Paper 3252-2015

How to Use SAS® for GMM Logistic Regression Models for Longitudinal Data with Time-Dependent Covariates

Katherine Cai, Jeffrey Wilson, Arizona State University

ABSTRACT

In longitudinal data, it is important to account for the correlation due to repeated measures and time-dependent covariates. Generalized method of moments can be used to estimate the coefficients in longitudinal data, although there are currently limited procedures in SAS® to produce GMM estimates for correlated data. In a recent paper, Lalonde, Wilson, and Yin provided a GMM model for estimating the coefficients in this type of data. SAS PROC IML was used to generate equations that needed to be solved to determine which estimating equations to use. In addition, this study extended classifications of moment conditions to include a type IV covariate. Two data sets were evaluated using this method, including re-hospitalization rates from a Medicare database as well as body mass index and future morbidity rates among Filipino children. Both examples contain binary responses, repeated measures, and time-dependent covariates. However, while this technique is useful, it is tedious and can also be complicated when determining the matrices necessary to obtain the estimating equations. We provide a concise and user-friendly macro to fit GMM logistic regression models with extended classifications.

INTRODUCTION

Longitudinal studies focus on a particular outcome measured across time. While studies investigating changes over time are very useful, multiple measurements collected on a single subject can result in correlated observations. The correlation can affect the standard errors in the estimation processes as the variation within a particular subject is likely to be much smaller than the variation between subjects. Moreover, time-dependent covariates present some additional challenges in working with longitudinal modeling. In particular, some predictors can change over time due to feedback from the response, and need to be accounted for in the modeling process. In turn, the change in predictors can impact the response.

SAS currently offers procedures, which utilize the statistical methods of generalized estimating equations (GEE) and generalized linear mixed models (GLMM), to analyze longitudinal data with binary outcomes. A macro that performs generalized method of moments (GMM) logistic regression is presented, which can appropriately take into account the correlation between covariate values. The use of the GMM macro is illustrated and compared to the results to SAS PROC GENMOD and PROC GLIMMIX. The three methods are demonstrated through the analysis of a body mass index (BMI) and morbidity dataset collected in the Philippines.

THE DATA

The data were collected by the International Food Policy Research Institute in the Bukidnon Province in the Philippines. BMI and morbidity were measured for 370 children at three separate time points, separated by 4-month intervals. The purpose of this study was to predict morbidity for children over time based on various factors.

The dataset contains a total of 1,110 observations, with three different BMI measurements for each of the 370 children. For each of the children (labeled by childID), the visit number (time) and body mass index (BMI) were recorded. For each of the three visits, it was recorded whether the child was sick (sick = 1) or healthy (sick = 0) at the time of measurement. Although additional information was collected in the study, we wish to predict morbidity based on the visit number and the child's BMI.

The SAS data set was created using the following code:

```
data Morbidity;
   input childID BMI time sick;
   datalines;
   206   14.95059    1     0
```

```
206   15.01923    2     0
206   14.79053    3     0
407   17.02125    1     1
407   16.0064     2     1
407   18.08021    3     0
705   15.83377    1     0
705   15.39259    2     1
705   15.08541    3     0
…
50805 14.78102    1     1
50805 15.05843    2     0
50805 14.91299    3     0
;
run;
```

The complete dataset is available online (www.public.asu.edu/~jeffreyw/).

A CRUDE ANALYSIS

An initial, although naïve approach, could be to assess the effect of BMI on morbidity assuming that each of the 1,100 observations is independent. This would provide an estimate of the effect of BMI on morbidity, but ignores that there is correlation between measurements for the same child. Standard logistic regression is equivalent to using generalized estimating equations (GEE) with an independent correlation structure (Liang and Zeger 1986). These methods can be performed using the GENMOD procedure:

```
proc genmod data=Morbidity descending;
   class childID time(ref="1");
   model sick = BMI time / dist = bin link = logit;
   repeated subject = childID / within=time corr=indep;
run;
```

PROC GENMOD is used for analyses with generalized linear models, and can perform logistic regression by specifying a binomial probability distribution with the option DIST=BIN and a logit link with the option LINK=LOGIT in the MODEL statement. This is the distribution at each time-period. The REPEATED statement employs the generalized estimating equations approach, and CORR=INDEP indicates an independent within-time correlation matrix, which will be a 3x3 matrix. Table 1 displays the results for this analysis.

