GMM Estimator Covariance Structure for Time Dependent Covariates with Unbalanced Replication

Trent L. Lalonde*

Abstract

It is common to have misbalance in longitudinal data sets, i. e. an unequal number of repeated observations for different subjects. For longitudinal data sets with time-dependent covariates, the Generalized Method of Moments (GMM) is the preferred method for estimating parameters in a way that accounts for the associations between responses and time-dependent covariates. However, for unbalanced longitudinal data, current GMM techniques suggest the estimator covariance structure and weight matrix should be estimated using only complete-case subjects. This paper presents a method of analyzing unbalanced longitudinal data such that all subjects are accounted for in the estimator covariance structure and weight matrix, thus reducing bias in both the parameter estimates and standard errors. This paper presents a data example with count response and programs intended for binary and for count response data.

Key Words: longitudinal; unbalanced; non-continuous; time-dependent covariates; generalized method of moments

1. Introduction

Unbalanced longitudinal data occur often in practice. For example, attrition of subjects in any study over time leads to varying numbers of responses per subject. Records of patient length of stay vary according to the number of times patients are admitted. Self-report longitudinal surveys can easily lead to misbalance in the number of observations per subject. It is important to have trustworthy analysis methods that takes this misbalance into account.

When longitudinal data include time-dependent covariates, the Generalized Method of Moments (GMM) is the preferred method of estimation (Lai and Small (2007), Lalonde et al. (2013)). This method allows the researcher to account for the autocorrelation inherent in longitudinal data as well as the correlation associated with relationships between responses and changing covariates. However, current GMM implementations for unbalanced data use only full-data subjects for the GMM weight matrix and the estimator covariance structure (Lai and Small (2007)). For data situations in which only a few subjects are observed at all times, this approach will not use the data from most of the subjects in finding parameter estimates and in calculating standard errors.

In this paper the existing GMM method for unbalanced longitudinal data is extended so that all subjects are used in both the weight matrix and standard error calculations. We show that when the data from incomplete subjects are ignored, both the point estimates of parameters of interest as well as the standard errors are biased. This can lead to faulty interpretations and inferences. The rest of this paper is organized as follows. Section 2 presents a description of the existing GMM estimation methods for balanced and unbalanced longitudinal data with time-dependent covariates. Section 3 presents a modification to the existing methods that allows for all subjects to be incorporated into the weight matrix and covariance estimates within GMM. A data example is presented and analyzed using both

*Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO 80639

existing and modified GMM methods in Section 4. Section 5 includes some discussion and concluding remarks.

2. The Generalized Method of Moments for Longitudinal Data with Time-Dependent Covariates

A major advantage of a longitudinal study is its capacity to separate change over time within subjects and differences among subjects (Diggle et al. (2002)). However, when dealing with longitudinal data not only do the response variables change over time, but the predictor variables or covariates can also change over time. The treatment of time-dependent covariates in the analysis of longitudinal data allows strong statistical inferences about dynamical relationships and provides more efficient estimators than can be obtained using cross-sectional data (Hedeker and Gibbons (2006)).

The Generalized Method of Moments (GMM) approach for longitudinal data with fitted time-dependent covariates provides more efficient estimates than using Generalized Estimating Equations (GEE) with the independent working correlation under certain conditions (Lai and Small (2007)). Lai and Small (2007) maintained that the GEE approach for time-independent covariates is an attractive option as it provides consistent estimates under all correlation structures for subjects' repeated measurements. Pepe and Anderson (1994) suggested the use of the independent working correlation when using GEE with time-dependent covariates as a safe choice of analysis. However, when there are time-dependent covariates, some of the estimating equations combined by GEE with an arbitrary working correlation structure may not be valid, meaning some estimating equations applied will not have zero expected value. More so, when there are time-dependent covariates, Hu (1993) and also Pepe and Anderson (1994) have pointed out that the consistency of GEE estimators is not assured with arbitrary working correlation structures unless a key assumption is satisfied, that a subject's repeated measurements are independent (meaning the independent working correlation structure is appropriate). The GMM process proposed by Lai and Small (2007) for longitudinal data with time-dependent covariates follows.

