Paper 1822-2014

A Stepwise Algorithm for Generalized Linear Mixed Models

Nagaraj K. Neerchal, University of Maryland Baltimore County, Baltimore, MD
Jorge G. Morel, Procter and Gamble Company, Cincinnati, OH
Xuang Huang, University of Maryland Baltimore County, Baltimore, MD
Alain Moluh, University of Maryland Baltimore County, Baltimore, MD

ABSTRACT

Stepwise regression includes regression models in which the predictive variables are selected by an automated algorithm. The stepwise method involves two approaches, namely, backward elimination and forward selection. Currently, SAS® has several regression procedures capable of performing stepwise regression. Among them are REG, LOGISTIC, GLMSELECT and PHREG. PROC REG handles linear regression model but does not support a CLASS statement. PROC LOGISTIC handles binary responses and allows for logit, probit and complementary log-log link functions. It also allows for CLASS statements. The GLMSELECT procedure performs selections in the framework of general linear models. It allows for a variety of model selection methods, including the LASSO method of Tibshirani (1996) and the related LAR method of Efron et al. (2004). GLMSELECT supports a CLASS statement. PHREG is appropriate for proportional hazard survival regression. We present a stepwise algorithm for Generalized Linear Mixed Models for both marginal and conditional models. We illustrate the algorithm using data from a longitudinal observational study aimed to investigate parents' beliefs, behaviors, feeding practices that associate positively or negatively with indices of sleep quality.

KEYWORDS

PROC GLIMMIX

INTRODUCTION

Stepwise regression is a widely used variable selection method applicable to any predictive model building process. It is a combination of forward selection and backward elimination methods. Each step consists of both forward-selection, when variables are considered for being added to the model, as well as backward-elimination which examines variables for removal from the model. The F-statistic for a variable to be added into the model must be significant at the SLENTRY=level. The forward-selection phase is modified in the sense that variables already in the model do not necessarily stay there. A backward-elimination phase takes also place by removing variables already in the regression model if any variable does not produce a significant F-statistic at the SLSTAY=level. Inclusion and deletion of variables is done one at a time. Once a variable is removed by a backward-elimination step it remains removed. A forward-selection step can be followed by a backward-elimination step. The stepwise selection process concludes if no further effect can be added to the model.

Draper, Guttman, and Kanemasu (1971) have pointed out that the traditional implementations of forward, backward, and stepwise selection methods are based on sequential testing with specified entry (SLENTRY) and stay (SLSTAY) significance levels. However, it is known that the "F-to-enter" and "F-to-delete" statistics do not follow an F-distribution. In spite of these

complexities, these regression selection methods remain as useful tools for building working regression models in the presence of several predictors. Careful use of variable selection methods still has its place in modern data analysis. Currently, SAS® provides an option to implement stepwise variable selection in REG, LOGISTIC, GLMSELECT and PHREG. PROC REG handles linear regression model but does not support a CLASS statement. PROC LOGISTIC handles binary responses and allows for logit, probit and complementary log-log link functions. It also allows for CLASS statements. The GLMSELECT procedure performs selections in the framework of general linear models. It allows for a variety of model selection methods, including the LASSO method of Tibshirani (1996) and the related LAR method of Efron et al. (2004). GLMSELECT supports a CLASS statement. PHREG is appropriate for proportional hazard survival regression.

We propose a stepwise algorithm for Generalized Linear Mixed Models (GLMM) which relies on the GLIMMIX procedure. The algorithm is intended mainly as a model selection tool and does not include hypothesis testing, testing of contrasts, and LS-means analyses. It does require that the user have some familiarity with the syntax of PROC GLIMMIX. In the following sections we provide the arguments needed for implementing the algorithm, two examples, and an extension based on linear trends applicable to longitudinal studies. A key feature of our algorithm is the way interactions of main effects with the trend coefficients are examined for inclusion and removal from the model at each step.

ARGUMENTS OF THE STEPWISE ALGORITHM

Our proposed algorithms follow the recommendation of using F-ratio and associated p-values for variable selection. Since we are targeting applications that involve random effects, trends in a logistic link, we follow the order a GLMM when writing the necessary statements in GLIMMIX, that is, DATA=SAS-data-set, EMPIRICAL<=CLASSICAL>, METHOD<=RSPL>, etc… The arguments available in our MACRO are listed in **Table 1** below.

