Paper 1025-2017

GMM Logistic Regression with Time-Dependent Covariates and Feedback Processes in SAS™

Kyle M. Irimata, Arizona State University; Jeffrey R. Wilson, Arizona State University

ABSTRACT

The analysis of longitudinal data requires a model which correctly accounts for both the inherent correlation amongst the responses as a result of the repeated measurements, as well as the feedback between the responses and predictors at different time points. Lalonde, Wilson and Yin (2013) developed an approach based on generalized methods of moments (GMM) for identifying and using valid moment conditions to account for time dependent covariate in longitudinal data with binary outcomes. However, the model developed using this approach does not provide information regarding the specific relationships that exist across time points. We present a SAS™ macro which extends the work of Lalonde, Wilson and Yin by utilizing valid moment conditions to estimate and evaluate the relationships between the response and predictors at different time periods. The performance of this method is compared to previously established results.

INTRODUCTION

Longitudinal studies often involve the collection of repeated measurements on the same subjects over time. In many cases, the effect of the covariates on the outcome can change over the course of the study. For example, in a study of patient readmission to a hospital across a number of visits, researchers may collect information on covariates such as the number of diseases affecting the patient. In these studies, we would expect the response (hospital readmission) to vary across time and would model this change as a function of the covariates. However, we may also expect the number of diseases to vary across the time points and thus may consider treating it as time dependent. This time dependence can also result in a lagged effect on the outcome in which the level of the covariate at previous time points has a carryover effect on the outcome at future time points. The repeated measurements may also result in a feedback process from the outcome onto the covariate across time points.

A number of approaches have been discussed to address the correlation that arises as a result of the associations present in longitudinal data. Generalized estimating equation (GEE) are often used when the clustering in longitudinal data leads to violations in the assumption of independence (Zeger, Liang 1986; Liang, Zeger 1992). These models provide inference for a population average of the outcome using a working correlation matrix to account for the dependencies in the outcomes. However, Pepe et al (1994) and Hu (1993) found that GEE may not achieve consistency in the presence of time dependent covariates. One method for adapting GEE to time dependent covariates is through the use of a lagged effect with an independent working correlation matrix. However, these models require the strong assumption that the responses for each individual across time are independent, which may not always be appropriate.

In the case of time dependent covariates, generalized method of moments (GMM) can be preferred to GEE (Lai and Small 2007). Lalonde, Wilson and Yin provided a GMM approach to incorporate only valid moment conditions in estimating GMM regression parameters (2014). We provide an extension of this approach to explicitly model the relationships between the outcome and the covariates across time and present a SAS macro implementation for fitting these models which is flexible and applicable to the analysis of binary outcome data with varying numbers of time points. Two illustrations are considered and comparisons are made to the GMM estimates provided by the Lalonde, Wilson and Yin approach.

GENERALIZED METHOD OF MOMENTS

Suppose our data has repeated observations taken across T time points on N subjects with measurements on J different covariates. We denote these observations as $(y_{it}, \boldsymbol{x_{it}})$ for the $i = 1, \dots, N$ subjects, at time $t = 1, \dots, T$. The covariates are contained within the vector $\boldsymbol{x_{it}}$, which is comprised of the J covariates indexed by $j = 1, \dots, J$. The outcomes y_{is} and y_{kt} are assumed to be independent when $i \neq k$,

but not necessarily when $i = k$ and $s \neq t$. Thus, observations from different subjects are independent, but measurements at different times taken on the same subject may not be independent.

Lai and Small (2007) presented a GMM approach to fit a marginal model for longitudinal data with continuous outcomes to account for time dependent covariates. They made use of the moment condition

$$\mathrm{E}\left[\frac{\partial \mu_{\mathrm{is}}(\boldsymbol{\beta})}{\partial \beta_{\mathrm{j}}}\{\mathrm{y_{it}} - \mu_{\mathrm{it}}(\boldsymbol{\beta})\}\right] = 0$$

where $\mu_{\mathrm{it}}(\boldsymbol{\beta})$ is the expected value of y_{it} based on $\boldsymbol{x_{it}}$ and where $\boldsymbol{\beta}$ is the vector of parameter values for the marginal regression on y_{it}. Further, we assume that the generalized linear model is appropriate in relating the outcome to the covariates. This assumption on the marginal distribution ensures that this moment condition will hold when the response and the covariates are from the same time point, or equivalently when $s = t$. However, this equation is not guaranteed to hold when $s \neq t$.

