Paper 2661-2018

Using SAS® to Estimate Lagged Coefficients with the %partitionedGMM Macro

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

ABSTRACT

Longitudinal data often includes time-dependent covariates, which must be accounted for with an appropriate model. Although a number of models have been proposed to analyze time-dependent covariates, most approaches constrain the effect of each covariate on the outcome to be constant across time. Irimata, Broatch, and Wilson (2018) introduced a partitioned generalized method of moments (GMM) model that used only valid moment conditions to estimate the differing relationships within longitudinal data. This model provides insight into potential lagged effects of a given covariate on the response in a later time period. Each regression coefficient is estimated using moment conditions corresponding to the respective time period. Irimata and Wilson (2017) presented a SAS® macro for fitting this partitioned GMM model for binary outcomes using SAS/IML® software. We extended the %partitionedGMM macro to allow for either continuous or binary outcomes. In this paper, we also expanded this macro to fit time-independent covariates. The performance and use of this macro are demonstrated through the analysis of two examples—one with a continuous outcome and one with a binary outcome.

INTRODUCTION

Longitudinal data are commonly encountered in many disciplines, such as in health research. These data involve repeated observations collected on the same set of subjects or units over time. For instance, in the National Longitudinal Study of Adolescent to Adult Health (Add Health), data are collected on a set of adolescents across the United States across four waves (Harris & Udry, 2016; NIPORT, 2011). In this study, outcomes such as obesity status are expected to vary across the collection waves. Due to the repeated measurements, the effect of time-dependent covariates, such as physical activity level or depression level, on the outcome may also be expected to vary as time progresses. For instance, a child's low physical activity level at the first collection wave will have an effect on his or her obesity status at the first wave, and may also have an effect on his or her obesity at future time-periods.

The carry-over effects of time-dependent covariates have often been addressed using lagged models. These approaches incorporate previous and current observations of the predictor variables in the model to account for autocorrelation (Keele & Kelly, 2005). In the case of longitudinal data, the generalized estimating equations (GEE) model developed by Zeger and Liang (1986) has been applied as a lagged model with an independent working correlation structure (Schildcrout & Heagerty, 2005). One drawback to this approach is that the independent working correlation structure may utilize moment conditions which are not valid into the estimation.

More recently, generalized method of moments (GMM) models have been introduced to account for the issues posed by time-dependent covariates. Lai and Small (2007) proposed a GMM model which relied on a covariate classification scheme to utilize valid moment conditions in obtaining parameter estimates. This approach ensured that all available information was employed, while invalid information was excluded in the analysis. Specifically, they developed the Type I, Type II and Type III covariates, which each incorporate a varying number of moment conditions. Lalonde, Wilson and Yin (2014) presented a GMM model which utilized a hypothesis test to separately evaluate each moment condition for validity. They utilized bivariate correlations to evaluate each moment condition, and were thus able to integrate all available valid information in developing regression models. However, both these approaches grouped all valid moment conditions for a given predictor, and thus estimated only a single regression parameter to characterize an overall effect of each time-dependent covariate. In practice, this overall estimate may not provide sufficient insight into the complex relationships inherent to longitudinal studies.

Irimata, Broatch and Wilson (2018) improved upon the GMM and lagged modeling frameworks and proposed the Partitioned GMM model to incorporate lagged coefficients, while also utilizing only valid moment conditions to ensure efficiency. They utilized either the Type II covariate presented by Lai and Small (2007), or the individual tests for validity presented by Lalonde, Wilson and Yin (2014) to select the moment conditions used in fitting the regression model. However, rather than using an overall estimate of the effect of each time-dependent covariate, they partitioned the moment conditions based on the relationship they represented. Moment conditions corresponding to observations of the covariate and response in the same time-period were used to estimate a cross-sectional regression parameter. Moment conditions obtained from observations on the covariate at one time-period prior to the outcome were used to estimate a one time-period lagged regression parameter, and similarly for larger time-period lags.