Table 1: Arguments of the Stepwise GLMM algorithm

Argument	Options*
DATA=	Your SAS-data-set
METHOD=	RSPL, MSPL, RMPL, MMPL, LAPLACE, QUAD<(quad-options)>
EMPIRICAL=	CLASSICAL|HC0, DF|HC1, MBN<(mbn-options)>, ROOT|HC2, FIRORES|HC3, FIROEEQ<(r)>, NONE, None, none
CLASS=	CLASS variables
Y= (**)	Your response variable (Dependent Variable)
VARLIST=	Linear Predictors (include your CLASS variable)
DIST= (***)	BETA, BINARY, BINOMIAL|BIN|B, GAMMA|GAM, NEGBINOMIAL| NEGBIN|NB, POISSON|POI|P
LINK=	IDENTITY|ID, LOG, LOGIT, PROBIT, LOGLOG
SLENTRY=	Specifies the significance level of the F-statistic for entering an effect/variable into the model during the FORWARD step
SLSTAY=	Specifies the significance level of the F-statistic for an effect/variable staying in the model during the BACKWARD elimination step
RANDOM=	_RESIDUAL_, INT, Your own factors
SUBJECT=	Your own effects

Table 1: Arguments of the Stepwise GLMM algorithm	
TYPE= (****)	Your own covariance-structure
HEADING=	Your own heading

(*) For more details see PROC GLIMMIX documentation
(**) Accepts <Y (ref="0">
(***) Has not been tested with the MULTINOMIAL distribution
(****) Allows statements like "TYPE=VC GROUP=Your own effect"

Note that the user has the flexibility of fitting a marginal model (RANDOM=_RESIDUAL_) or a conditional model (RANDOM=INT). The user can select the estimation method (RSPL, LAPLACE, QUAD, etc.) and choose an empirical covariance matrix for the fix effects along with any small sample bias correction.

TWO EXAMPLES

We will illustrate the use of the algorithm using data from a longitudinal observational study aimed to investigate parents' beliefs, behaviors, feeding practices that associate positively or negatively with indices of sleep quality. Participants in the study are new-born babies who were followed up for a whole year with multiple observations during the year. New-born babies were recruited from 10 demographic locations. Sleep records and mothers' diaries were employed to monitor sleep/wake patterns.

Examples of the sleep pattern dependent variables in the study were:

- Duration of Nights (hrs)
- Time to Onset (hrs)
- Efficiency
- Total Sleep Duration
- Total Wake Duration
- Number Wake Intervals
- Etc…

Some of the parents' beliefs, behaviors and feeding practices measured in the study were:

- Last Night Baby TOOK A BATH before Bedtime
- Last Night Baby HAD CLOSE TIME WITH PARENTS before Bedtime
- Last Night Baby HAD FOOD before Bedtime
- Last Night SLEPT WITH LIGHTS ON
- Last Night Baby FELL ASLEEP IN SWING OR STROLLER OR WITH FAMILY
- Etc…

The purpose of this paper is to exemplify the results of the algorithm using a real data set. However, details of the study are not disclosed as they will be part of forthcoming publications that will report the clinical outcomes of the study. To that end, the two dependent variables we use for illustrative purposes will be referred as “Y1” and “Y2”. Similarly, the independent variables used in the stepwise regressions will be denoted as X1, X2,…,X10. All independent variables were binary taking the values Yes/No.

Table 2 shows the syntax for invoking the stepwise algorithm for GLMMs “glimmixstep” for the dependent variable “Y1” with predictors X1, X2,…,X10. In this first example the underlying distribution is Normal with identity link function. An empirical covariance matrix for the fixed effects with MBN option (see Section 9.10 in Morel and Neerchal, 2012; and Morel, Bokossa and Neerchal, 2003) is used at each step of the stepwise procedure. Subjects are identified by the CLASS variable “Subject_id”. We allow different variance components for each of the demographic locations. This is denoted by the class variable “Dem_loc”. The “slentry” and “slstay” values were respectively 0.15 and 0.10.