Lai and Small (2007) discussed the use of a set of classifications for time dependent covariates. They identified Type I, Type II and Type III covariates, which can be used to include or exclude certain moment conditions, based on the type of covariate. A Type I covariate is one in which the above moment condition holds for all s and t. A Type II covariate is one in which this moment condition holds for $s \geq t$, but does not hold for all $s < t$. A Type III covariate is one in which the above moment condition does not hold for any $s > t$. They showed that incorporating this additional information regarding the moment conditions provided marked improvements over GEE.

Lalonde, Wilson and Yin (2014) extended this approach and identified the Type IV covariate, which is unique from the other three covariate types. Beyond this additional covariate type, they proposed a method for selecting and incorporating valid moment conditions when fitting generalized method of moments. This approach did not require the identification of covariate types, but instead investigated the relationships between the outcome and each covariate across time.

Since the moment conditions when $s = t$ are all assumed to be valid, consider the $T(T-1)$ moment conditions for the cases $s \neq t$ which must be evaluated for validity. Lalonde, Wilson and Yin (2014) showed that for $s \neq t$, the moment condition

$$\mathrm{E}\left[\frac{\partial \mu_{\mathrm{is}}(\boldsymbol{\beta})}{\partial \beta_{\mathrm{j}}}\{\mathrm{y_{it}} - \mu_{\mathrm{it}}(\boldsymbol{\beta})\}\right] = 0$$

is equivalent to

$$Corr\left[\frac{\partial \mu_{\mathrm{is}}(\boldsymbol{\beta})}{\partial \beta_{\mathrm{j}}}\{\mathrm{y_{it}} - \mu_{\mathrm{it}}(\boldsymbol{\beta})\}\right] = 0$$

Thus, each moment condition can be tested by evaluating the correlation between the residual at time t, denoted by e_t and the covariate at time s, which we denote by ρ_{x_s, e_t}. Based on previous work (Fisher 1928; Lalonde, Wilson and Yin 2014), we have that under $H_0: \rho_{x_s, e_t} = 0$ the sample correlation coefficient $\hat{\rho}_{x_s, e_t}$ is asymptotically distributed as normal; therefore each correlation and thus the respective moment condition can be evaluated using a standard normal test statistic given by

$$z_{ts}^{*} = \frac{\hat{\rho}_{x_s, e_t}}{\sqrt{\hat{\mu}_{22}/N}}$$

Where $\hat{\rho}_{x_s, e_t} = \frac{\sum_{i=1}^{N} e_{it} X_{isj}}{\sqrt{\sum_{i=1}^{N} e_{it}^2 \sum_{i=1}^{N} X_{isj}^2}}$ and the variance is given by $\hat{\mu}_{22} = \frac{1}{N}\sum_{i=1}^{N} e_{it}^2 x_{ijs}^2$. This test can be conducted as usual, with a user-defined significance level, α.

Once the valid moment conditions are identified, the GMM regression parameters can be obtained by solving an objective function based on a matrix composed of valid moment conditions and a weight matrix (Lalonde, Wilson and Yin 2014). The GMM estimators for the regression parameters is the vector $\boldsymbol{\beta}$ that minimizes the objective function

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{GMM}} = argmin_{\boldsymbol{\beta_0}} \boldsymbol{G_n}(\boldsymbol{\beta_0})^{\boldsymbol{T}} \boldsymbol{W_n}(\boldsymbol{\beta_0}) \boldsymbol{G_n}(\boldsymbol{\beta_0})$$

where $G_n(\boldsymbol{\beta_0})$ is a reshaped vector of valid moment conditions, summed across all N subjects, and $W_n(\boldsymbol{\beta_0})$ is the matrix of weights. The vector of valid moment conditions for the i[th] subject in the study is given by $\boldsymbol{g_i}$, which is composed of the elements $\frac{\partial \mu_{is}(\boldsymbol{\beta_0})}{\partial \beta_j}\{y_{it} - \mu_{it}(\boldsymbol{\beta_0})\}$.