For each subject i of a longitudinal study, consider a response vector $\mathbf{Y}_i = [Y_{i1} \dots Y_{iT}]^T$, where T is the number of observations for subject i, let $\mathbf{Z}_i$ indicate the associated design matrix for the time-independent covariates, and let $\mathbf{X}_i$ indicate the associated design matrix for the time-dependent covariates for subject i. Given a vector of valid estimating expressions $\mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})$ for each subject, such that $\mathrm{E}[\mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0}$, the GMM process is as follows (Hansen (1982), Lai and Small (2007), Hansen (2007), Lalonde et al. (2013)). For an average vector of valid moment conditions,

$$\mathbf{G}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}),$$

and an optimal weight matrix,

$$\hat{\mathbf{W}} = \mathbf{V}^{-1} = \left[\mathrm{Cov}(\mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}_0))\right]^{-1},$$

where $\hat{\boldsymbol{\beta}}_0$ indicates initial estimates of the mean parameters, the GMM estimators $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the quadratic form

$$Q(\boldsymbol{\beta}) = \mathbf{G}(\boldsymbol{\beta})^T \hat{\mathbf{W}} \mathbf{G}(\boldsymbol{\beta}).$$

The number of components of the vector of valid moment conditions for each subject depends on both the "type" of time-dependent covariate and also on the number of replications

for that subject (Lai and Small (2007)). A time-dependent covariate $\mathbf{x}_j$ is said to be of Type I if

$$\mathrm{E}\left[\frac{\partial \mu_{is}}{\partial \beta_j}(Y_{it} - \mu_{it})\right] = 0$$

for $s = 1, \ldots, T$ and $t = 1, \ldots, T$. In this case all combinations of times for the derivative and the residual terms in the moment conditions have zero expected value. Type I time-dependent covariates tend to arise in situations in which the response components Y_{it} are independent, or if all subjects are exposed to a common random effect. A time-dependent covariate $\mathbf{x}_j$ is said to be of Type II if

$$\mathrm{E}\left[\frac{\partial \mu_{is}}{\partial \beta_j}(Y_{it} - \mu_{it})\right] = 0$$

for $s \geq t$ and $t = 1, \ldots, T$. In this case the moment conditions have zero expected value only when the time for the derivative term is at or beyond the time for the residual term. Type II time-dependent covariates tend to arise in time-series models in which covariate values x_{is} affect future response components Y_{it}, but response values do not affect future covariate values. A time-dependent covariate $\mathbf{x}_j$ is said to be of Type III if it is not Type I or Type II, or if

$$\mathrm{E}\left[\frac{\partial \mu_{is}}{\partial \beta_j}(Y_{it} - \mu_{it})\right] \neq 0$$

for some $s > t$. In this case the product is only assured to have zero expected value if both times are equivalent. Type III time-dependent covariates tend to arise in situations with feedback from both covariates and responses. A time-dependent covariate $\mathbf{x}_j$ if said to be of Type IV if

$$\mathrm{E}\left[\frac{\partial \mu_{is}}{\partial \beta_j}(Y_{it} - \mu_{it})\right] = 0$$

for $s \leq t$ and $t = 1, \ldots, T$. Here the moment conditions have zero expected value only when the time for the derivative term is less than or equal to the time for the residual term. Similar to Type II time-dependent covariates, Type IV time-dependent covariates tend to arise when responses affect future covariate values, but covariates do not affect future response values. Types of time-dependent covariates can be selected by the researcher, or the individual valid moment conditions can be selected using the data though the Extended Classification process (Yin et al. (2013)).

