Paper 10260-2016

SAS® Macro for Generalized Method of Moments Estimation for Longitudinal Data with Time-Dependent Covariates

Katherine Cai, Jeffrey Wilson, Arizona State University

ABSTRACT

Longitudinal data with time-dependent covariates is not readily analyzed as there are inherent, complex correlations due to the repeated measurements on the sampling unit and the feedback process between the covariates in one time period and the response in another. A generalized method of moments (GMM) logistic regression model (Lalonde, Wilson, and Yin 2014) is one method to analyze such correlated binary data. While GMM can account for the correlation due to both of these factors, it is imperative to identify the appropriate estimating equations in the model. Cai and Wilson (2015) developed a SAS macro using SAS/IML to fit GMM logistic regression models with extended classifications. In this paper we expand the use of this macro to allow for continuous responses and as many repeated time points and predictors as possible. We demonstrate the use of the macro through two examples, one with binary response and another with continuous response.

INTRODUCTION

In many research studies, it is common to observe subjects over a period of time and collect multiple measurements during the study period. Longitudinal studies are useful for understanding long-term effects and reveal information about time-variant factors. However, these studies can be challenging to evaluate due to the time-dependent covariates and the underlying intra-subject correlation.

While understanding how factors vary over time can be important to the study, time-dependent covariates can present many challenges in the data analysis. Some predictors can change over time due to feedback from the response, and similarly the response may change due to feedback from the predictor. For example, a study on heart health may examine the amount of exercise the subject participates in. While this is an appropriate covariate for studying heart health, subjects with poor heart health are less likely to exercise which could in turn worsen their health. Wilson and Lorenz (2015) present a graphical representation of this feedback process.

In addition, longitudinal data often have intra-class correlation present. Since measurements collected on the same subject are more likely to be similar compared to measurements collected from other subjects, there is an inherent correlation between the data points. Correlation in the data needs to be accounted for with an appropriate modeling technique, or the statistical tests can produce invalid results.

Statistical modeling methods including generalized estimating equations (GEE) and generalized linear mixed models (GLMM) are commonly implemented to assess longitudinal data. GEE has a flexible correlation structure that allows the user to estimate the correlation within a subject at various time points. In a GLMM, the random effect can be used to estimate variation between subjects. SAS offers the procedures PROC GENMOD and PROC GLIMMIX to perform GEE and GLMM, respectively, as well as PROC MIXED for repeated continuous data. While both methods are able to assess longitudinal data, Lalonde, Wilson, and Yin (2014) proposed a GMM logistic regression model that can be used to analyze correlated binary data and take into account different types of time-dependent covariates. Lai and Small (2007) first developed an approach but grouped the valid moments. Cai and Wilson (2015) developed a SAS macro to fit GMM logistic regression models. We present updates to the %GMM macro which extend the capability of the macro to evaluate cases with many predictors and measurements at more than three time points. In addition, the GMM macro can now be used for continuous and binary data by performing linear and logistic regression.

REGRESSION MODEL FOR LONGITUDINAL DATA WITH TIME-DEPENDENT COVARIATES

While generalized method of moments is a common technique for obtaining model estimates (Hansen 1982), Lai and Small (2007) first demonstrated the use of GMM to obtain parameter estimates for data with time-dependent covariates. Zhou, Lefante, Rice, and Chen (2014) implemented a similar approach but made use of certain matrices to identify the different types of groupings of the moments for each covariate. Covariates were classified into one of three types of time-dependent covariates, types I, II and III. A type I covariate is a covariate for which all measurements collected over the time period are independent, or do not have a time dependence. Type I covariates are not dependent on prior measurements of the outcome. The type II covariate depends on previous values of the covariate, such as seen with autoregressive models. A typical example would be evaluating stock prices, where predictions of price depend on previous conditions in the company. Type III covariates depend on previous values of the outcome, such as in the heart health example mentioned above. Lalonde, Wilson, and Yin (2014) extended the classifications of this method to include a type IV covariate. For type IV covariates, future responses are not affected by the previous covariate process. This category takes into account covariates that may have immediate but not long term feedback.

