistical modeling and the HPREG procedure, see Cohen and Rodriguez (2013). Also, you can find review papers about regression modeling in SAS/STAT procedures (Rodriguez 2016) and SAS Viya procedures (Rodriguez and Cai 2018).

This section focuses on penalized methods in the REGSELECT procedure in SAS Viya and presents three examples that are related to the limitations of LASSO shown in the section "Elastic Net Selection."

Example 1: Analyzing Baseball Data Containing a Small Number of Observations

The **Sashelp.Baseball** data set contains salary and performance information for Major League Baseball players who played at least one game in both the 1986 and 1987 seasons. The salaries are from the 1987 season (Time Inc. 1987), and the performance measures are from 1986 season (Collier Books 1987). You can load the **Sashelp.Baseball** data set into your CAS session by using your CAS engine libref named mycas with the following DATA step:

```
data mycas.baseball;
    set sashelp.baseball;
 run;
```

Suppose you want to investigate whether you can model the players' salaries from the 1987 season by using performance measures from the previous season. Also, instead of all players' salaries, you are interested only in the catchers' salaries. You can use the following statements to perform LASSO selection for the **Baseball_Catchers** data:

```
ods graphics on;
proc regselect data = mycas.baseball(where=(Position='C'));
    partition roleVar = league(train='National' validate = 'American');
    class division;
    model logSalary = division nAtBat nHits nHome nRuns nRBI nBB yrMajor crAtBat
                      crHits crHome crRuns crRbi crBB nOuts nAssts nError;
    selection  method = lasso(choose=validate);
run;
```

The PARTITION statement assigns observations to training and validation roles on the basis of the values of the input variable **league**. The CHOOSE=VALIDATE option in the SELECTION statement selects the model that yields the smallest average square error (ASE) value for the validation data.

Figure 2 shows the model selected by the LASSO. Figure 1 shows that 15 observations were used for training and the number of effects is 19. This implies that you are unlikely to obtain a good result by using LASSO selection because

the number of effects is greater than the number of observations. In this case, you can use elastic net selection by submitting the following statements:

```
proc regselect data=mycas.baseball(where=(Position='C'));
    partition roleVar = league(train='National' validate = 'American');
    class division;
    model logSalary = division nAtBat nHits nHome nRuns nRBI nBB yrMajor crAtBat
                      crHits crHome crRuns crRbi crBB nOuts nAssts nError;
    selection  method = elasticnet(choose=validate);
run;
```

Figure 1 Basic Information about **Baseball_Catchers** Data

The REGSELECT Procedure

Number of Observations Read	40
Number of Observations Used	30
Number of Observations Used for Training	15
Number of Observations Used for Validation	15
Number of Observations Used for Testing	0

Class Level Information

Class	Levels	Values
Division	2	East West

Dimensions	
Number of Effects	19
Number of Parameters	19

Figure 2 Details of the Model Selected by LASSO for **Baseball_Catchers** Data

Selected Model by LASSO

Root MSE	0.66348
R-Square	0.68718
Adj R-Sq	0.51340
AIC	9.03005
AICC	25.03005
SBC	-3.72165
ASE (Train)	0.26413
ASE (Validate)	0.30621

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	5.093821
nHits	1	0.001115
nRuns	1	0.006507
nBB	1	0.004172
CrAtBat	1	0.000237
nAssts	1	0.000372

You can specify the L2= suboption of the METHOD=ELASTICNET option in the SELECTION statement if you have a good estimate of the ridge regression parameter. If you do not, you can omit the L2= suboption; in this case, PROC REGSELECT estimates this regression parameter according to the criterion specified in the CHOOSE= option, by

trying a series of candidate values for the ridge regression parameter. The optimal regression parameter is set to the value that achieves the minimum validation ASE.

The "Elastic Net Selection Summary" table in Figure 3 shows that the optimal L2 value is 0.1468 and the minimal validation ASE is 0.1797. Figure 4 shows the details of the model that the elastic net method selects by using the optimal L2 value. If you compare the ASE values for the model in Figure 2 and the model in Figure 4, you can see that the elastic net method selects a model that has lower values of both training and validation ASEs than the LASSO method.

