A Plan To Implement the Limu Model By a Slight Modification of the Current MIXED Procedure

Mark X. Norleans, Trilogy Consulting Corporation, Waukegan, Illinois, USA

Abstract

The limu model, proposed by Xie and Norleans, is a generalized mixed linear model that is useful for analyzing independent or correlated responses on an arbitrary scale. The model is mathematically defined as the first-order Taylor-series expansion of the generalized mixed linear model applied to observed responses. To summarize data with mean responses and measure the precision of the summarization, the multivariate normal distribution is used to represent the expected frequencies of the linearized responses and the maximum likelihood techniques are used to estimate the parameters in the mean and variance functions. The model includes the generalized linear model of Nelder and Wedderburn and the generalized estimating equations of Liang and Zeger as special cases if the Fisher scoring algorithm is used to solve the normal equations.

The limu model is fitted by iterating through a maximization step and an updating step. SAS® PROC MIXED may be expanded to implement the limu model by iteratively calling the Newton-Raphson routine followed by the updating step. Realization of this plan will allow users to analyze data with a much richer class of regression models. Due to the difficulties to update the variance functions, the modification may not be easily implemented at the user level. However, the idea of the limu model is demonstrated with a macro program for fitting the generalized linear model of Nelder and Wedderburn by iteratively calling PROC GLM or PROC REG followed by an updating.

Introduction

For the purpose of analyzing how much variation of responses can be explained by factors under study, SAS® PROC GLM and PROC MIXED may be convenient if we use the normal or multivariate normal distribution. However, for the same purpose, SAS does not have a convenient procedure if we prefer to using distributions other than the normal to model independent or correlated responses on an arbitrary scale. Statistical techniques for the analysis of variance on arbitrary scales have been developed. These techniques are the generalized linear model (GLM) of Nelder and Wedderburn[1], the generalized estimating equations (GEEs) of Liang and Zeger[2] and, recently, the limu model of Xie[3] and Norleans [4]. The limu model, which includes the GLM and GEEs as special cases, was developed from an idea of Schall[5] and justified under the principle of estimation of R. A. Fisher[6,7,8]. The limu model is as flexible as the generalized mixed linear model yet avoids the mathematical and computational difficulties for fitting that model.

The aim of this article is to introduce the limu model and propose a simple plan to implement the limu model by a slight modification of the current SAS
The Limu Model

Let $Y_i = (y_{i1}, ..., y_{in})^T$ denote the response vector of subject $i$, $X_i$ and $Z_i$ denote the fixed and random design matrices, and $\beta$ and $b_i$ denotes the corresponding coefficients. From an idea of Schall, the limu model is defined as

$$U_i = X_i \beta - Z_i b_i + \text{diag}(g'(\mu_i))(Y_i - \mu_i)$$

and $b_i \sim N(0, D)$, $\text{cov}(b_i, Y_i) = 0$, where $U_i$ denotes the updated response vector computed from the response vector $Y_i$ and the current estimates of $\beta$ and $b_i$; $N(0, D)$ denotes the multivariate normal distribution with mean vector 0 and covariance matrix $D$;

$$\text{diag}(g'(\mu_i)) = .$$

and $g$ denotes the link function for the scale of interest, such as the log, logit, probit and so on.

The mean and variance functions of the updated responses can be easily computed: $E(U_i) = X_i \beta$ and $V_i = \text{var}(U_i) = Z_i D Z_i^T + \text{diag}(g'(\mu_i)) - \text{var}(Y_i) \cdot \text{diag}(g'(\mu_i))$.

In (3), the covariance matrix $\text{var}(Y_i)$ may be conveniently parameterized based on hypothesis about the covariance structure. To be consistent with the generalized estimating equations of Liang and Zeger, $\text{var}(Y_i)$ is defined as

$$\text{var}(Y_i) = A_i^{1/2} \cdot R \cdot A_i^{1/2},$$

where $A_i = \text{diag}(\text{var}(y_{i1}) = v_i(\mu_i))$ for subject $i$ and $R$ corresponds to the working correlation matrix of Liang and Zeger. Note that elements of $R$ are not necessarily within $[-1, +1]$.

Comparing to the generalized mixed linear model,

$$g[E(Y_i) = \mu_i] = X_i \beta - Z_i b_i, \quad b_i \sim N(0, D),$$

the limu model (1) is mathematically the first-order Taylor series expansion of $g(Y_i)$. For statistical evaluation of observed data, however, the limu model (1) is not in any sense an approximation to the generalized mixed linear model (5). The particular mode of statistical evaluation for which the limu model is designed is to compute summary quantities and measure their precision with variance or the Fisher information. Comparing the summary quantities to the measure of their precision affords a quantitative evaluation of observed data.

