Paper Number: 190-29

Fitting Longitudinal Mixed Effect Logistic Regression Models with the NLMIXED Procedure

Peter H. Van Ness, John O'Leary, Amy L. Byers, Terri R. Fried, Joel Dubin,
Program On Aging, Yale University School of Medicine, New Haven, CT

Abstract

The NLMIXED procedure fits nonlinear mixed models; it is also useful for fitting linear mixed models having non-Gaussian error distributions. We recently used the NLMIXED procedure to perform longitudinal logistic regressions in which a random intercept was included to induce a compound symmetry covariance structure for repeated measures on individual subjects. We devised a macro to automate fitting this model.

The NLMIXED procedure requires writing out regression equations, declaring parameter names, and providing initial parameter estimates. To get accurate initial parameter estimates for our NLMIXED models we fit Generalized Estimating Equations (GEE) models with the GENMOD procedure. Our macro automates the procedure of fitting GEE models, entering their parameter estimates as the initial values for NLMIXED models, and then fitting the NLMIXED models. The macro does this recursively so that for most multivariable models only one new parameter estimate is entered into an NLMIXED model at a time. This step-by-step approach to model fitting increases the probability of successful convergence of the optimization procedure; however, it also means that fitting a model with, for example, five variables requires ten regression models—five GEE and five NLMIXED models. The macro makes this work simpler and more user-friendly.

Introduction

Logistic regression models for correlated data can be fit in several ways using SAS® statistical software. In a recent contribution to the 27th SAS® Users Group International Conference Oliver Kuss described and illustrated several such methods (Kuss, 2002). Like Kuss we are predisposed to modeling correlated binary data with the NLMIXED procedure because it provides improved maximum likelihood (ML) estimates relative to approximate ML estimates yielded by the GLIMMIX macro, and because, unlike the GENMOD procedure, it allows for the explicit modeling of random effects (*SAS/ STAT® User's Guide, Version 8*, 2000). Another drawback to the GLIMMIX approach is that its estimating method, penalized quasi-likelihood, has been shown—unless corrections are added—to yield biased results for binary outcomes in some circumstances (Breslow & Clayton, 1993; Breslow & Lin, 1995).

The NLMIXED procedure is not without limitations. Unlike the MIXED and GENMOD procedures it lacks a REPEATED statement and so has limited capacities for modeling the covariance structure of correlated data. In modeling longitudinal data in which there is not a high degree of serial correlation this limitation may not be serious. When a random effect is used for the intercept a compound symmetry covariance structure is

induced and this provides a reasonable fit to the data in many applied problems. An additional random time effect might improve the fit in certain cases and we are currently developing the capacity to handle models with two random effects. Preliminary modeling of the data using the GENMOD procedure and its REPEATED statement can help determine the best covariance structure.

Although a model properly fit with the NLMIXED procedure will generally yield improved ML estimates, in some cases the optimization procedure does not converge. No reliable estimates are then produced. Convergence is especially problematic for the NLMIXED procedure because it involves two complex algorithms. The first one, by default adaptive Gaussian quadrature, integrates out the random effects and thereby yields an approximation to the likelihood that is then optimized to yield ML estimates. The default for this second procedure is the quasi-Newton algorithm. SAS® provides alternatives to these procedures; however, the alternatives are generally less robust and useful only in special circumstances. Convergence with the NLMIXED procedure is considerably potentiated by the requirement that initial parameter estimates be provided. NLMIXED shares this requirement with the NLIN procedure and it is motivated by the fact that the nonlinear models—those for which the NLMIXED procedure was primarily designed—are more difficult to estimate than linear models. Provision of precise initial parameter estimates promotes convergence because—to use a spatial metaphor—it reduces the distance in the regression space over which the estimating algorithms must travel. Thus, it reduces opportunities for obstacles to convergence. By providing reasonably precise initial parameter estimates the macro also promotes successful convergence in the sense of converging to a global as opposed to a local maximum.

In the context of analyzing treatment preferences of a seriously ill elderly cohort (Fried, Bradley, Towle, & Allore, 2002; Fried, Bradley, & Towle, 2002), we have written a macro that fits longitudinal mixed effect models with NLMIXED in a way that minimizes its limitations. By first fitting generalized estimating equations (GEE) models with the GENMOD procedure we provide reasonably precise initial parameter estimates for the NLMIXED procedure. The code for the GENMOD procedure offers a second benefit in the early stages of model selection: simple changes of the code specify different correlation structures. This enables researchers to easily generate models with alternative correlation structures and thereby determine whether, for example, the compound symmetry assumption is appropriate for their data. The macro we have written automates model fitting and facilitates model selection. However, no macro can mechanically provide the combination of analysis and judgment required for expert model selection.