Although the GMM model discussed by Lalonde, Wilson and Yin (LWY) incorporates information from the time dependent covariates, it does not provide insight into the individual relationships that exist across time. We propose the use of an extension to the LWY approach to estimate each of these relationships. For the case of three time points, the partial GMM can be fit according to the model

$$g(\mu_{it}) = \beta_0 + \beta_j^{tt} X_{i,t} + \beta_j^{|s-t|=1} X_{i,t-1} + \beta_j^{|s-t|=2} X_{i,t-2}$$

where β_0 is the intercept, β_j^{tt} represents the effect of the j[th] predictor on the outcome in the same time point. The coefficients $\beta_j^{|s-t|=1}$ and $\beta_j^{|s-t|=2}$ denote the effect of the j[th] covariate on the outcome across a one or two point lag, respectively. Although this model is written with respect to three time points, this approach can be easily extended to data sets with more or less time points.

SAS MACRO

The partial GMM can be fit in SAS using the general macro call which mimics the arguments used in the %***GMM*** macro introduced by Cai and Wilson (2015; 2016):

```
%partialGMM(ds=, file=, timeVar=, outVar=, predVar=, idVar=, alpha=);
```

The first argument *DS* is used to specify the location of the dataset, while the second argument *file* is used to reference the SAS file (.sas7bdat) to be analyzed. The next four arguments are used to identify specific variables in the data set which will be used in fitting the partial GMM. The *timeVar* argument identifies the variable name for the time points. The variables *outVar* and *predVar* identify the binary outcome variable and set of covariates, respectively. Multiple covariates can be analyzed and specified in the predVar statement, where each covariate should be delimited by a space. The *idVar* argument takes the subject identification variable. The last argument, *alpha*, refers to the significance level at which the correlations between the residuals and covariates will be tested for evaluating validity of the moment conditions.

As an example of the syntax for this macro, consider the call to the %***partialGMM*** macro for the Medicare example discussed in the following section. The corresponding call is given by:

```
%partialGMM(ds='C:\Users\Documents\IML',
   file=Medicare,
   timeVar=time,
   outVar=biRadmit,
   predVar=NDX NPR LOS DX101,
   idVar=PNUM_R,
   alpha=0.05);
```

This macro relies on a number of base SAS procedures such as PROC LOGISTIC and PROC GENMOD to obtaining residuals for the logistic regression model, as well as appropriate starting values for optimization based on the GEE estimates. Identification of valid moment conditions as well as estimation of the model parameters and respective analyses are conducted primarily in PROC IML. Newton-Raphson optimization is used to identify the parameter values which minimize the objective function. The %***partialGMM*** macro returns parameter estimates for each of the covariates (and the respective lags), along with estimates of the standard deviation, Z-value and P-value for the hypothesis $H_0: \beta_j = 0$. The lags are referenced by concatenating an underscore and a lag count for each of the covariates, where the lag count begins at 0 for the covariate relationships within the same time point (also called current time point) and end at $T - 1$. For example, for a generic variable 'X' with measurements taken at 3 time points, the macro will return estimates for X_0, representing the effect of X at the current time point, X_1, denoting the effect of X at a one time point lag and X_2, representing the effect of X at a two time point lag. The

macro will also produce notes if any covariate relationships cannot be estimated. This macro is available online at http://www.public.asu.edu/~jeffreyw/.

DATA EXAMPLE

MEDICARE

To illustrate the use of the %***partialGMM*** macro, we first analyzed Medicare data extracted from the Arizona State Inpatient database (Lalonde, Wilson and Yin 2014). The outcome of interest was hospital readmission for the same condition. Thus, the response in this data set is binary representing whether or not the following visit for a given subject occurred within 30 days for the same condition. This data set contains information on 1,625 patients aged 65 and over, who were admitted to the hospital exactly four times, thus yielding three measurements per subject. In addition to the outcome, the data also included four covariates, representing number of diseases (NDX), number of procedures (NPR), length of stay (LOS) and presence of coronary atherosclerosis (DX101).