In this paper, we present the %partitionedGMM macro in SAS to fit the Partitioned GMM model. We provide discussions on the capabilities and functionalities of this macro and consider applications to both the Add Health dataset, and to a clinical study of depression scores.

%PARTITIONEDGMM SAS MACRO

The Partitioned GMM model can be fitted in SAS using the %partitionedGMM macro, which utilizes the general call:

```
%partitionedGMM(ds=.,
    file=,
    timeVar=,
    outVar=,
    predVarTD=,
    idVar=,
    alpha=0.05,
    predVarTI=.,
    distr=bin,
    optim=NLPCG,
    MC=LWY);
```

The first argument *ds* is optional and specifies the location of the dataset to be analyzed. The *file* argument can be used to reference either the SAS file (.sas7bdat) in conjunction with the *ds* statement, or in lieu of this argument references the SAS dataset in the current SAS session. The *timeVar* argument specifies which variable in the dataset denotes the time-periods. The *outVar* argument specifies the name of the binary or continuous outcome variable in the dataset. The argument *predVarTD* specifies the time-dependent covariate or covariates to be used in the model, where multiple time-dependent covariates should be separated by a space. The *idVar* argument specifies the subject identification variable. The *alpha* argument denotes the significance level at which the hypothesis test for validity will be evaluated, if using the Lalonde, Wilson and Yin method. By default, this argument is set to 0.05. The optional argument *predVarTI* specifies the time-independent covariate or covariates to be included in the analysis, where, similar to *predVarTD*, multiple covariates should be separated by a space. The *distr* argument specifies the distribution of the outcome variable and takes the value of *bin* by default, denoting a binary outcome, while changing this argument to *normal* will specify a continuous outcome. The *optim* argument is used to select which nonlinear optimization algorithm will be used, with the default being *NLPCG* for the conjugate gradient method. The SAS manual provides a discussion on alternative nonlinear optimization algorithms in IML, each of which are incorporated in the %partitionedGMM macro using the same naming system. Finally, the *mc* argument is used to select whether the Lalonde, Wilson and Yin approach to identifying valid moment conditions (*LWY*), or the Lai and Small (*LS*) approach with a Type II covariate should be used. The Lalonde, Wilson and Yin approach is utilized by default.

The %partitionedGMM macro provides parameter estimates, standard deviations, Z-values and p-values for the model. Estimates for the intercept, and any time-independent covariates are provided at the top of each output table. The parameter estimates are then grouped based on the relationships being estimated. Cross-sectional parameter estimates, corresponding to the case when the covariate and the response are observed in the same time-period are labeled as VariableName_0, denoting a zero time-period difference between the covariate and the outcome. One time-period lagged parameters are given by VariableName_1, providing insight into the effect of the covariate onto the outcome across a one time-period lag. Parameters pertaining to larger time-period lags are named analogously.

For illustration, sample output for this macro is given in Figure 1, with one time-independent covariate for *race*, and two time-dependent covariates *feelingscale* and *activityscale*. This model was fit using:

```
%partitionedGMM(ds='C:\Users',
      file=Addhealth,
      timeVar=Wave,
      outVar=BMI,
      predVarTD=feelingscale activityscale,
      idVar=ID,
      alpha=0.05,
      predVarTI=race,
      distr=bin,
      optim=NLPCG,
      MC=LS);
```

In this case, the intercept and the time-independent covariate *race* are provided first in the output table. The cross-sectional parameter estimates for the two time-dependent covariates are denoted by *feelingscale_0* and *activityscale_0*. The table also provides results for each of the time-dependent covariates up to a three time-period lag, which is the maximum possible with a dataset composed of four time-periods.

