Estimating Parameters for Binary Data with Time-Dependent Covariates Using the Generalized Method of Moments

Maryann N. Shane
University of Northern Colorado, 501 20th Street, Greeley, CO 80639

Abstract

Generalized Method of Moments (GMM) has been used with longitudinal data models with time-dependent covariates. Lai and Small (2007) applied the two-step GMM (2SGMM) to continuous longitudinal data with time-dependent covariates to improve efficiency over the independent Generalized Estimating Equations (GEE) approach (Zeger & Liang, 1986; Liang & Zeger, 1986). The continuously updating GMM (CUGMM) procedure, outlined by Hansen (1982), was discussed in the JSM 2012 topic contributed session, Correlated Logistic Regression with Time-Dependent Covariates (Lalonde et al, 2012), and has been applied to binary longitudinal data with time-dependent covariates. When a binary response and binary time-dependent covariate are present, convergence failure has been observed for 2SGMM.

This paper presents the use of 2SGMM for binary longitudinal data in the presence of at least one binary time-dependent covariate. The GMM algorithm fails to converge. Possible reasons and explanations, including two-way tables of the data, are offered.

Key Words: Generalized Method of Moments, binary covariates, binary response, time dependent covariates, convergence failure, quasi-complete separation

1 Introduction

Correlated binary response data are oftentimes observed in biomedical studies. This type of data must be handled with caution due to their different properties from uncorrelated (independent) or continuous data. In modeling this type of response, the Generalized Estimating Equations (GEE) approach by Zeger and Liang (1986) has become the most common and most popular. The GEE process takes into account the inherent correlation in the response over repeated observations and allows the researcher to specify the nature of this correlation (e.g., Compound Symmetry, 1^{st} Order Autoregressive). One of the main attractions behind GEE is that regardless of misspecification of the working correlation structure, the GEE algorithm always yields consistent parameter estimates. Additionally, the GEE approach does not require full specification of the response distribution. It only requires information involving the mean-variance relationship of the response, hence only assuming a quasi-likelihood instead of the full likelihood (Liang and Zeger, 1986; Zeger and Liang, 1986).

However, GEE may become problematic when special types of covariates that are associated via a feedback loop (i.e., time-dependent covariates) enter the picture. As suggested by Pepe and Anderson (1994) and Fitzmaurice (1995), the independent working correlation structure is the "safest" option when applying the GEE approach in modeling marginal means. When these time-varying covariates are present in modeling the marginal mean, they create clusters and groups, which result in non-diagonal covariance structures. A crucial concept behind Independent GEE is the assumption of a diagonal covariance structure; thus, the presence of time-dependent covariates almost always violates this key assumption behind the Independent GEE process (Pepe and Anderson, 1994). Moreover, when a time-dependent covariate is present and the Independent GEE approach is applied, substantial loss of efficiency can be observed as the correlation inherent in the response over repeated observations increases (Fitzmaurice, 1995). For these reasons, the analysis of correlated response data involving time-dependent covariates can become tricky and very difficult quickly.

In comparison to using the Independent GEE approach, the use of Generalized Method of Moments (GMM) estimation improves efficiency when time-dependent covariates are present. The results from Lai and Small's (2007) simulation study show that GMM estimators are more efficient than Independent GEE estimators when time-dependent covariates (TDC) of types I or II are involved and equally as efficient as Independent GEE estimators when a TDC of type III is present. For this reason, GMM estimation is a superior method when compared to Independent GEE in estimating parameters of a longitudinal study, especially when time-dependent, or time-varying, covariates are observed.

Our example involves data obtained from a hospital database in Arizona, with the interest to model hospital readmission within 30 days of discharge. Each patient was observed three times, and information including age, whether or not the patient suffered a stroke, whether or not (s)he suffered a coronary thrombosis, whether or not (s)he had pneumonia, and hospital length of stay were recorded. The response variables is binary: yes or no to readmission within 30 days. A logistic regression model with both binary and continuous TDCs will be applied, and we will use the GMM estimation process. When binary TDCs are present, GMM has a tendency to fail in algorithm convergence.

This paper is organized as follows: in Section 2, brief reviews of TDCs and the GMM process are given; in Section 3, a proposed model for our example data, GMM algorithm convergence failure, and possible explanations are discussed; suggestions for future research, as well as ongoing research, are offered in Section 4.