Figure 3 Elastic Net Selection Summary for **Baseball_Catchers** Data

Elastic Net Summary

ENstep	L2	Number Effects In	Validation ASE
1	0.00000000	6	0.3062
2	0.00000001	6	0.3062
3	0.00000001	6	0.3062
4	0.00000002	6	0.3062
5	0.00000003	6	0.3062
6	0.00000005	6	0.3062
7	0.00000007	6	0.3062
8	0.00000010	6	0.3062
9	0.00000015	6	0.3062
10	0.00000022	6	0.3062
	.	.	.
	.	.	.
	.	.	.
41	0.03162278	9	0.1993
42	0.04641589	9	0.1803
43	0.06812921	9	0.1866
44	0.10000000	9	0.1854
45	0.14677993	9	0.1797*
46	0.21544347	9	0.2985
47	0.31622777	4	0.5786
48	0.46415888	4	0.6210
49	0.68129207	4	0.6675
50	1.00000000	4	0.7458

*** Optimal Value Of Criterion**

Figure 4 Details of the Model Selected by Elastic Net for **Baseball_Catchers** Data

Selected Model by Elastic Net

Root MSE	0.60490
R-Square	0.82666
Adj R-Sq	0.59553
AIC	6.17483
AICC	61.17483
SBC	-4.45272
ASE (Train)	0.14636
ASE (Validate)	0.17967

Figure 4 *continued*

Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	4.435995
nHits	1	0.001148
nRuns	1	0.007537
nBB	1	0.006697
YrMajor	1	0.028748
CrAtBat	1	0.000118
CrHits	1	0.000383
nAssts	1	0.003097
nError	1	0.017681

You can also perform adaptive elastic net selection on the data when parameter estimates have different scales, as you can observe from the "Parameter Estimates" table in Figure 4. The following statements perform adaptive elastic net selection for the **Baseball_Catchers** data:

```
proc regselect data=mycas.baseball(where=(Position='C'));
    partition roleVar = league(train='National' validate = 'American');
    class division;
    model logSalary = division nAtBat nHits nHome nRuns nRBI nBB yrMajor crAtBat
                      crHits crHome crRuns crRbi crBB nOuts nAssts nError;
    selection  method = elasticnet(adaptive choose=validate);
run;
ods graphics off;
```

The "Adaptive Elastic Net Selection Summary" table in Figure 5 shows that the optimal L2 value is 0.0215 and the minimal validation ASE is 0.1661. Figure 6 shows that adaptive elastic net selection can produce better results, in the sense that the selected model is more sparse and produces lower values of training and validation ASEs.

Figure 5 Adaptive Elastic Net Selection Summary for **Baseball_Catchers** Data

Elastic Net Summary			
ENstep	**L2**	**Number Effects In**	**Validation ASE**
1	0.00000000	6	0.2780
2	0.00000001	5	0.1736
3	0.00000001	5	0.1736
	.	.	.
	.	.	.
	.	.	.
37	0.00681292	6	0.1952
38	0.01000000	5	0.1789
39	0.01467799	6	0.1667
40	0.02154435	7	0.1661*
41	0.03162278	7	0.1688
42	0.04641589	6	0.1986
43	0.06812921	6	0.2385
	.	.	.
	.	.	.
	.	.	.
48	0.46415888	4	0.5973
49	0.68129207	4	0.6637
50	1.00000000	4	0.7061
* **Optimal Value Of Criterion**			

Figure 6 Details of the Model Selected by Adaptive Elastic Net for **Baseball_Catchers** Data

Selected Model by Adaptive Elastic Net

Root MSE	0.46566
R-Square	0.86303
Adj R-Sq	0.76031
AIC	-1.35829
AICC	22.64171
SBC	-13.40194
ASE (Train)	0.11565
ASE (Validate)	0.16609

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	4.159782
nRuns	1	0.015687
nBB	1	0.003749
YrMajor	1	0.027027
CrAtBat	1	0.000256
nAssts	1	0.002826
nError	1	0.029585