Parameter	Estimate	Standard Error	Z-Value	p-Value
Intercept	-0.3021	0.8446	-0.36	0.7206
BMI	-0.0362	0.0545	-0.66	0.5072
Time (2)	-0.3382	0.1527	-2.21	0.0268
Time (3)	0.0993	0.1487	0.67	0.5045

Table 1. Parameter Estimates for GEE with an Independent Correlation Structure

Standard logistic regression assumes independence between observations and suggests that BMI is not a significant predictor in determining a child's morbidity (p-value = 0.5072). Time is shown to have a statistically significant effect on sickness. For time (2), the 95% confidence interval is (-0.6375, -0.0389), indicating that the odds of a child being sick decreased on the second visit compared to the first visit.

THE LOGISTIC REGRESSION MODEL FOR LONGITUDINAL DATA WITH TIME-DEPENDENT COVARIATES

Lai and Small (2007) demonstrated the use of the GMM approach to obtain estimates when analyzing data with time-dependent covariates. They classified time-dependent covariates into one of three types. For a particular patient, with measurements made at different times, a covariate is type I if all of the

measurements are independent. Type I covariates are not dependent on prior measurements of the outcome. A covariate is type II if the outcome depends on previous values of the covariate. These types of predictors are commonly seen in autoregressive models. Type III covariates depend on the previous values of the outcome. Lalonde, Wilson, and Yin (2014) expanded this method by adding a type IV covariate. A covariate is type IV if future responses are not affected by the previous covariate process. This occurs in cases where there is a covariate relationship, but feedback is not provided to the response process.

This method relies on the correlation between the residuals (response measured at a time t) and the covariates at a particular time (covariate measured at a time s). In order to obtain the valid conditions, Lalonde, Wilson, and Yin (2014) suggested that if there is no correlation between the residuals and the covariate at two different times, then the corresponding moment condition is valid. Multivariate integration is used to perform multiple comparison tests to determine if the correlation is significantly different from zero, and thus identifies a valid moment condition.

The GMM estimates are obtained using the valid moment conditions. The GMM estimators for the model parameters minimize a quadratic objective function based on the valid moment conditions, a weight matrix calculated from the inverse of the covariance, and initial parameter estimates obtained from GEE. The minimization is performed using Newton-Raphson optimization in SAS/IML. A logistic regression model with time-dependent covariates is fit using these GMM estimates.

THE GMM MACRO

Two macro calls are required to perform GMM logistic regression. The first macro, %MVINTEGRATION was adapted from the SAS/IML program written by Alan Genz and Frank Bretz (version date 6/21/00). %MVINTEGRATION calculates multivariate normal probabilities by applying a randomized lattice rule on the transformed integral (Genz 1992, Genz 1993). This macro requires that the reference library (REFLIB) where a SAS Catalog of the IML modules is stored, is specified in the macro call. Sample code to call %MVINTEGRATION is shown:

```
%MVIntegration(reflib="C:\Users\Documents\Code");
```

The second macro call to %GMM identifies the covariate types and performs generalized method of moments logistic regression. The %GMM macro has 8 required input values. DS is used to specify the location of the dataset, and the FILE input references the SAS file name (extension .sas7bdat). REFLIB refers to the reference library where the SAS Catalog of the IML modules is stored, and should contain the same file location as was provided to %MVINTEGRATION. TIMEVAR refers to the name of the variable in the SAS file that specifies time as “1”, “2”, etc. Currently %GMM can process up to three measurement periods. OUTVAR specifies the variable name of the response variable, and PREDVAR specifies the variable name(s) of the covariate(s). IDVAR refers to the variable name of the subject identification variable, which will be used to identify multiple measurements from the same subject. ALPHA is the significance level for identifying significant correlation between the residual and covariate. A sample call to %GMM is shown:

```
%GMM(ds='C:\Users\Documents\Data',
         file= Morbidity,
         reflib="C:\Users\Documents\Code ",
         timeVar=time,
         outVar=sick,
         predVar=BMI,
         idVar=childID,
         alpha=0.05);
```

In the Philippines data set, BMI was the only covariate investigated to predict morbidity. However, %GMM allows for additional covariates to be included in the model. Multiple predictors can be specified in the macro call using the syntax:

```
predVar= BMI age gender,
```

