User-specified likelihood expressions using NLMixed and the GENERAL statement

Melissa McTernan, University of California, Davis
Shelley Blozis, University of California, Davis

ABSTRACT

SAS PROC NLMixed was developed for the estimation of nonlinear mixed models, including both multilevel and longitudinal data. Nonlinear mixed models are models for which a random coefficient may enter an equation nonlinearly. PROC NLMixed may be used to estimate models for which a response follows a normal, binomial, or Poisson distribution, as well as several other distributions. For ones not included in the standard options, the program includes an option for the user to express a likelihood function. This naturally allows for more complex distributions, including joint distributions between variables that follow possibly unique distributions. Thus, this is an advantage to researchers because it allows for a great variety of response functions. This paper describes steps to developing a likelihood function and corresponding NLMixed syntax to estimate the model. Empirical examples are provided, including specification of a joint longitudinal model for a normal and binomial response.

INTRODUCTION

Mixed-effects models have become a major statistical framework in many areas of research because of their ability to handle multilevel and longitudinal data. SAS PROC NLMixed is a procedure that is often used to analyze such data and includes options for responses that follow a normal, binomial, or Poisson distribution. For other distributions, users are able to specify any likelihood function within the GENERAL statement of the NLMixed procedure. Thus, NLMixed is a very flexible procedure that can be broadly applied to complex distributions. The purpose of this paper is to demonstrate the flexibility of NLMixed with the GENERAL statement using real complex data.

To fit a model in NLMixed, a user will typically specify which probability distribution should be used to obtain maximum likelihood estimates for his or her data in the MODEL statement. This decision should be based on the shape of the observed response distribution (assuming that the response distribution is representative of the population). If the data is not appropriately modeled by any of the probability distributions built into the SAS NLMixed procedure, the user can manually specify a likelihood function.

To utilize this approach it is important to understand the relationship between a probability distribution function and a likelihood function. A probability function describes the probability of a particular event or value x, conditional on the parameters of the distribution. For example, a Normal Gaussian distribution is a symmetric bell-shaped distribution defined by parameters mu and sigma, or mean and standard deviation, respectively. The probability distribution will look slightly different with different values of these parameters. Given a particular set of parameter values, the distribution function will be defined and it will be possible to estimate the probability of a specific value x.

A likelihood distribution function is one that describes the likelihood of a particular set of parameters given the data. That is, the height of the curve of this function for any corresponding value on the x-axis (which will be made up of possible parameter estimates) relates to the likelihood of that value, given the data. NLMixed uses the maximum likelihood algorithm, which attempts to maximize the likelihood function to find the most likely parameter value given the data.

In this paper we hope to demonstrate the flexibility of the NLMixed procedure with the GENERAL model statement with an application to a joint model. The GENERAL statement is a useful tool to make the NLMixed procedure more flexible. By specifying two separate likelihood functions and then combining them in the GENERAL statement, one can use NLMixed to test joint models. This example is that of a two-part mixed-effects model for longitudinal data (Olsen & Schafer, 2001; Tooze, Grunwald & Jones, 2002). A two-part mixed-effects model can be used for semi-continuous data, or data that has a high proportion of zeros but is otherwise continuous. As the name suggests, the model is performed in two simultaneous parts for any outcome y. The first part models the log odds that y > 0 at a given time point with a logistic regression analysis. The second part models the magnitude of y, given that y > 0 with a standard mixed-effects model for continuous data. A random component is added to each model part and these random effects are allowed to covary across time.

EXAMPLE 1

For this example, we use semi-continuous data from the National Study of Daily Experiences (NSDE; Almeida, 2007) for which time-use data was collected over eight consecutive days, via a telephone survey. The variable used for this paper is one measuring how many hours of chores individuals performed on any given night of the week. A
dichotomous indicator variable for whether the day of the week was a weekday or a weekend day was included in the analysis. Many individuals reported that they spent 0 hours on chores, both during the week and the weekend, suggesting that a two-part mixed-effects model (Olsen & Schafer, 2001) might be appropriate to accommodate the pile-up of zero values in an otherwise continuous set of data.

There is no standard option in PROC NLMIXED for a joint model, but a user may use the GENERAL likelihood statement to perform such an analysis. Syntax for two versions of the joint model is included below. First, we test a two-part mixed-effects model with a normal distribution to model the non-zero values. Secondly, we perform another joint model that models the positive responses with the generalized gamma distribution. Liu et al. used a similar model also using NLMIXED in their analysis of medical cost data (Liu, Cohen, Strawderman & Shih, 2010). In both examples, au0 and bu0 are the random intercepts of the first and second model parts, respectively.

**Example 1:**

```
PROC NLMIXED DATA=time MAXITER=20000 QPOINTS=10 GCONV=1e-10;
PARMS b0=2.3502 b1=-0.2690
   a0=-14 covab=.6
   sda=-1 sdb=1;
*LL1;
y=a0+au0+a1*wkend;
expy=exp(y);
p=expy/(1+expy);
ll1 = log((1-p)**(1-chore_u)) + log(p**(chore_u));
IF chore_u=0 THEN L=ll1;
*LL2;
IF chore_u=1 THEN DO;
  mu=b0+bu0+b1*wkend;
  pi = arcos(-1);
  lla = log(1/(sqrt(2*pi*s2e)));
  llb=-(chore-mu)**2/(2*s2e);
  ll2=lla+llb;
  L=ll1+ll2;
END;
*COMBINE;
MODEL chore ~ GENERAL(L);
RANDOM au0 bu0 ~ NORMAL([0,0],[sda*sda, covab, sdb*sdb]) SUBJECT=ID;
RUN;
```

**EXAMPLE 2**

Example 2 further demonstrates the flexibility of this approach with a two-part mixed-effects model with a logistic link for the first part and a generalized gamma distribution for the second part. See the syntax below. Note the similarities with Example 1. The changes occur in the section of code labeled “LL2” where the likelihood function for the continuous portion, part 2, is specified.