2 TDCs and the GMM Process

2.1 Time-Dependent Covariates

Oftentimes in medical and longitudinal data situations, current and future covariates are associated via a feedback loop. Consider a patient with hypertension (high blood pressure). Whether or not the doctor prescribes medication depends on the patient's blood pressure reading at today's clinic visit. The patient's blood pressure when (s)he returns to the doctor's office at a future date may depend on whether or not the doctor prescribed him/her medication at the previous visit. These special types of covariates that involve feedback are referred to as time-dependent covariates (TDCs) and introduce correlation among variables over time. As mentioned earlier, this correlation must be accounted for when constructing longitudinal models.

There are four types of time-dependent covariates, and distinctions are made based on their relationship to the response variable(s). These four types of TDCs are defined in the following subsections.

2.1.1 *Type I TDCs*

When the current covariate affects only the current response, it is considered a type I time-dependent covariate. This type of TDC involves no feedback.

By definition, a type I TDC is a covariate that satisfies the expression

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

for all s, t, such that s = 1,...,T and t = 1,...,T,

and $\beta_{\mathbf{0}}$ is the vector of true parameters, and β_j is the j^{th} predictor.

There is no association between the derivative term and the residual term for any combination of s, t, such that s = 1,...,T and t = 1,...,T.

2.1.2 *Type II TDCs*

When the current covariate affects both the current response and some future response, it is a type II TDC. This type of TDC involves feedback between the current covariate and future response.

By definition, a type II TDC is a covariate that satisfies the expression

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

for all combinations of s, t, such that $s \geq t$.

An association exists between the current derivative and future residual terms.

2.1.3 *Type III TDCs*

Type III TDCs involve feedback loops such that the current response affects the covariate at some future time, and the current covariate affects the response at some future time.

By definition, a type III TDC is a covariate that satisfies the expression

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

for some $s>t$.

No association exists between current residual and current derivative terms. When this type of TDC is present in analyses, the independent working correlation structure works best.

2.1.4 *Type IV TDCs*

Type IV TDCs can be thought of as the "opposite" of type II TDCs. A feedback loop exists such that the current response is associated with the current covariate, and the current response also affects the covariate at some future time.

By definition, a type IV TDC is a covariate that satisfies the expression

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

for all combinations of s, t, such that $s\leq t$.

An association exists between the current residual and future derivative terms.

2.2 Generalized Method of Moments

In dealing with correlated binary response data, the Generalized Estimating Equations (GEE) approach by Liang and Zeger (1986) has become the most common method. In handling models with correlated response, GEE always yields consistent parameter estimates. This is true regardless of whether or not the correct working correlation structure is specified. More specifically, Independent GEE will always yield consistent parameter estimates, even when time-varying covarites (i.e., time-dependent covariates) are present. However, there are drawbacks to the Independent GEE approach, as outlined by Pepe and Anderson (1994) and Fitzmaurice (1995), including its loss of efficiency as correlation between responses increases.

In comparison to using Independent GEE, the use of Generalized Method of Moments (GMM) estimators improves efficiency when time-dependent covariates of type I or type II are involved in the analysis. Results of Lai and Small's (2007) simulation study show that the use of GMM estimators does not change the efficiency of parameter estimates when time-dependent covariates of type III are involved in the analysis. Thus, GMM estimators are equally as efficient as GEE estimators when the working correlation structure is correctly specified, and they are more asymptomatically efficient than GEE estimators when the working correlation structure is misspecified. For this reason, GMM estimation is a superior method when compared to Independent GEE when estimating parameters of a longitudinal study.

The Generalized Method of Moments process assumes no response distribution and therefore is not likelihood or quasi-likelihood based. Instead, GMM estimates are obtained using moment conditions, which are products of derivative and residual terms for the i^{th} subject at times s and t:

$$\sum_{i=1}^{N} \frac{\partial \mu_{is}(\beta_0)}{\partial \beta_j} \left\{ y_{it} - \mu_{it}(\beta_0) \right\}, \;\; s = 1,...,T \text{ and } t = 1,...,T.$$

The first term in the summation is the derivative of the mean for the i^{th} subject with respect to the j^{th} predictor at time s. The second term is the residual for the i^{th} subject at time t. β_0 is the vector of true parameter values. (In practice, we may not know β_0, in which case we use something like Independent GEE to obtain "probable" parameter estimates.)