Example 2: Analyzing Simulation Data with Grouping Effect

This simple simulation example is taken from the original elastic net paper (Zou and Hastie 2005), which shows how elastic net performs group selection as opposed to the LASSO. The following DATA step code generates the **Grouping** data set:

```
data mycas.Grouping;
  drop i j;
  array x{6} x1-x6;
  array z{2} z1-z2;
    do i=1 to 100;
      do j = 1 to 2;
         z{j} = 20*ranuni(1);
      end;
      y = z1 + 0.1*z2 + rannor(1);
      x1 = z1 + 0.25*rannor(1);    x2 = -z1 + 0.25*rannor(1);    x3 = z1 + 0.25*rannor(1);
      x4 = z2 + 0.25*rannor(1);    x5 = -z2 + 0.25*rannor(1);    x6 = z2 + 0.25*rannor(1);
      output;
    end;
run;
```

By construction, the response vector y is generated by

$$y = z_1 + 0.1z_2 + \epsilon, \quad \epsilon \sim N(0, 1)$$

where two independent "hidden" factors (z_1 and z_2) are generated from a uniform distribution in the range of 0 to 20,

$$z_1, z_2 \sim \text{Uniform}(0, 20)$$

Also, the observed predictors $(x_1, x_2, \ldots, x_6)$ are generated from the "hidden" factors (z_1, z_2) in the following way,

$$x_1 = z_1 + \epsilon_1 \qquad x_2 = -z_1 + \epsilon_2 \qquad x_3 = z_1 + \epsilon_3$$
$$x_4 = z_2 + \epsilon_4 \qquad x_5 = -z_5 + \epsilon_5 \qquad x_6 = z_2 + \epsilon_6$$

where $\{\epsilon_i\}_{i=1}^{6}$ are independent identically distributed $N(0, 1/16)$. A total of 100 observations were generated from this model. The variables x_1, x_2, x_3 and x_4, x_5, x_6 form two groups whose identifying factors are z_1 and z_2, respectively.

The within-group correlations are almost 1 and the between-group correlations are almost 0. A good selection procedure would identify the variables x_1, x_2, x_3 (z_1 group) together.

You can use the following statements to perform variable selection via the LASSO method and elastic net with fixed $L2 = 0.5$:

```
ods graphics on;
proc regselect data = mycas.Grouping;
    model y = x1-x6;
    selection  method = lasso plots = coefficients;
run;

proc regselect data = mycas.Grouping;
    model y = x1-x6;
    selection  method = elasticnet(L2=0.5)  plots = coefficients;
run;
ods graphics off;
```

The PLOTS=COEFFICIENTS option in the SELECTION statement produces the results in Figure 7 and Figure 8, which show the coefficient progression plots that are generated by the LASSO and elastic net selection, respectively. You can see that in elastic net selection, the variables x_1, x_2, and x_3 join the model as a group long before the other group members x_4, x_5, and x_6, whereas in LASSO selection the group selection is not clear. Also, the elastic net solution path is smoother and more stable than the LASSO path.