Table 2: Syntax for invoking the Stepwise algorithm for GLMMs Dependent Variable “Y1”

```
%glimmixstep( data=Sleep, method=RSPL, empirical=MBN,
            class=Dem_loc Subject_id X1 X2 X3 X4 X5 X6 X7 X8 X9 X10,
            Y=Y1,
            varlist=X1 X2 X3 X4 X5 X6 X7 X8 X9 X10,
            dist=normal, link=identity,
            slentry=0.15, slstay=0.10, random=int,
            subject=Subject_id, type=VC group=Dem_loc,
            heading=Dependent Variable Y1 );
```

Table 3 provides the results of the stepwise regression on “Y1”. LS-means computations were performed after the selection of the model. The predictors “X5”, “X2” and “X4” clearly show a positive association with “Y1”. On the other hand, the predictors “X7” and “X1” show a negative association with the response.

Table 3: Stepwise Regression Results for the Dependent Variable “Y1” Distribution=Normal, Link=Identity Subject=Subject_id, Type=VC, Group=Dem_loc

Questionnaire Item	Group/Comparison	Means/Difference	Standard Error	P-Value
X5	Yes	8.07	0.08595	
	No	7.43	0.05164	
	Yes vs. No	0.64	0.08280	<.01
X2	Yes	7.88	0.06222	
	No	7.62	0.06519	

Table 3: Stepwise Regression Results for the Dependent Variable "Y1" Distribution=Normal, Link=Identity Subject=Subject_id, Type=VC, Group=Dem_loc				
Questionnaire Item	**Group/Comparison**	**Means/Difference**	**Standard Error**	**P-Value**
	Yes vs. No	0.26	0.05470	<.01
X7	Yes	7.65	0.07131	
	No	7.85	0.05570	
	Yes vs. No	-0.19	0.05589	<.01
X4	Yes	7.89	0.04957	
	No	7.61	0.08725	
	Yes vs. No	0.27	0.08300	<.01
X1	Yes	7.68	0.06578	
	No	7.82	0.06785	
	Yes vs. No	-0.13	0.06789	0.05

The stepwise regression results on the dependent variable "Y2" are provided in **Table 4**. In this example the distribution is Poisson with logarithmic link function. It is appropriate to use the Poisson distribution since the dependent variable is a count. The syntax is similar to the one depicted in **Table 2**. As in the previous example, demographic locations are allowed to have different variance components. We requested LAPLACE method for maximum likelihood estimation. LS-means computations were performed after the selection of the model using the inverse link function. Thus, rates and ratios of rates are reported. P-values correspond to the F-tests associated to the questionnaire items in the model. Only the predictor "X10" showed a significant increase in the counts of "Y2" for responders in the YES group. All the other predictors in the model showed a decrease for responders in the in the YES group.

Table 4: Stepwise Regression Results for the Dependent Variable "Y2" Distribution=Poisson, Link=Log Subject=Subject_id, Type=VC, Group=Dem_loc				
Questionnaire Item	**Group/Comparison**	**Rates/Ratio**	**Standard Error**	**P-Value**
X2	Yes	2.49	0.1028	
	No	2.68	0.1092	
	Yes vs. No	0.93	0.01548	<.01
X5	Yes	2.40	0.1186	
	No	2.79	0.1035	

Table 4: Stepwise Regression Results for the Dependent Variable "Y2" Distribution=Poisson, Link=Log Subject=Subject_id, Type=VC, Group=Dem_loc				
Questionnaire Item	**Group/Comparison**	**Rates/Ratio**	**Standard Error**	**P-Value**
	Yes vs. No	0.86	0.02995	<.01
X9	Yes	2.37	0.1697	
	No	2.82	0.06108	
	Yes vs. No	0.84	0.05791	0.01
X3	Yes	2.53	0.1053	
	No	2.64	0.1067	
	Yes vs. No	0.96	0.01619	0.02
X4	Yes	2.50	0.09984	
	No	2.67	0.1201	
	Yes vs. No	0.94	0.02628	0.02
X10	Yes	2.63	0.1056	
	No	2.54	0.1082	
	Yes vs. No	1.03	0.02092	0.09

A MODIFIED ALGORITHM FOR DETECTING LONGITUDINAL TRENDS BASED ON SUCCESIVE VISITS

We modified the proposed algorithm to allow entering or removing independent variables if the longitudinal trends over time associated to the levels of the independent variables were significantly different or not. Longitudinal trends (linear trends) are based on the standardized babies' ages at the time the longitudinal observations took place. For a given predictor, say for instance "X1", the longitudinal trends associated with the factor "X1" are calculated from an analysis with "babies' ages" as a covariate, "X1" as the main effect, and the interaction "babies' ages-by-X1". If an interaction term enters into the model, so does its main effect. Interaction terms in the model are always accompanied with their corresponding main effects since all models considered in the modified algorithm are hierarchical. If an interaction term that is already in the model is removed from the model, its main effect would still remain in the model.