The "Two-Step GMM" (TSGMM) process starts with an initial estimate of the weight matrix $\hat{\mathbf{W}}$ using consistent estimates $\hat{\boldsymbol{\beta}}_0$, usually using Independent GEE. Then updated estimates $\hat{\boldsymbol{\beta}}$ are obtained by minimizing the quadratic form $Q(\boldsymbol{\beta})$ (Lai and Small (2007)). Hansen (1982) and also Hansen (2007) argued that GMM estimators are asymptotically normal and unbiased with variance-covariance structure

$$\Sigma = \left(\mathrm{E}\left[\frac{\partial \mathbf{g}_i}{\partial \boldsymbol{\beta}}\right]^T \mathbf{V}^{-1} \mathrm{E}\left[\frac{\partial \mathbf{g}_i}{\partial \boldsymbol{\beta}}\right]\right)^{-1},$$

which can be consistently estimated using

$$\hat{\Sigma} = \left(\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \hat{\mathbf{g}}_i}{\partial \boldsymbol{\beta}}\right)^T \hat{\mathbf{V}}^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \hat{\mathbf{g}}_i}{\partial \boldsymbol{\beta}}\right)\right)^{-1},$$

evaluating the derivatives at $\hat{\boldsymbol{\beta}}$, and using a consistent estimator $\hat{\mathbf{V}}^{-1}$ for $\mathbf{V}^{-1}$. Optimal efficiency is attained using the consistent estimator

$$\hat{\mathbf{V}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)\mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)^T.$$

For unbalanced replication in a longitudinal study, Lai and Small (2007) have recommended adjusting the vector of average valid moment conditions to include only contributing subjects. Thus each component of the vector of averaged valid moment conditions,

$$\left(G(\boldsymbol{\beta})\right)_j = \frac{1}{|I_j|}\sum_{i \in I_j} \mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}),$$

where I_j is the set of all subjects whose observation times allow a valid moment condition contribution to be calculated for the j^{th} component of $\mathbf{G}(\boldsymbol{\beta})$. Calculation of the weight matrix is made using only full-data subjects,

$$\hat{\mathbf{V}}_F = \frac{1}{N_F}\sum_{i \in I_F} \mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)\mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \hat{\boldsymbol{\beta}}_0)^T,$$

where I_F is the collection of subjects observed at all times, and N_F is the number of such subjects. However, if this weight matrix is used within the TSGMM process, both the quadratic form $Q(\boldsymbol{\beta})$ and the standard errors will be calculated using only full-data subjects. This ignores the contributions from incomplete-data subjects in all steps except for the calculation of average valid moment conditions, and may lead to bias in parameter estimates and standard errors.

3. A Modified GMM Method for Unbalanced Data

The Generalized Method of Moments (GMM) approach for parameter estimation within a Generalized Linear Model with time-dependent covariates requires minimization of the quadratic form, $Q(\boldsymbol{\beta}) = \mathbf{G}(\boldsymbol{\beta})^T\hat{\mathbf{W}}\mathbf{G}(\boldsymbol{\beta})$, where the vector $\mathbf{G}(\boldsymbol{\beta})$ represents an average of valid moment conditions across all subjects. In the balanced data case, this equates to the sum of all moment conditions for each component of the vector $\mathbf{G}(\boldsymbol{\beta})$, divided by the number of subjects N. In the unbalanced data case, subjects who are not observed at time t would not contribute to any moment conditions involving time t. An appropriate average would include only the subjects with valid moment conditions to contribute to each parameter estimate. Define I_j to be the set of all subjects whose observation times allow a valid moment condition contribution to be calculated for the j^{th} component of $\mathbf{G}(\boldsymbol{\beta})$. According to Lai and Small (2007), the j^{th} component of $\mathbf{G}(\boldsymbol{\beta})$ for unbalanced data can be written

$$\left(G(\boldsymbol{\beta})\right)_j = \frac{1}{|I_j|}\sum_{i \in I_j} \mathbf{g}_i(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\beta}).$$

However, instead of calculating the inverse weight matrix $\hat{\mathbf{V}}$ using only the subjects whose observations allow valid moment condition contributions for all parameters, construct $\hat{\mathbf{V}}$ to use information from each subject that contributes to each respective weight matrix component. Define I_{jk} to be the set of all subjects whose observation times allow a valid moment condition contribution to be calculated for both the j^{th} and k^{th} components of $\mathbf{G}(\boldsymbol{\beta})$. Then define each component of the inverse weight matrix $\hat{\mathbf{V}}$,

$$\left(\hat{V}\right)_{jk} = \sum_{i \in I_{jk}} \frac{(g_i)_j(g_i)_k}{|I_{jk}|}.$$

In this way each component of $\hat{\mathbf{V}}$ represents an average over all subjects contributing to both moment conditions involved in any weight matrix component. If subject i does not contribute to either one of the j^{th} or k^{th} components, then subject i does not contribute to the weight matrix estimate involving these two moment condition components. For data sets in which a relatively low number of subjects are observed for all times, this allows for more subjects to contribute to variance-covariance calculations than using only subjects who have been observed for all times.