Figure 7 LASSO Coefficient Progression for Simulation Data with Grouping Effect

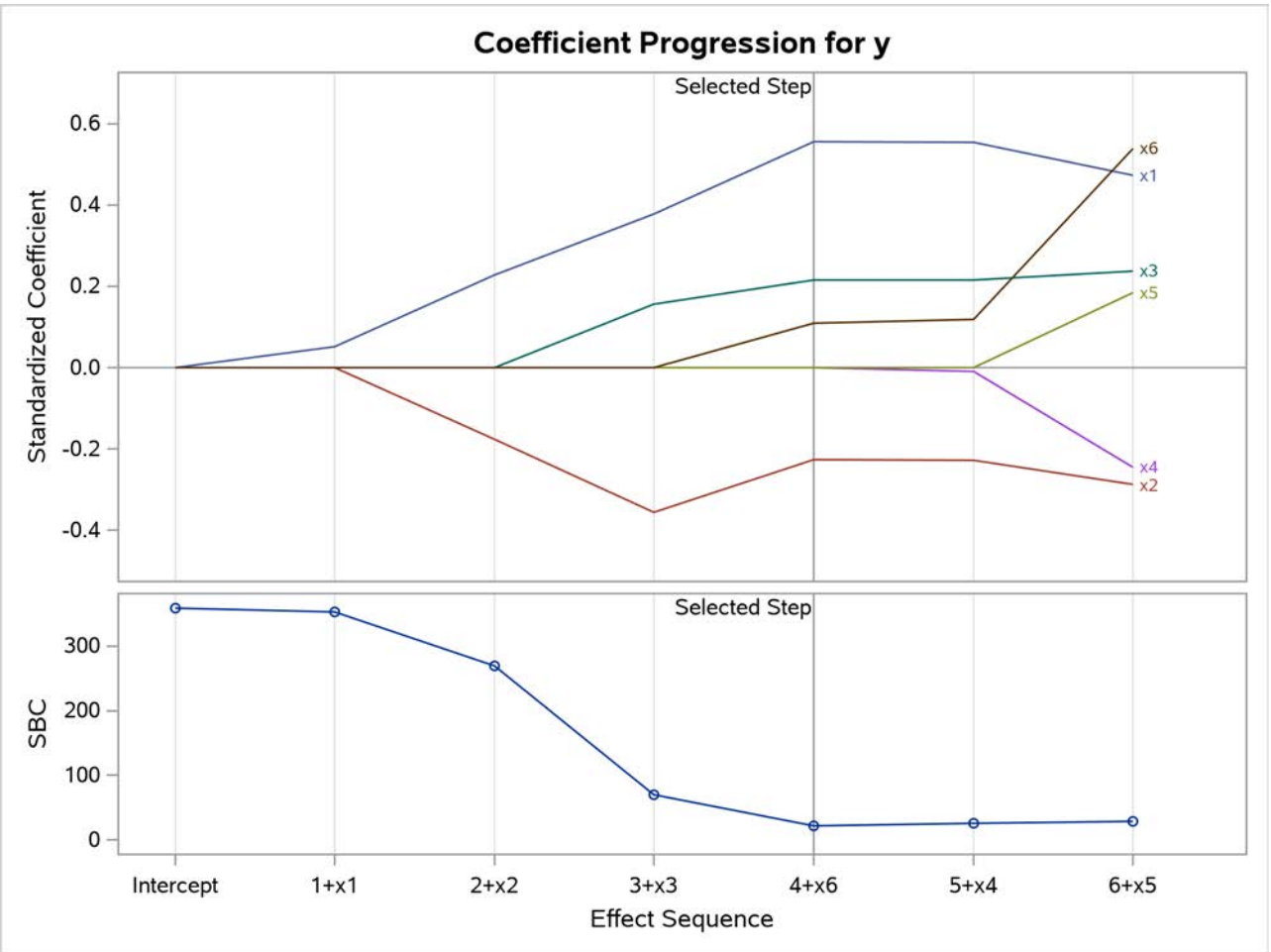
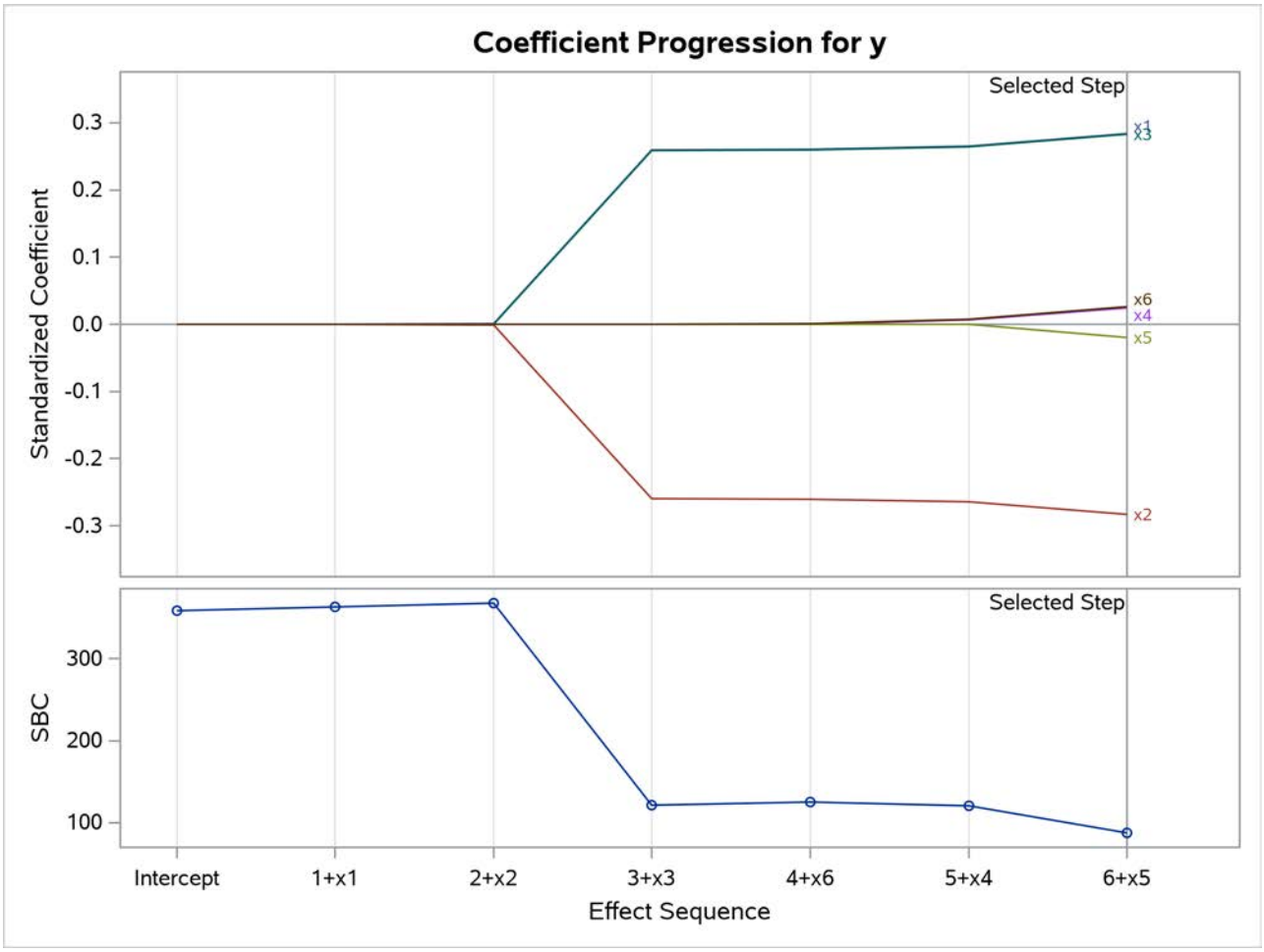