In the GMM approach proposed by Lalonde, Wilson, and Yin (2014), the valid moment conditions are identified and then the GMM estimates are obtained from the valid moment conditions. Moment conditions are considered valid if no significant correlation exists between the residuals and the covariate at two different time points t and s. The method employs multivariate integration to perform multiple comparison tests to examine the correlation between the different residuals from a model based on responses from a particular time and covariates from a different time. Correlations identified not to be significantly different from zero signify valid moment conditions. The valid moment conditions are used to produce the GMM estimates by minimizing a quadratic objective function with a weight matrix based on the covariance.

THE GMM REGRESSION MACRO

Cai and Wilson (2015) presented a SAS macro using SAS/IML to perform GMM logistic regression. That macro facilitated the analysis of longitudinal data measured at three time points with binary responses to implement the GMM logistic regression (Lalonde, Wilson, and Yin 2014). The macro output includes the correlation tests between the residuals and the covariate(s) and identifies the valid moment conditions. The parameter estimates, standard errors, test statistics, and p-values are also produced. Cai and Wilson provided a comparison of the GMM method to GEE and GLMM for a longitudinal morbidity study in the Philippines with time-dependent covariates. While the macro provided a GMM method to appropriately account for time-dependent covariates, it was limited to analyzing only three time points and did not allow the user to specify non-binary responses.

The SAS macro can now fit GMM linear and GMM logistic regression with time-dependent covariates. GMM linear regression is appropriate when the outcome variable is continuous, and follows a normal distribution, while GMM logistic regression should be implemented when the outcome is binary. The %GMM macro option DISTR= specifies the regression type, with the options “normal” for linear regression and “bin” for logistic regression, specified without quotation marks. The two GMM regression models for longitudinal data with time-dependent covariates perform similarly. The difference is in the weights and the link function. For continuous data, the weights are $\frac{1}{\mu}$ where $\mu = X\beta$ is the mean. For binary data, the data are weighted as $\mu(1-\mu)$ with mean $\mu = \frac{exp(X\beta)}{1+exp(X\beta)}$.

The %GMM macro has also been expanded to analyze data with longitudinal data collected at more than 3 time points. While missing data are often a limitation in longitudinal studies, the macro requires that complete records are provided for every subject at all time points. If a user would like to analyze data that have missing values, the subset of the subjects with complete records can be evaluated or missing data imputation methods can be used to prepare the data before calling the %GMM macro.

The %GMM macro requires the macro %MVINTEGRATION to be initialized. %MVINTEGRATION was presented in Cai and Wilson (2015), and is used to calculate multivariate normal probabilities. This macro

was adapted from a SAS/IML program (Alan Genz and Frank Bretz, Genz 1992, Genz 1993). In order to call the %MVINTEGRATION macro, the user needs to specify a file path (REFLIB) to store a SAS Catalog of IML modules. After the macro has been run once, the catalog will be permanently stored in the file location until the user removes the file. The %MVINTEGRATION macro can be called, using the command:

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

After the SAS Catalog of IML modules is stored in the specified file location, the %GMM macro to perform GMM regression with a continuous outcome can be called by the code:

```
%GMM(ds="C:\Users\Documents\Data",
          file=Data,
          reflib="C:\Users\Documents\Code",
          timeVar=time,
          outVar=y,
          predVar=x1 x2 x3,
          idVar=name,
          alpha=0.05,
          distr=normal);
```

The option DS= is used to specify the file location of the stored SAS data set. The file name of the SAS data set is specified in FILE= option, without a file extension. The variables in the dataset to be analyzed are specified with the options ending in "VAR." TIMEVAR is the name of the time variable in the file which records the visit number or time period that the measurement was taken in. OUTVAR is the outcome variable of interest, and PREDVAR is any specified number of covariates, separated by a space, to be analyzed as predictors of the outcome. IDVAR is a numeric or character identification variable that is used to group observations from the same subject, such as a name or ID number. ALPHA= is the significance level for the test of correlation between a covariate and the residual. The option DISTR=normal specifies that linear regression should be performed.