Analysis of Partitioned GMM Estimates				
	Estimate	StdDev	Zvalue	Pvalue
Intercept	-2.873152	0.1397386	-20.5609	0
RACE_	0.2452929	0.0888125	2.761919	0.0057463
FEELINGSCALE_0	0.4045083	0.09575	4.2246296	0.0000239
ACTIVITYSCALE_0	-0.081868	0.0313539	-2.611092	0.0090253
FEELINGSCALE_1	0.4528897	0.0740415	6.1166974	9.553E-10
ACTIVITYSCALE_1	-0.048962	0.0217798	-2.248041	0.0245736
FEELINGSCALE_2	0.8141972	0.0753706	10.802584	0
ACTIVITYSCALE_2	0.0767468	0.0230516	3.3293435	0.0008705
FEELINGSCALE_3	0.5488687	0.089897	6.105528	1.0246E-9
ACTIVITYSCALE_3	0.0265718	0.0269722	0.9851561	0.3245474

Figure 1. Sample Output for the %partitionedGMM Macro

DATA EXAMPLE

We analyzed two data examples to illustrate the use of the %partitionedGMM macro. The first example focused on a binary outcome representing obesity status in children, using data from Add Health (Harris & Udry, 2016). The second example utilized data from a clinical study of depressed inpatients (Reisby et al., 1977). We fitted the Partitioned GMM model to each data first using the Lai and Small (LS) Type II covariate, and secondly using the Lalonde, Wilson and Yin (LWY) method for individual moment testing. For comparison, we also provide the results of a lagged GEE model with an independent working correlation structure obtained using the GENMOD procedure in SAS.

ADD HEALTH

The Add Health data were collected beginning in 1994-1995, and followed a cohort of students. This group of students were subsequently interviewed at three later waves, producing a total of four measurements. This data contained information on 2,712 students at all four time-periods. The outcome of interest for this analysis was a binary indicator representing obesity status, based on the child's BMI at each wave. The data included four time-dependent covariates representing depression level (feelingscale), hours spent watching television (TVHRS), physical activity level (activityscale), as well as an indicator variable representing whether the child was a social alcohol drinker (alcohol). In addition, there was information on one time-independent covariate denoting whether the child was white or nonwhite (race). The data were first analyzed using the %partitionedGMM macro with the Lai and Small Type II covariate. This model was fit using:

```
%partitionedGMM(ds='C:\Users',
     file=Addhealth,
     timeVar=Wave,
     outVar=BMI,
     predVarTD=feelingscale tvhrs activityscale alcohol,
     idVar=ID,
     alpha=0.05,
     predVarTI=race,
     distr=bin,
     optim=NLPCG,
     MC=LS);
```

The output from this macro call is provided in Figure 2.

Analysis of Partitioned GMM Estimates				
	Estimate	StdDev	Zvalue	Pvalue
Intercept	-3.078897	0.1469666	-20.94964	0
RACE_	0.0733223	0.093983	0.780166	0.4352932
FEELINGSCALE_0	0.3851641	0.0969778	3.9716725	0.0000714
TVHRS_0	0.0145563	0.0019572	7.4373134	1.028E-13
ACTIVITYSCALE_0	-0.058127	0.0307419	-1.890809	0.0586498
ALCOHOL_0	-0.059623	0.0739677	-0.806067	0.4202044
FEELINGSCALE_1	0.3157231	0.0812432	3.8861475	0.0001018
TVHRS_1	0.0020571	0.0016679	1.233384	0.2174325
ACTIVITYSCALE_1	-0.02769	0.0215086	-1.287411	0.1979512
ALCOHOL_1	0.0785131	0.059543	1.3185941	0.1873049
FEELINGSCALE_2	0.6613397	0.086432	7.6515645	1.976E-14
TVHRS_2	0.0129445	0.0020089	6.4436617	1.166E-10
ACTIVITYSCALE_2	0.0693825	0.0229166	3.0276052	0.002465
ALCOHOL_2	0.0675177	0.0629146	1.0731645	0.2831973
FEELINGSCALE_3	0.4171074	0.1029439	4.0517948	0.0000508
TVHRS_3	0.0121165	0.0026031	4.6547024	3.2445E-6
ACTIVITYSCALE_3	0.0188998	0.0279744	0.6756094	0.4992887
ALCOHOL_3	0.0176148	0.0765853	0.2300016	0.8180905

Figure 2. Partitioned GMM (LS) Output for the Add Health Analysis

The data were also analyzed using the %partitionedGMM macro with the Lalonde, Wilson and Yin approach with the macro call:

```
%partitionedGMM(ds='C:\Users',
     file=Addhealth,
     timeVar=Wave,
     outVar=BMI,
     predVarTD=feelingscale tvhrs activityscale alcohol,
     idVar=ID,
     alpha=0.05,
     predVarTI=race,
     distr=bin,
     optim=NLPCG,
     MC=LWY);
```

The results of this macro call are given in Figure 3.