Figure 8 Elastic Net Coefficient Progression with L2=0.5 for Simulation Data with Grouping Effect



Example 3: Analyzing Heart Data with Significant Correlation

The **Sashelp.Heart** data set provides results from the Framingham Heart Study. Figure 9 displays the variables in the data set. Before you perform the model selection, you can load the **Sashelp.Heart** data set into your CAS session, apply a log transformation to the variable **AgeAtStart**, and assign observations to the training, validation, and testing roles by using the following DATA step:

```
data heart;
   set sashelp.heart;
   LogAgeAtStart = LOG(AgeAtStart);
   x = 10*ranuni(1);
   if x>5 then Role = 'TRAIN';
   else if x<3 then Role = 'VAL';
   else Role = 'TEST';
   drop x;
run;

data mycas.heart;
   set work.heart;
run;
```

Figure 9 **Heart** Data Set—Framingham Heart Study

The CONTENTS Procedure

Variables in Creation Order

#	Variable	Type	Len	Label
1	Status	Char	5	
2	DeathCause	Char	26	Cause of Death
3	AgeCHDdiag	Num	8	Age CHD Diagnosed
4	Sex	Char	6	
5	AgeAtStart	Num	8	Age at Start
6	Height	Num	8	
7	Weight	Num	8	
8	Diastolic	Num	8	
9	Systolic	Num	8	
10	MRW	Num	8	Metropolitan Relative Weight
11	Smoking	Num	8	
12	AgeAtDeath	Num	8	Age at Death
13	Cholesterol	Num	8	
14	Chol_Status	Char	10	Cholesterol Status
15	BP_Status	Char	7	Blood Pressure Status
16	Weight_Status	Char	11	Weight Status
17	Smoking_Status	Char	17	Smoking Status

Figure 10 displays the correlation matrix for the predictors of the **Heart** data. You can see some significant correlation between the predictor variables, where the highest correlation is 0.79673 (between **Diastolic** and **Systolic**).

Figure 10 Correlation Matrix for the **Heart** Data

The CORR Procedure

Pearson Correlation Coefficients, N = 5039
Prob > |r| under H0: Rho=0

	Height	Weight	Diastolic	Systolic	MRW	Smoking	Cholesterol
Height	1.00000	0.52329 <.0001	-0.01202 0.3937	-0.07059 <.0001	-0.13088 <.0001	0.28753 <.0001	-0.07869 <.0001
Weight	0.52329 <.0001	1.00000	0.32904 <.0001	0.26185 <.0001	0.76633 <.0001	0.09147 <.0001	0.07311 <.0001
Diastolic	-0.01202 0.3937	0.32904 <.0001	1.00000	0.79673 <.0001	0.38671 <.0001	-0.06463 <.0001	0.18307 <.0001
Systolic	-0.07059 <.0001	0.26185 <.0001	0.79673 <.0001	1.00000	0.36175 <.0001	-0.09206 <.0001	0.19845 <.0001
MRW **Metropolitan Relative Weight**	-0.13088 <.0001	0.76633 <.0001	0.38671 <.0001	0.36175 <.0001	1.00000	-0.12301 <.0001	0.13676 <.0001
Smoking	0.28753 <.0001	0.09147 <.0001	-0.06463 <.0001	-0.09206 <.0001	-0.12301 <.0001	1.00000	-0.01273 0.3664
Cholesterol	-0.07869 <.0001	0.07311 <.0001	0.18307 <.0001	0.19845 <.0001	0.13676 <.0001	-0.01273 0.3664	1.00000

Suppose you want to investigate whether you can model **AgeAtStart** by using the medical measures of the patients. You can use the following statements to perform variable selection for the **Heart** data via the LASSO and elastic net:

```
proc regselect data = mycas.heart;
    partition roleVar=Role(train='TRAIN' validate='VAL' test='TEST');
    model LogAgeAtStart = Height Weight Diastolic Systolic MRW Smoking Cholesterol;
    selection  method = lasso(choose=VALIDATE);
run;
```

```
proc regselect data = mycas.heart;
    partition roleVar=Role(train='TRAIN' validate='VAL' test='TEST');
    model LogAgeAtStart = Height Weight Diastolic Systolic MRW Smoking Cholesterol;
    selection  method = elasticnet(choose=VALIDATE);
run;
```

Figure 11 shows that there are 2,535 observations for training, 1,513 observations for validation, and 991 observations for testing. Figure 12 and Figure 13 show the parameter estimates and the fit statistics of the models that are selected by the LASSO and elastic net, respectively. Figure 14 shows that the optimal L2 value is 0.1. You can find that elastic net outperforms LASSO in the sense that both validation and test ASEs from the elastic net selection are smaller.