The %GMM macro can be used to perform GMM logistic regression using the call:

```
%GMM(ds="C:\Users\Documents\Data",
          file=Data,
          reflib="C:\Users\Documents\Code",
          timeVar=time,
          outVar=y,
          predVar=x1 x2 x3,
          idVar=name,
          alpha=0.05,
          distr=bin);
```

The macro call is similar to the former version, but the option DISTR=bin specifies that logistic regression should be used since the outcome OUTVAR is a binary variable.

The macro output includes the results of a GEE analysis using PROC GENMOD, and output for correlation tests, and GMM parameter estimates. The matrix R4OUT displays the correlation between the residuals and the covariate at each of the time points. P4OUT displays the p-values for the correlation test for the moment conditions. By default, the intercept and time indicator variables are treated as Type I covariates and are not included in the correlation test. The TYPE4OUT matrix contains binary indicators to specify valid moment conditions, where 1 represents a valid moment condition and 0 is an invalid moment condition. Moment conditions are considered invalid if the correlation between the residual and the covariate is not significant. The number of rows in the TYPE4OUT matrix is equal to the number of time points, and the number of columns is equal to $t(k+1+(t-1))$ where t is the number of time points and k is the number of predictors specified in PREDVAR.

The final %GMM output are the matrices OUTMTX and BETAVEC. OUTMTX contains the estimates and significance test information for each of the variables. The columns in OUTMTX contain the estimate (Estimate), standard deviation (StdDev), test statistic (Zvalue) and the p-value (Pvalue). There are k+1+(t-1) rows in OUTMTX, where the first row is the intercept term, the middle k terms are the predictors specified in PREDVAR and the last (t-1) rows represent the time indicator variables from $t_2, \ldots, t_t$. The variable names are included on the row in OUTMTX. Specific examples including the complete %GMM macro output are included in Cai and Wilson (2015).

OBESITY DATA: LINEAR REGRESSION

GMM linear regression is demonstrated with obesity data collected through the Add Health study, which is a social and behavioral longitudinal study conducted for adolescents in grades 7-12 in the United States. Data were collected on the adolescents over four waves, which occurred over a period of 15 years beginning in 1994.

We evaluated a subset of the data, for children that were not missing any of the indicators of interest and for subjects that participated in all four waves (Rhodes, Fang, and Wilson 2016). Our data contain information for 2,777 adolescents over four time points, for a total of 11,108 records. The children ID's (ID), the wave number (wave), height in inches (height), weight in pounds (weight), age (age), average number of hours spent watching television each week (tvhrs), and smoking status at the time of the measurement (smoking). For the variable smoking, smoking=1 represents that the adolescent smokes, and smoking=0 represents that the adolescents does not smoke.

The SAS data set was created as follows:

```
data Obesity;
    input ID wave height weight age tvhrs smoking;
    datalines;
    57101310 1 72 152 19 33 1
    57101310 2 71 168 19 50 1
    57101310 3 71 194 25 94 1
    57101310 4 70 251 32 35 1
    57109625 1 62 115 14 14 1
    57109625 2 63 120 15 15 0
    57109625 3 64 135 20 21 1
    57109625 4 63 149 27 15 1
      …
    99719976 1 65 162 15 20 1
    99719976 2 66 175 15 30 1
    99719976 3 63 141 20 25 1
    99719976 4 64 166 28 8  1
    ;
run;
```

The complete dataset is available online (www.public.asu.edu/~jeffreyw/). In order to study obesity, the body mass index was calculated for each adolescent during each wave. The variable BMI was created from the variables height and weight as:

```
data Obesity;
    set Obesity;
    BMI = (weight*.45)/((height*0.025)**2);
run;
```

ANALYSIS USING GEE

The obesity data can be analyzed using generalized estimating equations with an unstructured correlation matrix. GEE can be performed in SAS using PROC GENMOD, as shown below:

```
proc genmod data=Obesity descending;
class ID wave(ref="1") smoking;
model bmi = age tvhrs smoking wave;
repeated subject = ID / within=wave corr=un corrw;
run;
```

The CLASS statement is used to specify categorical variables, and the REF statement on the wave variable changes the reference time to the first measurement instead of the fourth measurement which is automatically selected by SAS. The MODEL statement is used to specify the outcome and covariates. The repeated measurement structure is within the time (WAVE) by subject (identified by ID). We select an unstructured (CORR=UN) correlation matrix to allow the correlations between each of the four time points to vary. The GEE model results from SAS are displayed in Output 1.