The macro produces SAS output for the multivariate integration, type matrices, GEE estimates, optimization using Newton Raphson, and the final GMM parameter estimates. Output 1 and Output 2

display the output related to the correlation tests for moment conditions. Output 1 displays the output tables R4OUT and P4OUT. R4OUT is the correlation between the residuals and BMI measurements at each of the three time points. P4OUT contains the p-values for each of the correlation tests for the moment conditions. Note that R4OUT and P4OUT only display output for the specified PREDVAR covariates. The intercept and time indicator variables are treated as type I.

r4out		
-0.000223	-0.032616	-0.081368
-0.050525	0.0023455	-0.048459
0.001922	0.0281022	0.0586089

p4out		
0.996578	0.5741719	0.1379108
0.3109452	0.9660497	0.3980495
0.9688526	0.5828711	0.2579849

Output 1. Output from %GMM

Output 2 displays the TYPE4OUT table which displays a 1 for valid moment conditions and a 0 for invalid moment conditions. Invalid moment conditions occur when the correlation between the residual and the covariate at the specified time is not significantly different from 0. For the TYPE4OUT table, columns 1-3 correspond to the intercept. Columns 7-9 and columns 10-12 correspond to the time indicator variables t2 and t3, respectively. The remaining columns 4-6 are for the three time measurements of the covariate BMI.

Type4out												
	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12
ROW1	1	0	0	1	1	0	0	0	0	0	0	0
ROW2	0	1	0	0	1	0	0	1	0	0	0	0
ROW3	0	0	1	1	1	1	0	0	0	0	0	1

Output 2. Output from %GMM

The results shown in Output 1 and Output 2 are used to determine the covariate type and the valid moment conditions to use in GMM estimation. Table 2 summarizes the SAS output and displays the correlation, p-value, and validity indicator information for the BMI covariate. A validity of 1 indicates that the correlation between the residuals and BMI is not significantly different than 0, and thus is a valid moment conditions. The validities marked with asterisks have corresponding p-values that suggest valid moment conditions by individual tests, but do not meet the requirements by a multiple test.

Correlation		**BMI**		
		s = 1	**s = 2**	**s = 3**
Residuals	**t = 1**	-0.0002	-0.0326	-0.0814
	t = 2	-0.0505	0.0023	-0.0485
	t = 3	0.0019	0.0281	0.0586
p-Value				

Residuals	**t = 1**	0.9966	0.5742	0.1379
	t = 2	0.3109	0.9660	0.3980
	t = 3	0.9689	0.5829	0.2580
Validity				
Residuals	**t = 1**	1	1	0*
	t = 2	0*	1	0*
	t = 3	1	1	1

Table 2. Correlation Test Results for BMI Moment Conditions

After the covariate types are determined, the valid moment conditions are used to determine the method of moments parameter estimates. The following output, included in Output 3, displays the estimate, standard deviation, test statistic, and p-value for each GMM parameter. OUTTITLE lists the column names for each of the columns in OUTMTX. OUTMTX contains the GMM parameter estimates and tests for significance. BETAVEC is a summary of the GMM estimates, which is equivalent to the first column of OUTMTX. The first estimate provided is for the intercept term. The last two estimates are for the variables "t2" and "t3," which are indicator variables created by the %GMM macro to represent times 2 and 3, respectively. %GMM assumes that measurement time 1 is the reference category, and thus has an estimate of 0.

Outtitle

Estimate	StdDev	Zvalue	Pvalue

Outmtx

0.5337492	0.4593011	1.1620901	0.2451989
-0.098186	0.0263665	-3.72389	0.0001962
-0.250055	0.0400893	-6.237433	4.448E-10
0.194842	0.1301101	1.4975164	0.1342589

betavec

0.5337492	-0.098186	-0.250055	0.194842

Output 3. Output from %GMM

Thus, for the morbidity data collected in the Philippines, the parameter estimates are obtained.