This data was analyzed using the %***partialGMM*** macro, where each of the four covariates were treated as time dependent. The partial GMM model was fit using the call:

```
%partialGMM(ds='C:\Users\Documents\IML',
    file=Medicare,
    timeVar=time,
    outVar=biRadmit,
    predVar=NDX NPR LOS DX101,
    idVar=PNUM_R,
    alpha=0.05);
```

Based on this analysis, we can see that NDX has a significant effect on readmission at the current time point and at a one time point lag. NPR has a moderately significant effect on readmission at the current time point and a significant effect at a one time point lag. LOS had a significant effect in the current time point and a moderately significant effect on readmission at a one time point lag. DX101 was not found to be significant at any time point. We note that the macro produced no results for LOS at a two point lag. Since there were no valid moment conditions for this covariate at a two point lag, this relationship cannot be estimated. The output from this macro call are included in Figure 1.

Analysis of Partial GMM Estimates				
	Estimate	StdDev	Zvalue	Pvalue
Intercept	-0.486973	0.1201859	-4.051835	0.0000508
NDX_0	0.0661016	0.0155742	4.2443094	0.0000219
NPR_0	-0.036581	0.0194239	-1.8833	0.0596598
LOS_0	0.0475609	0.0070442	6.7517682	1.461E-11
DX101_0	-0.068819	0.0930831	-0.739332	0.4597053
NDX_1	-0.045744	0.0117394	-3.896613	0.0000975
NPR_1	-0.003627	0.0220895	-0.164203	0.8695714
LOS_1	0.0124338	0.0067243	1.8490775	0.0644466
DX101_1	-0.033011	0.1043142	-0.316456	0.7516566
NDX_2	0.0233436	0.013723	1.7010574	0.0889322
NPR_2	-0.025618	0.0287561	-0.890858	0.3730053
DX101_2	-0.093735	0.1410271	-0.664657	0.5062698

Figure 1. Partial GMM Estimates for the Medicare Data Analysis

In addition to the GMM parameter estimates, the macro also produces a note regarding which parameters were estimated. As noted previously, LOS cannot be estimated at the two time lag. An example of this output is included in Figure 2.

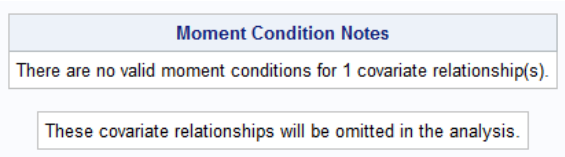

Figure 2. Moment Condition Notes for the Medicare Data Analysis

As a comparison, we also fit the generalized estimating equations model, as well as the GMM approach discussed by Lalonde, Wilson and Yin (2014). Although this is the same data set analyzed in their work, we omit the indicators for time and refit these models without the indicators; thus the estimates vary slightly from their original results. We can see that the GEE and LWY-GMM approach produce similar results. We compared these models to the parameter estimates developed using the Partial GMM approach for the current time point and saw that the parameter estimates are similar for the current time point and that the significance of the predictors are also nearly the same. However, both of these approaches produce only one parameter estimate per covariate, while the Partial GMM provides further insight into the individual relationships that may exist across time, as noted previously. In particular, we saw that there were significant lagged effects for some of the covariates, which cannot be evaluated using GEE or the LWY-GMM. The results of these analyses are included in Table 1.

	LWY-GMM		GEE		Partial GMM (Current time)	
Parameter	**Estimate**	**p-value**	**Estimate**	**p-value**	**Estimate**	**p-value**
Intercept	-0.6143	<.0001	-0.5936	<.0001	-0.487	<.0001
NDX	0.0567	0.000287	0.065	<.0001	0.0661	<.0001
NPR	-0.0239	0.2038	-0.0184	0.3274	-0.03656	0.0597
LOS	0.0463	<.0001	0.0304	<.0001	0.04756	<.0001
DX101	-0.0479	0.606371	-0.1022	0.2661	-0.0688	0.4597

Table 1. Comparison of LWY-GMM, GEE and Partial GMM models

PHILIPPINES

We also analyzed data collected in the Bukidnon Province of the Philippines by the International Food Policy Research Institute to further illustrate the use of the %***partialGMM*** macro (Bhargava 1994). This study investigated the relationship between morbidity and body mass index (BMI) and includes information on 370 children, each with three observations taken across time. The binary outcome in this data was morbidity status and we investigated two covariates representing BMI and age, both of which were treated as time dependent. This model was fit using the macro call:

```
%partialGMM(ds='C:\Users\Documents\IML',
    file=philippb,
    timeVar=time,
    outVar=sick,
    predVar=bmi age,
    idVar=childid,
    alpha=0.05);
```

We found that BMI was significant in the current time point, at the one time lag as well as at a two time point lag. Age was also shown to be significant within the current time point, as well as at a one time lag. The output produced by the macro is included in Figure 3.