Figure 11 Basic Information about **Heart** Data

The REGSELECT Procedure

Number of Observations Read	5209
Number of Observations Used	5039
Number of Observations Used for Training	2535
Number of Observations Used for Validation	1513
Number of Observations Used for Testing	991

Dimensions	
Number of Effects	8
Number of Parameters	8

Figure 12 Details of the Model Selected by LASSO for **Heart** Data

Selected Model by LASSO

Root MSE	0.17488
R-Square	0.22272
Adj R-Sq	0.22118
AIC	-6297.40980
AICC	-6297.36547
SBC	-8799.38210
ASE (Train)	0.03051
ASE (Validate)	0.03147
ASE (Test)	0.02913

Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	3.395598
Height	1	-0.003008
Systolic	1	0.002577
MRW	1	0.000241
Smoking	1	-0.001584
Cholesterol	1	0.000886

Figure 13 Details of the Model Selected by Elastic Net for **Heart** Data

Selected Model by Elastic Net

Root MSE	0.17498
R-Square	0.22184
Adj R-Sq	0.22031
AIC	-6294.56225
AICC	-6294.51793
SBC	-8796.53456
ASE (Train)	0.03054
ASE (Validate)	0.03144
ASE (Test)	0.02909

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	3.430009
Height	1	-0.003326
Systolic	1	0.002387
MRW	1	0.000398
Smoking	1	-0.001595
Cholesterol	1	0.000857

Figure 14 Elastic Net Selection Summary for **Heart** Data

Elastic Net Summary

ENstep	L2	Number Effects In	Validation ASE
1	0.00000000	6	0.0315
2	0.00000001	6	0.0315
3	0.00000001	6	0.0315
	.	.	.
	.	.	.
	.	.	.
41	0.03162278	6	0.0315
42	0.04641589	7	0.0314
43	0.06812921	7	0.0314
44	0.10000000	6	0.0314*
45	0.14677993	8	0.0314
46	0.21544347	8	0.0315
47	0.31622777	8	0.0315
48	0.46415888	8	0.0317
49	0.68129207	8	0.0319
50	1.00000000	8	0.0322

*** Optimal Value Of Criterion**

SUMMARY

This paper summarizes the variable selection methods—in particular, the regularization selection methods for linear regression modeling—and the related SAS/STAT and SAS Viya procedures. Also, it provides several examples to demonstrate how you can use the REGSELECT procedure, available in SAS Viya, to perform variable selection by using penalized regression methods. Although the results of the examples in the paper show that the elastic net method performs better than the LASSO method, you need to keep in mind that in practice, no single method consistently outperforms the rest. Furthermore, there are no universally best defaults for the tuning parameters in penalized regression methods. However, depending on your goal, an informed and judicious choice of these features can lead to models that have better predictive accuracy or models that are more interpretable. You should also experiment with different combinations of the options to learn more about their behavior.

REFERENCES

Cohen, R., and Rodriguez, R. N. (2013). "High-Performance Statistical Modeling." In *Proceedings of the SAS Global Forum 2013 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings13/401-2013.pdf.

Collier Books (1987). *The 1987 Baseball Encyclopedia Update*. New York: Macmillan.

Efron, B., Hastie, T. J., Johnstone, I. M., and Tibshirani, R. (2004). "Least Angle Regression." *Annals of Statistics* 32:407–499. With discussion.

Günes, F. (2015). "Penalized Regression Methods for Linear Models in SAS/STAT." In *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels.pdf.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag.

Rodriguez, R. N. (2016). "Statistical Model Building for Large, Complex Data: Five New Directions in SAS/STAT Software." In *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings16/SAS4900-2016.pdf.

Rodriguez, R. N., and Cai, W. (2018). "Regression Model Building for Large, Complex Data with SAS Viya Procedures." In *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2033-2018.pdf.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B* 58:267–288.

Time Inc. (1987). "What They Make." *Sports Illustrated* (April 20): 54–81.

Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101:1418–1429.

Zou, H., and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B* 67:301–320.

Zou, H., and Zhang, H. H. (2009). "On the Adaptive Elastic-Net with a Diverging Number of Parameters." *Annals of Statistics* 37:1733–1751.

ACKNOWLEDGMENTS

The author is grateful to Weijie Cai for his contributions to the manuscript.

CONTACT INFORMATION

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

Yingwei Wang
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Yingwei.Wang@sas.com