Parameter	Estimate	Standard Error	Z-Value	p-Value
Intercept	0.5628	0.4490	1.25	0.2100
BMI	-0.1003	0.0256	-3.91	<0.0001
Time (2)	-0.2485	0.0379	-6.56	<0.0001
Time (3)	0.1994	0.1305	1.53	0.1266

Table 3. Parameter Estimates for GMM Logistic Regression

The initial analysis compared to the GMM logistic regression model indicates a difference in the significance of BMI. While the standard logistic regression model showed that BMI was not a statistically significant predictor of sickness, the GMM model shows that BMI is a highly significant predictor (p-value

< 0.0001). This difference is captured by accounting for the correlation within BMI measurements for each patient, since BMI is a time-dependent covariate. Time 2 is also shown to be highly significant in predicting morbidity, which is consistent to the standard logistic regression results.

COMPARISON OF METHODS

Logistic regression using a GEE model, assuming an unstructured covariance structure, as well as a mixed model were assessed as alternatives to the GMM logistic regression method.

GENERALIZED ESTIMATING EQUATIONS

SAS PROC GENMOD can model correlated data by using the REPEATED statement. This option implements generalized estimating equations to account for the presence of clustering. GEE can be used to produce a population averaged model, which is comparable to the GMM technique. In this example, clustering or correlation is due to the repeated measurements on the children (indicated by childID). The within patient correlation structure can be specified using the CORR option. The call to PROC GENMOD is displayed:

```
proc genmod data=Morbidity descending;
   class childID time(ref="1");
   model sick = BMI time / dist = bin link = logit;
   repeated subject = childID / within=time corr=un corrw;
run;
```

Similar to the initial analysis, the options DIST=BIN and LINK=LOGIT are provided to specify a logistic regression model. DIST specifies the distribution of the outcome variable and LINK is the generalized linear model link function. The REPEATED statement indicates the GEE approach, and CORR=UN specifies an unstructured within-time correlation matrix. The option CORRW displays the working correlation matrix between the three time point measurements in the output.

Parameter	Estimate	Standard Error	Z-Value	p-Value
Intercept	-0.4196	0.8374	-0.50	0.6163
BMI	-0.0286	0.0541	-0.53	0.5974
Time (2)	-0.3372	0.1527	-2.21	0.0272
Time (3)	0.0999	0.1487	0.67	0.5018

Table 4. Parameter Estimates for GEE with an Unstructured Correlation Structure

While the GEE model takes into account the correlation between measurements on the same child, this model produces a similar result to the initial analysis. Standard logistic regression, which is equivalent to having an independent correlation structure, and GEE with an unstructured correlation structure, both suggest that BMI does not have a statistically significant effect on predicting morbidity. In both analyses, the only significant predictor is time 2. For the GEE model, the 95% confidence interval for time 2 is (-0.6364, -0.0380). The negative coefficient for time 2 suggests that at time 2 the odds of being sick are lower than the odds of being healthy.

From the SAS PROC GENMOD output, the working correlation matrix can be specified.

	Time 1	Time 2	Time 3
Time 1	1.000	0.1742	0.1268
Time 2	0.1742	1.000	0.1039
Time 3	0.1268	0.1039	1.000

Table 5. Working Correlation Matrix for GEE with an Unstructured Correlation Structure

Although GEE is an appropriate model to take into account repeated measurements, it did not provide reliable results for a longitudinal model including time-dependent covariates. Consistency of the GEE method is not assured when using time-dependent covariates, unless a key assumption is satisfied (Pepe

1994, Hu 1992). This assumption will be met and consistency will be guaranteed if using an independent working correlation matrix.

GENERALIZED LINEAR MIXED MODEL

The GLIMMIX procedure can be used to model generalized linear mixed models in SAS. Random effects can be evaluated for the intercept term or for another covariate. In contrast to using GEE in PROC GENMOD, the GLIMMIX procedure produces a subject-specific model. While both methods are options to perform logistic regression for correlated data, the interpretation will vary. The code for PROC GLIMMIX is shown:

```
proc glimmix data=Morbidity;
   class childID time(ref="1");
   model sick = BMI time / dist=bin link=logit solution;
   random intercept / subject = childID;
run;
```

Similar to PROC GENMOD, the options DIST=BIN and LINK=LOGIT are specified in the MODEL statement to specify logistic regression. The RANDOM statement is used to specify any random effect terms that should be included in the model. Since only a random intercept is being considered, INTERCEPT is included after the RANDOM statement. If a random slope was being considered, the name of the variable would follow the RANDOM statement. SUBJECT is used to specify the subject identification variable which will differentiate between repeated measurements. In the morbidity example, children have multiple BMI measurements, so SUBJECT = childID is used. The SOLUTION option in the MODEL statement provides the estimate(s) of the random effect(s) in the output. In this example, the SOLUTION option will present the estimate of the random intercept due to variation between children.