When time-dependent covariates are present, only those moment conditions that are considered to be "valid" are used in the GMM estimation process. Selection of these "valid" moment conditions depends on the types of time-dependent covariates that are present. As given in sections 2.1.1 through 2.1.4, the four types of time-dependent covariates work for specific combinations of time points s and t. These combinations of s and t determine which moment conditions are "valid" for the different types of TDCs. For example, for balanced longitudinal data with $T = 3$ (such as in our example), consider a T by T, or 3 by 3 matrix of products of derivative and residual terms for the j^{th} predictor:

$$\begin{array}{ll}
\text{time:} & \begin{array}{ccc} \qquad t=1 \qquad & \qquad t=2 \qquad & \qquad t=3 \qquad \end{array} \\
\begin{array}{c} s=1 \\ s=2 \\ s=3 \end{array} &
\begin{pmatrix}
\frac{\partial \mu_{i1}}{\partial \beta_j}\left\{y_{i1} - \mu_{i1}\right\} & \frac{\partial \mu_{i1}}{\partial \beta_j}\left\{y_{i2} - \mu_{i2}\right\} & \frac{\partial \mu_{i1}}{\partial \beta_j}\left\{y_{i3} - \mu_{i3}\right\} \\
\frac{\partial \mu_{i2}}{\partial \beta_j}\left\{y_{i1} - \mu_{i1}\right\} & \frac{\partial \mu_{i2}}{\partial \beta_j}\left\{y_{i2} - \mu_{i2}\right\} & \frac{\partial \mu_{i2}}{\partial \beta_j}\left\{y_{i3} - \mu_{i3}\right\} \\
\frac{\partial \mu_{i3}}{\partial \beta_j}\left\{y_{i1} - \mu_{i1}\right\} & \frac{\partial \mu_{i3}}{\partial \beta_j}\left\{y_{i2} - \mu_{i2}\right\} & \frac{\partial \mu_{i3}}{\partial \beta_j}\left\{y_{i3} - \mu_{i3}\right\}
\end{pmatrix}
\end{array}$$

For a type I TDC, all 9 elements in the matrix are "valid" moment conditions because the expectation of the moment condition equals 0 for all s and t, s = 1,...,T and t = 1,...,T. For a type II TDC, elements in the lower triangular are "valid" moment conditions because the expectation of the moment condition equals 0 for all combinations of s and t where $s \geq t$:

$$\begin{pmatrix}
\frac{\partial \mu_{\mathbf{i1}}}{\partial \beta_{\mathbf{j}}}\left\{\mathbf{y_{i1}} - \mu_{\mathbf{i1}}\right\} & \frac{\partial \mu_{i1}}{\partial \beta_j}\left\{y_{i2} - \mu_{i2}\right\} & \frac{\partial \mu_{i1}}{\partial \beta_j}\left\{y_{i3} - \mu_{i3}\right\} \\
\frac{\partial \mu_{\mathbf{i2}}}{\partial \beta_{\mathbf{j}}}\left\{\mathbf{y_{i1}} - \mu_{\mathbf{i1}}\right\} & \frac{\partial \mu_{\mathbf{i2}}}{\partial \beta_{\mathbf{j}}}\left\{\mathbf{y_{i2}} - \mu_{\mathbf{i2}}\right\} & \frac{\partial \mu_{i2}}{\partial \beta_j}\left\{y_{i3} - \mu_{i3}\right\} \\
\frac{\partial \mu_{\mathbf{i3}}}{\partial \beta_{\mathbf{j}}}\left\{\mathbf{y_{i1}} - \mu_{\mathbf{i1}}\right\} & \frac{\partial \mu_{\mathbf{i3}}}{\partial \beta_{\mathbf{j}}}\left\{\mathbf{y_{i2}} - \mu_{\mathbf{i2}}\right\} & \frac{\partial \mu_{\mathbf{i3}}}{\partial \beta_{\mathbf{j}}}\left\{\mathbf{y_{i3}} - \mu_{\mathbf{i3}}\right\}
\end{pmatrix}$$

Following this logic, the 3 diagonal elements are the "valid" moment conditions for type III TDCs, and the 6 elements in the upper triangular are the "valid" moment conditions for type IV TDCs because $s \leq t$.