Analysis of Partitioned GMM Estimates				
	Estimate	StdDev	Zvalue	Pvalue
Intercept	-3.026927	0.1512413	-20.01389	0
RACE_	0.2222988	0.0956948	2.3229967	0.0201793
FEELINGSCALE_0	0.5021873	0.103889	4.8338821	1.339E-6
TVHRS_0	0.0153684	0.0019934	7.7096675	1.266E-14
ACTIVITYSCALE_0	-0.1648	0.0328923	-5.010291	5.4348E-7
ALCOHOL_0	0.0104057	0.0782997	0.1328958	0.8942758
FEELINGSCALE_1	0.5815377	0.1046988	5.554389	2.7858E-8
TVHRS_1	0.0035101	0.0020637	1.7008165	0.0889774
ACTIVITYSCALE_1	-0.094535	0.0249006	-3.796499	0.0001468
ALCOHOL_1	0.0465212	0.0648791	0.7170454	0.4733461
ACTIVITYSCALE_2	0.1803	0.0208798	8.6351564	0
ALCOHOL_2	0.2953309	0.0678876	4.3502934	0.0000136
ACTIVITYSCALE_3	0.1585346	0.0213563	7.4233068	1.141E-13

Figure 3. Partitioned GMM (LWY) Output for the Add Health Analysis

For comparison, we fitted the lagged GEE model with an independent working correlation structure to the Add Health data and contrasted these results to the Partitioned GMM models. We saw that the results of the two Partitioned GMM approaches varied slightly. This discrepancy is due to the different moment conditions utilized in estimating each model. Notably, the Partitioned-LWY model did not produce parameter estimates for some parameters, such as depression level at a two time-period lag. Because each moment condition is separately tested for validity, in some cases certain relationships are not estimable. We also saw that the lagged GEE model varied as compared to either Partitioned GMM model. These differences were as expected, as the GEE approach does not incorporate any type of restriction regarding the validity of moment conditions utilized in obtaining the regression model. A comparison of these three approaches are provided in Table 1.

		Partitioned-LS		Partitioned-LWY		Lagged-GEE	
		Est.	**p-Value**	**Est.**	**p-Value**	**Est.**	**p-Value**
	Intercept	-3.076	**<.001**	-3.025	**<.001**	-2.526	**<.001**
	Race	0.074	0.433	0.222	**0.020**	0.067	0.456
Cross-sectional	**Depression**	0.384	**<.001**	0.501	**<.001**	0.137	0.166
	TV Hrs	0.015	**<.001**	0.015	**<.001**	0.013	**<.001**
	Activity	-0.059	0.057	-0.165	**<.001**	-0.144	**<.001**
	Alcohol	-0.060	0.414	0.010	0.895	-0.124	0.064
Lagged one period	**Depression**	0.315	**<.001**	0.582	**<.001**	0.290	**<.001**
	TV Hrs	0.002	0.216	0.004	0.089	0.004	**0.046**
	Activity	-0.028	0.197	-0.095	**<.001**	-0.021	0.350
	Alcohol	0.078	0.189	0.046	0.476	0.025	0.670
Lagged two periods	**Depression**	0.661	**<.001**	-	-	0.692	**<.001**
	TV Hrs	0.013	**<.001**	-	-	0.010	**<.001**
	Activity	0.069	**0.002**	0.180	**<.001**	0.075	**0.001**
	Alcohol	0.068	0.283	0.295	**<.001**	0.008	0.893
Lagged three periods	**Depression**	0.417	**<.001**	-	-	0.432	**<.001**
	TV Hrs	0.012	**<.001**	-	-	0.012	**<.001**
	Activity	0.019	0.493	0.158	**<.001**	-0.009	0.766
	Alcohol	0.017	0.822	-	-	0.057	0.466