A model is not completely defined until an algorithm for estimating the parameters is also defined. The limu model is fitted with an iterative algorithm consisting of a maximization step and an updating step. Given the updated responses $U_i$ computed with the observed responses and the estimates from the previous iteration, the maximum likelihood (ML) technique is used to estimate $\beta$, $D$ and $b_i$ by using the multivariate normal distribution

$$U_i \sim N(X_i \beta, V_i),$$

to represent the expected frequencies of $U_i$. Using the multivariate normal distribution is particularly convenient because we only want to use the mean responses and their variance to evaluate observed data and the multivariate normal distribution affords convenient parameters to represent these two quantities. Then, given the current estimates of $\beta$ and $b_i$, $U_i$ is updated with

$$U_i = X_i \beta - Z_i b_i - \text{diag}(g'(\mu_i))(Y_i - \mu_i),$$

where

$$\mu_i = g^{-1}(X_i \beta \cdot Z_i b_i).$$

If we highlight the definition of the limu model, it is
easy to see that limu, which stands for linearization, maximization and updating, fully characterizes the model.

**Special Cases**

If we omit the random terms in (1) and applying the Fisher scoring method to solve the score equations from the multivariate normal distribution for the fixed regression coefficients \( \beta \), the iteration equations are exactly those of the generalized estimating equations (GEEs) of Liang and Zeger if the Fisher scoring method is also used to solve the GEEs. Of course, if \( R=I \), the iteration equations are also those of the generalized linear model of Nelder and Wedderburn. However, the advantage of the limu model over those special cases is that the covariance parameters and the "overdispersion" parameter can be naturally defined in the model and are estimated with the maximum likelihood techniques. By using the maximum likelihood techniques, possible inconsistent estimates under the criterion of asymptotic consistency are excluded.

**Implementation Plans**

By using the multivariate normal distribution, the entire maximum likelihood techniques for the normal mixed linear model are directly applicable. We may use the Newton-Raphson or Fisher scoring method and the Henderson mixed model equations; or, we may simply use the EM algorithm. The only modification to these techniques is that after each iteration, \( U \) must be updated with (7) and (8). In addition, if we prefer, the restricted maximum likelihood estimation can also be used at the maximization step although for statistical evaluation of observed data, the restricted maximum likelihood estimates by no means contain more information than their maximum likelihood counterparts.

The simplest implementation is perhaps by making use of the current numerical techniques for the normal mixed model such as those used in SAS PROC MIXED or BMDP® 5V. If we use the Newton-Raphson or Fisher scoring algorithm at the maximization step, the current SAS PROC MIXED can be easily expanded. The only modification required is to replace the original response variable with the updated variable \( U \) and after each iteration to update the \( U \) and \( \text{var}(U) \). At the user level, only one additional option needs to be added to allow choosing the scale. The benefit after this trivial modification is that SAS users may use this expanded MIXED procedure to fit a much richer class of regression models, including the GLM and GEEs.

**Summary Statistics**

Given the estimates of \( \beta \) and \( D \),

\[
\text{var}(\beta) = \left( \sum_{i=1}^{n} X_i (V_i^{-1} X_i)^{-1} \right) \tag{9}
\]

Alternatively, we may define

\[
\text{var}(\beta) = \sum_{i=1}^{n} X_i V_i^{-1} \tilde{V}_i \tilde{V}_i^T \left( \sum_{i=1}^{n} X_i V_i^{-1} X_i \right)^{-1} \tag{10}
\]

where

\[
\tilde{V}_i = Z_i \tilde{O}_i Z_i^T + \text{diag}(g'(\mu_i)) \tilde{A}_{i}^{-1/2}(r_i') \tilde{A}_{i}^{-1/2} \text{diag}(g'(\mu_i)), \tag{11}
\]

and \( r_i \) denotes the Pearson residual vector for subject \( i \). For making a probability statement on the null hypothesis: \( H_0: \beta = h \), define
where $C$ and $h$ are constant matrix and vector, respectively. A p-value may be obtained by comparing $F$ with an $F$ distribution with $\text{rank}(C)$ and $n-p$ degrees of freedom, $p$ being the column rank of the design matrices $X$.

**Fitting GLM With PROC GLM**

The proposed modification to PROC MIXED cannot be done at the user level because the covariance matrix cannot be updated outside PROC MIXED. To illustrate the idea, however, a simple method to fit the generalized linear model (GLM), a special case of the glm model, is designed. The method is to call iteratively PROC GLM to regress the updated responses $U$ on regressor or regressors then to update the $U$ and the weight $W = \text{var}^{-1}(U)$ after each call of PROC GLM.