Table 5 shows the syntax for invoking "glimmixstep_trends" for the dependent variable "Y1" with predictors X1, X2,…,X10. Estimation method is RSPL. Subjects are identified by the CLASS variable "Subject_id". The "slentry" and "slstay" values we used were respectively 0.30 and 0.25. We set the value of 0.25 to allow an interaction to stay in the model in accordance of Bancroft's (1968) recommendation for testing equality of slopes. Note that the equality of slopes test corresponds to the test of interaction in an ANCOVA model. In invoking the modified algorithm, the reader can notice the similitude with the first algorithm. Note also the use of the GEE sandwich estimator along with the MBN option at each step of the stepwise algorithm. Note that the variable "SAge", which is the standardized babies' ages at the time the

longitudinal observations were taken, is not passed as an argument, but has to be in the data set.

Table 5: Syntax of invoking the Stepwise algorithm for testing different slopes with GLMMs(*)

```
%glimmixstep_trends( data=Sleep, method=RSPL, empirical=MBN,
      class=Areas Subject_id X1 X2 X3 X4 X5 X6 X7 X8 X9 X10,
      Y=Y1,
      varlist=X1 X2 X3 X4 X5 X6 X7 X8 X9 X10,
      dist=normal, link=identity,
       slentry=0.30, slstay=0.25, random=int,
       subject=Subject_id, type=VC,
      heading=Stepwise Regression Dependent Variable Y1 - Linear Trends );
```

(*) Standardized babies' ages ("SAge"), is not passed as an argument, but has to be in the data set.

We report in **Table 6** the results of the algorithm for finding significantly different trends. LS-means and construction of plots, were done after the selection of the model. Questionnaire items showing a positive longitudinal association across time are "X9 (Group=NO)", "X6 (Group=NO)", "X3 (Group=YES)", "X4 (Group=YES)', "X2 (Group=NO)", "X7 (Group=YES)", and "X1 (Group=YES)". Depictions of the linear trends of some selected predictors are provided in Figures 1 though 3.

Table 6: Stepwise Linear Trends Results for the Dependent Variable "Y1"
Dist=Normal, Link=Identity
Subject=Subject_id, Type=VC

Questionnaire Item	F-value	P-Value Equality of Slopes	Group	Slope	Standard Error	P-Value Significance Slope
X9	15.28	<.01	Yes	-0.18	0.1554	0.25
			No	0.42	0.0635	<.01
X6	2.34	0.13	Yes	0.04	0.1229	0.72
			No	0.20	0.0803	0.01
X3	2.29	0.13	Yes	0.16	0.0942	0.10
			No	0.09	0.0920	0.34
X4	2.20	0.14	Yes	0.18	0.0905	0.05
			No	0.06	0.1055	0.54

Table 6: Stepwise Linear Trends Results for the Dependent Variable "Y1"
Dist=Normal, Link=Identity
Subject=Subject_id, Type=VC

Questionnaire Item	F-value	P-Value Equality of Slopes	Group	Slope	Standard Error	P-Value Significance Slope
X2	1.89	0.17	Yes	0.09	0.0920	0.34
			No	0.16	0.0953	0.10
X7	1.70	0.19	Yes	0.16	0.0981	0.10
			No	0.08	0.0934	0.39
X1	1.51	0.22	Yes	0.16	0.0951	0.10
			No	0.09	0.0940	0.35

Some interactions have clear interpretations. Consider for instance the predictor "X4". **Figure 1** depicts the estimated linear trend. Based on this analysis it is clear the "benefit" for responders in the YES-group. Other interactions are more difficult to interpret and might have to be referred to experts on sleep behaviors. That is the case of the predictor "X3". This interaction is depicted in **Figure 2**. Clearly the linear trends cross each other making its interpretation more difficult. Care must be exercised in interpreting some of the other interactions. For instance, consider the stepwise results associated with the predictor "X1". The slopes associated to the groups YES (0.16) and NO (0.09) are significantly different (P-value=0.22). Since the slope associated to the YES-group is larger than that associated to the NO-group, one might conclude that babies in the YES-group have a longer values of "Y1" across the duration of the study. That does not seem to be the case if one examines the longitudinal trends depicted in **Figure 3**. Note that within the range of the duration of the study the linear trend for the babies in the NO-group dominates that of babies in the YES-group. Therefore, babies in the NO-group are showing longitudinally an "improving" trend over that of babies in the YES-group. By the end of the year the effect of the factor "X1" becomes undistinguishable.