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		17.7317	2.8368	12.1717	23.2917	6.25	<.0001
age		0.3166	0.1817	-0.0395	0.6727	1.74	0.0814
tvhrs		0.0022	0.0056	-0.0089	0.0132	0.38	0.7008
smoking	0	-0.0213	0.2019	-0.4170	0.3743	-0.11	0.9159
smoking	1	0.0000	0.0000	0.0000	0.0000	.	.
Wave	2	0.4800	0.0825	0.3183	0.6416	5.82	<.0001
Wave	3	2.4877	1.0628	0.4047	4.5706	2.34	0.0192
Wave	4	2.5602	2.3631	-2.0713	7.1917	1.08	0.2786
Wave	1	0.0000	0.0000	0.0000	0.0000	.	.

Output 1. PROC GENMOD Output for Obesity Data

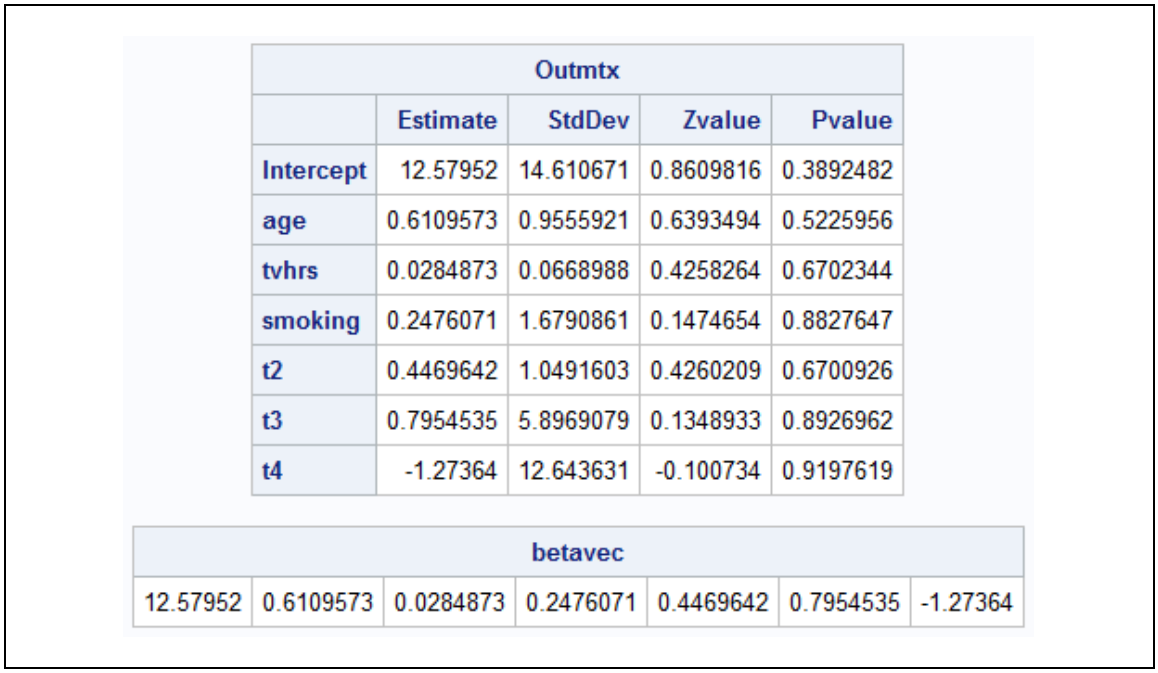
At the 5 percent significance level, age (age), the number of hours of TV watched (tvhrs), and smoking status are not significant. Although age is not a significant predictor at the 5 percent significance level, the p-value does indicate that there is a trace of significance.