The GMM process uses all valid moment conditions for each of the j covariates and combines them into the vector $\mathbf{g}(\mathbf{y_i}, \mathbf{x_i}, \beta_\mathbf{0})$. These $\mathbf{g}$ vectors are averaged over all N subjects to obtain G_N:

$$G_N(\beta) = \sum_{i=1}^{N} \mathbf{g}(\mathbf{y_i}, \mathbf{x_i}, \beta_\mathbf{0})$$

To obtain the Generalized Method of Moments parameter estimates, we construct a quadratic form Q_N:

$$Q_N(\beta) = G_N(\beta)^T \, W_N \, G_N(\beta)$$

and minimize this quadratic form. Here, W_N is a weight matrix, whose optimal choice was shown to be the inverse of the covariance matrix of the moment conditions $\mathbf{g}(\mathbf{y_i}, \mathbf{x_i}, \beta)$ by Hansen (1982).

3 The GMM Approach for Models with TDCs

3.1 Correlated Binary Response Model with Binary TDCs

A logistic regression model was constructed to model the rate of readmission within 30 days of discharge, where readmission is a binary variable. Each patient was observed 3 times, and model predictors include age, hospital length of stay (LOS), whether or not the patient suffered a stroke, whether or not the patient suffered a coronary thrombosis, and whether or not the patient had pneumonia. For the continuous time-dependent covariates, age is a type I TDC and hospital length of stay is a type II TDC. All three binary TDCs are of type III. The Generalized Method of Moments estimation technique was applied to construct the model:

$$\begin{aligned} logit(p) = \beta_0 + \beta_1 X_{it,Age} + \beta_2 X_{it,LOS} + \beta_3 X_{it,Coronary} \\ + \beta_4 X_{it,Pneumonia} + \beta_5 X_{it,Stroke} \end{aligned}$$

where p = probability of readmission within 30 days,
and $X_{it,Stroke}$, $X_{it,Coronary}$, and $X_{it,Pneumonia}$ are indicator variables.

3.2 GMM Algorithm Convergence Failure

3.2.1 *Binary TDCs and GMM Non-Convergence*

When applying the two-step Generalized Method of Moments (2SGMM) approach outlined by Lai and Small (2007) to these data, the algorithm failed to converge and no parameter estimates were obtained.

It is suspected that the GMM algorithm non-convergence is due to quasi-complete separation of the binary variables. The GMM process relies on moment conditions rather than on a set of likelihood equations. Recall that these moment conditions are products of derivative and residual terms for the i^{th} subject. For a Bernoulli response, this product simplifies to:

$$\mu_i \left(1 - \mu_i\right)$$

Further, GMM estimation involves computations that take the product of these moment conditions with the covariate:

$$\mu_i \left(1 - \mu_i\right) X_{it}$$

For the binary predictors, X_{it} assumes value 1 when the condition is present, thus returning the familiar product: $\mu_i \left(1 - \mu_i\right)$. However, when the condition is not present, X_{it} becomes 0, wiping out this entire product.

Moreover, when the value of the binary covariate is 0, the moment condition simplifies to 0, leaving very few moment conditions to be used in the GMM algorithm, thus resulting in convergence failure.

3.2.2 *Quasi-Complete Separation*

The algorithm used in the Generalized Method of Moments estimation process failed due to quasi-complete separation of the binary variables.

The issue with quasi-complete separation is most apparent when examining two-way tables of the response variable with the binary covariate.

The following are two-way tables of the response variable, Readmission, with each of the binary time-dependent covariates. Overall, 9325 patients, approximately 71% of our sample, were readmitted, and 3830 patients, approximately 29% of our sample, were not readmitted.

Table 1: Incidence of Readmission vs. Incidence of Stroke

	Stroke		
Readmission	No (0)	Yes (1)	sum
Not readmitted	3752	78	3830
Readmitted	9058	267	9325
sum	12810	345	13155

Table 1 illustrates that only 345 patients suffered a stroke out of a total of 13155 patients and only 267 of those who suffered a stroke were readmitted. These counts are relatively small compared to the total 13155 patients – proportionally, they are barely 2% and 3% of the total 13155 patients, respectively. 12810 patients did not experience a stroke, which amounts to over 97% of the total 13155 patients, and 9058 of these patients who were readmitted did not experience a stroke, an approximate 70% of all patients in the study.