4. Example: Transitional Housing Facility

As an example of the implementation of the adjusted GMM for unbalanced longitudinal data with time-dependent covariates, an analysis of client length of stay in a transitional housing facility (THF) is presented. The THF of interest is designed to provide temporary housing to families in need, and is located in the urban center of a small midwestern city with about 100,000 residents. The THF services only families that include children, and imposes a limit of 60 days for each family's stay. Monthly data have been recorded from January, 2006 through December, 2010, representing 415 families. Data for each family are recorded at the end of each month, including variables such as family size, education level, and employment status. In addition to housing, the THF staff provide services to leading family members to help their transition from living within the THF to maintaining a sustainable lifestyle outside of the facility. Two of these services are case management hours and job training. Case management hours represent the number of hours spent with a staff member of the THF, covering topics such as home food preparation, health care options, and basic financial strategies to help keep the family out of need. Job training programs provide advice and practice for common available local jobs, and is recorded in the data as a binary indicator of participation.

It is of interest to model the probability that a family will stay for longer than two weeks (14 days) in the facility in any given month. If THF management can predict that a family is likely to stay for less than two weeks, then monthly service for the space used can double in a given month. Management of the THF would like to know whether the number of children, employment status, education level, case management hours, and job training have an impact on the probability of staying for less than two weeks. In modeling probability of a stay of longer than two weeks, logistic regression will be the appropriate generalized linear model, with the probability of staying longer than two weeks treated as the mean response. Because data are recorded monthly, each family is observed multiple times and therefore a longitudinal binary logistic regression model is applied. Appropriate estimation methods will need to account for the autocorrelation in the multiple responses from the same families. Since the participation in case management and job training can fluctuate across months, these two variables are treated as time-dependent covariates. Thus the GMM will be applied to account for the potential correlation between responses and these two changing predictors. According to the definitions of Lai and Small (2007) and Yin et al. (2013), both case management hours and job training will be treated as "Type II" time-dependent covariates.

A family that stays for 60 days can have data for up to three different months. But not all families stay for an entire 60 days and not all families are observed across three different months, leading to an unbalanced data set. A small number of families have been allowed to stay in the THF for a second time, leading to a small number of families that have data for four, five, or six months. However, there is a very small number of families that have stayed in the THF during 6 different months, the maximum number of observations possible for

Table 1: Results of Existing GMM Analysis

	Existing GMM		
	Parameter Estimate	Standard Error	p-Value
Number of Children	0.1769	(0.000165)	< 0.0001
Unemployed	0.7936	(< 0.000001)	< 0.0001
Education Level	0.1110	(0.000059)	< 0.0001
Case Management	-1.0379	(0.000348)	< 0.0001
Job Training	-0.3864	(0.000007)	< 0.0001

Table 2: Results of Modified GMM Analysis

	Modified GMM		
	Parameter Estimate	Standard Error	p-Value
Number of Children	0.0644	(0.000550)	0.0030
Unemployed	0.5630	(0.000409)	< 0.0001
Education Level	-0.0537	(0.001895)	0.0109
Case Management	-0.9896	(0.000145)	< 0.0001
Job Training	-0.5918	(0.006368)	< 0.0001

a single family. In fact, fewer only 4 of the 415 observed families have data for a full 6 months. Therefore we will consider the misbalance to be severe for analysis of the THF. The following logistic regression model will have parameters estimated using both the existing GMM approach from Section 3 as well as the modified GMM of Section 4.

$$\text{logit}(\pi) = \beta_0 + \beta_1(\text{Children}) + \beta_2(\text{Unemployed}) + \beta_3(\text{Ed Level}) + \beta_4(\text{Case Hours}) + \beta_5(\text{Job Training}),$$

where π represents the probability of staying longer than two weeks. Table 1 presents the parameter estimates, standard errors, and p-values for the existing GMM method for unbalanced longitudinal data. Table 2 presents the same for the modified GMM method proposed in Section 4 of this paper.