Analysis of Partial GMM Estimates				
	Estimate	StdDev	Zvalue	Pvalue
Intercept	0.6462784	0.5241257	1.2330598	0.2175534
BMI_0	-0.044054	0.0191171	-2.304426	0.0211988
AGE_0	-0.021064	0.0070557	-2.985407	0.002832
BMI_1	-0.032355	0.0091918	-3.520039	0.0004315
AGE_1	0.0051446	0.0057537	0.8941264	0.3712543
BMI_2	0.0211082	0.0049901	4.2299885	0.0000234
AGE_2	0.0030977	0.0044567	0.695061	0.4870171

Figure 3. Partial GMM Estimates for the Philippines Data Analysis

For this analysis, we were able to estimate all covariate effects, which is also noted in the output from the macro. This output is included for illustration in Figure 4.

Moment Condition Notes
All covariate relationships will be evaluated.

Figure 4. Moment Condition Notes for the Medicare Data Analysis

CONCLUSION

There are a number of associations that arise as a result of the repeated measurements in longitudinal studies. Many methods have been proposed to help account for these associations, such as generalized estimating equations, or generalized method of moments. In many cases, relationships can exist across time points between the covariate and the response which cannot be accounted for by a single regression parameter.

We provide the %partialGMM macro in SAS, which explicitly accounts for these relationships across time. This macro incorporates only valid moment conditions to estimate logistic regression parameters, and also provides appropriate p-values for hypothesis testing. The effect of each covariate on the response is evaluated at the current time point, as well as for lags up to $T - 1$. Looking forward, this macro will be extended to accommodate continuous response data as well as time independent covariates.

REFERENCES

Bhargava, A. 1994. Modelling the health of Filipino children. Journal of the Royal Statistical Society, Series A; 157(3): 417-432.

Cai, K. and Wilson, JR. 2015. How to Use SAS® for GMM Logistic Regression Models for Longitudinal Data with Time-Dependent Covariates. SAS Global Forum: Paper 3252-2015.

Cai, K. and Wilson, JR. 2016. SAS® Macro for generalized method of moments estimation for longitudinal data with time-dependent covariates. SAS Global Forum: Paper 10260-2016.

Fisher, RA. 1928. The general sampling distribution of the multiple correlation coefficient. Proceedings of the Royal Society of London: 121(788):654-673.

Hu, FC. 1993. A statistical methodology for analyzing the causal health effect of a time dependent exposure from longitudinal data. Harvard School of Public Health: ScD dissertation.

Lai, TL. and Small, D. 2007. Marginal regression analysis of longitudinal data with time-dependent covariates: A generalized method-of-moments approach. Journal of the Royal Statistical Society, Series B 69, no.1:79-99.

Lalonde, TL, Wilson, JR, and Yin, J. 2014. GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications, Statist. Med., 27, 4756–4769.

Liang, K-Y. and Zeger, SL. 1986. Longitudinal data analysis using generalized linear models. Biometrika. 73: 13-22.

Pepe, MS. and Anderson, GL. 1994. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics. 23: 939-951.

Zeger, SL. and Liang, KY. 1992. An Overview of methods for the analysis of longitudinal data. Statistics in Medicine 11, no.14-15:1825-1839

ACKNOWLEDGMENTS

This work is funded in part by the National Institutes of Health Alzheimer's Consortium Fellowship Grant, Grant No. NHS0007. The content in this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CONTACT INFORMATION

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

Kyle Irimata
Arizona State University
kirimata@asu.edu

Jeffrey Wilson
Arizona State University
jeffrey.wilson@asu.edu