Parameter	**Estimate**	**Standard Error**	**t-Value**	**p-Value**
Intercept	-0.3780	0.8372	-0.45	0.6518
BMI	-0.0325	0.0537	-0.61	0.5453
Time (2)	-0.3448	0.1695	-2.03	0.0423
Time (3)	0.1019	0.1612	0.63	0.5276
Random Intercept	**Estimate**	**Standard Error**	**Z-Value**	**p-Value**
Subject = childID	0.4700	0.1548	3.04	0.0012

Table 6. Parameter Estimates for Generalized Linear Mixed Models with a Random Intercept

The generalized linear mixed model produces comparable results to the initial analysis and the GEE model. The analysis again fails to identify BMI as a significant predictor of morbidity for children. Similar to the previous models, time 2 is found to be mildly significant with a p-value of 0.0423.

The mixed model produced similar results to the GEE model, and did not provide a reliable result for the time-dependent covariate BMI. While this is another method that can be used for the analysis of longitudinal data, it differs in interpretation from GEE. The mixed model is a subject-specific model, and cannot be used to discuss the aggregate response for the population. This differs from the GMM logistic regression model which is a population averaged model.

CONCLUSION

Time-dependent covariates can create many challenges in data analysis due to the response feedback present in the data. While current methods are able to address repeated measurement issues in longitudinal data, many are limited in appropriately handling time-dependent covariates. Based on research performed by Lalonde, Wilson, and Yin (2014), a %GMM macro was developed to perform GMM logistic regression. This method incorporates valid moment conditions by checking for significant correlation between the residuals and covariates. The GMM estimates are produced using initial estimates from the GEE model, and then performing an optimization with the valid moment conditions

using Newton-Raphson Optimization. Using the morbidity data collected in the Philippines, the %GMM macro results was compared to various other methods.

This study showed that standard logistic regression, GEE with an unstructured correlation matrix, and GLMM produced similar results. All cases failed to identify BMI as a statistically significant predictor of morbidity. Standard logistic regression does not appropriately handle the repeated observations nor does it address the time-dependent covariates, since many of the valid moment conditions are left out. The GEE method is not assured to have consistent estimators when using time-dependent covariates, and GLMM similarly is unable to produce appropriate estimates and also cannot be used to address predictions for the aggregate population.

REFERENCES

Genz, Alan. 1992. "Numerical Computation of Multivariate Normal Probabilities."*Journal of Computational and Graphical Statistics*, 1(2):141-149.

Genz, Alan. 1993. "Comparison of Methods for the Computation of Multivariate Normal Probabilities." *Computing Science and Statistics*, 25:400-405.

Hu, Fu-Chang. 1992. "A Statistical Methodology for Analyzing the Causal Health Effect of a Time-Dependent Exposure from Longitudinal Data." Harvard School of Public Health Dissertation.

Lai, Tze Leung and Dylan Small. 2007. "Marginal Regression of Longitudinal Data with Time-Dependent Covariates: a Generalized Method-of-Moments Approach." *Journal of the Royal Statistical Society, Series B*, 69(1):79-99.

Lalonde, TL, Wilson, JR, and J. Yin. 2014. "GMM Logistic Regression Models for Longitudinal Data with Time-Dependent Covariates and Extended Classifications." *Statistics in Medicine*, 33(27):4756-4769.

Liang, Kung-Yee and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, 73(1):13-22.

Pepe, Margaret Sullivan and Garnet L. Anderson. 1994. "A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data." *Communications in Statistics – Simulation and Computation*, 23(4):939-951.

Zeger, Scott L. and Kung-Yee Liang. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." *Biometrics*, 42(1):121-130.

CONTACT INFORMATION

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

Katherine Cai
Arizona State University
katherine.cai@asu.edu

Jeffrey Wilson
Arizona State University
jeffrey.wilson@asu.edu
www.public.asu.edu/~jeffreyw/

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.