**Example 2.**

```
PROC NLMIXED DATA=time MAXITER=20000 QPOINTS=10 GCONV=1e-10;
PARMS b0=2.3502 b1=-0.2690
   a0=-14 covab=.6
   sda=-1 sdb=1
d0=-.5 dl=.5;
*LL1;
y=a0+au0+a1*wkend;
expy=exp(y);
p=expy/(1+expy);
ll1 = log((1-p)**(1-chore_u)) + log(p**(chore_u));
IF chore_u=0 THEN L=ll1;
*LL2;
IF chore_u=1 THEN DO;
  mu=b0+bu0+b1*wkend;
  sig=exp((d0+dl*wkend)/2);
  eta=abs(k)**(-2);
c=sign(k)*(log(chore)-mu)/sig;
```
User-specified likelihood expressions using NLMIXED and the GENERAL statement, continued

templ=eta*log(eta)-log(sig)-.5*log(eta)-lgamma(eta);
L=log(p)+templ+c*sqrt(eta)-eta*exp(abs(k)*c);
END;
*COMBINE;
MODEL chore ~ GENERAL(L);
RANDOM au0 bu0 ~ NORMAL([0,0],[sda*sda, covab, sdb*sdb]) SUBJECT=ID;
RUN;

RESULTS

AIC values, -2LL values and selected parameter estimates for the examples above are included below in Table 1. Note that the model in the second example that uses a generalized gamma distribution to model the positive responses has a lower AIC value, indicating that it might have better fit. A researcher might want to use other means to confirm this result, such as plotting the data and observing the residuals.

<table>
<thead>
<tr>
<th>Positive Value Distribution</th>
<th>Fit Statistics</th>
<th>Parameter Estimates (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (Ex. 1)</td>
<td>-2LL=25749</td>
<td>a0 = 2.306 (0.08)</td>
</tr>
<tr>
<td></td>
<td>AIC=25765</td>
<td>a1 = -0.042 (0.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b0 = 2.309 (0.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b1 = -0.270 (0.17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sda = -1.628 (0.08)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sdb = 1.080 (0.03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>covab = 0.616 (0.11)</td>
</tr>
</tbody>
</table>

| Generalized Gamma (Ex. 2)   | -2LL=16644    | a0 = 2.302 (0.08)       |
|                             | AIC=16667     | a1 = -0.029 (0.28)      |
|                             |               | b0 = 0.654 (0.02)       |
|                             |               | b1 = -0.111 (0.07)      |
|                             |               | sda = -1.623 (0.08)     |
|                             |               | sdb = 0.487 (0.01)      |
|                             |               | covab = 0.279 (0.04)    |

Table 1. Fit statistics and selected parameter estimates for the two-part mixed-effects models in Examples 1 and 2.

Overall, both forms of the two-part model indicate that people are less likely to perform chores on the weekends. Further, for those who do engage in chores on the weekends, less time will be spent on chores than on the weekdays. We can also derive cross-part correlations from the results in Examples 1 and 2. This parameter is one of the most motivating reasons for performing a two-part mixed-effects model and represents the relationship between the random effects of the two model parts. In Example 1, this correlation, \( \rho \), is \( \frac{\text{covab}}{\text{sda}\times\text{sdb}} = \frac{.2795}{1.6231\times1.6006} = .550 \). Under the model in Example 2, \( \rho = \frac{\text{covab}}{\text{sda}\times\text{sdb}} = \frac{.2795}{1.6231\times1.6006} = .354 \). Thus, both models suggest that there is a moderate positive correlation between the propensity for an individual to engage in chores and amount of time that individual will spend performing chores. Other parameter estimates also appear to give consistent interpretations across the two models.
CONCLUSION

As the examples above demonstrate, the NLMIXED procedure is a great tool for nonlinear mixed-effects-models that can be applied to a broad range of data when utilized with the GENERAL statement. In the examples, a two-part mixed-effects model is applied to an outcome in two forms. In both examples, a logistic link is used for the first model part. This part, part 1, utilizes logistic regression to estimate the odds of engaging in chores. In Example 1 the positive values are then modeled using a normal distribution. In Example 2 the positive values are modeled with a generalized gamma distribution. PROC NLMIXED and use of the GENERAL statement allow for estimation of these complex joint models, when they are not offered through the standard options in NLMIXED. Though this paper demonstrates the use of two specific joint models, the intention is to demonstrate the flexibility of the NLMIXED procedure with the GENERAL likelihood statement, which will allow for the application of a wide range of models to many complex data.

REFERENCES


CONTACT INFORMATION

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

Name: Melissa McTernan
Enterprise: University of California, Davis
Address: 202 Cousteau Place, Suite 250
City, State ZIP: Davis, CA 95616
E-mail: mlmcternan@ucdavis.edu

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

Other brand and product names are trademarks of their respective companies.