Table 1. Comparison of the Partitioned GMM models and Lagged GEE for the Add Health Study

DEPRESSION STUDY

The depression data were first examined by Reisby, et al (1977), who focused on the relationships between Imipramine (IMI) and Desipramine (DMI) levels in depressed patients. The data contained information on 52 unique inpatients, with measurements taken at each of four weeks. The outcome of interest was continuous and denoted changes in Hamilton Depression Scores. Both IMI and DMI were log transformed, and were treated as time-dependent. We also included a time-independent variable denoting each patient's gender. The data were analyzed using the %partitionedGMM macro with the Lai and Small Type II covariate using:

```
%partitionedGMM(ds='C:\Users',
     file=Reisby,
     timeVar=time,
     outVar=hamdelt,
     predVarTD=lnimi lndm,
     idVar=id,
     alpha=0.05,
     distr=normal,
     predVarTI=sex,
     optim=NLPCG,
     MC=LS);
```

The results of this macro call are provided in Figure 4.

Analysis of Partitioned GMM Estimates				
	Estimate	StdDev	Zvalue	Pvalue
Intercept	7.4540624	5.4893469	1.3579142	0.1744909
SEX	-3.040089	2.7976357	-1.086664	0.2771855
LNIMI_0	-0.659902	0.4074934	-1.619418	0.1053574
LNDM_0	-1.749253	1.2874029	-1.358746	0.1742271
LNIMI_1	-0.844463	0.2487781	-3.394441	0.0006877
LNDM_1	0.6264409	0.1468082	4.2670714	0.0000198
LNIMI_2	0.3525457	3.4115299	0.1033395	0.9176935
LNDM_2	-1.051899	2.7791249	-0.3785	0.7050592
LNIMI_3	0.2462623	5.9638145	0.0412928	0.9670625
LNDM_3	-0.450138	5.1025537	-0.088218	0.9297033

Figure 4. Partitioned GMM (LS) Output for the Depression Analysis

The data were also analyzed using the %partitionedGMM macro with the Lalonde, Wilson and Yin approach to evaluate the validity of moment conditions using:

```
%partitionedGMM(ds='C:\Users',
     file=Reisby,
     timeVar=time,
     outVar=hamdelt,
     predVarTD=lnimi lndm,
     idVar=id,
     alpha=0.05,
     distr=normal,
     predVarTI=sex,
     optim=NLPCG,
     MC=LWY);
```

For this analysis, all moment conditions were identified as valid. The results for this macro call are given in Figure 5.

Analysis of Partitioned GMM Estimates				
	Estimate	StdDev	Zvalue	Pvalue
Intercept	7.4540624	5.4893469	1.3579142	0.1744909
SEX	-3.040089	2.7976357	-1.086664	0.2771855
LNIMI_0	-0.659902	0.4074934	-1.619418	0.1053574
LNDM_0	-1.749253	1.2874029	-1.358746	0.1742271
LNIMI_1	-0.844463	0.2487781	-3.394441	0.0006877
LNDM_1	0.6264409	0.1468082	4.2670714	0.0000198
LNIMI_2	0.3525457	3.4115299	0.1033395	0.9176935
LNDM_2	-1.051899	2.7791249	-0.3785	0.7050592
LNIMI_3	0.2462623	5.9638145	0.0412928	0.9670625
LNDM_3	-0.450138	5.1025537	-0.088218	0.9297033

Figure 5. Partitioned GMM (LWY) Output for the Depression Analysis

For comparison, we fitted the lagged GEE model with an independent working correlation matrix. Since all moment conditions were identified as valid, the Partitioned-LS and Partitioned-LWY models both utilized the same sets of moment conditions. Thus, the results of these two Partitioned GMM approaches were identical. The lagged GEE model produced results that were different from those of the Partitioned GMM model. This discrepancy again illustrates the effect of utilizing invalid moment conditions in fitting the regression model. The results of these three models are provided in Table 2.