A SAS macro program is written and is available upon request. The program is very simple. First, PROC GLM is used directly on the original response variable to get initial estimates. On the basis of the initial estimates, the updated response variable $U$ and $\text{var}(U)$ are computed. Then an iteration routine is set up to toggle between PROC GLM with the response variable $U$ and the weight $\text{var}^{-1}(U)$ and the updating step until the iteration series converges or reaches a pre-specified number of iterations. The only quantities required by the updating step from PROC GLM are the predicted values.

Compare to PROC LOGISTIC, this macro has the following advantages. First, the overdispersion parameter is naturally included and estimated with the maximum likelihood technique. Because the "overdispersion" phenomenon is not a peculiar nature of data but due to the lack of an appropriate parameter in the binomial or Poisson model to represent the residual information, it is highly recommended to include the "overdispersion" parameter in the model. Nevertheless, if we prefer to a model without this parameter, we may print out the matrix $(X^tWX)^{-1}$ to obtain the variances for the regression coefficients by using option $i$. The second advantage is that with the $\text{class}$ statement, categorical regressors can be easily handled without having to code them to dummy variables. Although then there will be no meaningful regression coefficients, the $F$-statistics and the p-values based on the "type-III" sum of squares are useful. Finally, this macro has the potential to be expanded to model responses on an arbitrary scale. We only need to define the mean and variance functions and then add them to the $\%\text{linker}$ module. The only limit with this macro is that data in the table format are not handled correctly. Although the regression coefficients will be correctly computed, the degrees of freedom for computing p-values are not computed correctly.

An example in the documentation of PROC LOGISTIC was re-run with this macro. If the data set is current, then we simply specify two necessary statements:

```
MODEL REMISS=CELL SMEAR INFIL LI BLAST TEMP /SOLUTION I;
LINK LOGIT;
```

If, say, CELL is categorical, then we only need to add the $\text{CLASS}$ statement: \text{CLASS} CELL;: The program converges after 9 iterations and the PROC GLM at the final iteration prints the following results.
<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELL</td>
<td>1</td>
<td>0.80716465</td>
<td>0.80716465</td>
<td>0.94</td>
<td>0.3432</td>
</tr>
<tr>
<td>SMEAR</td>
<td>1</td>
<td>0.51779930</td>
<td>0.51779930</td>
<td>0.60</td>
<td>0.4276</td>
</tr>
<tr>
<td>INFIL</td>
<td>1</td>
<td>1.97525253</td>
<td>1.97525253</td>
<td>2.31</td>
<td>0.1444</td>
</tr>
<tr>
<td>BLAST</td>
<td>1</td>
<td>0.22767272</td>
<td>0.22767272</td>
<td>0.27</td>
<td>0.6116</td>
</tr>
<tr>
<td>TEMP</td>
<td>1</td>
<td>2.18372814</td>
<td>2.18372814</td>
<td>2.55</td>
<td>0.1260</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>1</td>
<td>74.2569</td>
<td>0.95</td>
<td>0.3533</td>
<td>78.1389</td>
</tr>
<tr>
<td>CELL</td>
<td>1</td>
<td>46.1401</td>
<td>0.97</td>
<td>0.3413</td>
<td>47.5225</td>
</tr>
<tr>
<td>SMEAR</td>
<td>1</td>
<td>43.8558</td>
<td>0.77</td>
<td>0.4475</td>
<td>56.6301</td>
</tr>
<tr>
<td>INFIL</td>
<td>1</td>
<td>-46.9592</td>
<td>-0.78</td>
<td>0.4459</td>
<td>60.3995</td>
</tr>
<tr>
<td>LI</td>
<td>1</td>
<td>-124.4716</td>
<td>-1.60</td>
<td>0.1260</td>
<td>77.9421</td>
</tr>
</tbody>
</table>

The diagonal elements of $(X'WX)^{-1}$ are (7130.70, 2637.52, 3745.35, 4260.54, 8.44, 6.32, 7094.83).

### Ending Remarks

Theimu model is useful for evaluating observed data with mean responses on an arbitrary scale and their Fisher information. The model can be implemented by making use of current numerical techniques for the maximum likelihood estimation with the multivariate normal distribution. Particularly, it appears to be easy to expand the current SAS PROC MIXED so that users can analyze data with a much richer class of regression models.

SAS is registered 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.

Author’s e-mail addresses:
- Mark_X_Norieans@sbphrd.com or
- xnorleans@aol.com

### References