ANALYSIS USING GMM

To use generalized method of moments instead of generalized estimating equations, the user can use %GMM. Example code to call the macro for the obesity data set is shown below:

```
%GMM(ds="C:\Users\Katherine\Data",
          file=Obesity,
          reflib="C:\Users\Katherine\Code",
          timeVar=wave,
          outVar=bmi,
          predVar=age tvhrs smoking,
          idVar=ID,
          alpha=0.05,
          distr=normal);
```

This code similarly evaluates BMI using age, the number of hours of TV watched, and smoking status with repeated measurements by wave. Since BMI is a continuous outcome, we specify the option DISTR=NORMAL. The parameter estimates for the GMM technique are displayed in the SAS output in Output 2.



Outmtx

	Estimate	StdDev	Zvalue	Pvalue
Intercept	12.57952	14.610671	0.8609816	0.3892482
age	0.6109573	0.9555921	0.6393494	0.5225956
tvhrs	0.0284873	0.0668988	0.4258264	0.6702344
smoking	0.2476071	1.6790861	0.1474654	0.8827647
t2	0.4469642	1.0491603	0.4260209	0.6700926
t3	0.7954535	5.8969079	0.1348933	0.8926962
t4	-1.27364	12.643631	-0.100734	0.9197619

betavec

12.57952	0.6109573	0.0284873	0.2476071	0.4469642	0.7954535	-1.27364

Output 2. %GMM Output for Obesity Data

Using generalized method of moments and accounting for the time-dependency of the covariates, we found that none of the variables are significant predictors of BMI. While these parameter estimates are similar to the parameter estimates using GEE, we found that none of the predictors are good predictors of BMI. The primary difference is the change of the effect of age from appearing to be moderately significant using GEE to clearly not significant using GMM.

MORBIDITIY DATA: LOGISTIC REGRESSION

To illustrate the use of the %GMM macro with logistic regression for a binary response, we will investigate children's health data collected by the International Food Policy Research Institute in the Bukidnon Province in the Philippines. In the study, the age, gender, BMI, and morbidity status were collected for 370 children. Data were collected over three separate time periods, separated by 4-month intervals.

Each of the 370 children had information recorded for all three time points, for a total of 1,100 observations. The outcome, morbidity status, was recorded as a binary variable which indicates whether the child was sick (sick = 1) or healthy (sick = 0) at the time of measurement. Other factors collected included the age in months (age), gender (gender), and body mass index (BMI). Children were assigned an identification number (childID), and the visit number for each measurement was recorded (time).

The SAS data set was created using the following code:

```
data Morbidity;
   input childID BMI time sick gender age;
   datalines;
   206   14.95059    1     0     0     59.27
   206   15.01923    2     0     0     63.40
   206   14.79053    3     0     0     66.83
   407   17.02125    1     1     0     17.5
```

```
    407   16.0064     2     1     0     21.67
    407   18.08021    3     0     0     25.07
    705   15.83377    1     0     1     62.6
    705   15.39259    2     1     1     66.77
    705   15.08541    3     0     1     70.17
    …
    50805 14.78102    1     1     0     34.33
    50805 15.05843    2     0     0     38.93
    50805 14.91299    3     0     0     42.33
    ;
run;
```

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

ANALYSIS USING GEE

The Philippines morbidity dataset can be analyzed using GEE with an unstructured correlation matrix, as demonstrated with the SAS PROC GENMOD code shown below:

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

We are interested in determining the impact of age, gender, and BMI on the child's sickness status over the three time points. The GEE model results are displayed in Output 3.