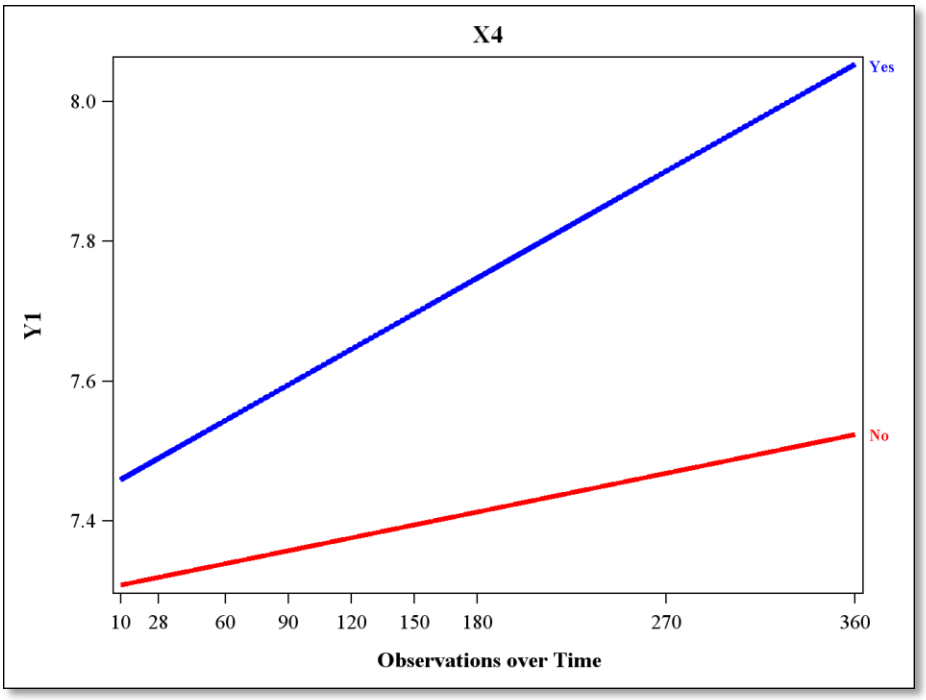

Figure 1: Estimated Linear Trends – Response "Y1"– Predictor "X4"

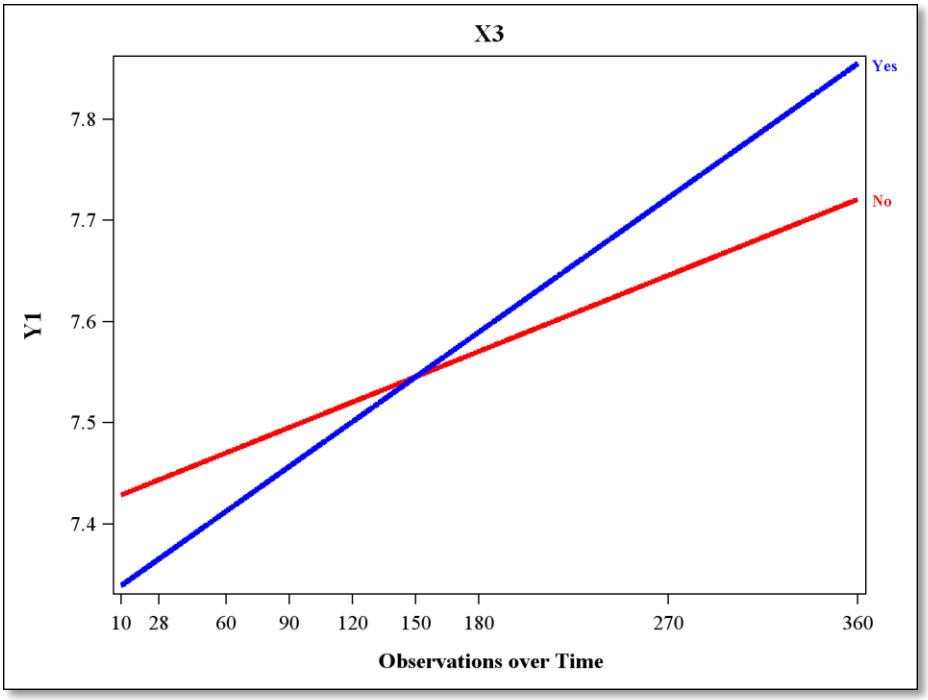

Figure 2: Estimated Linear Trends – Response "Y1"– Predictor "X3"