Also, by taking advantage of simple commands from the SAS® ODS Output System, the macro uses the output or final parameter estimates of an NLMIXED model as the input or initial parameter estimates for all but one of the variables in a subsequent model. This subsequent model is minimally more complex in that it contains only one new variable (or one set of dummy variables) that uses parameter estimates from the more dissimilar GENMOD model instead of the very similar NLMIXED model. For instance, in processing the final variable in a five-variable model, the first four variables will have initial parameter estimates supplied by a previous application of PROC NLMIXED and

only the fifth and final variable will have an initial parameter estimate supplied by the GENMOD model. This process assumes additivity—it assumes that the parameter estimates from the four-variable NLMIXED procedure do not substantially change when the fifth variable is added. In our experience this process works well. The step-by-step procedure in which GENMOD and NLMIXED models are alternately fit until one reaches the multivariable model of interest considerably promotes convergence and goodness-of-fit. When the additive assumption does not seem to apply, the appropriate interaction term can be entered as an additional variable.

Since the GENMOD procedure does not allow for random effects, it does not estimate the value of the standard deviation of the random intercept in the model fit by the NLMIXED procedure. An estimate of this term is required as one of the initial parameter estimates for a random effect NLMIXED model. If one is not provided by the analyst SAS uses a default value of 1. The macro provides the option of using initial estimates other than 1 and thus of using a process of trial and error for identifying an initial value that will allow the model to converge. Once it has converged, the resultant NLMIXED estimate of the random effect standard deviation can be used in subsequent applications of the NLMIXED procedure.

Technical Details

The syntax for the macro is the following:

%macro mixed_long_logit (_depvar, _indepvars, _dummyvars, _sdest, _libref, indata=);

_Depvar contains the binary outcome variable, _indepvars includes the independent variables, and indata specifies the input data set. No commas and only single spaces occur between the listed independent variables. _Libref is an optional argument used if permanent output data sets are desired from PROC NLMIXED. _Dummyvars, also optional, specifies one or more dummy variables that are added to the model in combination with other variables. The dummy variable(s) alert the macro to which variable(s) to delay entering into the model until it (they) can enter together with other dummy variables sharing the same reference group. _Sdest is the initial estimate of the standard deviation of the random effect for the intercept.

Macro Call Example:

%mixed_long_logit (health, age chf cancer income, chf, mydata,
indata=libindata.myfile);

In this case age separately enters the model first, the dummy variables chf and cancer (for which copd is the reference group) are next added together, and income enters last. The macro first calculates the number of independent variables by using the SAS® macro WORDS (*FAQ #1617*) and this count determines how many times the main processing loop will execute. During each iteration the next selected independent variable from _indepvars is added to macro variable INDEPVARSTR, is substituted for the value of

variable INDEPVARLAST, and is checked against _dummyvars to decide if it should enter the models. If confirmed, a call is made to PROC GENMOD and a data set containing GEEEmpEst is written to GM_OUT. Based on whether PROC GENMOD accepted one or more independent variables in the initial or subsequent execution, a macro (GETBETAINIT, GETBETALAST, or GETBETADUMMY) using the SYMPUT routine, searches through GM_OUT for the parameter estimates that will be used as an initial beta value(s) for PROC NLMIXED. Using several macro string functions the variable ETASTR and PARMSTR are created to build the syntax for statements used in each application of PROC NLMIXED.