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		1.3843	0.9629	-0.5029	3.2714	1.44	0.1505
age		-0.0216	0.0051	-0.0317	-0.0116	-4.21	<.0001
gender	0	-0.1362	0.1496	-0.4294	0.1569	-0.91	0.3624
gender	1	0.0000	0.0000	0.0000	0.0000	.	.
bmi		-0.0874	0.0565	-0.1982	0.0235	-1.55	0.1223
time	2	-0.2540	0.1583	-0.5642	0.0562	-1.61	0.1085
time	3	0.2698	0.1574	-0.0386	0.5783	1.71	0.0864
time	1	0.0000	0.0000	0.0000	0.0000	.	.

Output 3. PROC GENMOD Output for Morbidity Data

Using GEE revealed that age is a highly significant predictor of morbidity at the 5 percent significance level, while gender and BMI are not. The time points are not significant predictors of the child's sickness status.

ANALYSIS USING GMM

In order to analyze the data while appropriately accounting for time dependence, the user can use the GMM macro. The macro code to call %GMM for the morbidity data is shown below.

```
%GMM(ds="C:\Users\Katherine\GMM",
          file=Morbidity,
          reflib="C:\Users\Katherine\Code",
          timeVar=time,
          outVar=sick,
          predVar=age gender bmi,
          idVar=childid,
          alpha=0.05,
          distr=bin);
```

This code produces a comparable analysis to the GEE analysis shown above. The option DISTR=BIN signifies that the outcome variable (OUTVAR) sick is binary and implements GMM logistic regression.

Outmtx

	Estimate	StdDev	Zvalue	Pvalue
Intercept	-0.261352	0.1655737	-1.578466	0.1144587
age	-0.016081	0.0036902	-4.357574	0.0000132
gender	0.1183705	0.1111145	1.0653021	0.2867392
bmi	-0.007441	0.0038598	-1.927714	0.0538907
t2	-0.240173	0.0293117	-8.193766	2.22E-16
t3	0.187518	0.0351342	5.337191	9.4398E-8

betavec

-0.261352	-0.016081	0.1183705	-0.007441	-0.240173	0.187518

Output 4. %GMM Output for Morbidity Data

The GMM analysis produces slightly different results compared to GEE. Although age is still a highly significant predictor, BMI is moderately significant with a p-value very close the 5 percent threshold. The user should use caution in interpreting the effect of BMI. GMM is able to better account for the time dependent nature of the variable BMI. In Cai and Wilson (2015), we demonstrate the difference between the GEE and GMM results by analyzing the morbidity data with BMI as the only predictor of sickness status. In GEE, BMI was not a significant predictor of sickness while using GMM, BMI was a highly significant predictor of sickness.

CONCLUSION

GMM regression is a powerful technique that accounts for time-dependent covariates and intra-class correlation. We present updates to the %GMM macro (Cai and Wilson 2015) to improve the usability of the macro for various longitudinal studies by increasing the number of time points that can be evaluated and extending the macro to perform linear and logistic regression.

REFERENCES

Cai, K. and Wilson, JR. 2015. “How to Use SAS for GMM Logistic Regression Models for Longitudinal Data with Time-Dependent Covariates.” Proceedings of the SAS Global Forum 2015 Conference.

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.

Hansen L.P. 1982. “Large Sample Properties of Generalized Method of Moments Estimators.” Econometrica, 50:1029-1054.

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, TL and D. 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, K and S. Zeger. 1986. “Longitudinal Data Analysis Using Generalized Linear Models.” Biometrika, 73(1):13-22.

Pepe, MS and GL 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.

Rhodes R, Fang D, Wilson JR (2016) GMM Logistic Regression Model for Obesity with time-dependent Covariates- Working Paper: Department of Economics Arizona State University

Wilson, JR. and KA Lorenz. Modeling Binary Correlated Responses using SAS, SPSS and R. New York: Springer International Publishing, 2015.

Zeger, SL. and KY Liang. 1986. “Longitudinal Data Analysis for Discrete and Continuous Outcomes.” Biometrics, 42(1):121-130.

Zhou, Y., Lefante, J., Rice, J., and S. Chen. 2014. “Using modified approaches on marginal regression analysis of longitudinal data with time-dependent covariates.” Statistics in Medicine, 33(19):3354-3364.

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/