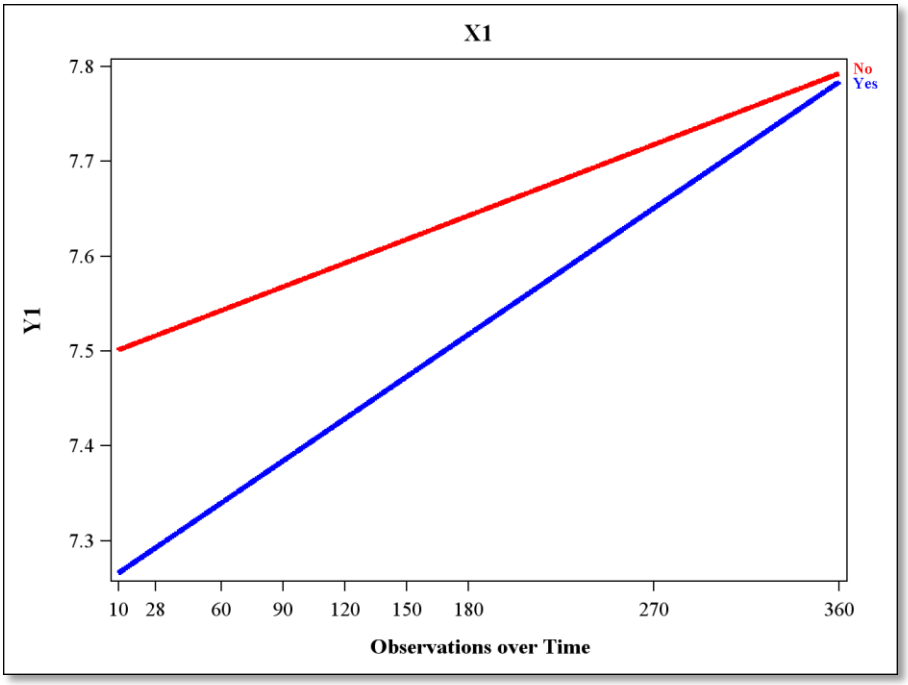

Figure 3: Estimated Linear Trends – Response "Y1"– Predictor "X1"

CONCLUSIONS

In this paper, we report the implementation of two stepwise variable selection algorithms in the context of GLMM, so they allow for random effects while specifying any distribution from the exponential family with customary link functions. The novelty of the second algorithm is that it allows specification of random effects while taking the corresponding interaction terms with the trend into account. Our algorithms use the F-statistics to determine inclusions and removals of variables in each step. We provide SAS® macros which incorporate our algorithms into the recent super star SAS® PROC GLIMMIX. The algorithms allow the use of empirical (robust, sandwich) covariance estimates of the fixed effects with the option of small sample bias corrections. The macros provided here can be further modified to expand the options available for variable selection. The use of macros are illustrate with two substantial examples.

REFERENCES

Bancroft T. A., 1968, **Topics in Intermediate Statistical Methods**, Volume One, Section 1.4, the Iowa State University Press, Ames, Iowa.

Draper, N. R., Guttman, I., and Kanemasu, H. (1971), "The Distribution of Certain Regression Statistics," *Biometrika*, 58, 295–298.

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

Morel, J. G. and Neerchal, N. K. (2012**), Overdispersion Models in SAS®**, SAS Press.

Morel, J. G, Bokossa, M. C. and Neerchal, N. K. (2003), “Small Sample Correction for the Variance of GEE Estimators”, Biometrical *Journal*, 4, 395-409.

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

CONTACT INFORMATION

If you are interested in a copy of the two stepwise algorithms, or if you have any comments, suggestions or recommendations, please feel free to contact Jorge G. Morel at:

Name: Jorge G. Morel, Ph.D
Organization: The Procter and Gamble Company
Address: 6280 Center Hill Avenue, MS S-M1
City, State ZIP: Cincinnati, OH 45224
Work Phone: (513) 945-0399
Fax:
Email: morel.jg@pg.com
Personal Email: statistics@gmail.com