		Partitioned-LS		Partitioned-LWY		Lagged-GEE	
		Est.	**p-Value**	**Est.**	**p-Value**	**Est.**	**p-Value**
	Intercept	7.454	0.174	7.454	0.174	4.897	0.231
	Gender	-3.040	0.277	-3.040	0.277	-0.502	0.780
Cross-sectional	IMI	-0.660	0.105	-0.660	0.105	-0.935	0.359
	DMI	-1.749	0.174	-1.749	0.174	-1.337	0.070
Lagged one period	IMI	-0.844	**<.001**	-0.844	**<.001**	-0.727	0.248
	DMI	0.626	**<.001**	0.626	**<.001**	0.389	0.490
Lagged two periods	IMI	0.353	0.918	0.353	0.918	0.569	0.271
	DMI	-1.052	0.705	-1.052	0.705	-1.103	**0.020**
Lagged three periods	IMI	0.246	0.967	0.246	0.967	0.814	0.549
	DMI	-0.450	0.930	-0.450	0.930	-1.050	0.350

Table 2. Comparison of the Partitioned GMM models and Lagged GEE for the Depression Study

CONCLUSION

Longitudinal data provide additional insight into subjects observed at multiple time-periods. However, the repeated measurements over time also introduce additional challenges such as those posed by time-dependent covariates. Though methods have been introduced to account for this correlation, many of these approaches produce only one parameter to represent an overall effect of the covariate on the response over time. Other models, which introduce additional parameters to represent the effect of the covariate on the outcome over time rely on potentially invalid moment conditions and thus can produce inefficient or biased estimates.

We provide the %partitionedGMM macro in SAS to fit the Partitioned GMM model introduced by Irimata, Broatch and Wilson (2018). This macro extends previous implementations by providing support for the use of either the Lalonde, Wilson and Yin approach to testing moment conditions, or the Type II covariate proposed by Lai and Small. Further, this macro allows for the fitting of either binary or continuous outcomes, while also taking into account time-independent covariates.

REFERENCES

Harris, K. M., & Udry, J. R. (2016). *National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]*. Retrieved from: http://doi.org/10.3886/ICPSR21600.v17

Irimata, K. M., Broatch, J., & Wilson, J. R. (2018). Partitioned GMM Logistic Regression Models for Longitudinal Data. *Manuscript Submitted for Publication*.

Irimata, K. M., & Wilson, J. R. (2017). *GMM Logistic Regression with Time-Dependent Covariates and Feedback Processes in SAS.* Paper presented at the SAS Global Forum 2017, Orlando, FL.

Keele, L., & Kelly, N. J. (2005). Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables. *Political Analysis, 14*(2), 186-205. doi:10.1093/pan/mpj006

Lai, T. L., & Small, D. (2007). Marginal Regression Analysis of Longitudinal Data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 69*(1), 79-99.

Lalonde, T. L., Wilson, J. R., & Yin, J. (2014). GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications. *Stat Med, 33*(27), 4756-4769. doi:10.1002/sim.6273

NIPORT. (2011). *Bangladesh Demographic and Health Survey 2011*.

Reisby, N., Gram, L. F., Bech, P., Nagy, A., Petersen, G. O., Ortmann, J., . . . Christiansen, J. (1977). Imipramine: clinical effects and pharmacokinetic variability. *Psychopharmacology (Berl), 54*(3), 263-272.

Schildcrout, J. S., & Heagerty, P. J. (2005). Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. *Biostatistics, 6*(4), 633-652. doi:10.1093/biostatistics/kxi033

Zeger, S. L., & Liang, K. Y. (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:

Kyle M. Irimata
Arizona State University
Kyle.irimata@asu.edu