Notice that the standard error estimates are very small using this estimation technique. Correspondingly, all of the p-values show high levels of significance, suggesting that all variables have a meaningful impact on the probability of a length of stay less than two weeks. But these standard errors are based *only* on those observations from the 4 families that were observed for a full 6 different months.

Using the modified GMM analysis, both standard errors and parameter estimates change. The standard errors are no longer noticeably small and the p-values no longer show the same extreme significance, as seen using the existing GMM method. In fact, the standard error for unemployment status has increased by more than three orders of magnitude, the standard error for education level has increased by two orders of magnitude, and the standard error for job training has increased by three orders of magnitude. Notice also that the standard error for case management hours decreased. This shows that it is unreasonable

to expect a unidirectional change in standard errors when using all applicable subjects to calculate variance components. Instead, it should be expected that the standard errors calculated using the existing GMM method will themselves show great variation and cannot be trusted. It is expected that standard errors calculated using the modified GMM method will show less variation and thus be more reliable.

The parameter estimates have also changed from those of the existing GMM method. The effect of education level has changed direction but has lost significance at the 0.01 level. Most parameter estimates who a decreased magnitude using the modified GMM, although the job training parameter estimate shows an increase in magnitude. Like the standard errors, the effect on parameter estimates cannot be predicted. It is expected that the parameter estimates calculated with contributions to the weight matrix from all subjects using the modified GMM will be more reliable and show less bias than those calculated using the existing GMM.

5. Conclusions

In this paper an alternative generalized method of moments estimation method is presented for the case of unbalanced longitudinal data with time-dependent covariates. Existing methods use only full-data subjects for many steps of the calculation. This new method can greatly increase the number of subjects whose information is used in calculation the key quadratic form of the generalized method of moments as well as calculation of estimator standard errors. It is expected that use of data from as many subjects as possible will reduce the chance of committing errors in hypothesis tests and reduce bias in parameter estimates.

To exemplify the two estimation methods presented, an example data set was analyzed. In modeling the probability of a length of stay greater than two weeks, there are noticeable differences in the results between the existing method and the proposed method. The proposed method shows larger standard errors in some cases but smaller standard errors in others. Parameter estimates differ between the two methods.

Overall it is expected that the proposed method is associated with more reasonable standard errors, as the standard errors calculated using the existing method are based on very few subjects and thus are subject to a large amount of variation themselves. This is noted in the example by seeing that not all of the parameters show significance in the proposed method. It is also expected that the proposed method is associated with less bias in parameter estimates, although in the data example values are close. Future research will include a simulation study to evaluate the empirical bias, power and Type I error rates associated with parameter estimates and hypothesis tests for the two methods.

References

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of longitudinal data.* USA: Oxford University Press, second edition.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hansen, L. P. (2007). Generalized method of moments estimation. *Department of Economics, University of Chicago*, pages 1–14.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis.* Wiley - Interscience.

Hu, F.-C. (1993). A statistical methodology for analyzing the causal health effect of a time dependent exposure from longitudinal data. *Unpublished Sc. D. dissertation, Harvard School of Public Health, Department of Biostatistics*.

Lai, T. L. 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(1):79–99.

Lalonde, T. L., Nguyen, A. Q., Yin, J., Irimate, K., and Wilson, J. R. (2013). Modeling correlated binary outcomes with time-dependent covariates. *Journal of Data Science*, 11:715–738.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communcations of Statistical Simulation and Computation*, 23:939–951.

Yin, J., Wilson, J. R., and Lalonde, T. L. (2013). Correlated gmm logistic regression models with time-dependent covariates and valid estimating equations. *Submitted to Statistics in Medicine*.