Table 2: Incidence of Readmission vs. Incidence of Coronary Thrombosis

	Coronary Thrombosis		
Readmission	No (0)	Yes (1)	sum
Not readmitted	3330	500	3830
Readmitted	8268	1057	9325
sum	11598	1557	13155

Table 2 illustrates that only 1557 patients suffered a coronary thrombosis out of the total 13155 patients, and only 1057 of those who suffered a coronary thrombosis were readmitted. These counts are again relatively small compared to the total 13155 patients – proportionally, they are about 8% and 12%, respectively. 11598 patients did not experience a coronary thrombosis, which amounts to approximately 88% of the total 13155 patients, and 8268 of these patients who were readmitted did not experience a coronary thrombosis, an approximate 63% of all patients in the study.

Table 3: Incidence of Readmission vs. Incidence of Pneumonia

	Pneumonia		
Readmission	No (0)	Yes (1)	sum
Not readmitted	3472	358	3830
Readmitted	8581	744	9325
sum	12053	1102	13155

Table 3 illustrates that only 1102 patients had pneumonia out of the total 13155 patients, and only 744 of those who had pneumonia were readmitted. These counts are relatively small compared to the total 13155 patients – proportionally, they are approximately 8% and 5.65%, respectively. 12053 patients did not have pneumonia, which amounts to approximately 92% of the total 13155 patients, and 8581 of these patients who were readmitted did not have pneumonia, an approximate 65% of all patients in the study.

4 Discussion

4.1 Existing Solutions

Currently exisiting solutions to this issue involving binary time-dependent covariates are not ideal. One of the traditional approaches involves ignoring the time-dependence that is inherent across all observation times. By doing so, we ignore the correlation in the response across multiple observations, which results in less efficient model parameter estimates.

Another common approach in dealing with binary TDCs is to use an estimation process that accounts for correlation among the response, such as the Independent Generalized Estimating Equations approach, as introduced by Zeger and Liang (1986). Using Independent GEE will always result in consistent parameter estimates; however, as outlined by Pepe and Anderson (1994) and Fitzmaurice (1995), Independent GEE yields less efficient parameter estimates as the correlation increases across multiple observations. Thus, using Independent GEE may circumvent the issue of GMM algorithm non-convergence but may result in parameter estimates that are less efficient.

The least ideal method in dealing with binary time-dependent covariates is to omit them altogether from statistical analyses. Avoiding the use of binary TDCs may result in easier and simpler analyses but will result in loss of useful and important information. Ease in handling the data by avoiding the inclusion of binary TDCs may distort the information extracted from the data.

4.2 Extensions and Further Research

4.2.1 *Simulation Study*

Ongoing research includes a simulation study involving manipulation of the success probability of the binary time-dependent covariates in an attempt to coerce GMM algorithm convergence. Issues that result in algorithm convergence failure, such as quasi-complete separation, may be avoided if the success probability of the binary TDC is contained at a certain level. Our research currently involves finding some threshold for p_x, the probability of success of the binary time-dependent covariate, such that convergence or non-convergence of the GMM algorithm can be observed.

4.2.2 *Selection of Moment Conditions*

Another research that is currently being conducted to handle data with time-dependent covariates involves a new method in selecting “valid” moment conditions. Current selection of “valid” moment conditions is directly dependent on the type of time-dependent covariate being estimated. Rather than selecting these moment conditions based on the type of TDC, we propose an improvement that involves assigning weights to all moment conditions. This way, all $T \times T$ valid moment conditions are used in the GMM estimation process, leaving none out of the estimation process.

References

Fitzmaurice, G. M. (1995), “A Caveat Concerning Independence Estimating Equations with Multivariate Binary Data,” *Biometrics*, 51, 309–317.

Hansen, L. P. (1982), “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50(4), 1029–1054.

Lai, T. L., and Small, D. (2007), “Marginal Regression Anlaysis of Longitudinal Data with Time-Dependent Covarites: a Generalized Method-of-Moments Approach,” *Journal of the Royal Statistical Society*, Ser. B, 69, 79–99.

Lalonde, T. L., et al (2012), *Correlated Logistic Regression with Time-Dependent Covariates*. Paper presented at the Joint Statistical Meetings, San Diego, CA.

Liang, K. Y., and Zeger, S. L. (1986), “Longitudinal Data Analysis Using Generalized Linear Models,” *Biometrika*, 73(1), 13–22.

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,” *Communication Statistics*, 23(4), 939–951.

Zeger, S. L., and Liang, K. Y. (1986), “Longitudinal Data Analysis for Discrete and Continuous Outcomes,” *Biometrics*, 42, 121–130.