SAS® output is produced for both PROC GENMOD and PROC NLMIXED after each independent variable entering the model completes the main processing loop. For example, five predictor variables entering the macro would generate five sets of GENMOD/NLMIXED output, where the final set provides SAS®'s results for the full model with all five predictor variables. Following each execution of the NLMIXED procedure, data steps are used to add an odds ratio variable and other housekeeping variables to a summary output data set MACROLOGIT_OUT. There is one of these macro-constructed output data sets for each execution of NLMIXED. Illustrative (and imaginary) macro-constructed final output for the macro call example given above is shown here:

Output Example:

Name of Variable	Parameter	Estimate	OR_Calculation	P_Value
age	beta1	0.0174	1.02	0.321
chf	beta2	0.5784	1.78	0.045
cancer	beta3	1.1962	3.31	<0.001
income	beta4	-0.2589	0.77	0.072

Extended Applications

The macro described above can be easily extended to mixed effect generalizations of logistic regression. Sheu has shown how to specify the regression equation in the NLMIXED procedure for ordinal outcomes and how to adapt its general likelihood capacities to outcomes with multinomial distributions (Sheu, 2002). This allows one to fit what McCullough has called "proportional odds models" (Walker & Duncan, 1967; McCullagh, 1980). Slight changes to the macro we have written would allow one to fit proportional odds models according to the same step-by-step approach. Bender and Benner have shown how to use the SAS® data step to prepare binary outcome data for analysis in what Fienberg (Fienberg & Mason, 1978; Fienberg, 1980) has named "continuation ratio models" (Bender & Benner, 2000). In some cases no further changes to the macro are necessary in order to fit continuation ratio models; in other cases, for instance, where one is using an inverted backward or forward version of the model, it is necessary only to make alterations that change the signs of parameter estimates. These

ordinal regression models are increasingly finding application in areas of research whose measurements reflect grouped continuous data (for which the proportional odds model is appropriate) or contribute to scales having a sequentially developmental character (for which the continuation ratio model is especially applicable). Other extensions would permit specification of link functions and error distributions available to both the GENMOD and NLMIXED procedures. For instance, an extension of the macro that allows substituting the log link for the logit link and the Poisson distribution for the binomial distribution is possible.

Conclusion

The macro we have designed for fitting longitudinal mixed effect logistic regression models makes this process simpler and more user-friendly. Minor modifications in the code extend the utility of the macro to fitting mixed effect generalizations of logistic regression models for correlated data, e.g., the proportional odds and continuation ratio models. Further extensions are anticipated. Because the macro fits both a GEE and a mixed effect model for each variable (or group of dummy variables) computation time is often considerable in a multivariable model. To address this issue we have written a second macro that allows a single variable to be added to a preexisting model that has already been fit with the above-described macro and whose results have been saved as a SAS® output data set. Both of these macros for logistic regression models with a random intercept term are currently available; work on extensions of these macros in order to handle a second random effect is underway.

Acknowledgments

The authors thank William T. Gallo and Heather G. Allore for their help and encouragement. This work was supported in part by NIH Grant # P30AG21342 and in part by NIH Grant # AG19769.

References

Bender, R., and Benner, A. (2000) Calculating Ordinal Regression Models in SAS and S-Plus, *Biometrical Journal* 42:677-99.

Breslow, N.E., and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* 88:9-25.

Breslow, N.E., and Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion, *Biometrika* 82:81-91.

FAQ #1617. SAS® Customer Support [cited September 22, 2003]. Available from http://support.sas.com/faq/016/FAQ01617.html.

Fienberg, S. (1980) *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: M.I.T. Press.

Fienberg, S.E., and Mason, W.M. (1978) Identification and estimation of period-age-cohort effects in the analysis of discrete archival data, *Sociological methodology 1979*:1-67.

Fried, T.R., Bradley, E.H., and Towle, V.R. (2002) Assessment of patient preferences: integrating treatments and outcomes, *Journals of Gerontology Series B-Psychological Sciences & Social Sciences* 57:S348-54.

Fried, T.R., Bradley, E.H., Towle, V.R., and Allore, H. (2002) Understanding the treatment preferences of seriously ill patients, *New England Journal of Medicine* 346:1061-66.

Kuss, O. (2002) How to use SAS® for logistic regression with correlated data, *Proceedings of the 27th Annual SAS® Users Group International Conference (SUGI 27)* 261-27.

McCullagh, P. (1980) Regression models for ordinal data, *Journal of the Royal Statistical Society, Series B* 42:109-42.

SAS/ STAT® User's Guide, Version 8(2000). Cary, NC: SAS Institute Inc.

Sheu, C.-F. (2002) Fitting mixed-effects models for repeated ordinal outcomes with the NLMIXED procedure, *Behavior Research Methods, Instruments, & Computers* 34:151-57.

Walker, S.H., and Duncan, D.B. (1967) Estimation of the probability of an event as a function of several independent variables, *Biometrika* 54:167-79.